Annotating and Analyzing Semantic Relations between Texts

Von der Fakultät für Ingenieurwissenschaften,
Abteilung Informatik und Angewandte Kognitionswissenschaft
der Universität Duisburg-Essen
zur Erlangung des akademischen Grades

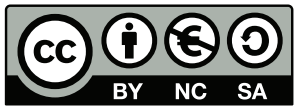
Doktor der Naturwissenschaften

genehmigte Dissertation

von

Darina Gold
aus
Kyiv

1. Gutachter: Prof. Dr. Torsten Zesch
2. Gutachter: Prof. Dr. Chris Biemann
Tag der mündlichen Prüfung: 01.07.2021



Abstract

In this thesis, we investigate machine computable and at the same time human-understandable representation dimensions of text that can subsequently be used to filter and display information. While texts can be represented individually e.g. using numeric dimensions such as sentence length or grammatical components, we focus on representation dimensions that express relations between pairs of text. Most of the herein researched relation dimensions are binary, meaning that the relations of interest either do or do not exist between a text pair.

Some dimensions are inherently defined as text-to-text relations e.g. textual entailment, paraphrases, contradiction, or semantic similarity. That is, there can be no paraphrase within one text, but it is a relation between a text pair.

While there has been much research on these dimensions individually, one of our contributions is the empirical research on the links between them. On the one hand, this provides us with a better understanding of each individual dimension. For instance, we find that although entailment, as well as paraphrases, exclude contradictions, text pairs not containing entailment are not necessarily contradictions, which has, however, been considered a given many previous works. On the other hand, our analysis has the potential of improving transfer learning by using corpora on one of the dimensions to automatize another. We find, i.a. that the most prominent assumed link between dimensions—bi-directional entailment being equivalent to paraphrases—does not always hold. However, in most cases it is true, meaning that transfer learning between these dimensions is possible.

As for dimensions that can also exist for individual pieces of text, we believe that some of them can also be better researched as relations between texts. By rating the sentiment of text in comparison to other texts instead of using a scale for each individual text, this has already been shown on the example of sentiment. Another contribution of this thesis is considering not only sentiment, but also specificity, as a relation. We find that specificity, just like sentiment, can be reliably annotated as a relation. Moreover, we find further potential parallels to sentiment regarding the operationalization of specificity—it can be more reliable annotated with an aspect, similar to the task of aspect-based sentiment.

A further contribution of this thesis is the research on the link between dimensions that are inherently a relation and the under-researched phenomenon of specificity. For instance, we hypothesize that the entailed text of an entailment pair has a lower specificity level than the entailing text, as the entailed text should not contain any additional information than already described in the entailing text. The analysis of links between the inherent relation dimensions and specificity helps us to deepen our understanding of this under-researched phenomenon and gives an incentive on how to improve its automation.

Finally, we present two potential applications using each dimension, namely heterogeneous multi-document summarization, and a more specific kind of summarization—user specific hotel review filtering.

Zusammenfassung

In dieser Arbeit untersuchen wir maschinenberechenbare und gleichzeitig vom Menschen verständliche Darstellungsdimensionen von Text, die anschließend zum Filtern und Anzeigen von Informationen verwendet werden können. Während Texte einzeln dargestellt werden können, z.B. unter Verwendung numerischer Dimensionen wie Satzlänge oder grammatikalischer Komponenten konzentrieren wir uns auf Darstellungsdimensionen, die Beziehungen zwischen Textpaaren ausdrücken. Die meisten der hier untersuchten Beziehungsdimensionen sind binär d.h., dass die Beziehungen zwischen einem Textpaar existieren oder nicht.

Einige Dimensionen sind per Definition Text-zu-Text-Beziehungen, z.B. textuelles Entailment, Paraphrasen, Widerspruch oder semantische Ähnlichkeit. So kann es keine Paraphrase innerhalb eines Textes geben, da es eine Beziehung zwischen einem Textpaar ist. Während diese Dimensionen einzeln jeweils viel erforscht wurden, ist einer unserer Beiträge die empirische Untersuchung der Verbindungen zwischen ihnen. Dies gibt uns einerseits ein besseres Verständnis für jede einzelne Dimension, so stellen wir zum Beispiel fest, dass sowohl Entailment als auch Paraphrasen Widersprüche ausschließen, Textpaare, die Entailment nicht enthalten, jedoch nicht unbedingt Widersprüche sind, was in vielen vorherigen Arbeiten jedoch als gegeben betrachtet wurde. Andererseits hat unsere Analyse das Potenzial, Transfer Learning zu verbessern, indem Korpora in einer der Dimensionen verwendet werden, um eine andere zu automatisieren. Wir finden, u.a. dass die prominenteste angenommene Verbindung zwischen Dimensionen—beidseitiges Entailment entspricht Paraphrasen—nicht immer gilt. In den meisten Fällen ist dies jedoch der Fall, was bedeutet, dass ein Transfer Learning zwischen diesen Dimensionen möglich ist.

Weiterhin zeigen wir, dass einige Dimensionen, die traditionell als Einzeldimensionen existieren, als Beziehungen besser erforscht werden können. Durch die Bewertung von Sentiment im Vergleich zwischen Texten anstelle einer Skala für jeinzelne Texte wurde dies bereits am Beispiel von Sentiment gezeigt. Wir finden, dass Spezifität genau wie Sentiment zuverlässig als Beziehung annotiert werden kann. Darüber hinaus finden wir weitere Parallelen zu Sentiment in Bezug auf die Operationalisierung der Spezifität - sie kann zuverlässiger mit einem Aspekt annotiert werden, ähnlich der Aufgabe der aspektbasierten Stimmungsanalyse.

Ein weiterer Beitrag dieser Arbeit ist die Erforschung des Zusammenhangs zwischen Dimensionen, die inhärent als Beziehung definiert sind, und dem bisher vernachlässigten Phänomen der Spezifität. Dies wird ermöglicht, indem die Spezifität als Beziehung betrachtet wird, wie im vorherigen Absatz beschrieben. Zum Beispiel nehmen wir an, dass der implizierte Text eines Entailment-Paares eine niedrigere Spezifität aufweist als der implizierende Text, da der implizierte Text keine zusätzlichen Informationen enthalten sollte, als bereits im implizierenden Text beschrieben. Die Analyse der Zusammenhänge zwischen den inhärenten Beziehungsdimensionen und der Spezifität hilft uns, unser Verständnis dieses unterforschten Phänomens zu vertiefen, und gibt einen Anreiz zur Verbesserung seiner Automatisierung.

Schließlich stellen wir zwei mögliche Anwendungen für jede Dimension vor, nämlich die heterogene Zusammenfassung mehrerer Dokumente und eine spezifischere Art der Zusammenfassung - das benutzerspezifische Filterverfahren von Hotelbewertungen.

Danksagung

Ich danke Torsten Zesch für die Chance an seinem Lehrstuhl zu promovieren und die wissenschaftliche sowie die moralische Unterstützung in diesen Jahren. Danke für deine Betreuung, deine kreativen Ideen, deine fördernde Kritik und auch die vielen spannenden Auslandserfahrungen.

Ebenso danke ich Chris Biemann für sein hilfreiches, detailliertes Feedback zu dieser Arbeit, aber insbesondere für seine großzügige und maßgebliche Unterstützung im Vorhinein, die mich erst dazu motiviert hat zu promovieren.

Ein großer Dank geht auch an meine Kolleg(inn)en—Andrea Horbach, Marius Hamacher, Piush Aggarwal, Ronja Laarmann-Quante, Semire Yekta, und Yuning Ding—für die langen Diskussionen, nützlichen Kommentare und Anregungen zu den Studien und letztendlich zur Dissertation selbst. Des Weiteren möchte ich meinen ehemaligen Social-Media-Büro-Kollegen, Michael Wojatzki und Tobias Horsmann, nicht nur für die diversen Disskusionen und Kommentare, sondern auch einfach für ihre Freundschaft danken. Ich danke außerdem den studentischen Hilfskräften, insbesondere Marie Bexte und Mara Ortmann, sowie Sarah Holschneider als freundliche Helferin, für die Annotationen ohne die die entsprechenden Studien nicht möglich gewesen wären.

Des Weiteren möchte ich meinen Eltern für ihren unermesslichen Beistand danken. Meine Eltern waren immer für mich da, ob mit einem warmen, leckeren Essen während einer Deadlinephase, mit Tipps zu Kommadozeilenbefehlen, oder als Kinderbetreuer. Und meiner Tochter danke ich dafür, dass sie mich auch nach den anstrengendsten Arbeitstagen innerhalb weniger Minuten zum Lachen bringt.

Der größte Dank geht aber an meinen Mann. Chris, ohne deine Geduld, Unterstützung, Ideen, akribische Hilfe beim Layout, Kommentare zum Text, Disskussionen zum Code, aber auch die neue Kaffeemaschine, die du mir geschenkt hast, hätte ich es bestimmt nicht geschafft.

Furthermore, I would like to give thanks for the international collaborations that helped me to widen my personal as well as my scientific horizon. Venelin Kovatchev, thanks for your ideas, discussions, time, and cooperation. It was fun working with you and I hope to continue our collaboration. Ido Dagan and his team, thanks for the possibility to visit and experience your lab—your feedback and insightful knowledge on relations between statements brought my work forward.

Diese Arbeit wurde unterstützt durch die Deutsche Forschungsgemeinschaft (DFG) unter grant No. GRK 2167, Research Training Group “User-Centred Social Media” (UCSM) und im Projekt “Argument-Based Decision Support for Recommender Systems"(ASSURE) unter grant No. ZE 915/6-1, als Teil des Schwerpunktprogramms “Robust Argumentation Machines” (RATIO) (SPP-1999). Außerdem wurde diese Arbeit von zwei Kurzstipendien des Deutschen Akademischen Auslandsdienstes (DAADs) für jeweils eine Reise nach Tunesien und Israel unterstützt.

Contents

List of Figures

List of Tables

Acronyms

ABSA aspect-based sentiment analysis
AMT Amazon Mechanical Turk

BWS best-worst scaling

CL Computational Linguistics
CV Cross Validation

EPT Extended Paraphrase Typology
ETPC Extended Paraphrase Typology Corpus

HIT Human Intelligence Task

IAA Inter-Annotator Agreement

MRPC Microsoft Paraphrase Corpus

NE Named Entitity
NLP Natural Language Processing
NYT New York Times

PI Paraphrase Identification
POS Part-of-Speech
PPW Perplexity Per Word

SHARel Single Human-Interpretable Typology for Annotating Meaning Relations
SHR Split-Half Reliability
SNLI Stanford Natural Language Inference
SVM Support Vector Machine

TPC Twitter Paraphrase Corpus

Introduction

The filtering of relevant information from the overflow of available textual data is a widely comprehensive issue. To filter relevant data is ubiquitous when dealing with digital online data: one needs it when looking for all kinds of information—film, television, music, books, news, web pages, or also reviews on the named items, to name only a few scenarios where filtering of redundant or unwanted information is required.

A more concrete example for the need of information filtering are users of hotel review websites. Mostly, a user is interested in specific properties of hotels, thus being in need for a filtering mechanism showing only reviews potentially containing these properties. By way of illustration, this application will be used as a running example in this thesis.

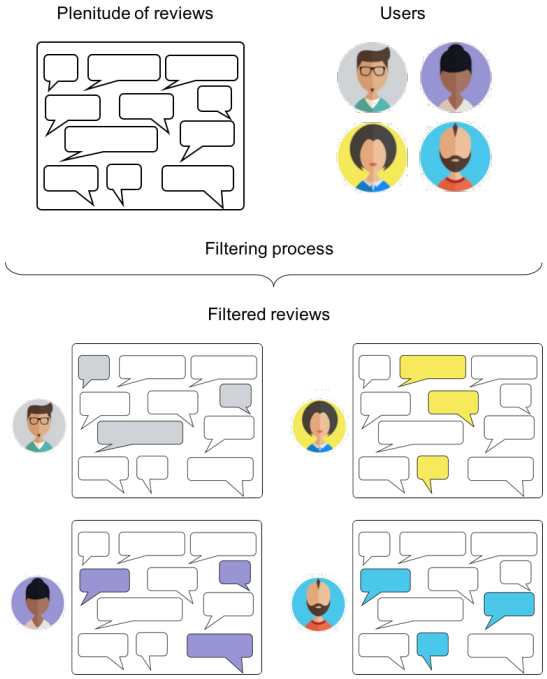


FIGURE 1: Illustration of user-specific hotel review filtering scenario

Figure 1 shows the user-specific filtering of reviews in a simplified way. The text bubbles represent statements. Statements are pieces of text containing information. In the application scenario of hotel reviews, the statements are reviews or pieces of such. Having a plenitude of statements on one or many topics, it is a difficult task to filter for relevant information. Furthermore, users have individual needs to which the filtering mechanism could be adjusted. For instance, a work traveler might be interested in a reliable and fast WiFi and may be less interested in reviews mentioning child friendliness. Another user might have other, more specific interests—a hard mattress, the noise level, or how quickly the water temperature

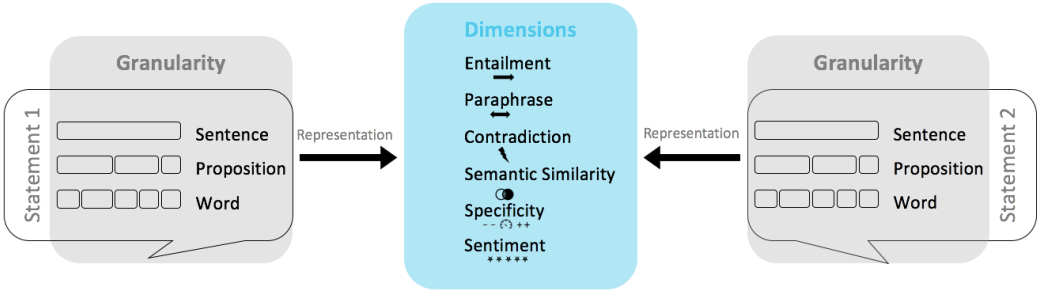


FIGURE 2: Illustration of relation dimensions (textual entailment, paraphrases, contradiction, semantic similarity, specificity, and aspect-based sentiment) and granularity levels (sentence, proposition, and word) used in this thesis

adjusts in the shower. Figure 1 represents these interest by differently colored backgrounds of the users. Furthermore, there are aspects of hotel reviews, e.g. avoiding redundancy, which is needed for all users. So, user-specific and user-unspecific aspects must be considered in the filtering process. As a result, each user should be shown a selection of filtered hotel reviews adjusted to her or his needs.

In order to be able to perform this filtering process, the textual information needs some kind of abstract representation. In Natural Language Processing (NLP) and Computational Linguistics (CL), there exists a plenitude of representation types for text. Herein, we examine these types and research new dimensions of representations with a special focus on explainability[1] of our representation. The representation needs to be explainable, as it is supposed to be shown to a user who needs to understand why she or he is shown exactly these statements. Currently, mathematical vectors, also known as *embeddings* are often used for text representations, as they perform well in most tasks. However, a representation in the form of a mathematical vector, even if it performs well, is not helpful in our case, as a semantic explanation is needed. Furthermore, we are interested in which dimensions might be helpful, which remains not human-understandable in a vector setting.

As shown in Figure 2, the focus of this thesis are dimension that are operationalized as relations between at least two pieces of text, in contrast to single dimensions. In this thesis, the dimensions are quite flat, as they represent binary relations. This means, each dimension shows whether the respective relation exists between the given pieces of text. This stands in contrast to the previously mentioned mathematical vectors which are highly dimensional. We argue that using relation dimensions helps in the filtering process, as it directly helps choosing some texts over the other and the filtering process is user comprehensible. Furthermore, we plead that textual entailment, paraphrases, contradiction, semantic similarity, specificity, and aspect-based sentiment are suitable dimensions. Figure 2 presents an abstract representation of all dimensions researched in this thesis.[2]

[1]Explainability is a neologism from the field of artificial intelligence and refers to results of artificial intelligence methods being understandable by humans. It stands in contrast to the "black box" principle.

[2]Throughout this thesis, textual entailment will be equivalent to entailment, semantic similarity to similarity, and aspect-based sentiment with sentiment.

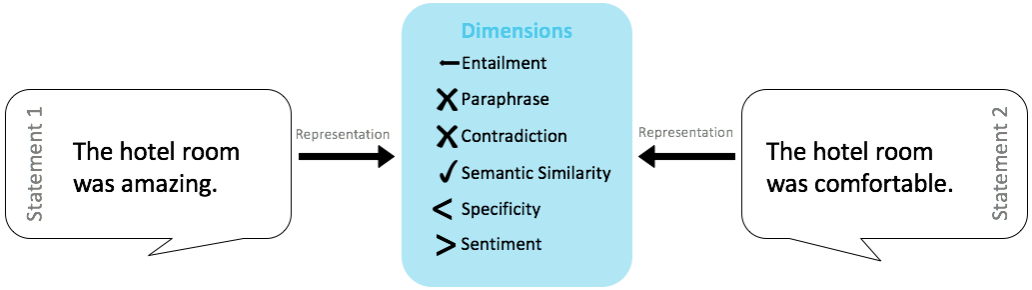


FIGURE 3: Concrete example of different relation dimensions between two sentences (*Statement 2* entails *Statement 1*; they are neither in a PARAPHRASE, nor in a CONTRADICTION relation; there is some SEMANTIC SIMILARITY; *Statement 2* contains more SPECIFICITY than *Statement 1*; *Statement 1* has a more positive SENTIMENT than *Statement 2*)

Contribution Overview

The main contribution of this thesis is operationalizing and uniting what we call relation dimensions, which have mostly been studied in isolation. On the one hand, we unite them in the operationalization, on the other hand we unite them by looking at them in parallel, thus being able to see links between different dimensions.

The operationalization is a further contribution by itself. While textual entailment, paraphrases, contradiction, and semantic similarity can only be relation dimensions—they only exist as relations between at least two statements. Seeing dimensions, which are usually not seen as relations—specificity and sentiment—as such is less researched. This is of special interest for the specificity dimension, which is less researched than the other dimensions and has hitherto not been regarded as a relation dimension.

Example 1 shows a concrete example of two statements on the granularity level of a sentence:

1 The hotel room was amazing.
2 The hotel room was comfortable.

EXAMPLE 1: Two statements on hotel rooms

(2) is more specific than (1), as it is not just a general positive judgment on the hotel room, but a reason of the positive judgment—the comfort. Concerning sentiment, (1) displays a more positive judgment on the hotel room than (2). The two statements are not paraphrases, as they do not display the same information. However, (1) textually entails (2): if the room is *amazing*, it has to be *comfortable* as well, but not the other way around—a comfortable room may have all practical devices, but lack the devices that would turn it into an amazing room. In the case of Example 1, the specificity may be found using textual entailment. The statements are not contradictions, as neither makes the other untrue. A visualization of Example 1 is shown in Figure 3.

Specificity and sentiment can be regarded both individually and in a relation. They can be changed to the other kind of dimension—from individual dimension to relation and back.

In the above paragraph, we have shown how they can act as relation dimensions. We are the first to regard specificity as a relation, thus making it an important contribution of this thesis and an innovation in the fields of CL and NLP. Most existing studies treat them as individual relations. In that case, a rating on one statement is performed. In Example 1, given a specificity scale of 0–4, where 0 denotes VERY GENERAL and 4 denotes VERY SPECIFIC, (1) could be given the specificity rating of 0 and (2) could be given a specificity rating of 2. Furthermore, given a sentiment rating on a scale of 0–4, where 0 denotes VERY NEGATIVE and 4 denotes VERY POSITIVE, (1) could be given the sentiment rating of 4 and (2) could be given a specificity rating of 3.

While most of the dimensions have been studied individually, another contribution of this thesis is the study of the links between different dimensions, which has been performed for only few of the discussed dimensions.

For instance, it is often assumed that statements in textual entailment, paraphrase, or contradiction relations also have at least some semantic similarity. Some of the dimensions have also been used in automatic annotation settings for other dimensions. However, this has not been done extensively—mostly only using one dimension to calculate another, ignoring that relation dimensions may be a dimension class that needs research on its own. Furthermore, hitherto the link between the other dimensions and and specificity has not been researched, as specificity has not been regarded as a relation dimension.

Another focus of the operationalization is the granularity on which the relations are researched. The granularities researched in this thesis are *sentence*, *proposition*, and *word*.

Application Scenario Example

Figure 4 illustrates a possible workflow using the herein researched dimensions for the filtering process. Furthermore, it shows how the different dimensions in this thesis can be used in the presented hotel review scenario. Given a plenitude of reviews, a user needs a filtering that fits her or his interests. In the first step, the statements are filtered according to the interest of the user. In the second step, the sentiment on the given aspect is added. Both steps together are represented by the dimension of aspect-based sentiment—meaning how positive or negative a statement is according to a given aspect. Consequently, the user now has reviews containing the aspect of interest with its rating. Although the amount of reviews has already been reduced, many statements containing aspect and sentiment of interest are still redundant, meaning that there are still many statements with the same textual information. In the third step, the statements with similar content are clustered into paraphrase clusters using the paraphrase dimension. In the fourth step, the statement with the right specificity level within the paraphrase cluster is chosen, leaving only one review per paraphrase cluster. Hence, the third and fourth steps together reduce the redundancy. In the filtering result, the different dimensions and the choice of of reviews is shown. This workflow is only one possible version of a filtering process, e.g. the paraphrase clustering could be replaced by entailment chains.

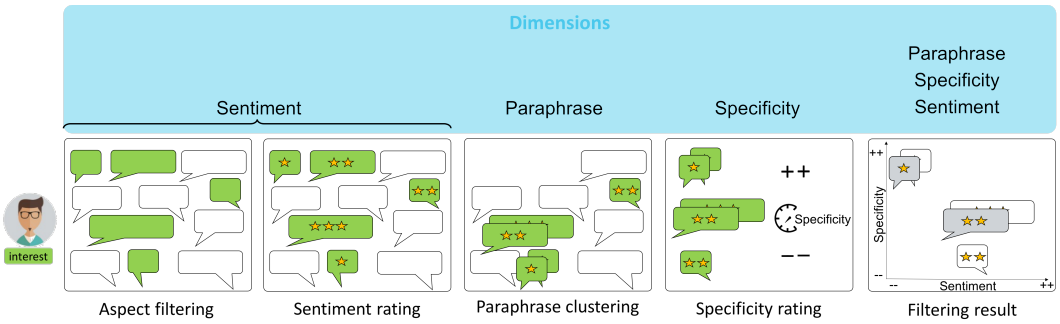


FIGURE 4: Illustration of a possible user-specific filtering process consecutively using the herein researched dimensions

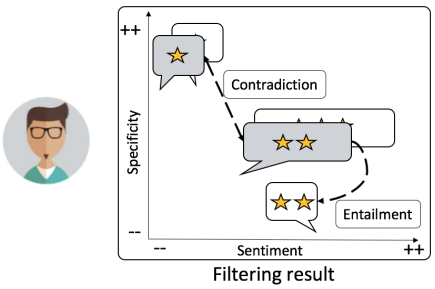


FIGURE 5: Possible dimension overview of the filtering result for the user showing filtered reviews and their the relation dimensions

The representation for the hotel reviews might include all possible dimensions in order to be better understandable for the user. Figure 5 shows such a possible display of a representation—two reviews that have been filtered to be displayed to the user. Their position in the different dimensions shows their specificity level, their sentiment polarity, and their semantic similarity. Dimensions are displayed in light blue. Both reviews have paraphrases, which are bundled together around the chosen review. Furthermore, both reviews have meaning relations, such as contradiction or entailment, to other reviews.

Studies Overview

In the following, we will shortly describe the studies that enabled our research on relation dimensions.

Annotation To empirically research the relation dimensions, their links, and operationalizations, we were in need of a new corpora with specific semantic manual annotations, as there was no previous work to base the analyses on. The topic of annotations in general, including its operationalization and evaluation, is discussed in Chapter 1.

Annotations were performed in the studies discussed in Chapter 4, Chapter 5, and Chapter 6. The operationalization and implementation of these studies is a contribution on its own. On the one hand, the operationalization mechanisms can be used in follow up studies. On the other hand, the resulting corpora enabled us to perform empirical research and, as they have been made freely available, can be used for further studies by the community.

Furthermore, we performed some automation experiments with the presented data using machine learning. The foundations of the automation are described in Chapter 2.

Representing Statements We are specifically interested in human-understandable dimensions of the representation that can be used for information filtering.

In Benikova et al. (2016), we examined existing statement representations, with a special focus on their expressiveness, meaning the human explainability. In this work, we are the first to describe a gap between computability, which is given by mathematical vectors, and expressiveness, which is given by *frames* for statement dimensions. This work is presented in detail in Chapter 3.

As we are interested in human-understandable representations, we find that *propositions* are a suiting granularity level, over which further dimensions such as *semantic roles* should be layered. Chapter 4 focuses discusses propositions as a granularity level for statement representation, including the studies performed in Benikova and Zesch (2017) and Gold and Zesch (2019). In Benikova and Zesch (2017), we examined the compositionality of similarity relations between statements on the example of paraphrases, which has not been done in previous studies. In Gold and Zesch (2019), we are the first to research how statement complexity is reflected in proposition extraction.

Relations between Statements The main study of this thesis was performed in Gold et al. (2019). In this study, we created and analyzed a corpus annotated with textual entailment, paraphrases, contradiction, semantic similarity, and specificity in parallel. As previously mentioned, specificity is a relatively newly researched dimension and has not been seen as a relation dimension. The links between the other dimensions and specificity is separately discussed in Chapter 6. Furthermore, the operationalization of specificity also discussed in this chapter.

Furthermore, we pioneered in the research of links between all of these dimensions. Based on the corpus developed in Gold et al. (2019), we took a closer look at the relations and their compositionality in Kovatchev et al. (2020). Both studies are presented in Chapter 5.

Sentiment in Statements We believe that the sentiment dimension might be helpful in tasks concerned with opinions, e.g. recommender systems. If a user is interested in a specific aspect of a product feature, statements containing this aspect and the sentiment towards it are useful. This filtering is shown in the first two steps in Figure 4. In Example 2, a user interested in the room might be more interested in seeing (2).

1 The breakfast was tasty.

2 The hotel room was comfortable.

EXAMPLE 2: Statements with different aspects

In Gold et al. (2018), we annotated and analyzed aspect-based sentiment of political debates, more specifically of the presidential debates between Hillary Clinton and Donald Trump.

We examined extreme statements of sentiment, namely hate speech as well. We were the first to examine non-linguistic factors affecting the manual semantic annotation of hate speech. In Benikova et al. (2017), we examined the influence of implicitness and explicitness on the perception of hate speech. In Wojatzki et al. (2018a), we examined the influence of group membership on the example of gender on the perception of hate speech. Contrary to Benikova et al. (2017) where we modeled sentiment as an individual dimension, in Wojatzki et al. (2018a), we regarded sentiment as a relation and not as an individual dimension, as was discussed previously. The sentiment dimension is discussed in Chapter 7. Hate Speech, as a vast topic on its own, is discussed in Chapter 8.

Publication Record

Representing statements

- **Darina Benikova and Torsten Zesch.** 2016. Bridging the gap between computable and expressive event representations in Social Media. In Proceedings of the Workshop on Uphill Battles in Language Processing: Scaling Early Achievements to Robust Methods. p. 6–10. Austin, TX, USA. `https://www.aclweb.org/anthology/W16-6002.pdf`
  Contributions: The research and the writing of the paper were entirely performed by Darina Gold (née Benikova) under the supervision of Torsten Zesch.
  Chapter:4

- **Darina Benikova and Torsten Zesch.** 2017. Same same, but different: Compositionality of paraphrase granularity levels. In Proceedings of the Recent Advances in Natural Language Processing (RANLP-2017). p. 90–96. Varna, Bulgaria. `http://acl-bg.org/proceedings/2017/RANLP%202017/pdf/RANLP014.pdf`
  Contributions: The research and the writing of the paper were entirely performed by Darina Gold under the supervision of Torsten Zesch.
  Chapter:4

- **Darina Gold and Torsten Zesch.** 2019. Divide and Extract – Disentangling Clause Splitting and Proposition Extraction. In Proceedings of the Recent Advances in Natural Language Processing (RANLP-2019). p. 399–408. Varna, Bulgaria. `https://www.aclweb.org/anthology/R19-1047/`
  Contributions: The research and the writing of the paper were entirely performed by Darina Gold under the supervision of Torsten Zesch.
  Chapter: 4

Relations between statements

- **Darina Gold, Venelin Kovatchev, and Torsten Zesch.** 2019. Annotating and analyzing the interactions between meaning relations. In Proceedings of the 13th Linguistic Annotation Workshop (LAW-2019). p. 26–36. Florence, Italy. https://sigann.github.io/LAW-XIII-2019/pdf/W19-4004.pdf
  Contributions: Darina Gold and Venelin Kovatchev contributed equally to this work under the supervision of Torsten Zesch.
  Chapters: 5, 6

- **Venelin Kovatchev, Darina Gold, and Torsten Zesch.** 2019. RELATIONS - Workshop on meaning relations between phrases and sentences. Gothenburg, Sweden. https://www.aclweb.org/anthology/W19-0800
  Contributions: The workshop was jointly organized by Venelin Kovatchev and Darina Gold under the supervision of Torsten Zesch.
  Chapter: 5

- **Venelin Kovatchev, Darina Gold, M. Antònia Martí, Maria Salamò, and Torsten Zesch.** 2020. Decomposing and Comparing Meaning Relations: Paraphrasing, Textual Entailment, Contradiction, and Specificity, In Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC-2020), p. 5782–5791. Marseilles, France. https://www.aclweb.org/anthology/2020.lrec-1.709.pdf
  Contributions: The annotation, including the iterative improvement of the guidelines, was performed jointly by Venelin Kovatchev and Darina Gold on the corpus developed in Gold et al. (2019). The analysis of the specificity dimension was mainly performed by Darina Gold. M. Antònia Martí, Maria Salamó, and Torsten Zesch supervised the study.
  Chapters: 5, 6

Sentiment in statements

- **Darina Benikova, Michael Wojatzki, and Torsten Zesch.** 2017. What does this imply? Examining the Impact of Implicitness on the Perception of Hate Speech. In Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology (GSCL-2017). p. 171–179. Berlin, Germany. https://link.springer.com/chapter/10.1007/978-3-319-73706-5_14
  Contributions: The concept idea and the guidelines for the manual paraphrasing process were developed by Darina Gold. The paraphrasing from implicit to explicit tweets, the distribution of the study and the analysis were jointly performed by Darina Gold and Michael Wojatzki. Darina Gold was the main contributor to the writing process of the paper. Torsten Zesch supervised this work.
  Chapter: 8

- **Darina Gold, Marie Bexte, and Torsten Zesch.** 2018. Corpus of Aspect-based Sentiment in Political Debates. In Proceedings of the Conference on Natural Language Processing (KONVENS). p. 89–99. Vienna, Austria. https://www.oeaw.ac.at/fileadmin/subsites/academiaecorpora/PDF/konvens18_11.pdf
  Contributions: The annotation schemata, the annotation, the analysis, and the automation experiment were jointly performed by Darina Gold and Marie Bexte. Darina Gold was the main contributor to the writing process of the paper. Torsten Zesch supervised this work.
  Chapter: 7

- **Michael Wojatzki, Tobias Horsmann, Darina Gold, and Torsten Zesch.** 2018. Do Women Perceive Hate Differently: Examining the Relationship Between Hate Speech, Gender, and Agreement Judgments, In Proceedings of the Conference on Natural Language Processing (KONVENS). p. 110–120. Vienna, Austria. https://www.oeaw.ac.at/fileadmin/subsites/academiaecorpora/PDF/konvens18_13.pdf
  Contributions: The idea of research group affiliation as an influencing factor of hate speech was developed by Darina Gold. The study, the experiment analysis, and the writing of the paper was performed by Michael Wojatzki, Tobias Horsmann, and Darina Gold, while Michael Wojatzki was the main contributor. Torsten Zesch supervised this work.
  Chapter: 8

Publications that did not make it into the thesis

- **Darina Benikova, Margot Mieskes, Christian M. Meyer and Iryna Gurevych.** 2016. Bridging the gap between extractive and abstractive summaries: Creation and evaluation of coherent extracts from heterogeneous sources. In Proceedings of International Conference on Computational Linguistics (Coling), p. 1039–1050. Osaka, Japan. https://www.aclweb.org/anthology/C/C16/C16-1099.pdf
  Contributions: The concept and the actual implementation of the summary creation process, were performed by Darina Gold. The analysis was performed by Darina Gold and Margot Mieskes, while Darina Gold was the main contributor. Darina Gold was the main contributor to the writing process of the paper. Margot Mieskes and Iryna Gurevych supervised this work.

- **Christian M. Meyer, Darina Benikova, Margot Mieskes and Iryna Gurevych.** 2016. MDSWriter: Annotation tool for creating high-quality multi-document summarization corpora. In Proceedings of Association for Computational Linguistics (ACL), p. 97–102. Berlin, Germany. https://www.aclweb.org/anthology/P/P16/P16-4.pdf#page=109
  Contributions: Darina Gold described the exemplary application of the tool in the paper. Margot Mieskes and Iryna Gurevych supervised this work.

Chapter 1

Manual Semantic Annotation of Statements

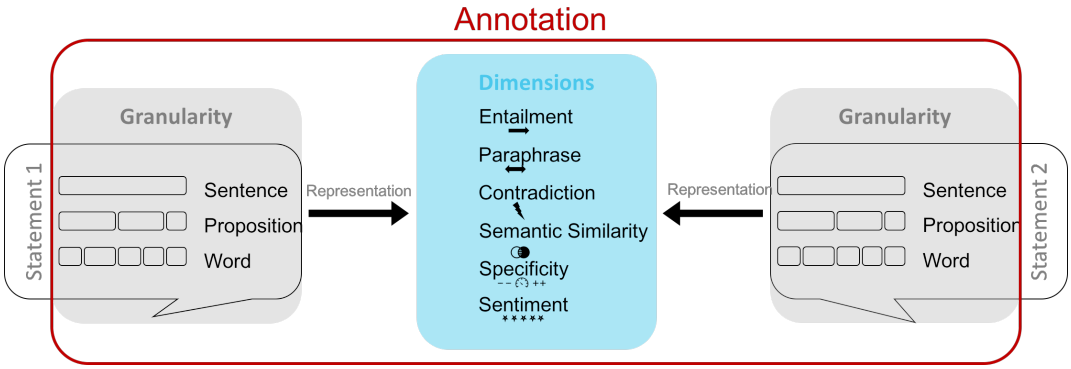


FIGURE 1.1: Illustration of importance of annotation in this thesis

Annotations are needed for any kind of corpus linguistic analysis or automation of linguistic features. In psychology (Guetzkow, 1950), but also in early days of Computational Linguistics (CL) (Krippendorff, 1995), the process of *annotation* was called *coding*. In some cases, the process is also called *rating* (Tinsley and Weiss, 1975).

According to Ide and Pustejovsky (2017), linguistic annotation developed from being used in small manual CL studies for testing or developing linguistic theories to also being used for automatic data increasingly available in ever-growing quantities in Natural Language Processing (NLP).

Bird and Liberman (2001) define *annotation* in the following way:

> 'Linguistic annotation' covers any descriptive or analytic notations applied to raw language data. The basic data may be in the form of time functions — audio, video and/or physiological recordings — or it may be textual. The added notations may include transcriptions of all sorts (from phonetic features to discourse structures), part-of-speech and sense tagging, syntactic analysis, 'named entity' identification, co-reference annotation, and so on.
> (p.23)

In this thesis, we solely focus on textual annotation on the semantic level. Furthermore, in this thesis, by annotation we mean manual annotation, i.e. human effort is involved. The

humans performing the annotation will be called *annotators* in the further course. In the illustrative example in Figure 1.1, we show all the levels annotated in this thesis. The annotation levels are the dimensions, which are also relations between two statements. Although we did not perform research on the topic of annotation on its own, it was vital to this thesis, as illustrated in Figure 1.1. In the course of our research, we found that our questions could not be answered with existing corpora. As no suitable annotations were present, we created several corpora according to the standards and measurements, which are discussed in this chapter.

There are many parameters playing into the annotation procedure and for each NLP task there are decisions to be made in order to fit the annotation procedure to the task at hand. However, there is a common procedure that is followed in each task, which is discussed in detail in Section 1.1. Basically, one or several annotators attach additional information to text according to given guidelines. Out of these annotations, final labels, referred to as *gold standard*, are formed. The quality of the annotation is measured in Inter-Annotator Agreement (IAA).

Depending on the task, different measurement scales and methods can be applied in the annotation process—e.g. in Part-of-Speech (POS) tagging, predefined classes are attached to individual words, whereas for other tasks, e.g. semantic similarity, two texts are compared and given a rating on a scale. Section 1.2 explains the different measurements in more detail.

Also depending on the task, there are many tools that can be used for annotation, which are reviewed in Section 1.3.

To evaluate the quality of an annotation, but also to determine the upper bound for an automatic annotation, the IAA is determined. This is done using IAA measures, which are discussed in more detail in Section 1.4 and Section 1.5.

1.1 Annotation Procedure

In every annotation task, there is the prior procedure of formalizing the phenomenon of interest in an operational procedure that has the aim to be reproducible. In order to be reproducible and consistent, guidelines describing the task and the annotation procedure are developed. Prior to the actual annotation, oftentimes an automatic pre-processing step is performed. In this step, unitizing tasks e.g. such as word or sentence segmentation, may be executed. A pre-processing step mostly reduces the annotation effort and also makes the annotation procedure more interesting, as it spares repetitive and easy tasks.

The annotation procedure can also differ according to how many people perform the annotation per item and what their background is. Depending on the study, the annotation procedure can be comparable to a psychological survey.[1]

In most cases, prior to the actual annotation, a pilot study is conducted in order to make a proof-of-concept for each step of the annotation. As it is mostly the case with pilot studies, their result helps to improve the setting of the actual study.

[1]Especially for phenomena that are not purely linguistic, the background of the annotator has an impact on the annotation. We examined the link between gender and gender-related hate speech in Wojatzki et al. (2018a), which is discussed in Chapter 7.

The annotation process itself consists of two tasks are often not treated separately—namely *unitizing*, and *coding* (also called *categorizing* or *classification*) (Guetzkow, 1950; Krippendorff, 1995). In the actual annotation process, unitizing and classification are mostly performed in one step, as it is difficult and time consuming to do them separately.

Evaluation, however, is performed separately, if the unitizing is not already given. An example of both steps for two annotators is shown in Figure 1.2 for two dimension annotations—entailment and paraphrase.

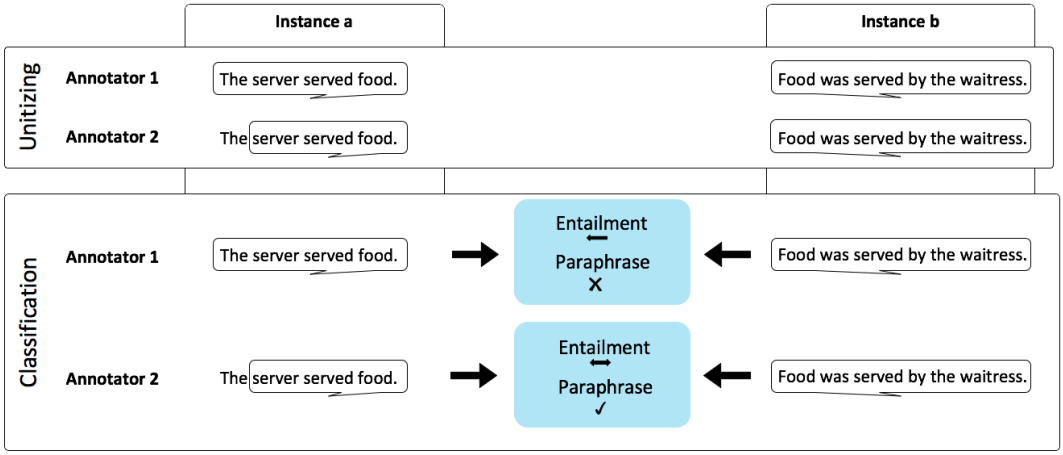


FIGURE 1.2: Examples of unitizing and classification annotation

Unitizing is the process of choosing the units that are to be annotated (Guetzkow, 1950; Krippendorff, 1995), e.g. in the case of POS, the unit are words. In some cases, as e.g. in finding spans in sentences that are paraphrases of each other, it is not that trivial (Vila et al., 2014; Kovatchev et al., 2018b). Figure 1.2 shows an example, where the annotators disagree on the unitizing - Annotator 1 considers the article in *Instance a* as part of the unit, whereas Annotator 2 does not. In *Instance b* the annotators agree.

Classification on the other hand is the process of assigning labels to the units chosen in the prior unitizing step (Krippendorff, 1995). In Figure 1.2, the annotators also disagree on the classification label - Annotator 1 regards *Instance a* to entail *Instance b*, while Annotator 2 annotated the instances to be in a bi-directional entailment and a paraphrase relation.

1.1.1 Guidelines

Guidelines are instructions on the annotations that are given to the annotators. Mostly the given rules are explained with the help of examples and counter examples.

The form of the guidelines has to suit the form of annotation and the kind of annotators. Two examples are the following: In the case of expert annotations, they are mostly developed in an iterative manner. In the case of crowdsourcing, the guidelines have to be very short and easily understandable to laypersons. All guidelines created in this thesis are publicly available[2] and are attached in the appendix (see Appendix A.1).

[2] https://github.com/MeDarina

1.1.2 Annotation Methods

Annotation methods differ in terms of the number and experience of the annotators. Those methods are not exclusive, but have individual characteristics.

Expert Annotation is a form where mostly linguists or computer linguists perform the annotation, using their knowledge of this field. In most cases, only few annotators are used in this form due to personnel costs. We used expert annotations in Benikova et al. (2016); Benikova and Zesch (2016, 2017); Gold et al. (2018); Gold and Zesch (2019) and Kovatchev et al. (2020).

Crowdsourcing is a form of annotation in which the phenomenon of "the wisdom of the crowd" is used. This means that many people, who are not experts in the given field, are asked to annotate. In contrast to the expert annotation, the guidelines have to be short and easy to understand to laypersons. For quality control or rather to prevent misuse of the annotation, qualification tests or so called "gold nuggets" are used. Gold nuggets are annotations where the correct answer is already known and very obvious, hence if annotators make several of these wrong, their answers are not used. Crowdsourcing is widely used in the fields of CL and NLP. There have been many descriptions on its implementation Munro et al. (2010); Kneißl (2014); Ide and Pustejovsky (2017) and tools or plug-ins Yimam et al. (2013); Bontcheva et al. (2014); Leemann et al. (2016). We used crowdsourcing in Gold et al. (2019) and Gold and Zesch (2019).[3]

Surveys are mostly used for annotations where not linguistic knowledge, but rather opinions or feelings of the annotators are needed, e.g. in our work this is the case for hate speech or sentiment. Hence, in surveys rarely expert annotators are used. Similar to crowdsourcing, many annotators are needed. We used surveys in Benikova et al. (2017) and Wojatzki et al. (2018a).

1.2 Measurements and Measurement Methods

Annotations can be performed using different kinds of measurements and measurement methods. By measurements, we mean the classes, also called labels or tags that are used for the annotation, e.g. in case of POS, possible tags are "noun" or "verb", in case of sentiment possible tags are "positive" or "negative". The whole set of all tags is called "tag set". Tags can also be annotated as a relation between two instances, e.g. one being "more negative" than the other. In all our classification annotations, we used pre-defined tag sets, meaning that we provided a list of possible tags. Furthermore, annotations can be performed on a scale, e.g. in the case of sentiment the annotation could be on how positive or how negative the annotated statement is.

[3] In Gold and Zesch (2019), we used crowdsourcing in the preliminary step and expert annotation in the main step.

Statement Pair	Textual Similarity Rating				
	0 no similarity	1 little similarity	2 some similarity	3 similar	4 textual equality
The service was good. The service was great.	☐	☐	☐	☐	☒
The service was good. The garden has flowers.	☒	☐	☐	☐	☐

TABLE 1.1: Exemplary rating scale choice

1.2.1 Simple Annotation Methods

By a *simple annotation method* we refer to studies where tags are assigned to one textual unit, as is the case in many classic annotation tasks, e.g. POS, Named Entitity (NE), or binary sentiment annotation. There are also more complex methods, where the units stand in relation towards each other, e.g. dependency or frame annotation.

1.2.2 Scaled Annotation Methods

Rating scales are a widely used method for the annotation of quantitative or qualitative data in many fields, including social sciences and NLP. A rating scale gives the annotator the choice of categorical or numeric labels that represent a characteristic that is to be measured on a given data set. In the case of NLP, an annotator might be asked to measure the similarity between two words on a scale of 1 to 5, with 1 representing *no similarity* and 5 representing *textual equality*.

Two exemplary pairs are given in Table 1.1. The first pair has nearly the same content, which is why it ranks high on the textual similarity rating. The second pair, on the contrary, shares no common information, which is why it ranks low on the textual similarity rating.

The Likert scale (Likert, 1932) is an exemplary rating scale. Proper Likert scales are both *bipolar*, *symmetric*, and *balanced*. Bipolar means that the scale has both positive and negative options. Symmetric means that there are as many negative as positive options. Balanced means that the distance between each candidate value is the same. Mostly, Likert scales do not present simple numbers, but have a verbal equivalent for each position. The example presented in Table 1.1 is a proper Likert scale.

Although rating scales are widely used, they suffer from several limitations (Schuman and Presser, 1996; Baumgartner and Steenkamp, 2001; Kiritchenko and Mohammad, 2017) that Kiritchenko and Mohammad (2017) summarize as *inconsistencies throughout annotators, inconsistencies within one annotator, scale region bias*, and *fixed granularity*. While the first two are self-explanatory, we will shortly explain the last two. A scale region bias describes the phenomena of an annotator being biased towards annotations in one specific region e.g. in the example of textual similarity, an annotator could tend to see no similarity between statement pairs. The issue of a fixed granularity mergers when an annotator feels restricted by the given scale and would like to place an annotation between two given labels e.g. in the

example of textual similarity, an annotator would like to annotate the similarity between 2 and 3.

Scales can be used in annotation of both individual statements e.g. the rating of sentiment or specificity, as well as in comparisons of two or more statements as shown in the example of textual similarity. In this thesis, we used a scale for comparatively rating textual similarity in Gold et al. (2019)[4].

1.2.3 Comparative Annotation Methods

By using comparative methods for annotation, it is tried to reduce the limitations discussed in Section 1.2.2 (Kiritchenko and Mohammad, 2017), as it is easier to annotate an object in comparison to another object on a given characteristic, such as *similarity*, *specificity* or a given emotion than attaching a scale label to an object.

Paired Comparisons is a simple comparison method in which annotators are presented with object pairs and asked to annotate which of them corresponds more to a given characteristic (Thurstone, 1927; David, 1963). The method is used in social, psychological, and political studies of preferences, attitudes, voting systems, social choice, and public choice, as well as in computer science for requirements engineering, and multi-agent artificial information systems (Ramík, 2020). Table 1.2 shows an example of a paired comparison annotation for specificity, in which the second statement is annotated as more specific than the first.

Statement	Choose the **more specific** statement
The hotel was good.	☐
The service in this hotel was attentive.	☒

TABLE 1.2: Exemplary paired comparisons choice

The annotations can then be converted to real-valued scores and rankings indicating the degree to which the given item is associated with the characteristic. However, in order to get these values and rankings, a large number of annotations—namely N^2, where N is the number of objects, is needed.

We use paired comparisons in Gold et al. (2019).

Best-worst scaling (BWS) was first proposed by Louviere and Gaeth (1987) as a more reliable method to identify extreme options Louviere et al. (2015). It was first used in marketing research under the name of *maximum difference scaling (maxdiff)*. The main idea behind the method is that annotators identify the *best* or *worst* option[5] amongst a set of at least three options. Table 1.3 shows an exemplary best-worst scaling (BWS) choice in the case of specificity, as described in Chapter 6.1.3.2.

[4] Although we annotate textual similarity by comparing two statements, it is not a comparative annotation as discussed in the next section. To do so, we would have had to compare several statement pairs to e.g. say which pair is textually most similar. We will discuss this in more detail in Chapter 5.

[5] The best and the worst is to be seen metaphorically and define two extremes of a subjective continuum, e.g. *tallest* and *shortest* could be the two extremes for a set of at least three persons.

Choose the **most specific** statement	Statement	Choose the **least specific** statement
☐	The hotel was good.	☒
☐	The service in this hotel was wonderful.	☐
☒	The service in this hotel was attentive.	☐
☐	Service was good.	☐

TABLE 1.3: Exemplary best-worst scaling choice

There are three different areas of BWS theory—*the object case*, *the profile case*, and *the multi-profile case* that differ in their complexity. In this thesis, only the first case, which is the "classic" case, is relevant. It requires a list of objects organized in subsets of at least three and a sample of annotators who select the best and the worst fitting choice in each subset. In the example in Table 1.3, the listed objects are *statements* and the best fitting choice is *most specific*, whereas the least fitting choice is *least specific*.

The theoretical framework behind BWS is the random utility theory. It assumes that people make errors, but that when choosing repeatedly, the frequency of choice give an indication of how strong they value the given objects. Hence, the choice how often object A is picked over object B indicates how much A is preferred over B. (Louviere et al., 2015)

In our example, it means the number of times *The service in this hotel was attentive.* is chosen over other objects in the list indicates how specific it is in comparison to them.

In NLP, BWS has been used for annotating relational similarity (Jurgens et al., 2012), word-sense disambiguation (Jurgens, 2013), word-sentiment intensity (Kiritchenko and Mohammad, 2016), emotion intensity (Mohammad and Bravo-Marquez, 2017), and the degree of support or opposition of statements (Wojatzki et al., 2018a,b). Typically, the set of objects in this discipline is 4. Working with 4-tuples is efficient, because the result of answering the two questions are five out of six item-item pair-wise comparisons. Each best-worst annotation consists of only two decisions, which is the most and least fitting utterances, compared to making a binary decision between each pair that could be created from a 4-tuple, i.e. for the statements A, B, C, and D, if A is selected as best, and D is selected as worst, then we know that A $>$ B, A $>$ C, A $>$ D, B $>$ D, and C $>$ D. Using this logic, not all possible permutations need to be annotated. Typically the number of quadruples is $1.5 * the\ number\ of\ statements$. All items can be efficiently organized in m 4-tuples e.g. by using the script provided by Kiritchenko and Mohammad (2016). This script ensures that the created tuples satisfy the following constraints:

- each 4-tuple occurs only once
- each statement occurs only once within a tuple (no duplicates)
- each statement appears approximately in the same number as tuples as other statements

Then, real-valued scores for each of the objects can be computed using a counting procedure by Orme (2009):

$$fittingscore(a) = \%\textit{most fitting}(a) - \%\textit{least fitting}(a) \tag{1.1}$$

Consequently, the score ranges from -1(least fitting) to 1 (most fitting).

The annotation reliability is usually calculated using the average Split-Half Reliability (SHR) over 100 trials (Kiritchenko and Mohammad, 2017). To do so, all tuple annotations are randomly split in two halves. From the two halves, scores are produced. The correlation of the two sets of scores is calculated.

Kiritchenko and Mohammad (2017) empirically show that BWS produces high-quality annotations using 1.5-2 times the number of objects in the set. In their study, Kiritchenko and Mohammad (2017) directly compared BWS against the rating scale method and showed that it produced significantly more reliable results.

We use BWS in Wojatzki et al. (2018a) (described in Section 8.2) for annotating hate speech and in an unpublished study for annotating specificity (described in Section 6.1.3.2). For both studies, we use the script by Kiritchenko and Mohammad (2016) for 4-tuple generation.

1.3 Annotation Tools

There exists a number of annotation tools and more are constantly created. They differ in many factors, e.g. being *web-based* or *off-line*, *collaborative* or *unilateral*, *multi-purpose* or *specific*. For instance, there are very simple off-line tools such as Excel, in which annotators can individually annotate their statements in a column dedicated to this cause. There are also complex online tools offering a simultaneous annotation and views that compare the annotations of different annotators such as WebAnno (Yimam et al., 2013). According to Biemann et al. (2017), web-based tools have the advantages of having a *lower training effort* (as they can be employed using a basic web-browser), the potential to *unlock a larger workforce*, and a *distributed annotation* (as annotators can work independently of each other).

Except for Excel, the tools used in this thesis are web-based and collaborative. Thus, and also because mostly annotation tools differ in their purpose and the annotation tasks they are able to serve, in this section we will only distinguish between *multi-purpose* and *specific* tools.

1.3.1 Multi-purpose tools

There are many tools for annotating within one document[6], such as WebAnno (Yimam et al., 2013), Anafora (Chen and Styler, 2013), CSNIPER (Eckart de Castilho and Gurevych, 2014), and UAM CorpusTool (O'Donnell, 2008). They often offer the possibility of curation, meaning the comparison of different annotations on the same data and their correction, as well as the calculation of IAA. According to Biemann et al. (2017), multi-purpose tools have the benefit of *enhanced flexibility* in terms of annotation layers and often offer an *all-in-one solution*, as infrastructures for e.g. annotator management, agreement computation, and project workflows can be re-used. In this thesis, we used WebAnno in Benikova and Zesch (2016, 2017) and Gold et al. (2018).

[6]This means that annotations on multiple layers can be performed within one document. However, it makes comparative or alignment annotations difficult, as these tools are mostly not built for this purpose.

1.3.2 Specific tools

Specific tools solve one or a very limited amount of tasks, e.g. *MMAX2* (Müller and Strube, 2006), which enables relation annotation such as co-reference within one document, *WordFreak* (Morton and LaCivita, 2003), which enables several annotation types (e.g. span and constituency annotation) within one tool, or *NITE XML toolkit* (Carletta et al., 2003), which enables video and transcription annotation. Biemann et al. (2017) assume that many specific tools result from "difficulties of adequately modeling the annotation data and implementing a sophisticated user interface on top of the data model"[p.232]. The selection of specific tools presented in more detail here will be limited to those that were used in this thesis. In this thesis, we need to annotate relations between two statements and parts of statements. This is possible, but very inconvenient in multi-purpose tools such as WebAnno[7], as e.g. the display of pairs, as well as the annotation on the sentence layer is not a designated feature. For performing the relation annotation as we need it, cross-document tools might be more fitting, as they are built to display several documents, or in our case statements, simultaneously. However, tools for cross-document annotation tasks are mostly limited to event and entity co-reference, e.g. CROMER (Girardi et al., 2014) or Callisto/EDNA (Day et al., 2008).

MDSWriter In one of the works not presented in this thesis, we developed a tool for creating manual extractive summaries, namely MDSWriter (Meyer et al., 2016). MDSWriter enables annotators to select, compare, and tag several texts simultaneously. It was developed for several different annotation tasks e.g. paraphrase detection and the selection of the best fitting statement out of bundle of paraphrases.

One task was to find text pieces with similar content, which basically is the task of paraphrase identification. Another task was to select the best fitting text piece for a summary given the similar text pieces. This is similar to selecting the most or least specific statement and thus could also be used for this task.

WARP-Text This tool (Kovatchev et al., 2018b) was developed to perform alignment annotations, such as semantic relations including paraphrases, entailment, contradiction, between two text pieces. WARP-Text supports multi-layer annotation and annotation on different granularity levels. We used it for the annotation in Gold et al. (2019) and Kovatchev et al. (2020).

1.4 Evaluation of Classifying Annotation

To show that an annotation is reliable and reproducible, an evaluation is performed using IAA measures.

Artstein and Poesio (2008) performed a survey of common IAA measures that are still used in state-of-the-art research. We will use their distinction between agreement measures—1) not chance-corrected agreement, 2) agreement between two annotators, and 3) agreement

[7] We performed a pairwise paraphrase annotation in Benikova and Zesch (2017).

between more than two annotators. Overall, this section strongly relies on and reflects the findings in Artstein and Poesio (2008).

1.4.1 Notation of Inter-Annotator Agreement

We will use Artstein and Poesio (2008)'s notation to explain the individual methods:

- The set of items, meaning units to which a category is assigned, is $\{i|i \in I\}$
- The set of categories, meaning tags that are assigned to items, is $\{k|k \in K\}$
- The set of annotators is $\{c|c \in C\}$
- A_o is observed agreement
- D_o is observed disagreement
- A_e is expected agreement
- D_e is expected disagreement
- $P(\cdot)$ is reserved for the probability of a variable, and $\hat{P}(\cdot)$ is an estimate of such probability from observed data.
- n with a subscript indicates the number of judgments of a given type:
  - n_k is the total number of items assigned by all annotators to category k
  - n_{ik} is the number of annotators who assigned item i to category k
  - n_{ck} is the number of items assigned by annotator c to category k

1.4.2 Agreement Without Chance Correction

The most simple measure to calculate agreement is %-agreement, also called *observed agreement*. It is identical to *accuracy*, which is the term used for evaluating automatic annotation. Scott (1955) defines it as

> "the percentage of judgments on which the two analysts agree when coding the same data independently.
> (Scott, 1955, p.23)

This is the number of items on which the annotators agree divided by the total number of items.

$$A_o = \frac{number\ of\ items\ the\ annotators\ agree\ on}{all\ items} \tag{1.2}$$

However, this measure does not correct for *chance agreement*, which is the probability of annotators agreeing by chance. In order to circumvent this issue, agreement measures using *chance correction* (taking chance agreement into account) are used.

1.4.3 Agreement between two Annotators with Chance Correction

The three best-known coefficients, S (Bennett et al., 1954), π (Scott, 1955), and κ (Cohen, 1960), and their generalizations, use the idea to consider expected agreement between annotators. S, π, and κ use Equation 1.3:

$$S, \pi, \kappa = \frac{A_o - A_e}{1 - A_e} \tag{1.3}$$

The agreement values range between -1 and 1, where 0 signifies chance agreement.

Observed agreement (see Equation 1.2) has the same value for all coefficients.

However, the notion of expected agreement varies. S, π, and κ assume the independence of the two annotators, meaning that the chance of two annotators agreeing on any given category k is the same. The difference lies within the assumption of the calculation of the chance of an annotator assigning an arbitrary item to a category (Artstein and Poesio, 2008).

Bennet's S is based on the assumption that if annotators were operating by chance alone, we would get a uniform distribution: i.e., for any two annotators c_m, c_n and any two categories k_j, k_l, $P(k_j|cm) = P(k_l|c_n)$, meaning that all categories are equally likely for this coefficient.

Scott's π is based on the assumption that if annotators were operating by chance alone, we would get the same distribution for each annotator, meaning for any two annotators c_m, c_n and any category k, $P(k|c_m) = P(k|c_n)$. That is, the random assignment of categories to items, by any annotator, is governed by the distribution of items among categories in the actual world.

Cohen's κ is based on the assumption that if annotators were operating by chance alone, we would get a separate distribution for each annotator. This means that this coefficient assumes that random assignment of categories to items is governed by prior distributions that are unique to each annotator, and which reflect individual annotator bias. An individual annotator's prior distribution is estimated by looking at her actual distribution.

The most common method to measure IAA agreement between two annotators is Cohen's κ. We use it in Benikova and Zesch (2017); Gold et al. (2018) and Gold et al. (2019).

1.4.4 Agreement between more than two Annotators

Sometimes agreement for more than two annotators is reported through each pair of two annotators or the range between all pairs of annotators. However, there are also generalized versions of the coefficients. A generalization of Scott's π is proposed by Fleiss (1971) and a generalization of Cohen's κ is proposed by Davies and Fleiss (1982). There are called multi-π and multi-κ by (Artstein and Poesio, 2008). In the studies presented in this thesis, we did not use the generalized versions of the coefficients, but presented the agreement between individual pairs of annotators, as is done in most other studies of this kind.

1.5 Evaluation of Unitizing Annotation

As previously noted, unitizing is the choosing of units that are to be annotated. Although it is often assumed that units of annotation are given, this is not the case when e.g. labeling

syntactic constituents as in parsing or chunking (Artstein and Poesio, 2008). If unitizing is the main task, it needs to be evaluated individually.

Krippendorff's α According to Artstein and Poesio (2008), there is only one coefficient for evaluating unitizing tasks—α_u, which is one variant of Krippendorff's α (Krippendorff, 1980, 2018). However, as Artstein and Poesio (2008) further point out, although broadly used in content analysis Mayring (2010), it is very complex and it is not used in this thesis. Hence, it will not be further explained here.

IAPTA-TPO In Kovatchev et al. (2020), we use two different versions of the IAA for Paraphrase Type Annotation-Total/Partial overlapping (IAPTA-TPO) coefficients, which were proposed by Vila et al. (2014) and refined by Kovatchev et al. (2018a). This coefficients were proposed to measure IAA of different paraphrase types and measures the agreement of both unitizing and classifying. The two IAPTA-TPO measures are *total*, measuring full agreement on both unitizing and classifying, and *partial*, measuring full agreement on classification and only partial agreement on unitizing.

It is an agreement between two annotators calculated through Precision, Recall and F1, assuming one annotator as gold standard. We use this measure in Kovatchev et al. (2020).

1.6 Summary

In this chapter, we explained the annotation foundation for all studies presented in this thesis. Using this basis, we operationalize the annotation of the different dimensions, enabling us to empirically answer our research questions. Throughout our studies, the basic procedure stays the same: we iteratively develop guidelines, which are then followed using different annotation methods—we make use of expert annotations, crowdsourcing, and surveys. The annotation are performed using the best-fitting tools, which may be either multi-purpose tools, or in the case of very specific needs—specific tools. Furthermore, we make use of different measurements and measurement methods in the operationalization. In this thesis, we focus on the advantages of comparative methods, which we use for all our dimensions, over scaled methods. Moreover, we experiment with different operationalizations of comparative methods on the example of specificity, which is discussed in detail in Chapter 6. To evaluate the operationalization and reproducibility of the study, we perform an evaluation in form of IAA using the best-fitting measure. Throughout this thesis, we will reference the terms, methods, measures, and tools described in this chapter.

In the next chapter, we will discuss how the results of the manual annotation can be automatized.

Chapter 2

Machine Learning

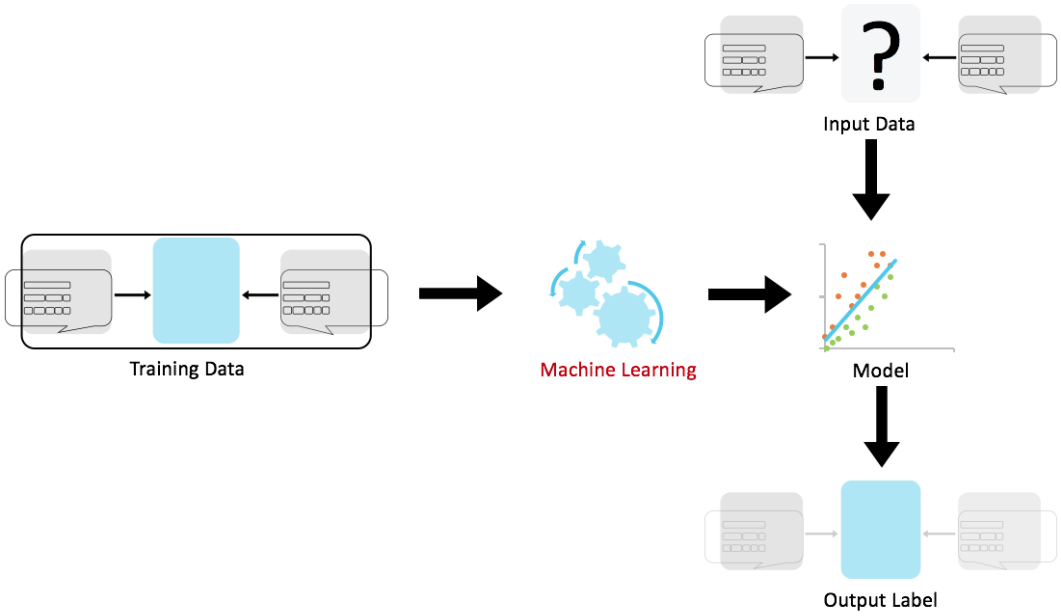


FIGURE 2.1: Illustration of how machine learning is used in this thesis

"Machine learning is the systematic study of algorithms and systems that improve their knowledge or performance with experience." (Flach, 2012, p.3). Using *features*, relevant domain objects are described to solve a *task*, which is an abstract representation of a problem regarding these domain objects. These tasks can be represented as *models* by mapping data-points from the objects to outputs (Flach, 2012).

The basic idea of machine learning is illustrated in Figure 2.1. The machine learning algorithm is gets *training data* as input and outputs a *model*, which given unseen data of a similar form of the input is able to predict a class or score. Basically, a machine learning algorithm basically is a function $y(x)$ that gets an input x and generates an output vector y, which is encoded in a pre-defined way (Bishop, 2006). The form of the function is determined during the *training phase*, based on the input *training data* that the algorithms uses (Bishop, 2006).

Machine learning algorithms can be divided into *supervised* and *unsupervised*. The distinction will be discussed in the following. In this chapter, we will focus on machine learning as it is used in Computational Linguistics (CL) or Natural Language Processing (NLP).

Supervised learning algorithms are applications in which the training data contains pairs of input and corresponding output vectors (Bishop, 2006). In CL or NLP, the desired outputs are given in form of manual semantic annotations, as described in the previous chapter. Basically, supervised learning algorithms output automated annotations. An algorithm could be given a great amount of any of the given examples for one task and would theoretically learn how to solve similar tasks automatically. For instance, given many paired comparison annotations for specificity, an algorithm would be able to state which sentence is more specific given a sentence pair. Although creating datasets of inputs and especially the desired outputs are a time- and work intensive process, supervised learning algorithms are easier to understand and to evaluate in comparison to unsupervised algorithms.

Unsupervised learning algorithms are applications which are given training data without corresponding output values Bishop (2006). On the one hand, that saves the laborious annotation process. On the other hand, these algorithms are usually harder to evaluate than supervised ones. (Müller, 2016) As this thesis focuses on the human understandable dimensions, as well as their operationalization and creation process, unsupervised learning algorithms, will not be further discussed in this thesis. However, they are an effective, powerful, valid, and state-of-the-art tool that is vastly used in the fields of CL and NLP. (Müller, 2016)

The goal of a machine learning algorithms is to predict correct outputs for unseen inputs, which is mostly evaluated by using a so-called *test set*. Evaluation methods and metrics be further discussed in Section 2.2.

2.1 Supervised learning for semantic relation dimensions

Supervised learning is used in a setting where the algorithm, given example pairs containing an input and the desired output, builds a model that is supposed to predict a certain output vectors from a given input vector (Bishop, 2006). The model may additionally get analytically derived *features*, which are potentially helpful for the learning process as input (see Section 2.1.1).

Furthermore, supervised machine learning issues are mainly divided in two types—*classification* (see Section 2.1.2) and *regression* (see Section 2.1.3)—depending on the desired output of the model. In classification, there are pre-defined classes, while in regression the output is a real numbered value.

2.1.1 Features

Mostly, the input that the model learns from is not just raw data, but also analytically derived features. According to Flach (2012), "a feature can be thought of as a kind of measurement that can be easily performed on any instance. Mathematically, they are functions that map from the instance space to some set of feature values called the *domain* of the feature." [p.38,39]. Simple and commonly used types of features are numeric or binary features, as well as other finite sets (Flach, 2012), e.g. Part-of-Speech (POS) labels, aspects for aspect-based sentiment, or dependency parse labels. For instance, in our study on the links between the various dimensions, we used all other dimensions as binary features to predict one of

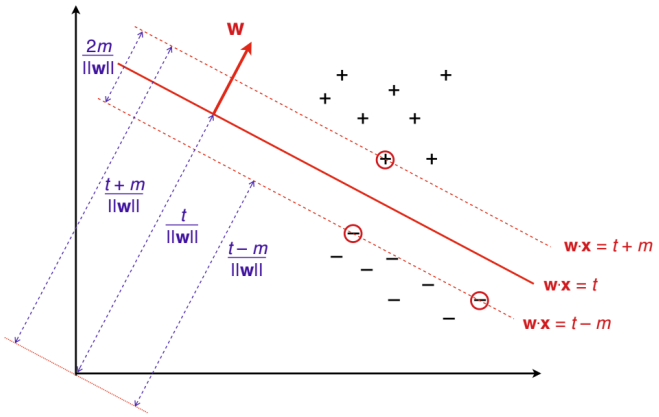
them. This means, that the model was given the information whether each of the other relation dimensions was present in the given statement pair and was supposed to learn the desired dimension from that (c.f. Section 5.1). In our study on aspect-based sentiment in political texts, we also equipped a supervised model with binary features in the form of a lookup whether words in the statements were contained in lists of positive or negative terms (c.f. Section 7.1). In feature engineering, the interaction or correlation between features has to be considered (Flach, 2012), e.g. in the case of our study on the links between relations the dimensions of *paraphrase* and *entailment* are strongly correlated, thus when predicting another dimension, these features might be amplified, while the effect of other features decreases. It depends on the task—or in this case which dimension is to be predicted—whether this is an issue, e.g. when predicting *contradiction*, it might be helpful, as both features strongly negatively correlate with this dimension.

2.1.2 Classification

In *classification*, a class label from a predefined list is predicted. In this thesis, this is the case for *entailment*, *paraphrase*, *specificity*, and *sentiment*. Furthermore, classification can be subdivided in *binary* and *multiclass classification*. In our study examining the links between various relation dimension, we use the therein annotated dataset in a supervised binary classification setting (c.f. Section 5.1)—which means that the classifier predicts whether the given relation exists between the input statement pair. In our study on political sentiment, the *sentiment* is predicted in a multiclass setting. In this case, the prediction is actually trinary: POSITIVE, NEGATIVE, and NEUTRAL (c.f. Section 7.1). In the same study, we also predict the aspect on which the sentiment is based on. Although this prediction may look like a multiclass classification, we actually also performed a binary classification for each aspect—meaning that a statement could be predicted having more than one aspect. This is an example of a *multilabel*, but not *multiclass classification*, as the aspect classification was not mutually exclusive. Popular supervised machine learning classifiers are *k-Nearest Neighbors*, *Naive Bayes*, *Decision Trees*, *Ensembles of Decision Trees*, and *kernelized Support Vector Machines*. (Müller, 2016) In this thesis, we made use of various forms of the latter classifier—Support Vector Machines (SVMs).

SVM This supervised learning method has been developed and gradually refined by Vapnik and collegues (Boser et al., 1992; Cortes and Vapnik, 1995; Drucker et al., 1997) based on Vapnik and Chervonenkis (1981). In this thesis, SVMs refers to kernelized SVMs for classification, also known as *support vector classification* (as opposed to *support vector regression*). The simple version of the linear SVM creates a hyperplane in two-dimensional space to perform a binary classification. (Flach, 2012)

In general, there is an infinite number of possible hyperplanes, which are decision boundaries, that separate the two classes. Intuitively, some are better suited than others, e.g. in Figure 2.2, the red dotted hyperplanes are very close to the cluster of one of the classes (depicted by $+$ for the *positive* class and $-$ for the *negative* class), which would define the



The geometry of a support vector classifier. The circled data points are the support vectors, which are the training examples nearest to the decision boundary. The support vector machine finds the decision boundary that maximises the margin $m/||\mathbf{w}||$.

FIGURE 2.2: Illustration and caption of maximum-margin hyperplane (Flach, 2012, p.212)

decision space between the classes in favor of one class. The distance from the nearest training examples of each class and the desired hyperplane should be maximized. The examples nearest to the decision boundary are called *support vectors*. The distance between the *support vectors* and the decision boundary is a *margin*. Hence, the task is to find a *maximum-margin hyperplane*. In a SVM, this *maximum-margin hyperplane* is defined as a linear combination of the *support vectors*. The *margin* is defined as $m/||w||$, m being the distance between the decision boundary and the support vectors (at least one of each class), as measured along w, a weight vector. Customarily, $m = 1$ is chosen. Maximizing the margin then means minimizing $||w||$, or rather $\frac{1}{2} * ||w||^2$, given that no training examples are inside the margin. (Flach, 2012)

In this way, the problem describes a quadratic optimization problem with the constraint of the training examples falling outside the margin, t being the decision threshold, which is equal for both classes (Flach, 2012):

$$w^*, t^* = \underset{w,t}{argmin} \frac{1}{2}||w||^2 \ \ subject\ \ to\ y_i(w * x_i - t) \geq 1, 1 \leq i \leq n \tag{2.1}$$

This margin is called a *hard-margin* Usually, this kind of quadric optimization problem is solved using the method of *Lagrange* multipliers (Flach, 2012), which we will not be further discuss herein.

However, real-world data is not necessarily, which does not satisfy the constraints discussed in Equation 2.1. To solve this issue, a so-called *slack-variables* ξ_i is introduced—one for each example. This allows some of the examples to lie within the margin or even at the other side of th decision boundary. These are called *margin errors*. This leads to the following *soft-margin* optimization problem (Flach, 2012):

$$w^*, t^*, \xi_i^* = \underset{w,t,\xi_i}{argmin} \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i$$
$$\textit{subject to } y_i(w * x_i - t) \geq 1 - \xi_i \textit{ and } \xi_i \leq 0, 1 \leq i \geq n \tag{2.2}$$

According to Flach (2012), "C is a user-defined parameter trading of margin maximisation against slack variable minimisation" [p.217]. To solve this optimization problem, again, the *Lagrange* method can be used.

However, there is training data that is not linearly separable, even using a *soft-margin*. To solve this issue, Boser et al. (1992); Cortes and Vapnik (1995) proposed to apply the *kernel trick*, also known as *kernel substitution*, and project the data in a higher dimensional space where the data points will become linearly separable by a hyperplane. The original space is called the *input space*, the transformed space is called the *feature space*. Conveniently, not all data points, but only the dot products need to be projected to the feature space. This reduces the computational load. There are many kernels that can be substituted. A commonly used one is the *Gaussian kernel*, based on the *Gaussian radial basis function*. The resulting algorithm is formally similar, except that every dot product is replaced by a nonlinear kernel function. (Flach, 2012)

By definition, an SVM is a binary classifier. However, many real world problem involve $K > 2$ classes. Thus, many methods have been proposed to combine multiple binary SVMs to perform multiclass classification. One common approach, know as the *one-versus-the-rest approach*, was proposed by Vapnik (1998). Therein, K separate SVMs are constructed. Each SVM is trained using the remaining $K-1$ classes as negative examples (Bishop, 2006). However, there are several issues with this approach, e.g. it deals badly with class imbalance and the scales from different classes are not necessarily the same (Bishop, 2006). There have been other approaches addressing these issues, e.g. Weston and Watkins (1998); Platt et al. (2000); Allwein et al. (2000), discussing these, however, is out of scope of this thesis.

SVMs can also be used in regression tasks (Drucker et al., 1997) and for unsupervised learning (Ben-Hur et al., 2001).

2.1.3 Regression

Regression problems, as opposed to classification problems, do not output class labels from a discrete set of classes, but they output a real number out of a given range. "A *function estimator*, also called a *regressor*, is a mapping $\hat{f} : \mathfrak{X} \to \mathbb{R}$. The regression learning problem is to learn a function estimator from examples $(x_i, f(x_i))$." (Flach, 2012, p.91). For instance, the automatic prediction of similarity between two statements is a regression problem, as the predicted output would be a real number between 1 and 5 (similar to the scaled annotation methods described in Section 1.2.2). The predicted output, or rather the to-be-predicted output variable, is often referred to as *dependent variable*. Furthermore, what in this thesis is referred to as *features* is also called *independent variables*. (Flach, 2012) The most common

form of regression analysis is *linear regression*. A univariate linear regression can be written in the following matrix form (Flach, 2012):

$$\underset{y}{\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}} = \underset{X^\circ}{\begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}} \underset{w}{\begin{pmatrix} a \\ b \end{pmatrix}} + \underset{\epsilon}{\begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}} \tag{2.3}$$

The notation is the following (Flach, 2012):

- y is a vector of the *predicted variables* (from the input/output pairs)
- X denotes an n-by-d matrix, describing n instances in rows containing $[d]$ features in the columns
- X° denotes an n-by-$(d+1)$ matrix, whose first column is all 1s. For convenience, the regression equation is mostly simplified to $y = Xw + \epsilon$, with X being the matrix as described above
- w is the *parameter vector* and contains the *regression coefficients*, or *estimated effects*, a and b; the number of its rows depend on the number of of columns in X
- ϵ denotes an *error variable*, also known as *disturbance term*

The univariate case is also the simplest kind of linear regression, there is also *bivariate* and *multivariate linear regression*. However, we will not further explain these, as we did not make use of regression in this thesis. Nevertheless, it should be noted that to automatize the annotation of textual similarity described in Section 5.1 or the level of offensiveness as described in Section 8.1, regression could have been used.

2.2 Evaluation

According to Flach (2012), "[a]n important thing to keep in mind with all these machine learning problems is that they don't have a 'correct' answer." [p.18]. Some of the reasons are mistakes or subjectiveness in the supposedly correct training output labels, errors in the features, or the signals in the features might be too weak. In order to get some idea of how well an algorithm performs on unseen data in terms of output performance (in this case we do not mean in terms of runtime or memory usage), we need methods and metrics. In order to prevent overfitting or underfitting and to get an overall idea of how the model performs on unseen data, the data that the model is trained on is not the same it is evaluated on (Flach, 2012). The issues of overfitting and underfitting are explained in the next paragraph. There are to ways to address this issue—a *simple train-test split* or *cross-validation* (Flach, 2012). While these methods are applied only to supervised algorithms, the metrics (see Section 2.2.4) are used to evaluate unsupervised methods as well.

2.2.1 Overfitting and Underfitting

As previously mentioned, a model is evaluated on how well it is able to generalize on unseen data. The two extremes that occur when it is not able to do so are *overfitting* and *underfitting*. Figure 2.3 shows both issues in an example for a regression problem. The Xs depict the

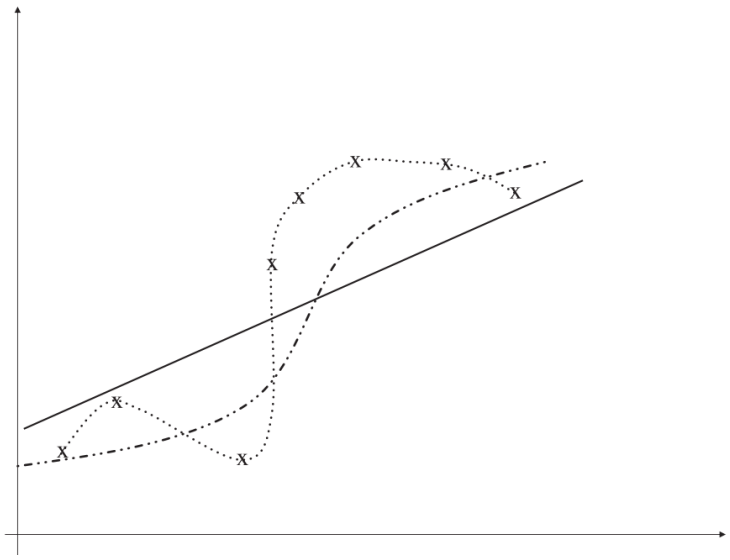


FIGURE 2.3: Illustration of overfitting, underfitting, and well-fitting on a regression issue (Japkowicz and Shah, 2011, p.133)

training data. The dotted curve depicts an *overfitted* function, as it passes through every data point and thus is very specific to the training data. The drawn through line depicts an *underfitted* function, as it is a very coarse approximation of the data. The dash-dotted curve depicts a desirable solution. To find a balance between these two extremes, a *bias-variance* analysis can be performed. (Japkowicz and Shah, 2011)

2.2.2 Simple train-test split

To create a portion of unseen data, before training, the data is split in data for training (training data) and data for testing (testing data). Mostly the split is performed in a proportion of 90% (train) and 10% (test). (Flach, 2012)

This way to measure model performance is mostly used in shared tasks, where the test set remains unknown to the participants at least until the final evaluation of the participants.

2.2.3 Cross Validation

Another way to create a portion of unseen data, which is often preferred in the case of data shortage, is the the so-called *Cross Validation (CV)*, where a random train-test split is performed several times. The data is partitioned in k folds and the model is trained on $k - 1$ folds, one set being set aside for testing. This procedure is repeated k times, so that each fold has been used as a test set. By repeating the procedure several times, a variance of the learning data is captured. Conventionally, CV is applied with $k = 10$. (Flach, 2012)

We make use of this evaluation technique, which is also referred to as 10-fold CV, in most of our studies.

		Predicted	
		+	-
Actual	+	TP	FN
	-	FP	TN

TABLE 2.1: A confusion matrix showing the terms True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN)

2.2.4 Metrics

The metrics explained in the following can be used for both supervised and unsupervised methods, given that there is a set of input/output examples. Then, using the desired and the predicted outputs, a *contingency table* or *confusion matrix* is built as a basis for the calculation of the following metrics.

Table 2.1 displays an example for a binary classification problem e.g. a binary sentiment classification. Similar to (a *contingency table* or *confusion matrix*, in this table each row refers to a desired output class, while each column refers to the classes predicted by the classifier. This table is filled with terms and not numbers, as would be the case for a *contingency table* or *confusion matrix*. The terms illustrated in the table are the following: *true positives (TP)*, *true negatives (TN)*, *false positives (FP)*, and *false negatives (FN)*. The terms *positive* and *negative* refer to the actually desired output and the terms *true* and *false* to the prediction of the classifier. Hence, the overall goal is to maximize the *true* outputs (TP and TN). However, depending on the task, one might put more emphasis on the one or the other.

Accuracy The most simple way to measure the performance is through *accuracy*, which is also discussed in Section 1.4.2. It is expressed in the following way:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.4}$$

However, this simple metric can be misleading for class imbalanced data sets, which is true for many data sets, including the ones presented in this thesis. This issue is best shown through the distribution of the majority class. For instance, in our study on the links between the relations, the majority class for the task of finding contradictions is .87, meaning that in 87% of the data there is NO CONTRADICTION. If the classifier learns that it is a sure guess to predict NO CONTRADICTION, the classifier does not really learn to predict contradiction, but it just learn the class distribution. Hence, there are better suited metrics addressing this issue, e.g. balanced accuracy or F-score, which will be explained in the following.

Precision Oftentimes, one does not need to evaluate the general performance of a classifier, but rather how precise it is—or how many of the predictions were actually relevant. This is important in e.g. tasks related to search engine requests—*how many of the shown results are*

actually relevant?. This metric is called *precision (P)* and is calculated in the following way:

$$Precision = \frac{TP}{TP + FP} \tag{2.5}$$

Recall The opposite case to *precision* is *recall* (R) and one tries to find how many relevant elements were found by the classifier:

$$Recall = \frac{TP}{TP - FN} \tag{2.6}$$

F-score *F-score*, or also called *F-measure*, is a metric that combines *precision* and *recall*. The traditional F-score is actually the $F_1 - score$, but the 1 is usually dropped. In this version, both precision (P) and recall (R) are weighted equally:

$$F_1 = \frac{P * R}{P + R} \tag{2.7}$$

The two other sometimes used F-measures—F_2, putting an emphasis on recall, and $F_{0.5}$, putting an emphasis on precision—are not further discussed here, as they are not used in this thesis.

2.3 Summary

In this chapter, we explained the (supervised) machine learning foundation for all studies presented in this thesis. More specifically, this chapter has shown how the manual annotations (the basics of which have been discussed in the previous chapter), can be used to train automatic systems to recreate the same annotation task on unseen data. In the studies discussed in this thesis, we mostly made use of SVMs—a linear model using the kernel trick to perform multiclass classification. Furthermore, this chapter has shown methods and metrics of evaluating machine learning model results. Now, having laid the annotation foundations in the previous chapter and the machine learning foundations in this one, the next chapter will discuss representations that are needed to annotate, analyze and learn the dimensions of interest in this thesis.

Chapter 3

Representing Statements

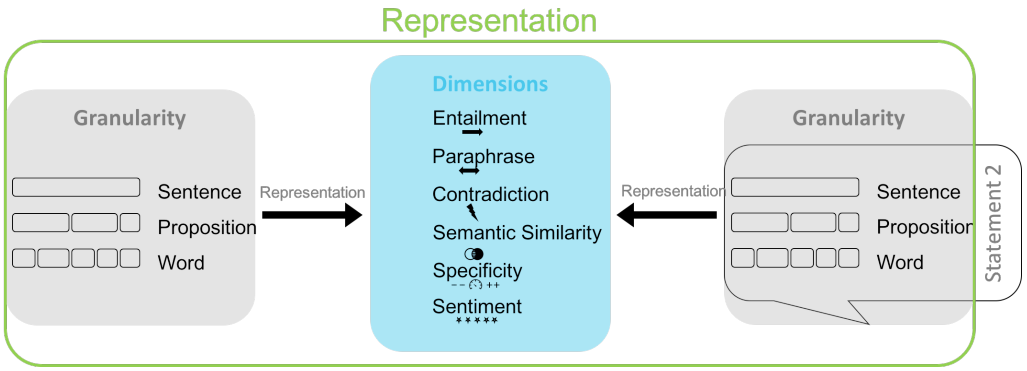


FIGURE 3.1: Illustration of role of representations in this thesis

According to Davis et al. (1993), knowledge representation is a substitute for a real-world entity that can be processed by computers. Furthermore, it should be able to generalize to many concepts, relationships, and real-world tasks (Neelakantan, 2017). Knowledge representation is a main challenge in the field of artificial intelligence, including natural language understanding (Neelakantan, 2017). As shown in the introductory example in Figure 3.1, representations have an encompassing role in this thesis. Basically, they are the result of annotation as discussed in the previous chapter.

Representation of text is easier to show on single words, but can theoretically be used for any text size. The most frequent representations in current Natural Language Processing (NLP) are word embeddings. In simplified terms, this representation shows words as coordinates in multi-dimensional space. Using vector arithmetics, word analogies can be solved. The most famous example for this was presented by Mikolov et al. (2013b) of representing word vectors as equations—"King - Man + Woman = Queen", (see Figure 1). In this example, words such as "king" and "queen" are represented as points in multi-dimensional space and their coordinates can be used for the calculation of their relationship. Although this kind of representation works for a great number of tasks, the representations highly depend on how they were created. However, creating such human-understandable is not always possible, which makes it difficult to improve its flaws. One disadvantage is that the dimensions in the space are not transparent, meaning it is unclear what a dimension denotes, e.g. one does not know whether the word vectors are in a synonymous or antonymous relation, which would be an important distinction in most applications. For instance, a transparent dimension

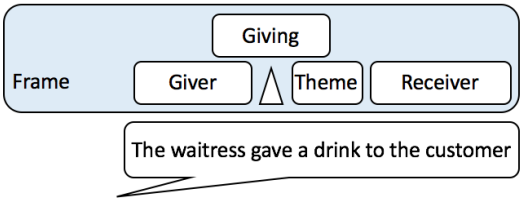


FIGURE 3.2: Exemplary frame representation

could be as simple as *Part-of-Speech (POS)* or *word length*, but also more complex such as *sentiment* or *topic*. The other disadvantage is that the dimensions are so numerous a human cannot process the representation as a whole without further processing steps (Panchenko, 2016).

There are also representations that are well understandable to humans, e.g. frames as proposed by Fillmore (1976). These built upon the argument level, i.e. the arguments are labeled with semantic roles, as shown in Figure 3.2. On this level, words are represented with their function within a statement. Although such representations are easily understandable to humans, they are difficult to scale, especially to unknown or new content.

The aim of this thesis is finding a representation that

1) has the dimensions needed to semantically process and comprehend the statements.
2) is of a formalism that is both robustly computable and human-understandable.

In order to address 1), we research semantic dimensions. In our understanding, dimensions are features of the given text. As shown in Figure 3.1 the dimensions researched in this thesis are textual entailment, paraphrases, contradiction, semantic similarity, specificity, and aspect-based sentiment. This is discussed in Section 3.1.

To address 2), we discuss which granularity and formalisms are best for our purpose. To address the granularity, we perform a study in Benikova and Zesch (2016), which is described in Chapter 4 in more detail. In this chapter, we theoretically discuss which representation format, building upon a chosen granularity, would best suit the purpose of being human-understandable. This is discussed in Section 3.2.

The two points strongly interact, as a format is needed to internally represent the statements, and the semantic dimensions are needed to display the semantics behind the given representation.

In this chapter, we will use made-up examples from the hotel domain in order to fit our overall exemplary application of user-specific hotel reviews. In the corresponding chapters, we will use real examples from the actual data.

3.1 Dimensions

A representations can have many dimensions. In our application scenario, these dimensions should be human-understandable and help in the selection of useful and non-redundant reviews. In this thesis, we focus on similarity relations, more specifically textual entailment,

paraphrases, contradiction, semantic similarity, as well as specificity, and aspect-based sentiment. As previously mentioned, the similarity relations can only be regarded as comparative dimensions between at least two statements by definition, while both specificity and sentiment can be seen as a comparative dimension between two statements, as well as an individual dimension of one statement. All dimensions are potentially helpful in filtering relevant information: While similarity dimension help to reduce redundancy or emphasize contradictions, specificity helps to choose statements have the correct amount of information and sentiment helps to filter for positive or negative reviews.

In this chapter, we will shortly discuss the definitions of the dimensions as used in this thesis. The corresponding chapters present the dimensions and their definitions in more detail: Chapter 5 discusses relations between the dimensions, Chapter 6 deals with specificity in detail, while Chapter 7 deals with sentiment.

3.1.1 Similarity relations

Similarity relations are dimensions that exist between at least two statements. Semantic similarity measures meaning similarity on a scale, while entailment, paraphrases, and contradiction are binary relations indicating a specific kind of similarity.

Entailment Textual entailment was introduced by Dagan and Glickman (2004) and was afterwards used in Recognizing Textual Entailment (RTE) tasks (Dagan et al., 2005, 2009). Textual Entailment is a directional relation between pieces of text in which the information of the *text* infers the information of the *hypothesis*, while the entailment in the other direction is not necessarily given (Dagan and Glickman, 2004).

In Example 3.1, the text shown in (1) entails the hypothesis shown in (2):

1 The waitress gave a drink to the customer.
2 There was a waitress.

EXAMPLE 3.1: Statement pair in entailment relation

Paraphrases The strict logical definition of paraphrases encompasses the relation between sentences or phrases that convey the same meaning in different words (Bhagat and Hovy, 2013). According to De Beaugrande and Dressler (1981), Hirst (2003), and Bhagat and Hovy (2013), the linguistic definition of paraphrases is broader than the strict logical one, allowing for approximate equivalence and not necessarily requiring synonymy. In Computational Linguistics (CL) and NLP, it the phenomenon is also often referred to as *quasi-paraphrase* or *near-synonymy* (Hirst, 2003; Bhagat and Hovy, 2013). The relation is symmetric (Gold et al., 2019).

The automatic detection of paraphrases is useful in tasks such as summarization, information extraction, plagiarism detection, machine translation, question answering, and natural language generation (Bhagat and Hovy, 2013).

In Example 3.2, (1) and (2) are paraphrases:

1 The waitress gave a drink to the customer.
2 The waitress served a drink to the customer.

EXAMPLE 3.2: Statement pair in paraphrase relation

Contradiction In Giampiccolo et al. (2008), a further task was added to the RTE challenge—the detection of contradiction between a sentence pair, or more specifically, the *text* contradicted the *hypothesis*. Following this definition, in this thesis, and also in the corpora developed herein—namely Gold et al. (2019) and Kovatchev et al. (2020), we regard contradiction as a relation between two statements that cannot be true at the same time. Example 3.3 shows two statements that contradict each other.

1 The waitress gave a drink to the customer.
2 The waitress did not serve.

EXAMPLE 3.3: Statement pair in contradiction relation

Semantic similarity According to Resnik (1995), *semantic similarity* is a special kind of *semantic relatedness*. Lin (1998) defines *semantic similarity* as "the ratio between the amount of information in the commonality and the amount of information in the description of the two objects" [p.299]. Furthermore, Lin (1998) states that the maximum similarity is reached when the compared text pieces are identical. Semantic similarity has often been used a proxy for the other relations in applications such as summarization (Lloret et al., 2008), plagiarism detection (Alzahrani and Salim, 2010; Bär et al., 2012), machine translation (Padó et al., 2009), question answering (Harabagiu and Hickl, 2006), and natural language generation (Agirre et al., 2013). Furthermore, it has also been regarded in relation to the other similarity relations, as e.g. paraphrases should be semantically similar. Given a scale from 0–4, where 0 denotes *no similarity* and 4 denotes *textual equality*, the statements in Example 3.2 would be rated as 4.

3.1.2 Specificity

Specificity is mostly regarded between noun phrases (Cruse, 1977; Enç, 1991; Farkas, 2002). However, Yager (1992) and Louis and Nenkova (2011) researched specificity on the sentence level. We are the first to explicitly see specificity of a statement in relation to another statement. In our definition, specificity is a relation between statements in which one phrase is more precise and the other more vague (Gold et al., 2019). In Example 3.4, Statement (1) is more specific than Statement (2) as it gives information on who gets served what:

1 The waitress gave a drink to the customer.
2 The waitress served.

EXAMPLE 3.4: Statement pair in specificity relation

If, however, we would see specificity as an individual dimension, given a scale from 0–4, where 0 denotes *very general* and 4 denotes *very specific*, Statement (1) in Example 3.4 could be rated as 3, while Statement (2) could be rated as 1.

3.1.3 Sentiment

Sentiment analysis deals with the computational treatment of opinion, sentiment, and subjectivity in text (Pang and Lee, 2008). According to Pang and Lee (2008), *sentiment analysis* and *opinion mining* denote the same field of study and can be regarded as sub-areas of subjectivity analysis. Sentiment analysis is mostly used for the automatic analysis of evaluative text, such as reviews (Pang and Lee, 2008). In the task of aspect-based sentiment, the sentiment is directed towards a given aspect. In Example 3.5, in Statement (2), there is a *positive* sentiment towards the aspect of *service*, while Statement (1) is rather *neutral*. However, sentiment can not only be seen as an individual dimension, but also comparatively e.g. whether one statement is more positive that the other. In Example 3.5, (2) is more positive than (1), as it explicitly mentions that the waitress is attentive and quick.

1 The waitress gave a drink to the customer.
2 The attentive waitress served a drink to the customer quickly.

EXAMPLE 3.5: Statement pair in sentiment relation

3.2 Survey on Representation Formalisms

In Benikova and Zesch (2016), we envision a project that bridges the gap between robustly computable, but less expressive argument level representations representations and *frame level representations* which are highly expressive, but not robustly computable.

It should be added we are aware of other representation forms, e.g. simple lexical representations or vectors. The most simple is the lexical representation, as it is just contains the word itself (a lexical representation may also consist of several words, e.g. n-grams, which are n-1 neighboring words, including the word itself and considering the word order). Lexical units, such as words, n-grams, phrases, or sentences can be represented as mathematical vectors of real numbers in a multi-dimensional space. Most recently, these kind of vectors are called word embeddings. They can be created using methods such as neural networks (Mikolov et al., 2013a), dimensionality reduction on the word co-occurrence matrix (Lebret and Collobert, 2013; Levy and Goldberg, 2014b), probabilistic models (Globerson et al., 2007), and

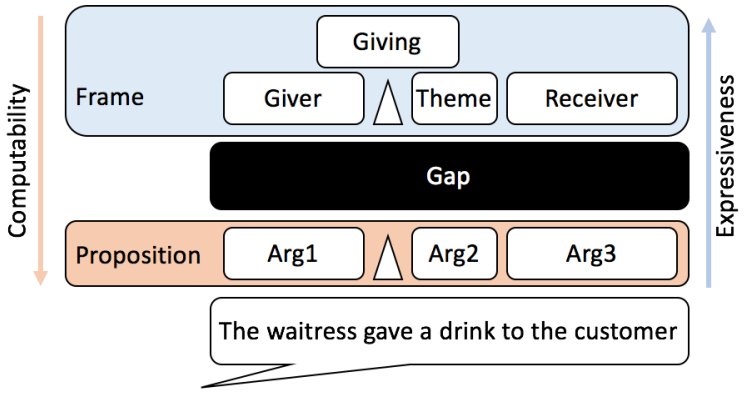


FIGURE 3.3: Representations of an exemplary statement on argument and frame level

representations using the context (Levy and Goldberg, 2014a), which are mostly n-grams. However, in Benikova and Zesch (2016), we focus on various predicate-argument structures, as they are human-readable and computable.

Problem description As described above, the aim of Benikova and Zesch (2016) was bridging the gap between robustly computable argument level representations and highly expressive frame level representations. Argument level representations are less expressive as they do not contain additional information on the arguments. Frame level representations are not robustly computable, as they collapse when receiving unknown data (either resulting from a genre that is different from the one they were trained on or when information that is in the frame data bank is mentioned). The distinction and the gap between the two main representation formalisms is presented in Figure 3.3. On the proposition level, the predicate "give" and all its arguments—"The waitress", "a drink", and "to the customer"—are identified. Additional semantic role labels are assigned to the arguments on the frame level—"The waitress" is GIVER, "a drink" is THEME, and "to the customer" is RECEIVER.

Solution Idea We envision a representation that enables the detection of the discussed dimensions. These operations are not only necessary to compress the amount of information, which is especially important in high-volume, high redundancy social media posts, but also for other tasks such as to analyze and understand statements efficiently.

Outcome A possible solution to bridge this gap by inducing distributional semantic clusters as labels in a frame structural representation. This could be achieved by building a robustly computable and expressive representation that is suited to perform the discussed operations by using social media domain specific clusters and topic labeling methods for the frame-labeling. The validity of a new representation and approach should be evaluated extrinsically and application-based. However, this work is theoretical and it presents a research gap and possible solutions rather than a technical solution.

3.2.1 Formalisms

We distinguish between representation formalisms on two levels: (i) proposition level, which can be robustly implemented more easily, and (ii) frame level, which is highly expressive.

Proposition level The most simple kind of predicate-argument structures are propositions. As this structure is the best applicable to our scenario, we researched it in more detail in Chapter 4. In short. it can be said that propositions are relational tuples from sentences or parts of sentences, consisting of a predicate and its arguments.

As shown in Figure 3.3, most argument representations consist of an event trigger (marked with a triangle in the figure), which is mostly a predicate, and its corresponding arguments (Banarescu et al., 2013; Kingsbury and Palmer, 2003). In the example shown in Figure 3.3, "The waitress gave a drink to the customer", "gave" is the PREDICATE, whereas the other elements are ARGUMENTS. The first argument is often reserved for the role of the SUBJECT, in this case "The waitress", while "a drink" and "to the customer" are regular arguments treated equally regardless of their syntactic role.

Propositions are the basis for frame-semantic representations. They are better computable and scale better to non-standard texts than structures that are built upon it, such as e.g. frames. In Gold and Zesch (2019), we studied the impact of sentence complexity on the computability of propositions and found that, unsurprisingly, more complex sentences are often computed incorrectly. Hence, frame-semantic structures that built upon propositions must be even more affected by this.

Frame level On this level, events are represented as frame structures such as proposed by Fillmore (1976) that built upon the argument level, i.e. the arguments are labeled with semantic roles. In CL and NLP, events (Krifka, 1989; Ritter et al., 2012; Agerri et al., 2014), frames (Pustejovsky, 1991), schemas (Krifka, 1989), abstract meaning representations (Zadeh, 1978), and corpus patterns (Hanks, 2004) are similar in the sense that these structures represent the content of text, mostly on a sentence or phrase basis.

The best known frame-semantic instantiations are FrameNet (Fillmore, 1976) and Abstract Meaning Representation (AMR). Semantic frames are "schematic representations of the conceptual structures and patterns of beliefs, practices, institutions, images, etc. that provide a foundation for meaningful interaction in a given speech community" (Fillmore et al., 2003, p. 235). A well-known frame-semantic tagger is SEMAFOR (Das et al., 2010). Figure 3.3 shows a frame-semantic representation of an exemplary sentence as it is proposed by FrameNet. In this example, "The waitress gave a drink to the customer", "gave" is the predicate, which triggers the frame of GIVING, identifying "The waitress" as GIVER, "a drink" as THEME, and "to the customer" as RECEIVER. FrameNet is a computational lexicography project based on this theory. It analyzes the word meaning by applying frames that underly their meaning and studying the syntactic properties of words (Fillmore et al., 2003). AMR is a whole-sentence semantic representation. AMRs are "rooted, directed, edge labeled, leaf-labeled graphs" (Banarescu et al., 2013, p. 179) that are human- and machine-readable. AMRs aim to have the same representation for complete paraphrases.

3.2.2 Challenges

The main challenge is to bridge the gap between argument and frame level representation, as shown in Figure 3.3. In this section, we will discuss the challenges of *performance of operations* and *coverage*.

Performance of operations Our goal is to develop a representation that is both computable even on noisy social media text or more specific a hotel review and expressive enough to support the extraction of all required dimensions. Example 3.6 shows two semantically equivalent sentences, meaning that these are in a paraphrase relation.

1 The waitress gave a drink to the customer.
2 The customer received a drink from the waitress.

EXAMPLE 3.6: Statement pair in more complex paraphrase relation

In this example, *receive* is the antonym of *give* and the roles of *Giver* and *Receiver* are inverted. On the proposition level, it remains a hard problem to establish the equivalence between the two sentences, while that would be easy on the frame level. However, getting to the frame level is an equally hard problem (Palmer and Sporleder, 2010). Palmer and Sporleder (2010) categorized and evaluated the coverage gaps in FrameNet (Baker et al., 2003).

Coverage Palmer and Sporleder (2010) categorized and evaluated the coverage gaps in FrameNet (Baker et al., 2003). Coverage, whether of undefined units, lemmas, or senses, is of special importance when dealing with non-standard text that contains spelling variations and neologisms that need to be dealt with.

In our opinion, the lack of undefined units is an especially problematic issue in social media texts. Furthermore, it may contain innovative, informal or incomplete use of frames, due to space restrictions such as presented by Twitter or by review platforms. Also by cause of space restrictions, which lead to a lack of context, and considering the variety of topics that is addressed in social media, it is more challenging to find a fitting frame out of an existing frame repository (Ritter et al., 2012; Li and Ji, 2016).

Giuglea and Moschitti (2006) and Mújdricza-Maydt et al. (2016) tried to bridge the gap by combing repositories on frame and proposition level and representing them based on ICL (Kipper et al., 2006). ICL, which are used in VerbNet (Kipper et al., 2006), are more fine-grained than classic Levin verb classes, formed according to alternations of the grammatical expression of their arguments (Levin, 1993). Classic Levin verb classes were used for measuring semantic evidence between verbs (Baker and Ruppenhofer, 2002).

However, these approaches also have to deal with coverage problems due to their reliance on manually crafted frame repositories.

3.2.3 Comparison of predicate-argument approaches

Table 3.1 shows different representation approaches and their features. The approaches in the first section are bottom-up, whereas the ones in the second section are top-down.

All approaches except for Ritter et al. (2012) and Agerri et al. (2014) use text-internal links, meaning relations between the words in the statement. While all top-down approaches except Concept Relation Concept (CRC) and Probabilistic Relational Universal Fuzzy (PRUF) use phrases in their representation, the bottom-up approaches all use words.

Some representations have ontological labels as representation constituents. Corpus Pattern Analysis (CPA) (Hanks, 2004) and UNICON (Wu and Palmer, 1994) use generic hypernyms such as *animate/inanimate* or *Person/Object* as ontology labels, whereas CRC (Velardi et al., 1991), FrameNet (Baker et al., 2003) and PRUF (Zadeh, 1978) use more specific terms for their labels, such as *Giver, Donor* or *Receiver, Recipient*. (Agerri et al., 2014) use links to a knowledge base to create more specific terms for their labels. However, the ontological labels assigned by (Velardi et al., 1991), FrameNet (Baker et al., 2003) and PRUF (Zadeh, 1978) are specific to individual frames or event formalisms, whereas those in the representations described by VerbNet (Schuler, 2005) and ACE (Doddington et al., 2004) are frame-independent, but also less specific.

Approach	**Relations**	**Spans**	**Abstraction Level**				**Extra-Propositional**			
	Text-Internal Links	Phrases	Term	Part-of-Speech	Argument	Ontology	Modality	Negation	Quantifier	Time
Dependency	•		•	•			•	•	•	•
(Ritter et al., 2012)			•	•	•					•
AMR	•		•	•	•		•	•	•	•
UNICON	•		•	•	•	•				
(Agerri et al., 2014)			•		•	•				•
PropBank	•	•	•	•	•		•	•		•
CPA	•	•	•	•	•	•				
CRC	•		•	•	•	•				
Frames	•	•	•	•	•	•	•	•	•	•
VerbNet	•	•	•	•	•	•	•	•	•	•
ACE	•	•	•	•	•	•	•	•	•	•
PRUF	•		•	•	•	•		•	•	•

TABLE 3.1: Comparison of representation features based on abstraction level distinction

Ontological structures in the representations may facilitate the computation of dimensions such as paraphrases, entailment, and contradiction. Furthermore, nearly all approaches use *POS* and *arguments* in their representation. Some approaches provide extra-propositional information, such as modality, negation, quantifiers, and time. We did not focus on these aspects in this thesis, although all of these are very important to understand the meaning of a statement. However, they are also very difficult to compute especially on non-standard text.

3.2.4 Approach

According to Modi et al. (2012), frame-semantic parsing conceptually consists of four stages:

- Identification of frame-evoking elements
- Identification of their arguments
- Labeling of frames
- Labeling of roles

We summarize these tasks in groups of two, namely *identification* and *labeling*, and discuss our approach towards them in the following subsections.

Identification of frame-evoking elements and their arguments We regard the first two tasks as tasks of the argument level, which we plan to solve with part-of-speech tagging and dependency parsing, by extracting all main verbs to solve the first task and considering all its noun dependencies as arguments in the second task. This is similar to the approach of Modi et al. (2012).

Labeling of predicates and their arguments Like Modi et al. (2012), we focus on the last two tasks, which we regard as tasks of the frame level. We observe this task under the aspect of fitting the realization of operation tasks as discussed earlier. As we only regard predicate frames and their arguments for the role labeling, we will use *predicate* as a term for the unlabeled form of *frame* and *argument* as the unlabeled form of *role*.

Pre-defined frame labels There have been attempts to bridge the gap on Social Media texts by projecting ontological information in the form of computed *proposition formalisms* on the proposition trigger on the argument level (Ritter et al., 2012; Li et al., 2010; Li and Ji, 2016) in order to solve the task of frame labeling. However, according to Ritter et al. (2012) the automatic mapping of pre-defined proposition formalisms is insufficient for providing semantically meaningful information on the proposition.

We aim to augment those approaches by inducing frame-like structures based on distributional semantics. Moreover, we want to use similarity clusters for the labeling of arguments in frames. We seek to compute the argument labels by the use of supersense tagging, similar to the approach presented by Coppola et al. (2009). They successfully used the WordNet supersense labels (Miller, 1995) for verbs and nouns as a pre-processing step for the automatic labeling of frames and their arguments.

Approaches using Levin classes, Intersective Levin Classes (ICL), or WordNet supersenses tackle the same tasks, namely labeling the frame and their corresponding roles. However, all of these suffer from the discussed coverage problem.

Clusters as labels To circumvent the coverage issue, there have been approaches using clusters similarly to frame labels. Directly labeling predicates and their arguments has been performed by Modi et al. (2012), who iteratively clustered verbal frames with their arguments.

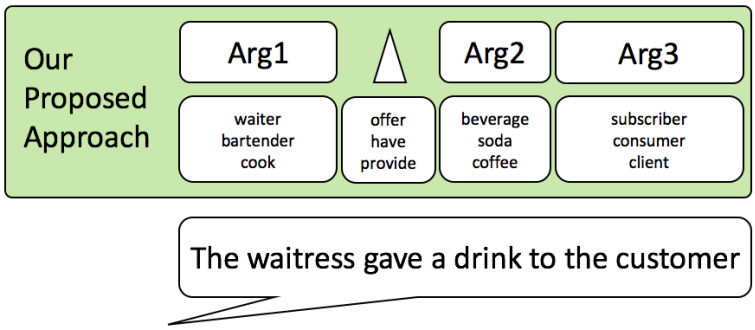


FIGURE 3.4: Representations of our approach to bridge the gap using the Twitter bi-gram model of JoBimViz (Ruppert et al., 2015)

As our main goal is to perform operations on proposition representations, we do not need human-readable frames as proposed by FrameNet, but a level that is semantically equivalent to it, thus our first goal is to compute domain specific clusters for the labeling.

In contrast to Modi et al. (2012), we plan to cluster the verbal predicates and the arguments separately. Although this might seem less intuitive, we believe that due to the difficulties with social media data, the structures of full frames are less repetitive and are more difficult to cluster. Thus, by dividing the two tasks of predicate and argument clustering, we hope to achieve better results in our setting.

Furthermore, in order to deal with the issues of the previously discussed peculiarities of the social media domain, we plan to train clusters on large amounts of tweets.

An example of our envisioned representation is shown in Figure 3.4, which was produced using the Twitter bi-gram model of JoBimViz (Ruppert et al., 2015). JoBimViz is an interactive visualization for distributional semantic graph-based models. The models use term similarities, similarities between context features, clustered word senses and their labeling with hypernym relations. One possible output is a list of similar terms in the context of the sentence (to each term in the sentence).

Figure 3.4 shows the clustering for finding the correct sense in the labeling task, for both the predicate and its arguments, e.g. “waitress” is similar to “waiter”, “bartender”, and “cook”. However, aiming at representations that are suited for dimensions such as paraphrases and entailment, the known problems of antonyms being in the same cluster needs to be solved. Similarly to Lobanova et al. (2010), who automatically extracted antonyms in text, we plan to solve this issue with a pattern-based approach.

Topic-clustered labels After succeeding in the clustering task, we plan to experiment with human-readable frame clusters. In contrast to using pre-defined WordNet supersenses and mapping these to frames, we want to solve the task of finding labels for the clusters by using supersenses computed from domain-specific clusters to directly label the frames and their arguments.

Our hypothesis is that by using more and soft clusters for the supersense tagging, the role labels of the proposition arguments become semantically richer, because more specific semantic information on the arguments and their context in the proposition is encoded.

Thus, we plan to use the supersense tagging by using an LDA extension, in which a combination of context and language features is used, as described by (Riedl, 2016).

3.2.5 Evaluation plan

We plan to evaluate our approach in an extrinsic, application-based way on a manual gold standard containing *proposition paraphrases*. In order to test how well our approach performs in comparison to state-of-the-art approaches of both *argument* and *frame representations*, such as Das (2014) or Li and Ji (2016) in the task of equivalence computation, we will compare the results of all approaches.

For this purpose, we plan to develop a dataset that is similar to Roth and Frank (2012), but tailored to the social media domain. They produced a corpus of alignments between semantically similar predicates and their arguments from news texts on the same proposition.

3.3 Conclusion on Representing Statements

In this chapter, we outlined dimension and formalisms that are of interest in this thesis. While we conducted many studies on the described dimensions (see Chapter 5, Chapter 6, Chapter 7), we did not follow up on our plan for the representation formalism above the proposition level as described in Section 3.2. One reason for this is that the studies on the dimensions were a focus in this thesis. The comparative method that we used to research them took up many, mostly time, resources. The main reason, however, will be explained in the next chapter, which focuses on propositions. Shortly it can be sad that the extraction of propositions from sentences of interest already poses an obstacle to automatic methods. Having shown that proposition are a complex topic on its own, they are the focus of the next chapter.

Chapter 4

Representing Statements: The Case for Propositions

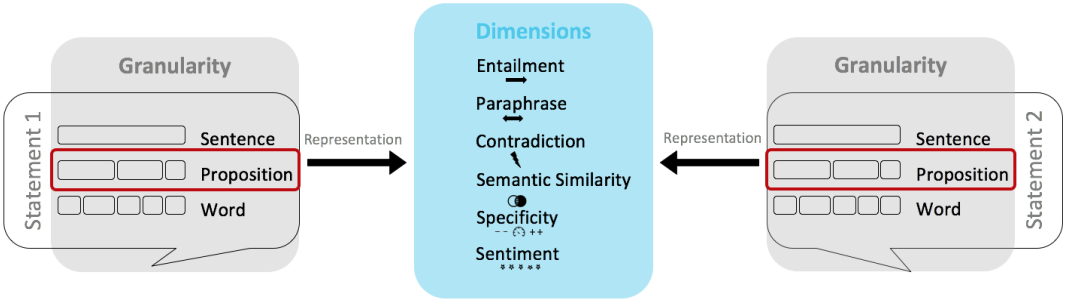


FIGURE 4.1: Illustration of proposition in this thesis

Propositions are relational structures that are extracted from sentences or phrases to be used as representations for a number of applications. Propositions are used in language understanding tasks such as relation extraction (Riedel et al., 2013; Petroni et al., 2015), information retrieval (Löser et al., 2011; Giri et al., 2017), question answering (Khot et al., 2017), word analogy detection (Stanovsky et al., 2015), knowledge base construction (Dong et al., 2014; Stanovsky and Dagan, 2016), summarization (Melli et al., 2006), and other tasks that need comparative operations, such as equality, entailment, or contradiction, on phrases or sentences.

In the previous chapter, we explain that propositions are not just a granularity level, but also an effective and computable representation of text. The previous chapter discusses its theoretical differences when compared to lexical and vector representations. Furthermore, it shows a comparison of predicate-argument structures, including structures underlying propositions, e.g. dependencies, and structures overlying propositions, such as frames. We state that propositions are more robustly computable than overlying structures such as frame level representations (see Section 3.2), which is why we research them in more detail. Although proposition extraction is more robustly computable, automatic systems still have issues.

In this thesis, propositions are used as one granularity level for representing statements. As shown in Figure 4.1, the granularity is placed between the sentence and the word level. Furthermore, the figure shows that the dimensions of interest can be annotated on this level. In Benikova and Zesch (2016), we analyze the compositionality of the three granularity levels

on the dimension of paraphrases, which has not been done before. Apart from verbatim paraphrases, two statements in a paraphrase relation are composed of different words. Analyzing paraphrases on the three different levels enabled us to see how paraphrases are composed e.g. to what extent a sentence paraphrase consists of proposition paraphrases.

In Gold and Zesch (2019), we are the first to show that sentence complexity has a big influence on proposition extraction performance. However, we also show that with or without considering sentence complexity, the proposition extraction systems perform proportionally—regardless of the complexity, the ranking of the systems does not vary much.

Definition Although propositions are widely used, there is no clear consensus on their exact definition. However, there is a consensus on its basis: Propositions are predicate-centered tuples consisting of the *predicate*, the *subject*, and other *arguments* such as *objects* and *modifiers* that are extracted from sentences. Apart from this basic definition of propositions there are no common guidelines and subsequently no gold standard defining a valid extraction (Stanovsky and Dagan, 2016; Niklaus et al., 2018). The different definitions differ in details such as the labeling of arguments and dealing with complex constructions, but also formalizations such as the notation. For our purposes, this plain definition is sufficient.

Figure 4.2 shows an example of a proposition representation using the basic definition. In Figure 4.2, "smiled" is the PREDICATE and the other elements are ARGUMENTS. The first argument is reserved for the role of the SUBJECT, in this case "The waitress", while "at her friend" and "now" are arguments, without further sub-specification.

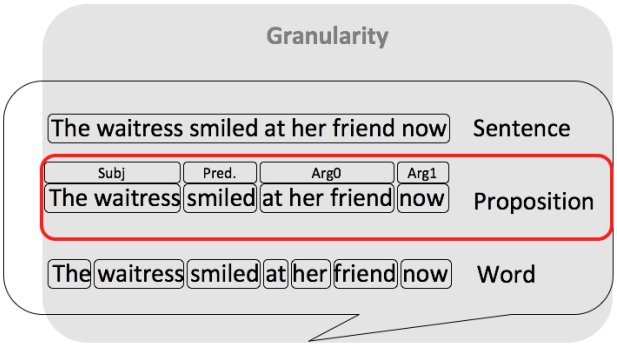


FIGURE 4.2: Example statement on all three granularity levels

In our studies, we focused on verbal predicates only. For further information regarding the handling of e.g. auxiliary verbs, negations, modals, see our annotation guidelines in Appendix A.1.1.3 and Appendix A.1.1.2.

In our definition, we allow for nested propositions i.e. propositions containing arguments that are propositions themselves, e.g. here, the sentence "I think their food is great" is split in two propositions—"I | think | their food is great" and "their food | is | great ". This definition is restrictive in that it asks for exactly two propositions in the given example. However, it is the representation that is needed to extract information from reviews, as it would help to reduce redundancies, e.g. by clustering sentences such as "Their food is great" and "I think their food is great". Furthermore, we are not interested in inferred information, e.g. "They | have | food".

Further little alteration e.g. considering special handling of conjunctions or interjections, will be discussed in the respective sections.

Influence of Sentence Complexity on Proposition Extraction Proposition extraction systems have performance issues on real data. We believe that one of the main issues of proposition extraction systems is sentence splitting, which needs to be performed on complex sentences. Ideally, sentences for proposition extraction would be *simple* i.e. contain one predicate only. In this case, the task would be to identify the (one) predicate and its arguments. However, sentences may contain several predicates, which demands for splitting the sentence into clauses containing only one predicate. To research the impact of sentence complexity on proposition extraction, in our study, we build the first proposition corpus differentiating between simple and complex sentences and evaluate state-of-the-art proposition systems on it (Gold and Zesch, 2019). To do so, we use a corpus of reviews and create simplified sentences using crowdsourcing in a preliminary step. In the next step, we produce a manual gold standard of propositions from these sentences. In the final step, we evaluate the performance of proposition extraction systems on our proposition corpus. The study is described in more detail in Section 4.1.

Compositionality of Granularity Levels In a further study (Benikova and Zesch, 2017), we research the compositionality of different granularity levels of statement representations using two already existing paraphrase corpora—the Microsoft Paraphrase Corpus (MRPC) (Dolan and Brockett, 2005) and the Twitter Paraphrase Corpus (TPC) (Xu et al., 2014). As shown in Figure 4.1, we use three different granularity levels—sentence, proposition, and word—in our thesis. In Benikova and Zesch (2017), the lowest level is not words, but individual proposition elements, meaning the predicate and its arguments. We annotate paraphrases on all three levels independently in order to research their compositionality. The study is described in more detail in Section 4.2.

4.1 Influence of Sentence Complexity on Proposition Extraction

In Gold and Zesch (2019), we study the impact of sentence complexity on the performance of proposition extraction systems.

Problem Description We have a focus on the influence of sentence complexity or the task of clause splitting that precedes the actual proposition extraction. Our main idea is that sentence complexity has a strong correlation with the performance of proposition extraction systems. To our knowledge, there has been no study on the impact of sentence complexity on proposition extraction system performance, meaning that we do not know how well proposition extraction works without the component of clause splitting that precedes it. Most system evaluations ignore sentence complexity and that splitting complex sentences is an obligatory preliminary step. Thus, they do not judge proposition extraction independently. The best systems for proposition extraction, which we intend to find, do not necessarily perform best

	Sentences					
	SIMPLE			COMPLEX		
	The waitress smiled at her friend now			We were quickly seated and served		
Systems	**Subject**	**Predicate**	**Other elements**	**Subject**	**Predicate**	**Other elements**
Allen	The waitress	smiled	at her friend | now	We	were quickly seated	
				We	served	
ClausIE	The waitress	smiled	at her friend now	We	were seated	quickly
	The waitress	smiled	now	We	were served	quickly
	her	has	friend	We	were seated	
				We	were served	
ReVerb	The waitress	now smiled at	her friend			
Stanford	waitress	smiled at	her friend	We	were	quickly seated
	waitress	now smiled at	her friend	We	were	seated
OLLIE	The waitress	now smiled at	her friend			
OpenIE	The waitress	smiled	now | at her friend	We	were served	
				We	were quickly seated	
BL1	The	waitress	smiled at her friend now	We	were	quickly seated and served
BL2	The waitress	smiled	at her friend now	We	were	quickly seated and served
Us	The waitress	smiled	at her friend | now	We	were seated	quickly
				We	were served	quickly

TABLE 4.2: Output of proposition extraction systems and our two baselines for a simple and a complex sentence

on complex sentences, as their performance may be strongly influenced by the clause splitting step. Thus, it is not trivial that the best-performing system on simple sentences is also the best on complex sentences. Table 4.2 shows the outputs of different systems used in the discussed comparisons.[1] By definition, all systems extract a predicate and its arguments. The complex sentence in the table shows an example of the issues that systems have with complex sentences in particular. As shown in Table 4.2, for the complex sentence, some systems do not give any output (ReVerb and OLLIE), while other systems make various other mistakes, e.g. not recognizing the second verb (Stanford).

Solution Idea In order to measure the impact of sentence complexity on system performance, we build a corpus, differentiating between simple and complex sentences. The corpus creation is depicted in Figure 4.3. Having done this, we are able to evaluate the performance of several systems based on this separation.

Outcome We show that sentence complexity has a measurable impact on proposition extraction performance of both humans and machines. Furthermore, we show that the ranking

[1] We list and very shortly describe the systems used in the comparison in Section 4.1.3.

of systems is similar among simple and complex sentences.[2] The main issues in complex sentences that we could identify were conditional and temporal clauses.

4.1.1 Related Work

There have been several comparisons of proposition systems (Gashteovski et al., 2017; Schneider et al., 2017; Stanovsky et al., 2018; Saha and Mausam, 2018; Niklaus et al., 2018). These will be discussed in the individual paragraphs of this subsection. In all described comparisons, the system of the respective authors is the best, which makes sense as it addresses the issue shown by the respective authors.

Gashteovski et al. (2017) aim at finding a system with minimal attributes, meaning that hedging[3] and attributes expressed e.g. through relative clauses or adjectives, can be optionally removed. Thus, they use recall and two kinds of precision in the evaluation in order to account for the feature of minimality. As the feature of minimality is not a focus of this thesis, these evaluation measures will not be discussed. Gashteovski et al. (2017) evaluates OLLIE (Mausam et al., 2012), ClausIE (Del Corro and Gemulla, 2013), and Stanford OIE (Angeli et al., 2015) against their own system.

Schneider et al. (2017) present a benchmark for analyzing errors in proposition extraction systems. Their classes are *wrong boundaries*, *redundant extraction*, *wrong extraction*, *uninformative extraction*, *missing extraction*, and *out of scope*. Their pre-defined classes do not map directly to sentence complexity, although *wrong boundaries* and *out of scope* would also be of some interest in an even more detailed error analysis. Schneider et al. (2017) perform and report a quantitative analysis of Stanford, OpenIE-4, ClausIE and PredPat (White et al., 2016) using different datasets. Their benchmark can be used for further systems and datasets. It can also be used for both quantitative and qualitative analysis. For the quantitative analysis described in their work, they use F_1-score, precision, and recall. For the qualitative analysis they use two human annotators. In their study, they cannot find a clear winner, as each systems performs best on a particular data set, but does not outperform others significantly on more than one set. In this way, they show that the performance of each system depends on the individual task and hence needs to be tested accordingly.

Stanovsky et al. (2018) evaluates ClausIE, PropS (Stanovsky et al., 2016), and Open IE-4 against their new system we will call *Allen* (Stanovsky et al., 2018) herein, using precision-recall, area under the curve, and F_1-score. They compare the individual proposition elements. For a proposition to be judged as correct, the predicate and the syntactic heads of the arguments need to be the same as the gold standard.

[2] This means the best-performing systems among simple sentences that are disentangled from the task of clause splitting, are also the best in complex sentences, where clause splitting also needs to be performed. This may mean that to find the overall best system, one does not need to classify between simple and complex sentences. However, it is necessary to find out whether sentence complexity is one problem of proposition extraction.

[3] In pragmatics, hedging is a textual construction that lessens the impact of an utterance. It is often expressed through modal verbs, adjectives, or adverbs, e.g. through “I believe that”, “isn’t it?”, “I’m not an expert, but”.

Learning-based	Rule-based	Clause-based	Capturing Inter-Proposition Relationships
TEXTRUNNER	ReVerb	ClausIE	OLLIE
WOE	Kraken	Stanford	OpenIE
OLLIE	Exemplar		CSD-IE
ReNoun	Props		NestIE
	PredPatt		MinIE
			Graphene

TABLE 4.3: Classification of proposition systems by Niklaus et al. (2018)

Niklaus et al. (2018) presented an overview of proposition extraction systems and classified them into the classic categories of learning-based, rule-based, and clause-based approaches, as well as approaches capturing inter-propositional relationships. Their classification of state-of-the-art systems is shown in Table 4.3. They described the specific problems each system tackles as well as gaps on the overall evolution of proposition extraction systems. The evolutionary process described by them can be seen in the classification from left to right—from learning-based systems to systems capturing inter-propositional relationships.

Saha and Mausam (2018) evaluate ClausIE, OpenIE-4, and CALMIE (a part of OpenIE) using precision. With the findings of this comparison, they introduce a new version of their system, OpenIE-5[4]. According to Saha and Mausam (2018) conjunctive sentences are one of the issues in proposition extraction, as conjunctions are a challenge to dependency parsers (Ficler and Goldberg, 2016) which proposition extraction systems are mostly built upon. Hence, Saha and Mausam (2018) built a system that automatically creates simple sentences from sentences with several conjunctions that are used for proposition extraction. For the proposition extraction of the simple sentences they used ClausIE and OpenIE. They evaluated their data using three different proposition datasets. The correctness of the extracted proposition from the original sentence were evaluated manually. In their study, simple sentences were sentences without conjunctions.

4.1.2 Corpus Creation of Propositions from Simple and Complex Sentences

We create a corpus to evaluate the performance of proposition extraction systems entangled with and disentangled from the task of clause splitting. Figure 4.3 gives an overview of the corpus creation process, while Table 4.4 gives examples as well as the statistics of each step.

The *input* described in the table is the portion of the Aspect Based Sentiment Analysis (ABSA) task (Pontiki et al., 2014) concerned with restaurant reviews within one aspect—*service*.[5] We use all 425 sentences that were annotated with this aspect.

[4] `http://knowitall.github.io/openie/`

[5] Online users' restaurant reviews are a fruitful domain for proposition extraction, as propositions extracted from reviews would be useful for several user-centered tasks, as they would allow to display only information pieces of interest.

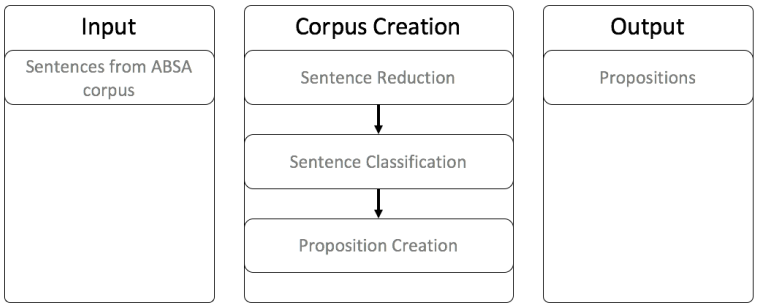


FIGURE 4.3: Corpus creation process of propositions from simple and complex sentences

Step	Method	Count	Example
Input	Download	425 sentences	The service was horrible – the waitress did not know and could not find out costs .
Sentence Reduction	Crowdsourced	2,181 reduced sentences	The service was horrible. The waitress did not know and could not find out costs
Proposition Creation	Expert annotation	2,195 propositions	The service | was | horrible The waitress | did not know | costs The waitress | could not find out | costs
Evaluation	Inter-annotator Agreement		

TABLE 4.4: Corpus creation process for comparison of proposition extraction from simple and Complex sentences

In a preliminary step, the *sentence reduction*, we produce a corpus of 2,181 REDUCED sentences (class distribution in Table 4.5) and 2,526 propositions. To examine the influence of sentence complexity, we classify the reduced sentences as either 1) SIMPLE sentences, meaning sentences with potentially just one proposition, and 2) COMPLEX sentences, meaning sentences with potentially multiple propositions.

Then, in the *proposition creation* step, we produce propositions from the reduced sentences using expert annotation and perform the *evaluation* it by calculating the Inter-Annotator Agreement (IAA).

Definition of Simple and Complex Sentences Quirk (1985) defines a *simple sentence* as a sentence consisting of exactly one independent clause that does not contain any further clause as one of its elements. Hence, a *complex sentence* consists of more than one clause. This is also the definition that we use in our study. By clearly distinguishing between 1) simple sentences, meaning sentences with potentially just one proposition, and 2) complex sentences, meaning sentences with potentially multiple propositions, we will be able to examine the impact of sentence complexity on proposition extraction.

Creating Reduced Sentences As a preliminary step, we created a gold corpus of 2,181 reduced sentences formed from originally 425 complex sentences. The guidelines for creating reduced sentences are attached in Appendix A.1.1.1. A *reduced sentence* is a sentence that contains only a portion of the original sentence. In Example 4.1, (1) is the original sentence, whereas (2) and (3) are possible reduced sentences from (1).

1 The server was cool and served food and drinks.
2 The server was cool.
3 The server served food.

EXAMPLE 4.1: Two reduced sentences (2, 3) from one complex sentence (1)

One reduced sentence was not allowed to be split in further reduced sentences, at least within the output of one worker, in this way trying to prevent nested structures. The intention behind this step was to create sentences with one proposition only. Hence, the guidelines contained rules such as decomposing conjunctive sentences or creating independent sentences from relative clauses. This step turned out to be more difficult than expected, as some sentences contained several factors that could be reduced, as will be discussed in the following. In Example 4.2, the original sentence shown in (1) cannot be split in both reduced sentence (2) and (4), (2) and (5), (3) and (4), or (4) and (5), as these reduced sentences would be overlapping. If one worker produced the reduced sentences (2) and (3), or (3) and (5), it was correct. However, neither reduced sentence (4) nor (5) is optimal, as it can be further split in general.

1 The service was horrible – the waitress did not know and could not find out costs.
2 The service was horrible.
3 The waitress did not know costs.
4 The service was horrible – the waitress did not know costs.
5 The service was horrible – the waitress could not find out costs.

EXAMPLE 4.2: Reduced sentences (2-5) from one complex sentence (1)

However, our guidelines insured that sentences were reduced in comparison to the original version, if a reduction was possible. In this way, we are able to create a sufficiently big set of simple and more complex sentences. We perform this preliminary step via crowdsourcing and evaluate it qualitatively.

We used Amazon Mechanical Turk (AMT) for crowdsourcing our data. We paid \$0.04 per Human Intelligence Task (HIT)[6] and \$0.01 for each extra reduced sentence to ensure that a sentence was reduced as far as possible. Each sentence was reduced by three workers.

To measure the quality of the crowdsourced reduced sentences, we chose 100 random reduced sentences together with their original sentence and evaluated their correctness using

[6] A HIT is an individual task for a worker on AMT

Complexity class	# of Occurrences
NO VERB	101
SIMPLE	1,648
COMPLEX	432
All	2,181

TABLE 4.5: Distribution of sentence complexity classes

Original Sentence	The server was cool and served food and drinks.
REDUCED	The server was cool and served food.
SIMPLE	The server was cool.
GRAMMAR	The server was.
INFERENCE	The server is good.

(A) Classification examples

Sentence Class	#
ORIGINALSIMPLE	20
REDUCED	66
SIMPLE	87
GRAMMAR	5
INFERENCE	12

(B) Classification distribution

TABLE 4.6: Classification of reduced sentences

the following non-exclusive categories: ORIGINALSIMPLE, REDUCED, SIMPLE, GRAMMAR, and INFERENCE.

We provide an exemplary sentence for each category, except for ORIGINALSIMPLE, as it means that the original is already a simple sentence, containing only one proposition which cannot be further reduced. 20 sentences in the random sample were categorized as being ORIGINALSIMPLE. However, some workers still tried to reduce some of these sentences—two of them were grammatically incorrect (GRAMMAR) and three fell into the class INFERENCE, meaning that their content was not explicitly mentioned in the original sentence, but was lexically inferred.

There were 66 REDUCED sentences, which means that the sentences have been successfully reduced to a sentence that is simpler that the original one. 60 of the REDUCED resulted in SIMPLE sentences, which means that they contained only one proposition after the reduction, and six were simpler than the original sentence, but contained more than one proposition.

We believe that the results are usable as is, as the error rate is quite low—only 17 of the reduced sentences in the random sample were incorrect (GRAMMAR and INFERENCE). Furthermore, we show that our reduction step was necessary to produce enough simple sentences for our experiment, as 80% of the random sample were originally complex.

Creating Propositions from Reduced Sentences To evaluate the performance of proposition extraction systems, we created a gold standard corpus for propositions from the reduced sentences. In this study, we follow the most simple possible annotation, similar to Stanovsky et al. (2018). We use the definition that was presented in the beginning of this chapter i.e. we extract propositions with one main verb and all arguments that are linked to it. In our notation, the first position of the proposition is the subject, the second is the predicate and

the order of the other elements is irrelevant.[7] The choice of definition will also be reflected in the performance of systems that do not adhere to our understanding of propositions. However, this does not necessarily cloud the performance comparison of simple and complex sentences, as we will still measure the influence of sentence complexity. Each sentence is processed by two annotators and the disagreements are curated in a subsequent step.

As the creation of propositions is not a trivial task, due to many different cases that need to be explained in the guidelines[8], this task should be performed by people who were trained longer than a crowdsourcing platform allows for. The full guidelines can be found in Appendix A.1.1.2. We produced proposition annotations in a double-annotation process by three graduate students.[9] The disagreements were curated by the author of this thesis. The result of the curation is the gold standard. The gold standard, all annotations, and the guidelines are available.[10]

	SIMPLE		COMPLEX		All	
	A1	Gold	A1	Gold	A1	Gold
A1	-	.80	-	.66	-	.76
A2	.71	.79	.53	.63	.66	.74
A3	.61	.66	.39	.48	.57	.62

(A) Agreement on propositions

	SIMPLE		COMPLEX		All	
	A1	Gold	A1	Gold	A1	Gold
A1	-	.90	-	.77	-	.86
A2	.85	.79	.70	.63	.81	.74
A3	.83	.83	.67	.70	.80	.80

(B) Agreement on proposition elements

TABLE 4.7: Inter-annotator agreement and curator agreement in %-agreement

To evaluate the proposition creation, we report %-agreement between annotators as well as between annotator and curator to evaluate the proposition creation step on proposition level (see Table 4.7A) and proposition element level (see Table 4.7B). Although it is difficult to interpret these results in comparison to other works and we are aware that %-agreement is ignorant of chance agreement, we believe that it is the best measure for this problem, as chance agreement is quite low in the case of this complex annotation problem and we need some measurement for comparing the system results. As previously described, there are no clear guidelines for propositions and also no manual gold datasets created explicitly for this purpose.

On the proposition level, we calculate the agreement on whole propositions. On the proposition element level, we calculate the agreement on individual elements of the propositions whilst taking their label (subject, predicate, or other element) into account.

Table 4.7A shows that the IAA on the proposition level is .39 and .53 on complex sentences and .61 and .71 on simple sentences. These agreement differences show that the underlying clause splitting is also difficult for humans.

[7] We are not interested in different types of objects and modifiers, similar to Stanford, OpenIE, and AllenNLP, and thus we do not discuss this information. For a better overview, we asked the annotators to present the other elements in their order of occurrence.

[8] The guidelines include explanations of what predicates, arguments, and nested propositions are. This in itself is not difficult. However, such instructions consume more time and need more training, as simple mistakes are made by untrained annotators. We saw this in a pilot set for this task is not included or discussed here due to space restrictions.

[9] The result is shown in Table 4.7. A1 annotated the whole set, while A2 and A3 annotated parts.

[10] https://github.com/MeDarina/review_propositions

The agreement with the curator is .05 to .19 higher than the IAA. This could be explained with the annotations of different annotators being actually complementary rather than contradictory e.g. one annotator forgot to split up an argument. The agreement on the proposition element level is .67 and .7 on complex sentences and .83 and .85 for simple sentences—nearly double of the whole proposition agreement. The difference in agreement between complex and simple sentences shows how much manual proposition annotation is affected by sentence complexity.

4.1.3 Evaluation of Proposition Extraction Systems

Our approach is similar to Saha and Mausam (2018), as we also evaluate the performance of proposition extraction systems. However, our main goal is to show the performance of proposition systems on their main task—the extraction of propositions—without the task of clause splitting. By showing the performance of both simple and complex sentences, we are furthermore able to show the impact of clause splitting.

Setup To identify the system that performs best when the issue of sentence complexity is removed, we use the herein produced corpus to analyze and evaluate the performance of various proposition extraction systems as used in evaluations by Stanovsky and Dagan (2016), Gashteovski et al. (2017), Saha and Mausam (2018), and Stanovsky et al. (2018). Hence, we will analyze proposition extraction performance using AllenNLP (Stanovsky et al., 2018), ClausIE (Del Corro and Gemulla, 2013), ReVerb (Fader et al., 2011), Stanford Open Information Extraction (Angeli et al., 2015), OLLIE (Mausam et al., 2012), and OpenIE-5[11].[12] Furthermore, we will provide two baseline systems in order to better compare the system performance.

We use agreement to measure the performance of systems. We measure full agreement, not just matching phrase heads, as performed by Stanovsky et al. (2018). Furthermore, we evaluate only agreement, as in our setup the argument or the predicate matching is what we are interested in, meaning we do not need precision and recall in our setting. In this way, our evaluation setup is similar to Saha and Mausam (2018), who also identified specific issues in proposition extraction systems. As in IAA, we calculate agreement on two levels: proposition and proposition element level. The results of the performance comparison is shown in Table 4.8.

Baselines We provide two baselines in order to better compare the systems. Both baselines create propositions with three elements at most: subject, predicate, and one other element. The first baseline (BL1) takes the first word as subject, the second word as predicate and the rest as one other element. The second baseline (BL2) is somewhat more engineered and uses

[11] `http://knowitall.github.io/openie/`

[12] We do not use MinIE (Gashteovski et al., 2017), as it is an extension of ClausIE containing an on demand removing information on polarity, modality, attribution, and quantifiers. As we are not interested in this information, we will use only ClausIE in our comparison. We use OpenIE-5, which implements the system described by Saha and Mausam (2018), instead of its older version used by Stanovsky and Dagan (2016) and Stanovsky et al. (2018)

POS-tags. It creates a proposition for each verb. All words before the verb are the subject and all words after the verb are one other element. Examples for the baselines are shown in Table 4.2.

System performance Table 4.8A shows that performance of proposition extraction on propositions is equally bad for both simple and complex sentences. Table 4.8B shows that performance on individual proposition elements is much better than on proposition level. Furthermore, the table shows that all systems, except ReVerb, perform much better on the simple sentences, which was expected.

Systems	SIMPLE	COMPLEX	All
Allen	.08	.09	.08
ClausIE	.06	.09	.07
ReVerb	.02	.02	.02
Stanford	.01	.01	.01
OLLIE	.03	.04	.03
OpenIE	**.09**	**.12**	**.09**
BL1	.00	.00	.00
BL2	.00	.00	.00

(A) System performance on propositions

Systems	SIMPLE	COMPLEX	All
Allen	.50	.40	.46
ClausIE	.37	.36	.36
ReVerb	.15	.14	.14
Stanford	.20	.09	.17
OLLIE	.24	.19	.22
OpenIE	**.51**	**.42**	**.47**
BL1	.05	.04	.05
BL2	.26	.24	.21

(B) System performance on elements

TABLE 4.8: System performance measured in accuracy

It is also noteworthy that although the performance of both baselines on whole propositions is 0, the performance of the second baseline on proposition elements is competitive. This shows that the task of proposition extraction can, to a big part, be solved by correct verb extraction. BL2 outperforms ReVerb, Stanford, and on simple and complex sentences also OLLIE. Note that BL2 performs slightly worse on all sentences, as these, additionally to the simple and complex sentences, also include sentences without a verb and this baseline is verb-based. This indicates that either the automatic systems have problems with the extraction of verbs or they have deeper issues, e.g. they do not extract from a lot of sentences, as is discussed in Section 4.1.4. The fact that the second baseline performs almost equally on both simple and complex sentences may show that correct verb extraction alone solves only a particular portion of proposition extraction. Other systems, especially the two best ones, perform about two times better on the simple sentences but then have a much bigger drop on the complex sentences. This reveals that sentence splitting has a bigger impact on better or probably more intelligent systems than on more simple systems.

On both levels, OpenIE is the best system, very closely followed by Allen whereas the other systems are left behind.

4.1.4 Analysis of System Performance

Identifying further problems except clause splitting could improve current proposition extraction systems. On the one hand there are sub-issues in clause splitting. On the other hand, there are issues besides clause splitting.

In the case of ClausIE and ReVerb, many further clauses and also arguments are cut, as these consist of a maximum of three elements, which makes the comparison difficult.

Systems	Miss.	Cond.	Temp.
Allen	.08	.13	.19
ClausIE	.06	.11	.13
ReVerb	.03	.00	.03
Stanford	.02	.00	.00
OLLIE	.04	.06	.02
OpenIE	.10	.19	.17

(A) System performance on propositions excluding sentences with missing propositions (Miss.), conditional clauses (Cond.), and temporal clauses (Temp.)

Systems	Miss.	Cond.	Temp.
Allen	.50	.57	.55
ClausIE	.38	.40	.38
Stanford	.26	.03	.14
ReVerb	.32	.00	.21
OLLIE	.31	.00	.20
OpenIE	.54	.53	.50

(B) System performance on proposition elements excluding sentences with missing propositions (Miss.), conditional clauses (Cond.), and temporal clauses (Temp.)

TABLE 4.9: System performance excluding sentences with specific issues

General issues We first manually examined some potential issues in the proposition extraction from simple sentences. After the manual analysis of potential issues, we calculated the system performance if the issue would be eliminated. One big issue we found is *missing propositions*, meaning that systems do not always extract propositions. Except for the missing propositions, there was no big difference in the system performance with or without the issue. Also, some systems have different models of propositions, which may also affect their performance. On the one hand, there are issues with previous steps, e.g. *negations* or *quantifiers* are ignored. On the other hand, there are issues with formatting, e.g. a different treatment of *prepositions* or conditionals.

Missing propositions One big issue is that proposition extraction systems often do not produce any extraction from a sentence. Unsurprisingly, this issue is bigger among the systems that do not perform well—namely ReVerb (58% of sentences do not have an extraction), Stanford (39%), and OLLIE (33%), whereas the better performing systems have much lower rates—Allen (3%), ClausIE (4%), and OpenIE (10%). In ReVerb, Stanford, and OLLIE, we could not find a clear reason why there are no extractions. In the case of Allen, there are only no extractions from sentences without verbs.[13] ClausIE and OpenIE have no extractions from sentences that are missing a verb or a subject. Additionally, OpenIE has no extractions from existential clauses.

In Table 4.9A, where we show the performances of systems on full propositions without the discussed issues, it is shown that systems perform slightly better when eliminating missing propositions from simple sentences. However, the improvement is clearer in Table 4.9B on the element level. Especially for the systems that had more missing propositions, namely Stanford, ReVerb, and OLLIE, the change is between .06–.17.

[13]These sentences are classified as neither simple nor complex, but are included in all.

Conjunctions As already stated by Saha and Mausam (2018), conjunctive sentences pose an issue to proposition extraction systems. In our case, we wanted to separate all conjunctive sentences in individual propositions, e.g. the sentence "The waitress smiled at her friend and at me." contains the propositions "The waitress | smiled | at her friend" and "The waitress | smiled | at me.". OpenIE and Stanford have the same rules on conjunctions, whereas Allen, ClausIE, and ReVerb keep the conjuncted elements together—from the previous sentence they would create one proposition—"The waitress | smiled | at her friend and me.".

Negations Stanford does not extract from negated sentences, while the rest can deal with negations. These specific problems are difficult to show in numbers, as they are rare—only about 7% of the sentences contained negations.

Prepositions OLLIE, ReVerb, and Stanford place the prepositions with the predicate, whereas all other systems as well as our gold standard place it with the associated argument, as is shown in the example in Table 4.2. For these cases we would need adjusted evaluations that ignore this difference.

Quantifiers Stanford ignores "every" in propositions.

Issues with complex sentences We looked at issues within complex clauses, namely conditional and temporal clauses.

Conditional clauses In some cases, Allen, ClausIE, OLLIE, and OpenIE extract the if-clause for the argument, but delete the "if", which leads to disagreements on both full proposition and proposition element level. For instance, the complex sentence "You can get a table if you get there early." should be split in the propositions "You | can get | a table | if you get there early" and "you | get | there | early'. OpenIE, however, extracts the second proposition correctly, but the first proposition is 'You | can get | a table". Comparing the performance on all complex clauses as shown in Table 4.8A to complex clauses without conditional clauses, as shown in Table 4.9A, all systems, except for ReVerb and Stanford, clearly perform better. Allen is better by .04 and OpenIE by .05, which shows that they have the biggest issues with conditional clauses. On proposition element level this becomes even clearer. Here, the three better systems, ClausIE, Allen, and OpenIE perform .04–.17 better without conditional clauses.

Temporal clauses Conceptually, Allen, OLLIE, and OpenIE extract temporal clauses correctly, but have some problems if the sentence is too long. Stanford cuts out the "when". For temporal clauses, the performance is similar to conditional clauses. The three better systems perform .06 -.11 better on full proposition level, and .02-.09 better on proposition element level. Stanford and OLLIE perform worse without the temporal clauses.

4.1.5 Conclusion on Influence of Sentence Complexity

We created a dataset with sentences classified as simple and complex. The dataset enabled us to research the performance of proposition extraction detached from the task of clause splitting. On the one hand, we showed that sentence complexity has a measurable impact on proposition extraction performance of both humans and machines. Hence, one step towards improving the performance of such systems, is the improvement of clause splitting. Furthermore, we believe that the performance of the original complex sentences, without the preliminary reduction step, would pose an even bigger problem to proposition systems, which implies that using these systems on real data could be problematic.

On the other hand, our study also showed that the ranking of systems is similar among simple and complex sentences. This means that the best-performing systems among simple sentences which are disentangled from the task of clause splitting, are also the best on complex sentences, where clause splitting also needs to be performed. This may mean that to find the overall best system, one does not need to classify between simple and complex sentences. However, to find that one of the problems of proposition extraction is sentence complexity, it is necessary.

Also, our intelligent baseline system was able to extract verbs, outperformed three of the systems. However, the better systems did not only perform much better, but they were also more affected by sentence complexity.

Additionally, we looked into further problems of proposition extraction systems. The main issues in complex sentences that we could identify were conditional and temporal clauses. While some of the issues are due to different proposition models, which e.g. may include or exclude keywords of such clauses such as “when”, “if”, or “unless”, this may also show the difficulties of parsing such sentences. Furthermore, it may be a clearer indicator on how to improve the systems.

4.2 Compositionality of Granularity Levels

Apart from propositions, statements can be seen on several levels of granularity. In our corpus study described in Benikova and Zesch (2017), we researched the compositionality of semantic relations on three granularity levels—namely sentence level, proposition level, and proposition element level. We do so on the example of the paraphrase relation. Madnani and Dorr (2010) discuss that paraphrases exist on several granularity levels, namely *sentences*, *phrases*, and *individual lexical items* (or *words*). The study performed in Benikova and Zesch (2017) will be discussed in this section. In Kovatchev et al. (2020), we performed another compositionality study, without a strict restriction on granularity levels, for all relations. This study is discussed in Chapter 5.2.

Problem Description We argue that working on the sentential level is not optimal for both machines and humans, and that it would be easier and more efficient to work on sub-sentential levels.

Annotating all relations would have been too costly in terms of time and money. The paraphrase dimension is most robustly annotatable (see Chapter 5). Therefore, we perform the compositionality study on this dimension.

Solution Idea By building a new corpus with paraphrase annotations on three different levels, we are able to quantify and analyze the difference between paraphrases on both sentence and sub-sentence level in order to show the significance of the problem. Figure 4.4 shows the corpus creation process. Our final corpus consists of 88 sentence pairs with 161 proposition pairs.

Outcome Although we could not prove that human annotation performance is better on the proposition level, the compositionality analysis shows that this level is the way to go when trying to find more complex paraphrases on a sub-sentential level. However, we must admit that our sample size is quite small and thus our findings may not generalize.

4.2.1 Related Work

Although many approaches in NLP are build on the assumption of semantic compositionality (Sammons et al., 2010), to our knowledge, there has been no explicit and empirical analysis of the paraphrase compositionality on different independently annotated levels prior to our work[14]. However, next to approaches on each individual level, there have been several approaches where different granularity levels have been annotated in one corpus.

Paraphrases on the Individual levels Paraphrase detection is mostly performed on the sentence level (Dolan and Brockett, 2005; Ganesan et al., 2010; Xu et al., 2014; White et al., 2015; Socher et al., 2011; Fernando and Stevenson, 2008). Propositions or other kinds of predicate-argument structures have been previously used in paraphrasing and closely related tasks (Roth and Frank, 2012; Xu et al., 2014; Shwartz et al., 2017; Li and Ji, 2016), as they are considered to contain the most salient information in a form that is easier to process than full sentences.

Paraphrases on Different Levels Cohn et al. (2008) performed an annotation on all three levels in parallel, by using existing sentential paraphrase corpora such as the MRPC and adding the other two layers upon those. In this study, the lowest paraphrase level is the proposition element level, which is seen in context of its sentence.

Classification of Paraphrases Cabrio and Magnini (2013) and Vila et al. (2014) classified paraphrases according to paraphrase classes and also classified lexically differing parts within the pairs according to the same classification. Similarly, Sammons et al. (2010) took existing textual entailment corpora that are classified according to classes including paraphrases and

[14] In Kovatchev et al. (2020), we also look at compositionality of paraphrases and other dimensions, but this work came after Benikova and Zesch (2017).

classified the *arguments* according to paraphrase classes.[15] In Kovatchev et al. (2020), we combine different existing typologies to classify several relations on sub-sentential level (see Chapter 5).

4.2.2 Annotatability of Paraphrases on Different Granularity Levels

It is likely that annotation of paraphrases on the different levels is of different difficulty. However, a comparison is challenging, as there are few studies and they are not comparable between levels.

SemEval 2015 Task 1 (Xu et al., 2014) was the task to find paraphrases and semantic similarity between tweets. The therein used data is based on the TPC, as is also part of our corpus. The IAA measured in terms of the F_1-score is .82. Assuming the tweets are roughly equivalent to 'sentences'—we can use this measurement for comparing on the sentence level. In MRPC, which our corpus is also partly based on, the IAA in terms of Cohen's κ was .62 on the sentence level (Dolan and Brockett, 2005), which the corpus was annotated on. For the phrase level, Cohn et al. (2008) report an F_1 IAA between .71 and .76. They also report IAA on the word level, which is between .74 and .79. In line with our hypothesis, IAA is higher on the word level than on the phrase level, but they did not compare their results to the sentence level. We cannot directly compare with the results from the Twitter dataset, as the definitions of the levels differ.

In our study, we annotate a single dataset on all three levels in order to gain insights on which level works best and possibly also how to break down the task of paraphrase detection.

4.2.3 Corpus Creation of Paraphrases on three Different Levels

We created the first corpus of paraphrases with parallel annotations on three granularity levels. The corpus creation process is shown in Figure 4.4, while Figure 4.5 shows exemplary annotations of a sentence on three levels.

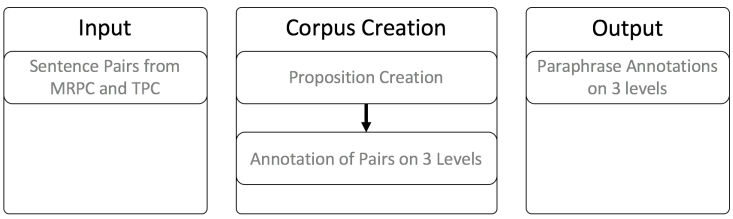


FIGURE 4.4: Corpus creation process of paraphrases on three granularity levels

Source Data We annotate paraphrases on existing sentential paraphrase corpora, such as the MRPC (Dolan and Brockett, 2005) and the TPC (Xu et al., 2014). Our dataset is based on 41 sentence pairs from the MRPC and 47 tweet pairs from the TPC. We choose these corpora because, (i) they have been widely used, which makes our approach comparable to others,

[15]Two sentences entailing each other are considered a paraphrase by many definitions (Rus et al., 2014; Hovy et al., 2013).

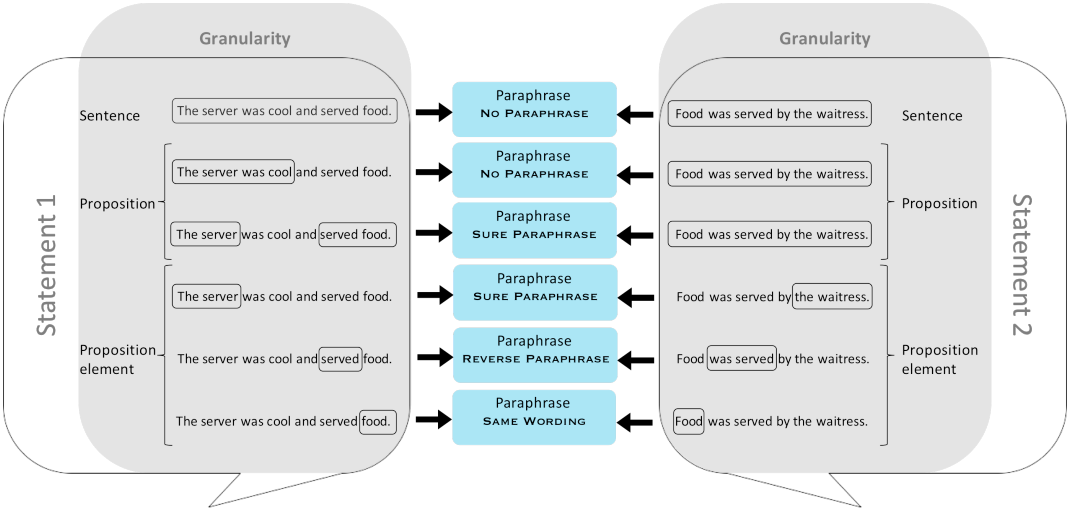


FIGURE 4.5: Paraphrase levels annotated in our model

(ii) they also contain negative examples of paraphrases, and (iii) the Twitter corpus contains non-standard data, which provides a good basis for the robustness of our model.

Annotation Setup The full set of sentence pairs was annotated by one annotator, the author, while 75% of the corpus where also annotated by a second annotator, a graduate linguistics student, in order to show the reproducibility via IAA. Figure 4.5 shows an example of the annotation on all three levels. In the first step, the annotators re-annotated the sentential paraphrases. This step is done in order to analyze the compositionality of the granularity levels and to compare the annotatability between them. In the second step, they annotate the proposition paraphrases in each sentence pair. In the third step, the proposition element paraphrases in each proposition pair are annotated. The sub-sentential tasks can be basically separated in two subtasks: (i) finding the propositions and (ii) aligning paraphrased propositions and elements.

Finding propositions For the propositions, we use the definition described in the introduction of this chapter. *Proposition elements* are verbal predicates and their arguments. The predicate is always one word. In general, arguments can span from lexical items to phrases, but have clear boundaries with regard to the verb. In this study, only verb-verb and argument-argument paraphrases are considered. Additionally to the definition above, in this study, we do not consider words that are not arguments of a verbal predicate such as conjunctions (e.g. *so*, *because*, *if*) or interjections (e.g. *oh*, *wow*, *hello*).

By using a dependency parser we simplify the task of finding the proposition with its verb-argument structure. We pre-annotate the first sentence of each pair with the Stanford Dependency Parser in the version provided by DKPro Core (Eckart de Castilho and Gurevych, 2014), using the Recurrent Neural Network (RNN) model with collapsed dependencies, as this parser worked best for our purpose. The difficulties of automatic proposition detection are discussed in Section 4.1. Unfortunately, the quality of the parsing output was insufficient and had to be manually corrected for both datasets. This was mainly due to arguments with

	Sentence	Proposition	Element
κ	.61	.55	.73
F	.91	.88	.93

TABLE 4.10: Inter-annotator agreement on the three granularity levels

very long spans in MRPC and large amounts of non-standard language in TPC. The proposition is then connected with its elements by marking the span between the elements. Thus, the annotators are shown the annotation of the proposition and the individual elements before performing the alignment annotation.

Aligning propositions and elements The annotation of paraphrases is performed by an alignment annotation between two instances on the same granularity level in a sentence pair. The alignment annotation is performed on each level independently, in this way reducing the bias of annotating similarly on all levels on purpose. If there is no paraphrase, the annotators do not perform any alignment. Figure 4.5 shows exemplary annotations on all three levels. Additionally to a simple alignment, we measure the confidence on the particular alignment. Hence, we distinguish between

- SAME WORDING
- SURE PARAPHRASE
- UNSURE PARAPHRASE

on all three levels.

Furthermore, there are special alignment labels on the *proposition element level* or more specifically the verb that we created for antonyms and passives—REVERSE PARAPHRASE[16]. In Figure 4.5, "served" and the passive form "were served" are annotated as REVERSE PARAPHRASE, as subject and object are switched.

Annotatability The results in Table 4.10 show that in general the IAA is rather high for a task of that difficulty. For the sake of comparability we also report F_1-score, but using chance-corrected measures like Cohen's κ is certainly more appropriate. κ on the sentence level is .01 lower than the original MRPC annotation. F_1-score is higher than in previous studies, but not directly comparable. For both measures, we do not observe the expected result that smaller units get higher agreement. While elements are clearly easier than sentences, propositions are even harder than sentences. As also found in the study described in the previous section Section 4.1, manual proposition annotation poses difficulties to annotators. This may be rooted in difficulties to clearly define propositions. As our sample size is rather small, no definitive conclusions should be drawn from these results.

[16]For verbs, we also have several spacial labels that catch the change in its semantics e.g. verbal negations, modal verbs and multi-word verbs. However, as this is not essential to paraphrases or their compositionality, we do not discuss this further. Details may be found in Appendix A.1.1.3.

4.2.4 Evaluation of Paraphrases on Different Granularity Levels

In the trivial case of two identical sentences, they are paraphrases and so are all the propositions they consist of. The same holds for each of the identical propositions from the two sentences where each of the proposition elements has a perfect match on the other side. However, there certainly are sentence paraphrases, where there is no such perfect overlap—and we are more interested in these cases. In these cases, it is an open question whether the issue of sentence paraphrases can be settled by only looking at the propositions or the issue of proposition paraphrases by looking at the proposition elements.

It is often hard to decide whether two sentences are indeed paraphrases due to only partially overlapping content as shown in Example 4.3:

1 The waitress gave a drink to the customer.
2 The waitress served a drink, as it's her job.

EXAMPLE 4.3: Example for arguable paraphrases

The decision of whether (1) and (2) are paraphrases is difficult, because only a part of (2) has the same content, whereas the rest is additional information.

Compositionality Using our newly created paraphrase annotations on the three granularity levels, we can now turn towards the question of compositionality. Similar to the study in the previous section, in this analysis we also differentiate between SIMPLE and COMPLEX sentences. We perform two analyses: first we check the compositionality between all three granularity pairings to empirically analyze whether paraphrases are compositional in general. For instance, in Example 4.3, on the sentence level, this complex sentence could be an UNSURE paraphrase. On the proposition level, there is a SURE paraphrase—"The waitress gave a drink to the customer" and "The waitress served a drink". On the proposition element level, there would be two SAME wording paraphrases—"The waitress" and "a drink"—and one sure paraphrase—"gave [...] to the customer" and "served".

Afterwards, we compare the differences of the higher classes in more detail in order to show the advantages of working on lower granularity levels.

All granularity levels Table 4.11 shows the results of the averaged percentage values between the paraphrase classes of two granularity levels. Tables 4.11A and 4.11B show that 67%-71% of SURE PARAPHRASE sentence pairs consist of SURE PARAPHRASE proposition pairs. Table 4.11B shows that simple sentence pairs that are not paraphrases do not contain any SAME WORDINGS or SURE PARAPHRASES on the proposition level. Especially when looking at the compositionality between the higher levels and elements, it is clear that there is a big lexical overlap, as SURE PARAPHRASES on the proposition level consist of 48% SAME WORDING element pairs. Furthermore, the figures show that proposition elements having the same wording are the most frequent label in each higher leveled paraphrase class.

		Sentence		
		SURE	UNSURE	NO
Proposition	SAME	11	0	0
	SURE	67	20	12
	UNSURE	22	20	0
	NO	0	60	88

(A) Simple sentences and propositions

		Sentence		
		SURE	UNSURE	NO
Proposition	SAME	13	5	19
	SURE	71	32	10
	UNSURE	10	37	24
	NO	6	26	48

(B) Complex sentences and propositions

		Proposition		
		SURE	UNSURE	NO
Element	Same	48	40	28
	SURE	46	12	17
	UNSURE	1	6	3
	NO	6	42	53

(C) Proposition and proposition elements

		Sentence		
		SURE	UNSURE	NO
Element	SAME	54	32	22
	SURE	40	31	15
	UNSURE	1	1	5
	NO	6	35	58

(D) Sentence and proposition elements

TABLE 4.11: Compositionality of the three granularity levels in percent

Although they are more often components of SURE PARAPHRASES on the higher levels, they are also present in higher leveled instances that are not paraphrases.

This means that although SURE PARAPHRASES are composed of SURE PARAPHRASES or SAME WORDING, these two labels are also present in instances that are not paraphrases, which may be due to the highly lexically overlapping construction of the source datasets, as discussed by (Rus et al., 2014). In any case, it means that only looking at the paraphrases on the lower levels is not sufficient to decide over paraphrases on the higher levels and other features need to be also considered, as pairs that are not paraphrases on the sentence and proposition level also contain 22% or 28% of proposition elements that are of the label SAME WORDING.

All tables show that both SURE PARAPHRASE and NO PARAPHRASE primarily consist of the identical labels on the lower levels, or in the case of the first possibly also of SAME WORDING i.e. it is most likely that if a paraphrase surely exists on a higher level, its lower-leveled components have the same paraphrase label (or are verbatim) and if there is no paraphrase on the higher level, there is also no paraphrase on the lower levels. This shows that paraphrases are compositional in most cases, especially when regarding simple or lexically highly-overlapping sentence pairs.

Proposition element paraphrases are nearly never UNSURE, meaning that insecurities about whether pairs are paraphrases are more frequent on the higher levels. UNSURE PARAPHRASES on the higher levels consist of different components, meaning that a clearer definition of paraphrases could improve the security on paraphrase annotation.

Sentence level vs. Proposition level To compare the differences of paraphrases on the upper two granularity levels, we consider three different cases, namely: 1) Same paraphrase label, 2) proposition paraphrase only, and 3) sentence paraphrase only.

Same paraphrase label This is the case of full compositionality, meaning that the paraphrase pair of the higher level consists of paraphrase pairs on the lower level that have the same label as the higher level.

The compositionality of sentences with one or with several propositions differs slightly, although most sentences consist of propositions with the same paraphrase label as the sentence. Table 4.11A shows that sentences with only one proposition have paraphrase labels differing from that of their proposition in 33% of the cases, of which 11% are SAME WORDING, which means that 78% of SURE PARAPHRASES consist of either SURE PARAPHRASE or SAME WORDING proposition pairs. In Example 4.4 we show a sentence pair that is a SURE PARAPHRASE on both sentence and proposition level.

---

1 The waitress gave a drink to the customer.
2 The waitress served a drink.

EXAMPLE 4.4: Example for SURE PARAPHRASE on both sentence and proposition level

Sentence pairs that are labeled as NO PARAPHRASE in 88% of the cases consists of proposition pairs that are also labeled as NO PARAPHRASE. In our source corpora, these sentence pairs are sometimes unrelated, as shown in Example 4.5. In this example, there is NO PARAPHRASE on any level.

---

1 The waitress gave a drink to the customer.
2 The hotel room was cozy.

EXAMPLE 4.5: Example for NO PARAPHRASE on any level

Proposition paraphrase only A proposition paraphrase only means that part of the sentence is a paraphrase of another part of the other sentence, but the full sentences are not paraphrases. Example 4.6 shows a statement pair that is NO PARAPHRASE on the sentence level but has a SAME WORDING on the proposition level—"You can get a table".

---

1 You can get a table only if you get there early
2 You can get a table.

EXAMPLE 4.6: Example for arguable paraphrases

Additionally to the finding of simple sentences having homogeneous labels with their propositions, Table 4.11B, shows that sentences with multiple propositions also contain propositions with differing labels. This shift is especially prominent in the case when complex sentences are NO PARAPHRASE, but 10% of them are SURE PARAPHRASE and 19%

are of SAME WORDING, which is also the previously discussed case of partially overlapping information.

Sentence paraphrase only This means that the full sentences are paraphrases of each other, but the propositions mentioned in them are distinct. This may occur especially in cases were the information in the sentence is not expressed through verb-argument structures as considered in this work, as e.g.

1 The waitress' professionalism made her serve the drink.
2 The waitress is professional.

EXAMPLE 4.7: Example for paraphrases not expressed in proposition form

In our dataset, there is no case of uni-directional paraphrases.

4.2.5 Conclusion on Compositionality of Granularity Levels

In this study, we have examined the compositionality of paraphrases on different levels by analyzing our newly produced corpus, which was manually annotated with paraphrases on three granularity levels—namely the sentence, proposition, and proposition element level. Although we could not prove that human annotation performance is better on the proposition level, the compositionality analysis shows that this level is the way to go when trying to find more complex paraphrases on a sub-sentential level. However, we must admit that our sample size is quite small and thus our findings may not generalize.

4.3 Conclusion on Proposition as a Representation

In Section 4.2, we showed that on the propositional level, we find more complex and more interesting paraphrases than on the sentence or word level. In our study, 29% of complex sentences which are not paraphrases contain propositions which are paraphrases. This means working on the sentence level one misses propositions of interest, which might be a big loss in applications using the relation dimensions. In our running example of user-specific hotel reviews, we might loose relevant information within the reviews or not cluster all paraphrases and thus not choose the best-fitting statement out of a paraphrase cluster when working on the sentence level. Furthermore, this study shows that complex sentences pose a bigger issue for relation annotation, as they potentially contain relations of interest on the proposition level. In Section 4.1, we find that complex sentences pose a bigger issue to proposition extraction than simple sentences. Therefore, in practice, complex sentences seem to pose an issue to relation annotation, regardless of the granularity level. Hence, sentence splitting needs to be improved to improve performance and research on both proposition extraction and relation annotation. However, even though error-prone in the proposition extraction, we would continue to work on the proposition level for relation annotation, as 1) simple sentences would be processed correctly anyway, and 2) the complex sentences from which

we could correctly extract propositions probably contain more relations of interest than when working on the sentence level.

In the next chapter, we will shift from proposition types to proposition dimensions. More specifically, the next chapter discusses relations between various semantic dimensions.

Chapter 5

Relations between Semantic Dimensions

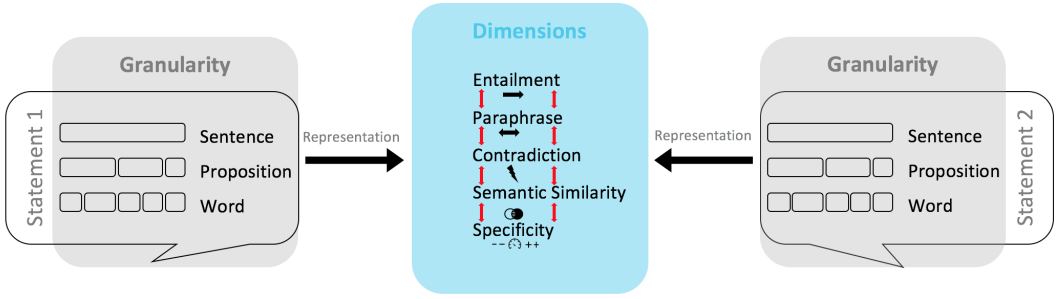


FIGURE 5.1: Illustration of relations between dimensions in this thesis

This section is concerned with the relations between semantic dimensions. The dimensions of interest in this chapter are textual entailment, paraphrases, contradiction, and semantic similarity. Their definitions are discussed in Chapter 3.[1] Although relations between individual pairs of the discussed dimensions have already been considered, e.g. between entailment and paraphrasing (Bosma and Callison-Burch, 2006; Madnani and Dorr, 2010; Androutsopoulos and Malakasiotis, 2010; Sukhareva et al., 2016; Kovatchev et al., 2018a) or semantic similarity and entailment (Castillo and Cardenas, 2010; Yokote et al., 2011; Marelli et al., 2014; Cer et al., 2017), there has been no work on several dimensions simultaneously and their relations between each other. The relations between all these dimensions have first been discussed in a workshop organized by us in Kovatchev et al. (2019a) and analyzed in Gold et al. (2019). The most prominent example for an assumed link between relations is that between entailment and paraphrasing. Bi-directional entailment is often regarded as paraphrasing (Madnani and Dorr, 2010; Androutsopoulos and Malakasiotis, 2010; Sukhareva et al., 2016). Bosma and Callison-Burch (2006) even used paraphrasing to solve entailment. Example 5.1 (which repeats the example in Table 1.2), shows two statements in a paraphrase relation. However, they are also in a bi-directional entailment relation. Furthermore, the statement pair in Example 5.1 has a high semantic similarity score.

[1] In the studies discussed in this chapter, we also studied the relation of the mentioned dimensions and specificity. However, as specificity is discussed individually in Chapter 6, the relations towards this dimension will be discussed therein.

1 The server served food.
2 Food was served by the waitress.

EXAMPLE 5.1: Statement pair in paraphrase relation

An analysis of these relations does not only give empirical insights on themselves, it may also help to automatically calculate or more efficiently represent the individual dimensions, e.g. the similarity score could be an indicator for an entailment relation.

Furthermore, empirically analyzing the relations between dimensions, we prove and disprove assumptions that have been posted but not empirically shown on the links between the relations. For instance, we found that in a small part of our bi-directional entailments were not paraphrases.

Links between Relations In Gold et al. (2019), we create and annotate the first corpus with all of the relations of interest in parallel. We present a corpus creating methodology which contains all the dimensions of interest. Furthermore, the methodology creates a corpus with more lexical diversity between the pairs than other paraphrase and entailment corpora. Having created the corpus, we analyze the relations between all the dimensions. This study is discussed in Section 5.1.

Compositionality of Relations In Kovatchev et al. (2020), we perform a study on all pairings from the corpus in Gold et al. (2019) where there was a relation dimension and further analyze the compositionality of dimensions. In this work, we successfully build a framework and typology for studying and processing multiple meaning relations. This framework facilitates the analysis and comparison of the different relations and may improve the transfer of knowledge between them. Furthermore, we show that the discussed relations have similar underlying linguistic and reasoning phenomena in the decomposition process. This study is discussed in Section 5.2.

5.1 Links between Relations

Problem Description Despite the interactions and close connection of these meaning relations, to our knowledge, prior to our work (Gold et al., 2019), there existed neither an empirical analysis of the connection between them nor a corpus enabling it.

Solution Idea We bridge this gap by creating and analyzing a corpus of sentence pairs annotated with all discussed meaning relations.

Outcome We empirically analyze the relations between dimensions and are able to confirm and contradict previously made assumptions on relations between dimensions. For instance, on the one hand we find that the assumption that bi-directional entailment is equal to paraphrase can be confirmed in most cases, but on the other hand, we found that a small part

of our bi-directional entailments were not paraphrases. An example from our actual data is shown in Example 5.2. [1] and [2] entail each other, as they make each other true. A human reading [1] would infer that [2] is most likely true, as if one needs to study a specific language, "Latin", to understand a specific book, "the Bible", then the specific book is most probably written in this language. Vive versa, a human reading [2] would infer that [1] is most likely true—if the book is written in that specific language, one needs to know that language to understand the book. However, [1] and [2] are not paraphrases, as they do not have the same content—[1] is about the requirement to read a book, the other one is the statement of the language that the book is written in.

1 Reading the Bible requires studying Latin.
2 The Bible is written in Latin.

EXAMPLE 5.2: Bi-directional entailment pair that is not a paraphrase

5.1.1 Related Work on Links between Relations

Although our work was the first to directly compare all the discussed meaning relations, there has been some work on the interaction between some of the discussed meaning relations, especially on the relation between entailment and paraphrasing, and also on how semantic similarity is connected to the other relations.

Our analysis finds that previously made assumptions on some relations (e.g. paraphrasing being bi-directional entailment (Madnani and Dorr, 2010; Androutsopoulos and Malakasiotis, 2010; Sukhareva et al., 2016)) are not necessarily right in a practical setting. In our corpus, we also find that contradictions are often perceived as dissimilar, although some previous work suggests otherwise (Marelli et al., 2014).

Interaction between Entailment and Paraphrases According to Madnani and Dorr (2010) and Androutsopoulos and Malakasiotis (2010), bi-directional entailment can be seen as paraphrasing. Furthermore, according to Androutsopoulos and Malakasiotis (2010), both entailment and paraphrasing are intended to capture human intuition. Kovatchev et al. (2018a) emphasize the similarity between linguistic phenomena underlying paraphrasing and entailment. There has been practical work on using paraphrasing to solve entailment (Bosma and Callison-Burch, 2006).

Interaction with Semantic Similarity Cer et al. (2017) argue that to find paraphrases or entailment, some level of semantic similarity must be given. Furthermore, Cer et al. (2017) state that although semantic similarity includes both entailment and paraphrasing, it is different, as it has a gradation and not a binary measure of the semantic overlap. Based on their corpus, Marelli et al. (2014) state that paraphrases, entailment, and contradiction have a high similarity score; paraphrases having the highest and contradiction the lowest of them.

There also was practical work using the interaction between semantic similarity and entailment: Yokote et al. (2011) and Castillo and Cardenas (2010) used semantic similarity to solve entailment.

Corpora with Multiple Semantic Layers There are several works describing the creation, annotation, and subsequent analysis of corpora with multiple parallel phenomena.

SICK is a corpus of around 10,000 sentence pairs that were annotated with semantic similarity and entailment in parallel (Marelli et al., 2014). As it is the corpus that is the most similar to our work, we will compare some of our annotation decisions and results with theirs.

Sukhareva et al. (2016) annotated subclasses of entailment, including *paraphrase, forward, revert*, and *null* on propositions extracted from documents on educational topics that were paired according to semantic overlap. Hence, they implicitly regarded paraphrases as a kind of entailment.

5.1.2 Corpus Creation

To analyze the interactions between semantic relations, a corpus annotated with all relations in parallel is needed. Hence, we develop a new corpus-creation methodology which ensures all relations of interest to be present. Figure 5.2 shows an overview of our generation, methodology, while Figure 5.3 provides examples and annotations taken from our actual corpus for each step. First, we create a pool of potentially related sentences. This part of our methodology differs significantly from the approach taken in the SICK corpus (Marelli et al., 2014). They do not create new corpora, but rather re-annotate existing corpora, which does not allow them to control for the overall similarity between the pairs. Second, based on the pool of sentences, we create sentence pairs that contain all relations of interest with sufficient frequency. This contrasts existing corpora on meaning relations that are tailored towards one relation only. Finally, we take a portion of the corpus and annotate all relations via crowdsourcing.

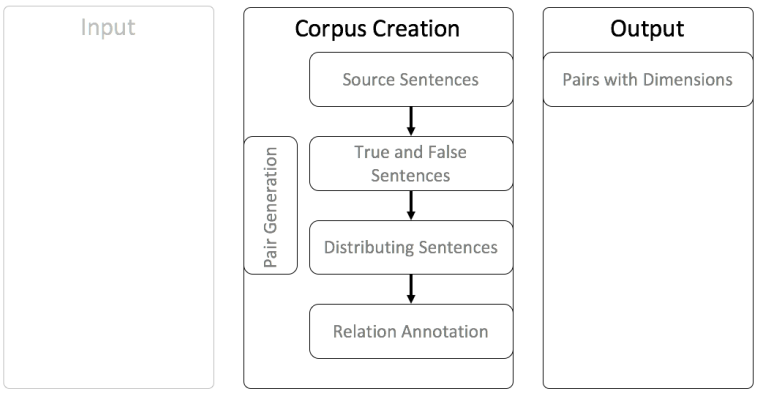


FIGURE 5.2: Corpus creation process of relation dimensions studied in this thesis

Sentence Pool In the first step, we create 13 sentences, henceforth *source sentences*, shown in Table 5.2. The sentences are on three topics: *education*, *technology*, and *language*. We

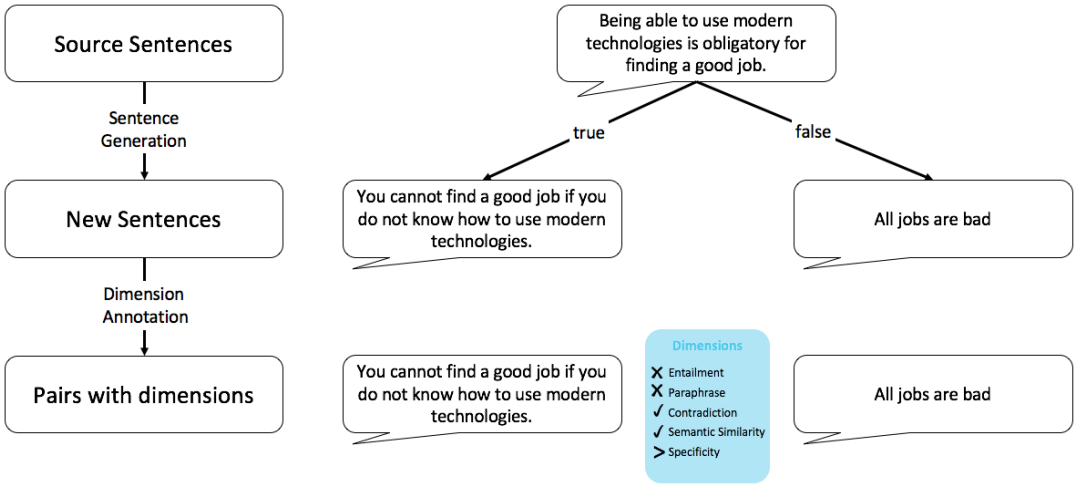


FIGURE 5.3: Workflow and examples for generating sentence pairs with semantic relations

choose sentences that can be understood by a competent speaker without any domain-specific knowledge and which due to their complexity potentially give rise to a variety of lexically differing sentences in the next step. Then, a group of 15 people[2], further on called *sentence generators*, is asked to generate *true* and *false* sentences that vary lexically from the source sentence.[3] Overall, 780 sentences are generated. The 13 *source sentences* are not considered in the further procedure. For creating the *true* sentences, we ask each sentence generator to

Being able to use modern technologies is obligatory for finding a good job.
Christian clergymen learn Latin to read the bible.
Getting a high educational degree is important for finding a good job, especially in big cities.
Going to school socializes kids through constant interaction with others.
In many countries, girls are less likely to get a good school education.
Learning a second language is beneficial in life.
Machines are good in communicating with people.
Machines are good in strategic games such as chess and Go.
Modern assistants such Cortana, Alexa, or Siri make our everyday life easier by giving quicker access to information.
New technologies lead to asocial behavior by e.g. depriving us from face-to-face social interaction.
One important part of modern education is technology, if not the most important.
Self-driving cars are safer than humans as they don't drink.
Speaking more than one language helps in finding a good job.

TABLE 5.2: List of given source sentences

create two sentences that are true and for the *false* sentences, two sentences that are false given one source sentence (that is considered to be true). This way of generating a sentence pool is similar to that of the textual entailment Stanford Natural Language Inference (SNLI) corpus (Bowman et al., 2015), where the generators were asked to create true and false captions for given images. Example 5.3 shows a true (T) and a false (F) sentence created from one source sentence (S).

[2]These people were graduate students, PhD students, and post-docs, who are non-natives in English. We did not control for any knowledge domain bias, as the main objective behind this step was creating sentences potentially containing the desired dimensions.

[3]The full instructions given to the sentence generators is in Appendix A.1.2.1.

S Getting a high educational degree is important for finding a good job, especially in big cities.
T Good education helps to get a good job.
F There are no good or bad jobs.

EXAMPLE 5.3: A true (T) and a false (F) sentence created from one true sentence (S)

Our intention was to create sentence pairs with a high potential of being in one of the relations of interest. The probability is high for the *true* sentences, as they are potentially similar, which is needed for entailment and paraphrases according to (Cer et al., 2017; Marelli et al., 2014). For *false* sentences we intended to have potential contradiction relations for the pairs.

Pair Generation We combine individual sentences from the sentence pool into pairs, as meaning relations are present between pairs and not individual sentences. To obtain a corpus that contains all discussed meaning relation with sufficient frequency, we use four pair combinations: 1) a pair of two sentences that are true given the same source sentence—*true-true*; 2) a pair of two sentences that are false given the same source sentence—*false-false*; 3) a pair of one sentence that is true and one sentence that is false given the same source sentence—*true-false*; 4) a pair of randomly matched sentences from the whole sentence pool and all source sentences—*random*.

From the 780 sentences in the sentence pool, we created a corpus of 11,310 pairs, with a pair distribution as follows: 5,655 (50%) *true-true*; 2,262 (20%) *false-false*, 2,262 (20%) *true-false*, and 1,131 (10%) *random*. We include all possible 5,655 *true-true* combinations of 30 true sentences for each of the 13 source sentences. For *false-false*, *true-false*, and *random* we downsample the full set of pairs to obtain the desired number, keeping an equal number of samples per source sentence. We chose this distribution because we are mainly interested in paraphrases and entailment, as well as their relation to specificity. We hypothesize that pairs of sentences that are both true have the highest potential to contain these relations. Taking the biggest portion from sentences that are both true has the potential to contain many pairs with these relations.

From the 11,310 pairs, we randomly selected 520 (5%) for annotation, with the same 50–20–20–10 distribution as the full corpus.

Relation Annotation We annotate all the relations in the corpus of 520 sentence pairs using Amazon Mechanical Turk (AMT). We select 10 crowdworkers per task, as this gives us the possibility to measure how well the tasks has been understood overall, but especially how easy or difficult individual pairs are in the annotation of a specific relation. In the SICK corpus, the same platform and number of annotators were used.

We chose to annotate the relations separately to avoid biasing the crowdworkers who might learn heuristic shortcuts when seeing the same relations together too often. We launched the tasks consecutively to have the annotations as independent as possible. This differs from

the SICK corpus annotation setting, where entailment, contradiction, and semantic similarity were annotated together.

The complex nature of the meaning relations makes it difficult to come up with a precise and widely accepted definition and annotation instructions for each of them. This problem has already been emphasized in previous annotation tasks and theoretical settings (Bhagat and Hovy, 2013). The standard approach in most of the existing paraphrasing and entailment datasets is to use a more generic and less strict definitions. A relatively loose definition of semantic equivalence is adopted in most empirically oriented paraphrasing corpora. For example, pairs annotated as PARAPHRASE in Microsoft Paraphrase Corpus (MRPC) "can differ in total information content, with an added word, phrase or clause in one sentence that has no counterpart in the other" (Dolan et al., 2004, 355-356).

We take the same approach towards the task of annotating semantic relations: we provide the annotators with simplified guidelines, as well as with few positive and negative examples. In this way, we believe that annotation is more generic, reproducible, and applicable to any kind of data. It also relies more on the intuitions of a competent speaker than on understanding complex linguistic concepts. Prior to the full annotation, we performed several pilot studies on a sample of the corpus in order to improve instructions and examples given to the annotators. In the following, we will shortly outline the instructions for each task.

Paraphrasing In Paraphrasing (PP), we ask the crowdworkers whether the two sentences have approximately the same meaning or not, which is similar to the definition of Bhagat and Hovy (2013) and De Beaugrande and Dressler (1981).

Textual Entailment In Textual Entailment (TE), we ask whether the first sentence makes the second sentence true. Similar to Recognizing Textual Entailment (RTE) Tasks (Dagan et al. (2005); Bentivogli et al. (2011)), we only annotate for forward entailment (FTE). Hence, we use the pairs twice: in the order we ask for all other tasks and in reversed order, to get the entailment for both directions. Backward Entailment is referred to as *BTE*. If a pair contains only backward or forward entailment, it is uni-directional (UTE). If a pair contains both forward and backward entailment, it is bi-directional (BiTE). Our annotation instructions and the way we interpret directionality is similar to other crowdworking tasks for textual entailment (Marelli et al., 2014; Bowman et al., 2015).

Contradiction In Contradiction (Cont), we ask the annotators whether the sentences contradict each other. Here, our instructions are different from the typical approach in RTE (Dagan et al., 2005), where contradiction is often understood as the absence of entailment.

Specificity In Specificity (Spec), we ask whether the first sentence is more specific than the second. To annotate specificity in a comparative way is new.[4] Like in textual entailment, we pose the task only in one direction. If the originally first sentence is more specific, it is forward specificity (FSpec), whereas if the originally second sentence is more specific than the first, it is backward specificity (BSpec).

Semantic Similarity For semantic similarity (Sim), we do not only ask whether the pair is related, but rate the similarity on a scale 0, denoting *completely dissimilar* to 5, denoting

[4]Louis and Nenkova (2012) labelled individual sentences as *specific*, *general*, or *cannot decide*.

identical.[5] Unlike previous studies Agirre et al. (2014), we decided not to provide explicit definitions for every point on the scale, because we believe that this instruction is quicker to read and the decision is thus made faster, which is important to AMT workers.

Annotation Quality To ensure the quality of the annotations, we include 10 control pairs, which are hand-picked and slightly modified pairs from the original corpus, in each task.[6] We discard workers who perform badly on the control pairs.[7]

Final Corpus For each sentence pair, we get 10 annotations for each relation, namely paraphrasing, entailment, contradiction, specificity, and semantic similarity. Each sentence pair is assigned a binary label for each relation, except for similarity. We decide that if the majority (at least 60% of the annotators) voted for a relation, the sentence pair gets the label for this relation.[8] For similarity, the label is the average value amongst annotators.

Table A.1 shows exemplary annotation outputs of sentence pairs taken from our corpus. For reasons of illustration, Pair #4 is repeated in Example 5.4:

1 The bible is in Hebrew.
2 Bible is not in Latin.

EXAMPLE 5.4: Sentence pair #4 from Table A.1

Example 5.4 contains two relations: forward entailment and forward specificity. This means that it has uni-directional entailment and the first sentence is more specific than the second. The semantic similarity of this pair is 2.7.

	%	κ	%✓	%✗	Control
Paraphrase	.87	.67	.83	.90	.98
Textual Entailment	.83	.61	.75	.89	.89
Contradiction	.94	.71	.84	.95	.95
Specificity	.80	.56	.81	.82	.89

TABLE 5.3: Inter-annotator agreement for binary relations
(✓denotes a relation being there; ✗denotes a relation not being there)

Inter-Annotator Agreement We evaluate the agreement on each task separately. For semantic similarity, we determine the average similarity score and the standard deviation for each pair. We also calculate the Pearson correlation between each annotator and the average

[5] In this way, the similarity between pairs is rated on a scale and not directly in a comparative way, as would be the case when e.g. using best-worst scaling (BWS). However, to operationalize it this way would have complicated the setting for the whole study, as all dimensions would have to be annotated in this way. Then, the comparison to other studies would be even more difficult.

[6] The control pairs are also available online at `https://github.com/MeDarina/meaning_relations_interaction`

[7] Only two annotators were discarded across all tasks. To have an equal number of annotations for each task, we re-annotated these cases with other crowdworkers.

[8] Although we are aware that a majority vote is not always an optimal solution, we believe that this a suitable option in our setting. However, we need a gold label for the analysis. In the corpus, we also provide the number of votes per dimension.

score for their pairs. We report the average correlation, as suggested by SemEval (Agirre et al., 2014) and SICK.

For all nominal classification tasks we determine the majority vote and calculate the % of agreement between the annotators. This is the same measure as used in the SICK corpus. Following the approach used with semantic similarity, we also calculated Cohen's κ between each annotator and the majority vote for their pairs. We report the average κ for each task.[9]

Table 5.3 shows the overall Inter-Annotator Agreement (IAA) for the binary tasks. We report:

- the average %-agreement for the whole corpus
- the average κ score
- the average %-agreement for the pairs where the majority label is YES
- the average %-agreement for the pairs where the majority label is NO
- the average % agreement between the annotators and the expert-provided *control labels* on the control questions

The overall agreement for all tasks is between .80–.94, which is quite good given the difficulty of the tasks. Contradiction has the highest agreement with .94. It is followed by the paraphrase relation, which has an agreement of .87. The agreements of the entailment and specificity relations are slightly lower, which reflects that the tasks are more complex. SICK report agreement of .84 on entailment, which is consistent with our result.

The agreement is higher on the control questions than on the rest of the corpus. We consider it the upper boundary of agreement. The agreement on the individual binary classes shows that, except for the specificity relation, annotators have a higher agreement on the absence of relation.

The average standard deviation for semantic similarity is 1.05. SICK report average deviation of .76, which is comparable to our result, considering that they use a 5 point scale (1–5), and we use a 6 point one (0–5). Pearson's r between annotators and the average similarity score is .69 which is statistically significant at $\alpha = .05$.

Distribution of Meaning Relations Table 5.4 shows that all meaning relations are represented in our dataset. We have 160 paraphrase pairs, 195 textual entailment pairs, 68 contradiction pairs, and 381 specificity pairs. There is only a small number of contradictions, but this was already anticipated by the different pairings. The distribution is similar to Marelli et al. (2014) in that the set is slightly leaning towards entailment[10]. Furthermore, the distribution of uni- and bi-directional entailment with our and the SICK corpus are similar: they are nearly equally represented.[11]

[9] We are aware that κ does not fit the restrictions of our task very well and also that it is usually not averaged. However, we wanted to report a chance-corrected measure, which is non-trivial in a crowd-sourcing setting, where each pair is annotated by a different set of annotators. Multi-π and multi-κ by (Artstein and Poesio, 2008) are not applicable due to the differing annotators.

[10] As opposed to contradiction. However, as contradiction and entailment were annotated exclusively, it is not directly comparable.

[11] In SICK, 53% of the entailment is uni-directional and 46% are bi-directional, whereas we have 44% uni-directional and 55% bi-directional.

	all	T/T	F/F	T/F	rand.
Paraphrase	31	49	27	2	6
Textual Entailment	38	60	36	2	2
Contradiction	13	0	10	56	0
Specificity	73	79	7	66	63
∅Semantic Similarity	2.27	2.90	2.39	1.32	0.77

TABLE 5.4: Distribution of dimensions within different pair generation patterns in percent (except for semantic similarity) (True (T), False (F))

Distribution of Meaning Relations with Different Generation Pairings Table 5.4 shows the distribution of meaning relations and the average similarity score in the differently generated sentence pairings. In the true/true pairs, we have the highest percentage of paraphrase (49%), entailment (60%), and specificity (79%). In the false/false pairs, all relations of interest are present: paraphrases (27%), entailment (36%), and specificity (72%). Unlike in true/true pairs, false/false ones include contradictions (10%). True/false pairs contain the highest percentage of contradiction (85%). There were also few entailment and paraphrase relations in true/false pairs. In the random pairs, there were only few relations of any kind. The proportion of specificity is high in all pairs.

This different distribution of phenomena based on the source sentences can be used in further corpus creation when determining the best way to combine sentences in pairs. In our corpus, the balanced distribution of phenomena we obtain justifies our pairing choice of 50–20–20–10.

Lexical Overlap within Sentence Pairs As discussed by Joao et al. (2007), a potential flaw of most existing relation corpora is the high lexical overlap between the pairs. They show that simple lexical overlap metrics pose a competitive baseline for paraphrase identification. Due to our creation procedure, we reduce this problem. In Table 5.5, we quantified it by calculating unigram and bigram BLEU (Papineni et al., 2002)[12] scores between the two texts in each pair from our corpus, MRPC and SNLI, which are the two most used corpora for paraphrasing and textual entailment. The BLEU score is much lower for our corpus that for MRPC and SNLI. Hence, our corpus creation methodology is less prone to the flaw of high lexical overlap than previous corpora.

	MRPC	SNLI	Our corpus
unigram	61	24	18
bigram	50	12	6

TABLE 5.5: Comparison of BLEU scores between the sentence pairs in different corpora

[12]Bilingual evaluation understudy (BLEU) is an algorithm that was originally developed for text quality evaluation of machine-translated text from one natural language to another. The main idea was to measure computer generated output by its correlation with good human output. It basically measures lexical overlap between two texts. BLEU's output is a number between 0 and 1. 1 represents identical texts, 0 represents less similar texts.

Relations and Negation Our corpus also contains multiple instances of relations that involve negations and also double negations. Those examples could pose difficulties to automatic systems (Tian and Breheny, 2016; Haase et al., 2019) and could be of interest to researchers that study the interaction between inference and negation. Pairs #1, #2, and #9 in Table A.1 are examples for pairs containing negation in our corpus.

5.1.3 Interactions between Dimensions

We analyze the interactions between the relations in our corpus in two ways. First, we calculate the correlation between the binary relations and the interaction between them and similarity. Second, we analyze the overlap between the different binary relations and discuss interesting examples.

Correlations between Dimensions We calculate correlations between the binary relations using the Pearson correlation. For the correlations of the binary relations with semantic similarity, we discuss the average similarity and the similarity score scales of each binary relation.

In Table 5.6, we show the Pearson correlation between the binary dimensions. For entailment, we show the correlation for uni-directional (UTE), bi-directional (BTE), and any-directional (TE). Paraphrases and any-directional entailment are highly similar with a correlation of .75. Paraphrases have a much higher correlation with bi-directional entailment (.70) than with uni-directional entailment (.20). Prototypical examples of pairs that are both paraphrases and textual entailment are pairs #1 and #2 in Table A.1. Furthermore, both paraphrases and entailment have a negative correlation with contradiction, which is expected and confirms the quality of our data. Specificity does not have any strong correlation with any of the other relations, showing that it is independent of those in our corpus.

	TE	UTE	BiTE	Cont	Spec	∅ Sim
Paraphrase (PP)	.75	.20	.70	-.25	-.01	3.77
(all) Textual Entailment (TE)		.57	.66	-.30	-.01	3.59
Uni-directional Textual Entailment (UTE)			-.23	-.17	-.04	3.21
Bi-directional Textual Entailment (BiTE)				-.20	-.01	3.89
Contradiction (Cont)					-.09	1.45
Specificity (Spec)						2.27

TABLE 5.6: Correlation between all relations

We look at the average similarity for each relation (see Table 5.6) and show boxplots between dimension labels and similarity ratings (see Figure 5.4). Table 5.6 shows that bi-directional entailment has the highest average similarity, followed by paraphrasing, while contradiction has the lowest.

Figure 5.4 shows plots of the semantic similarity for all pairs where each relation is present and all pairs where it is absent. The paraphrase pairs have much higher similarity scores than the non-paraphrase pairs. The same observation can be made for entailment. The contradiction pairs have a low similarity score, whereas the non-contradiction pairs do not have a clear

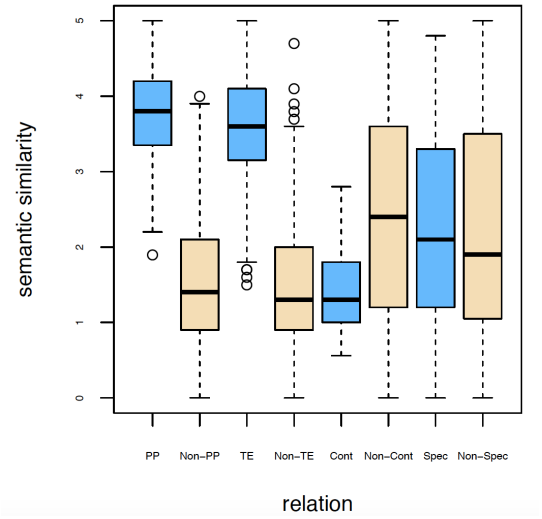


FIGURE 5.4: Similarity scores of sentences annotated with different dimensions

tendency with respect to the similarity score. In contrast to the other relations, pairs with and without specificity do not have any consistent similarity score.

Overlap of Relation Labels Table 5.7 shows the overlap between the different binary labels. Unlike Pearson correlation, the overlap is asymmetric—the percentage of paraphrases that are also entailment (UTE in PP) is different from the percentage of entailment pairs that are also paraphrases (PP in UTE). Using the overlap measure, we can identify interesting interactions between phenomena and take a closer look at some examples.

	PP	UTE	BiTE	Contra	Spec
In paraphrase (PP)		28	64	0	73
In uni-directional Textual Entailment (UTE)	52		-	0	73
In bi-directional Textual Entailment (BiTE)	94	-		0	72
In Contradiction (Contra)	0	0	0		63
In Specificity (Spec)	30	17	21	11	

TABLE 5.7: Distribution of overlap within dimensions in percent

In a more theoretical setting, bi-directional entailment is often defined as being paraphrases (Madnani and Dorr, 2010; Androutsopoulos and Malakasiotis, 2010; Sukhareva et al., 2016). This implies that paraphrases equal bi-directional entailment. In our corpus, we can see that only 64% of the paraphrases are also annotated as bi-directional entailment. An example of a pair that is annotated both as paraphrase and as bi-directional entailment is pair #10 in Table A.1. However, in the corpus we also found that 28% of the paraphrases are only uni-directional entailment, while in 8% annotators did not find any entailment. An example of a pair where our annotators found paraphrasing, but not entailment is sentence pair #5 in Table A.1, which is repeated in Example 5.5.

1 All around the world, girls have higher chance of getting a good school education.
2 Girls get a good school education everywhere.

EXAMPLE 5.5: Sentence Pair #5 from Table A.1

The agreement on the paraphrasing for this pair was 80%, the agreement on (lack of) forward and backward entailment was 80% and 70% respectively.[13] Although the information in both sentences is nearly identical, there is no entailment, as "having a higher chance of getting smth" does not entail "getting smth" and vice versa.

If we look at the opposite direction of the overlap, we can see that 52% of the uni-directional and 94% of the bi-directional entailment pairs are also paraphrases. This finding confirms the statement that bi-directional entailment is paraphrasing (but not vice versa).

There is also a small portion (6%) of bi-directional entailments that were not annotated as paraphrases. An example of this is pair #6 in Table A.1, repeated in Example 5.6:

1 Reading the Bible requires studying Latin.
2 The Bible is written in Latin.

EXAMPLE 5.6: Sentence pair #6 from table A.1

Although both sentences make each other true, they do not have the same content.

These findings are partly due to the more *relaxed* definition of paraphrasing adopted here. Our definition is consistent with other authors that work on paraphrasing and the task of paraphrase identification, so we argue that our findings are valid with respect to the practical applications of paraphrasing and entailment and their interactions. Neither paraphrasing nor entailment had any overlap with contradiction, which further verifies our annotation scheme and quality.

Machine Learning Experiment To empirically determine the degree in which one label can be inferred from the others, we used a linear Support Vector Machine (SVM) to predict the most likely binary label for one of the relations, using the labels of the other relations as features. Table 5.8 shows the average performance of the classifier over 10-fold Cross Validation (CV) and compares it with a simple majority baseline. In the case of paraphrasing and entailment, the SVM outperforms the majority baseline indicating that it can learn a meaningful dependence between the labels. In the case of contradiction or specificity, the results obtained by the SVM are not significantly different from the majority baseline.

5.1.4 Conclusion and Further Work on Links between Relations

In this study, we make an empirical, corpus-based study on interactions between various semantic relations. Our methodology for generating text pairs has proven successful in creating

[13] For the sake of completeness, we have to admit that for many of the other pairs that were annotated as PARAPHRASES, but not as ENTAILMENT, the agreement was low. So, probably the phenomenon of paraphrases without entailment is even more rare than found in this corpus.

	Binary	Majority
Paraphrase	.90	.69
Forward Textual Entailment	.90	.70
Backward Textual Entailment	.86	.71
Contradiction	.87	.87
Forward Specificity	.69	.65
Backward Specificity	.66	.69

TABLE 5.8: Predicting the binary label using the other labels as features

a corpus that contains all relations of interest. The IAA was good for all relations. By selecting different sentence pairings, we have obtained a balance between the relations that best suit our needs. This methodology can easily scale to much larger corpora and will be used in future research.

We provide empirical evidence that supports or rejects previously hypothesized connections in practical settings. The resulting corpus can be used to study individual relations and their interactions. We release a new corpus that contains all relations of interest and the corpus creation methodology to the community. The corpus can be used to further study dimension interactions or as a more challenging dataset for detecting the different relations automatically.[14]

It should be emphasized that our findings strongly depend on our decisions concerning the annotations setup, the guidelines in particular. When examining the interactions between the different relations, we found several interesting tendencies. We showed that paraphrases and any-directional entailment had a high correlation, high overlap, and a high semantic similarity. Almost all bi-directional entailment pairs are paraphrases. However, only 64% of the paraphrases are bi-directional entailment, indicating that paraphrasing is the more general phenomena, at least in practical tasks. Some of our most important findings are:

- there is a strong correlation between paraphrasing
- most paraphrases include at least uni-directional entailment
- paraphrases and bi-directional entailment are not equivalent in practical settings
- contradictions (in our dataset) are perceived as dissimilar

5.2 Compositionality of Relations

After studying the compositionality of relations on the example of paraphrasing, as discussed in Section 4.1, we studied the compositionality of all the discussed relations on the corpus described in Section 5.1. In this section, we present a methodology for decomposing and comparing multiple meaning relations (paraphrasing, textual entailment, contradiction, and specificity).

The methodology includes Single Human-Interpretable Typology for Annotating Meaning Relations (SHARel)—a new typology that consists of 26 linguistic and eight reason-based

[14] The full corpus, the annotation guidelines, and the control examples can be found at `https://github.com/MeDarina/meaning_relations_interaction`.

categories. We use the typology to annotate a corpus of 520 sentence pairs in English. Furthermore, we demonstrate that unlike previous typologies, SHARel can be applied to all relations of interest with a high IAA. We analyze and compare the frequency and distribution of the linguistic and reason-based phenomena involved in textual entailment, paraphrases, contradiction, and specificity. This comparison allows for a much more in-depth analysis of the workings of the individual relations and the way they interact and compare with each other.

Problem Description Recently, several researchers have argued that a single label such as TEXTUAL ENTAILMENT, paraphrase, or SEMANTIC SIMILARITY is not enough to characterize and understand the individual dimension (Sammons et al., 2010; Bhagat and Hovy, 2013; Vila et al., 2014; Cabrio and Magnini, 2013; Lopez-Gazpio et al., 2017; Benikova and Zesch, 2017; Kovatchev et al., 2018a). These authors demonstrate that the different instances of meaning relations require different capabilities and linguistic knowledge.

For example, the pairs in Example 5.7 and 5.8 are both examples of a PARAPHRASE. However, determining the relation dimension in Example 5.7 only requires lexical knowledge, while syntactic knowledge is needed for correctly predicting the relation dimension in Example 5.8.

1 Education is equal for all children.
2 Education is equal for all kids.

EXAMPLE 5.7: Paraphrase pair requiring lexical knowledge

1 All children receive the same education.
2 The same education is provided to all children.

EXAMPLE 5.8: Paraphrase pair requiring syntactic knowledge

This distinction cannot be captured by a single PARAPHRASE label. The lack of distinction between such examples can be a problem in error analysis and in downstream applications. Kovatchev et al. (2019b) empirically demonstrate that in the case of Paraphrase Identification (PI), the different *paraphrase types* are processed in a different way by automated PI systems.

Solution Idea A richer set of labels is needed to better characterize the complexity of meaning relations. We believe that a typology of TEXTUAL ENTAILMENT, PARAPHRASE, and SIMILARITY would capture the distinctions between the different instances of each relation.

In Kovatchev et al. (2020), we propose a new approach for the decomposition of textual meaning relations. Instead of focusing on a single dimension we demonstrate that Paraphrasing, Textual Entailment, Contradiction, and Specificity can all be decomposed to a set

of simpler and easier-to-define linguistic and reason-based phenomena. The set of *atomic* phenomena is shared across all relations.

For the purpose of decomposing the meaning relations we propose SHARel. With the goal of showing the applicability of the new typology, we also perform an annotation experiment using the SHARel typology.

Outcome We demonstrate that multiple dimensions can be decomposed using a shared typology. This is the first step towards building a single framework for analyzing, comparing, and evaluating multiple meaning relations. Such a framework has not only theoretical importance, but also clear practical implications. Representing every dimension with the same set of linguistic and reason-based phenomena allows for a better understanding of the nature of the relations and facilitates the transfer of knowledge (resources, features, and systems) between them. More specifically, it bundles different dimensions discussed in this thesis and enables a use of richer representations.

Furthermore, we annotate a corpus of 520 text pairs in English, textual entailment, paraphrases, contradiction, and specificity. The quality of the typology and of the annotation is evident from the high IAA. As shown in this annotation experiment, we are able to perform a quantitative comparison between the different meaning relations in terms of the types involved in each of them.

5.2.1 Related Work on Decomposition of Several Dimensions

The last several years have seen an increasing interest towards the decomposition of paraphrasing (Bhagat and Hovy, 2013; Vila et al., 2014; Benikova and Zesch, 2017; Kovatchev et al., 2018a), textual entailment (Sammons et al., 2010; LoBue and Yates, 2011; Cabrio and Magnini, 2013), and textual similarity (Lopez-Gazpio et al., 2017).

Decomposition of Single Dimensions Sammons et al. (2010) argue that in order to process a complex dimension such as textual entailment a competent speaker has to take several *inference steps*. This means that a meta-relation such as textual entailment, paraphrasing, or semantic similarity can be *decomposed* or broken down into such *inference steps*. These inference steps, traditionally called *types* can be either linguistic or reason-based in their nature. The linguistic types require certain linguistic capabilities from the speaker, while the reason-based types require common-sense reasoning and world knowledge.

The different authors working on decomposing meaning relations all follow a similar approach. First, they propose a typology—a set of *atomic* linguistic and/or reasoning types involved in the inference process of the particular meta-relation (paraphrasing, entailment, or similarity). Then, they use the atomic types in a corpus annotation and finally, they analyze the distribution and correlation of the types. The corpus based studies have demonstrated that different atomic types can be found in various corpora for textual entailment, paraphrasing, and semantic similarity research.

Kovatchev et al. (2019b) empirically demonstrated that the performance of a PI system on each candidate-paraphrase pair depends on the *atomic types* involved in that pair. That

is, they showed that state-of-the-art automatic PI systems process *atomic paraphrases* in a different manner and with a statistically significant difference in quantitative performance (Accuracy and F1). They show that more frequent and relatively simple types like LEXICAL SUBSTITUTION, PUNCTUATION CHANGES, and MODAL VERB CHANGES are easier across multiple automated PI systems, while other types like NEGATION SWITCHING, ELLIPSIS and NAMED ENTITY REASONING are much more challenging.

Similar observations have been made in the field of Textual Entailment. (Gururangan et al., 2018) discovered the presence of annotation artifacts that enable models that take into account only one of the texts (the hypothesis) to achieve performance substantially higher than the majority baselines in SNLI and Multi-Genre Natural Language Inference (MNLI). Glockner et al. (2018) showed that models trained with SNLI fail to resolve new pairs that require simple lexical substitution. Naik et al. (2018) create label-preserving adversarial examples and conclude that automated Natural Language Inference (NLI) models are not robust. Wallace et al. (2019) introduce universal triggers is, sequences of tokens that fool models when concatenated to any input. All these authors identify different problems and biases in the datasets as well as the systems trained on them. However, they focus on a single phenomenon and/or a specific linguistic construction. A typology-based approach can evaluate the performance and robustness of automated systems on a large variety of tasks.

Limitations of Decomposing Single Dimensions One limitation of the different decompositional approaches is that there exist many different typologies and each typology is created considering only one dimension (paraphrasing, textual entailment, textual similarity). This follows the traditional approach in the research on meaning relations: each dimension is studied in isolation, with its own theoretical concepts, datasets, and practical tasks.

In recent years, the *single relation* approach has been questioned by several authors (Androutsopoulos and Malakasiotis, 2010; Marelli et al., 2014), as described in Section 5.1.1.

However, to date, the joint research of meaning relations is limited only to the binary textual labels. There has been no work on comparing the different typologies and the way different relations can be decomposed. None of the existing typologies is fully compatible with multiple meaning relations, which further restricts the research in this area. We address this research gap in this work.

5.2.2 Shared Typology for Meaning Relations

The goal behind the **S**ingle **H**uman-Interpretable Typology for **A**nnotating Meaning **Rel**ations (SHARel) is to come up with a unified list of linguistic and reason-based phenomena that are required in order to determine the meaning relations that hold between two texts. The list of types should not be limited to texts that hold a specific single textual relation, such as textual entailment, paraphrasing, contradiction, and textual specificity.

Rather, the types should be applicable to texts holding multiple different relations.

1 All children *receive* the same education.
2 The same education *is received* by all kids.

EXAMPLE 5.9: Paraphrase pair requiring lexical knowledge

1 All children *receive* the same education.
2 The same education *is* **not** *received* by all kids.

EXAMPLE 5.10: Paraphrase pair requiring lexical and syntactic knowledge

In Example 5.9, the relation dimension at a textual level is paraphrasing, while in Example 5.10, the dimension is contradiction. In order to determine the dimension for both Example 5.9 and Example 5.10, a competent speaker or an automated system needs to make several inference steps. First, they have to determine that "kids" and "children" have the same meaning and the same syntactic and semantic role in the texts. Second, they need to account for the change in grammatical voice. In terms of typology, these inference steps involve two different types - SAME POLARITY SUBSTITUTION ("kids" - "children") and DIATHESIS ALTERNATION ("receive" - "is received"). In addition, in Statement (2) of Example 5.10, the human or the automated system needs to determine the presence and the function of the *negation* (**not**).

By successfully performing all necessary inference steps, the human (or the automated system) is able to determine that in the pair 3a-3b there is equivalence of the expressed meaning, while in the pair 8a-8b there is a logical contradiction. The required inference steps in the two examples are not specific to the textual label (paraphrasing or contradiction). The *types* are general linguistic or reason-based phenomena.

With the goal of addressing such situations, we propose a list of types that, following the existing theoretical research, can be applied to multiple meaning relations. We justify the choice of types for SHARel in the context of existing typologies.

The SHARel Typology Table A.2 in the Appendix shows the SHARel Typology and its 34 different types, organized in eight categories. The first six categories (morphology, lexicon, lexico-syntactic, syntax, discourse, other) consist of the 24 *linguistic* types. The two types in the *extremes* category (IDENTITY and UNRELATED) are neither linguistic, nor reason-based. The last category consists of the 8 *reason-based* types.

The distinction between linguistic and reason-based types is introduced by Sammons et al. (2010) and Cabrio and Magnini (2013) for textual entailment. The linguistic phenomena require certain linguistic capabilities from the human speaker or the automated system. The reason-based phenomena require world knowledge and common-sense reasoning.

For the linguistic types, we compared the existing typologies and decided to use the Extended Paraphrase Typology (EPT) (Kovatchev et al., 2018a) as a starting point. The authors of EPT have already combined various linguistic types from the fields of Paraphrasing and Textual Entailment, taking the work of Sammons et al. (2010), Vila et al. (2014), Cabrio and

Typology	Relation	Types	Linguistic	Reasoning	Hierarchy
Sammons et al. (2010)	TE, Cont	22	13	9	No
LoBue and Yates (2011)	TE, Cont	20	0	20	No
Cabrio and Magnini (2013)	TE, Cont	36	24	12	Yes
Bhagat and Hovy (2013)	PP	25	22	3	No
Vila et al. (2014)	PP	23	19	1	Yes
Kovatchev et al. (2018a)	PP	27	23	1	Yes
SHARel	TE, Cont, PP, Spec, Sim	34	24	8	Yes

TABLE 5.10: Comparing typologies of dimensions (Textual Entailment (TE), Paraphrase (PP), Contradiction, Semantic Similarity (Sim), Specificity (Spec))

Magnini (2013) into account. As such, the majority of the linguistic types that they propose are in principle applicable to both Paraphrasing and Textual Entailment.

We examined the types from EPT and made several adjustments in order to make the linguistic types fully independent of the textual relation:

- The PI-specific types ENTAILMENT and NON-PARAPHRASE were removed.
- We added UNRELATED type (#26) to the category *extremes* to capture information which is not related at all to the other sentence in the pair.
- We added ANAPHORA type (#16) in the syntax category.

For the reason-based types we studied the typologies of Sammons et al. (2010), LoBue and Yates (2011) and Cabrio and Magnini (2013). We combined similar types of the three typologies into more general types and reduced the original list of over 30 reason-based types to 4. For example, the NAMED ENTITY REASONING (#30) includes both reasoning about geographical entities and publicly known persons (those two were originally separated types).[15]

With respect to specificity, we propose a fine-grained token level annotation, which allows us to determine the particular elements in one sentence that are more (or less) specific than their counterpart in the other sentence. Ko et al. (2019a) demonstrated that specificity needs to be more linguistically and informational theoretically based to be more semantically plausible. This could partially be solved through a more fine-grained annotation of specificity, as it is performed in this study.

Hence, we also add a SPECIFICITY type to the *reasoning* category in order to determine when a segment in one of the sentences is more specific than a segment in the other one.

Table 5.10 lists some properties of the existing meaning relations. All typologies before SHARel were created only for one (or two) meaning relations. SHARel contains general

[15]The annotation guidelines and examples for all types can be seen at `https://github.com/venelink/sharel`.

types that are not specific to any particular dimension and can be applied to pairs holding Textual Entailment, Contradiction, Paraphrasing, Textual Specificity, or Semantic Textual Similarity meaning relation. SHARel follows the good practices of typology research and organizes the types in a hierarchical structure of eight categories and has a good balance between linguistic and reasoning types.

There are two main objectives that motivated this work:

- (1) To demonstrate that multiple meaning relations can be decomposed using a single, shared typology
- (2) To demonstrate some of the advantages of a shared typology of meaning relations.

Based on our objectives, we pose the following research question: Is it possible to use a single typology for the decomposition of multiple (textual) meaning relations?[16]

We address these objectives in a corpus annotation study by evaluating the quality of the corpus annotation by measuring the IAA.

5.2.3 Corpus Annotation

In order to determine the applicability of **SHARel** to all relations of interest, we carried out a corpus annotation on our already available corpus (Gold et al., 2019) that is described in Section 5.1.

We perform an annotation with the SHARel typology on all pairs from Gold et al. (2019) that have at least one of the following relations: forward entailment, backwards entailment, paraphrasing, or contradiction. We discard pairs that are annotated as UNRELATED. This is a typical approach when decomposing meaning relations. Sammons et al. (2010); Cabrio and Magnini (2013); Vila et al. (2014) only decompose pairs with a particular dimensions (entailment, contradiction, or paraphrasing).

After discarding the unrelated portion, the total number of pairs that we annotate with SHARel is 276. Prior to the annotation we tokenized each sentence using the NLTK[17] python library.

During the annotation process, our annotators go through each pair in the corpus. For each linguistic and reason-based phenomenon that they encounter, they annotate the type and the scope (the specific tokens affected by the type). We use an open source web-based annotation interface, called WARP-Text (Kovatchev et al., 2018b).

Each pair of texts was annotated independently by two trained expert annotators. In the cases where there were disagreements, the annotators discussed their differences in order to obtain the best possible annotation for the example pair.[18]

For calculating IAA, we use the two different versions of the IAPTA-TPO measures are explained in more detail in Section 1.5.

[16] In the original study, we also research the similarities and the differences between the (textual) meaning relations in terms of types. However, this was mainly the work of the first author and is thus only very shorty discussed in Section 5.2.4. The results concerning the specificity dimension and its differences compared to the other dimensions with regard to the typology were the work of the author of this thesis and will be discussed in Chapter 6.

[17] `https://www.nltk.org/`

[18] The annotation guidelines and the annotated corpus are available at `https://github.com/venelink/sharel`.

The agreement of our annotation can be seen in Table 5.11. We calculate the agreement on all pairs (all), and we also report the agreement for the pairs with the labels PARAPHRASE (PP), TEXTUAL ENTAILMENT (TE), and CONTRADICTION (Cont).

	TPO-Partial	**TPO-Total**
This corpus (all)	.78	.52
This corpus (PP)	.77	.51
This corpus (TE)	.77	.52
This corpus (Cont)	.75	.50
MRPC-A	.78	.51
Kovatchev et al. (2018a) (Non-PP)	.72	.68
Kovatchev et al. (2018a) (PP)	.86	.68

TABLE 5.11: Comparison of inter-annotator agreements of different corpora (Textual Entailment (TE), Paraphrase (PP), Contradiction (Cont))

To put our results in perspective, we compare our agreement with the one reported in MRPC-A Vila et al. (2014) and Kovatchev et al. (2018a). For Kovatchev et al. (2018a) the authors report both the agreement on the pairs annotated as paraphrases (pp) and as non-paraphrases (non-pp). To date, MRPC-A and Kovatchev et al. (2018a) are the only two corpora of sufficient size annotated with a typology of meaning relations. They also use the same inter-annotation measure to report agreement, so we can compare with them directly.

The overall agreement that we obtain (.52 Total and .78 Partial) is almost identical to the agreement reported for MRPC-A (.51 Total and .78 Partial) and slightly lower than the agreement reported for Kovatchev et al. (2018a) (.68 Total and .86 Partial).

Kovatchev et al. (2018a) detected a significant difference in the agreement between paraphrase and non-paraphrase pairs. In their annotation, the NON-PARAPHRASE includes mostly entailment and contradiction pairs and the lower agreement indicates that their typology is not well equipped for dealing with those cases. However in our corpus, we do not observe such a difference. Our annotation agreement is very consistent across all pairs indicating that SHARel is successfully applied to all relations of interest.

The consistently high agreement score indicates the high quality of the annotation. Even though our task and our typology are much more complex than those of Vila et al. (2014) and Kovatchev et al. (2018a), we still obtain comparable results.

In addition to calculating the inter-annotation agreement, we also asked the annotators to mark and indicate any examples and/or phenomena not covered by the typology. Based on their ongoing feedback during the annotation, we decided to introduce the ANAPHORA type. We re-annotated the portion of the corpus that was already annotated at the time when we introduced the new type.

Arriving at this point, we have demonstrated that it is possible to successfully use a single typology for the decomposition of multiple (textual) meaning relations, which answers the research question of this study.

5.2.4 Analysis of the Results

This subsection provides a summary of the original analysis of the distribution of types across all relations[19]. Table A.3 shows the relative frequencies in pairs that have paraphrasing, uni-directional entailment[20], or contradiction relations at textual level.[21] The similarities and common tendencies between entailment, paraphrasing, and contradiction clearly indicate that these relations belong within the same conceptual framework and should be studied and compared together. The results also suggest the possibility of the transfer of knowledge and technologies between these relations.

The differences between the textual meaning relations in terms of the involved types can help us to understand each of the individual relations better. This information can also be useful in the automatic classification of the different relations in a practical task.

5.2.5 Discussion on Compositionality of Relations

The quality of the annotation is attested by the high IAA as discussed in Sections 5.2.3. We also demonstrated that a shared typology, such as SHARel, is useful to compare different meaning relations in a quantitative and human interpretable way.

In this work, we provide a new perspective on the joint research into multiple meaning relations. Traditionally, the meaning relations have been studied in isolation. Only recently researchers have started to explore the possibility of a joint research and a transfer of knowledge. We propose a new framework for a joint research on meaning relations via a shared typology. This framework has clear advantages: it is intuitive to use and interpret; it is easy to adapt in practical setting—both in corpora creation and in empirical tasks; it is based on solid linguistic theory. The similarities and common tendencies between entailment, paraphrasing, and contradiction clearly indicate that these relations belong within the same conceptual framework and should be studied and compared together. The results also suggest the possibility of the transfer of knowledge and technologies between these relations. The differences between the textual meaning relations in terms of the involved types can help us to understand each of the individual relations better. We believe that our approach cannot only lead to a better understanding of the workings of the meaning relations, but also to improvements in the performance of automated systems.

The biggest challenge in the joint study of meaning relations is the limited availability of corpora annotated with multiple relations. The corpus that we used for our study is relatively small in size. It also has restrictions in terms of sentence length and the frequency of Named Entities. However, it is the only corpus to date annotated with all relations of interest.

Despite the limitations of the chosen corpus, the obtained results are promising. We provide interesting insights into the workings of the different relations, and also outline various

[19]The results presented in this subsection are the contribution of the first author of the collaborative paper and are thus not to be seen as a contribution of the author of this thesis.

[20]We discard the pairs that have bi-directional entailment to reduce the overlap with paraphrases (94% of the bi-directional entailment pairs are also paraphrases).

[21]The table also shows the type frequencies for the paraphrase portion of the Extended Paraphrase Typology Corpus (ETPC) (Kovatchev et al., 2018a) corpus, as it shares the majority of the linguistic types with SHARel and thus enables a comparison of the results.

practical implications. Kovatchev et al. (2019b) demonstrated that a corpus with a size of a few thousand sentence pairs can be successfully used as a qualitative evaluation benchmark. SHARel and the annotation methodology we used easily scale to such size of corpora. This opens up the possibility for a qualitative evaluation of multiple meaning relations as well as for an easier transfer of knowledge based on the particular types involved in the relations.

5.2.6 Conclusions and Future Work on Compositionality of Relations

In this work, we presented the first attempt towards decomposing multiple meaning relations using a shared typology. For this purpose we used SHARel—a typology that is not restricted to a single meaning relation. We applied the SHARel typology in an annotation study and demonstrated its applicability. We analyzed the shared tendencies and the key differences between textual entailment, paraphrasing, contradiction, and specificity at the level of linguistic and reason-based types.

Our work is the first successful step towards building a framework for studying and processing multiple meaning relations. We demonstrate that the linguistic and reasoning phenomena underlying the meaning relations are very similar and can be captured by a shared typology. A single framework for meaning relations can facilitate the analysis and comparison of the different relations and improve the transfer of knowledge between them.

As future work, we aim to use the findings and resources of this study in practical applications such as the development and evaluation of systems for automatic detection of paraphrases, entailment, contradiction, and specificity. We plan to use the SHARel typology for a general-purpose qualitative evaluation framework for meaning relations.

5.3 Conclusion on Relations between Semantic Dimensions

In this chapter, we presented two empirical, corpus-based studies on relations between semantic dimensions. In the first study (Gold et al., 2019), we created a corpus annotated with textual entailment, paraphrasing, contradiction, semantic similarity, and specificity in parallel. Furthermore, we analyzed the dimensions individually as well as their relations on the sentence level. We found evidence confirming (e.g. a strong correlation between paraphrasing and entailment and most paraphrases include at least uni-directional entailment) as well as objecting (e.g. contradictions (in our dataset) are perceived as dissimilar) previous assumptions on relations between dimensions.

Using the corpus from the first study, in the second study (Kovatchev et al., 2020), we presented a unifying typology that can be used for decomposing all the dimensions. Both studies showed empirical evidence of the benefit of working the described dimensions in parallel. The dimensions show clear relations between each other. These findings might be beneficial in both the automation of each individual dimension (e.g. using data that is available for one dimension only to train the others) as well as the qualitative evaluation (using the SHARel typology).

In this chapter, we did not discuss the relations towards the specificity dimension that was also annotated and analyzed in Gold and Zesch (2019). Moreover, we did not discuss the

decomposition of specificity that was performed in Kovatchev et al. (2020). This will be done in the next chapter, which focuses on specificity.

Chapter 6

Specificity of Statements

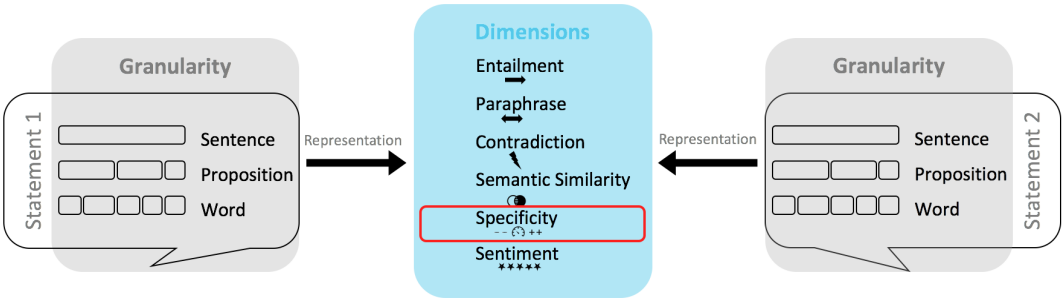


FIGURE 6.1: Illustration of specificity amongst the other relation dimensions in this thesis

Although it has been introduced nearly a decade ago, sentence specificity (Louis and Nenkova, 2011) is a phenomenon that has not been broadly researched in Computational Linguistics (CL) and Natural Language Processing (NLP). Hence, additionally to our work on specificity, we will provide a survey on several aspects of this dimension. In the introduction, we will discuss definitions of specificity. Likewise, we will address differences in its operationalization. Furthermore, we will show the importance of the concept by presenting applications of specificity. Then we will present studies that we performed with respect to specificity.

Definition *Specificity* can be defined as the opposite of generality or fuzziness. Yager (1992) defines specificity as the degree to which a fuzzy subset points to one element as its member. This means that more specific entities reference fewer elements than general entities. In the following, we will use *subsets*, *elements*, and *references* for explanation. However, these are abstract in the practical setting and just serve the purpose of illustration. An abstraction of that definition is shown in Figure 6.2. On the left, the possible subset of references is shown, which in this case is PERSONNEL. A more specific phrase is "attentive waitress", as it points to just some elements within the subset of WAITRESS in contrast to just "waitress", which points to all elements within the same subset.

Specificity was researched on the noun phrase level by identifying the level of how specific the reference in a noun phrase was (Reiter and Frank, 2010; Frawley, 2013). Louis and Nenkova (2011) introduced the task of identifying *general* and *specific* sentences.

We define specificity of statements, which is very similar to Yager (1992) in that we measure the degree to which a statement refers to one element in a subset.

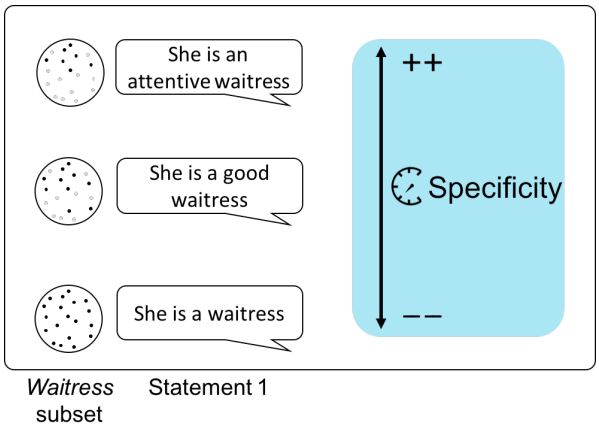


FIGURE 6.2: Statements on "waitress" on different specificity levels

---

1 She is an **attentive waitress**.
2 She is a **good waitress**.

EXAMPLE 6.1: Statement pair on "waitress" on different specificity levels

In Example 6.1, (1) refers to less elements in a subset than (2) because there are fewer "attentive waitresses" than "good waitresses", while a "good waitress" might have been inattentive (but friendly, accommodating, and informed). Figure 6.2 graphically illustrates this example. The circle on the left shows possible references to "waitress" from WAITRESS, on the lowest level of specificity, i.e. only a part of the waitresses can be referred to as "good waitresses", and only a part of these are "attentive waitresses".

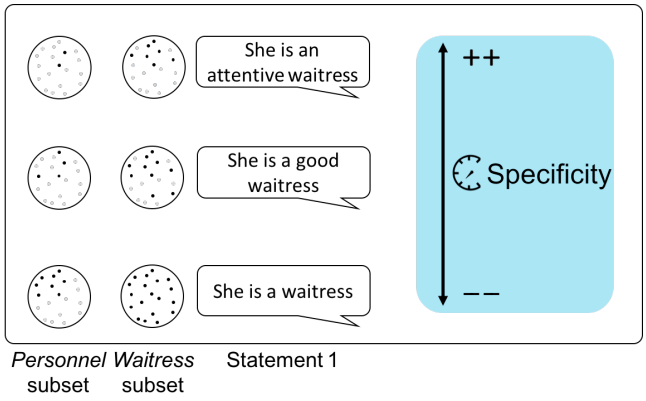


FIGURE 6.3: Statement specificity shown on different subsets

To measure the degree of specificity, the entities from the sentences do not need to refer to the same narrow fuzzy subset, i.e. the subset of WAITRESS, but can also be from a wider subset, e.g. the subset of HOTEL PERSONNEL. Figure 6.3 shows the references of the same statements within different subsets. The figure shows that in the subset of WAITRESS, "She is a waitress" refers to all elements, whereas in the subset of HOTEL PERSONNEL it refers to only some.

In Example 6.2, (1) is more specific than (2), as it refers to less entities in than "a receptionist" in the subset of HOTEL PERSONNEL.

1 She is an **attentive waitress**.
2 He is a **receptionist**.

EXAMPLE 6.2: Statements in the subset of HOTEL PERSONNEL

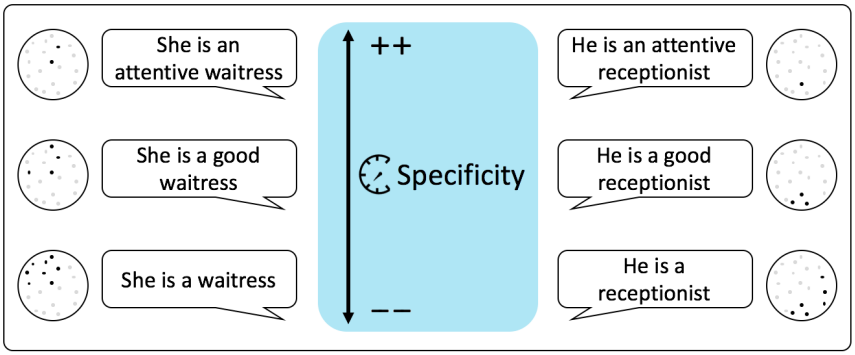


FIGURE 6.4: Statements in the subset of HOTEL PERSONNEL

Therefore, it is possible to comparatively rate specificity of two text pieces that refer to different entities i.e. in our example the entities "receptionist" and "waitress". This can be explained with both "receptionist" and "waitress" belonging to the same wide subset of PERSONNEL, making a comparison possible.[1] This example is illustrated in Figure 6.4.

However, these terms need to be related in terms of ontology or at least topic, otherwise it is very difficult to judge their specificity:

1 She is an attentive waitress.
2 The garden had flowers.

EXAMPLE 6.3: Statements without common subset

In Example 6.3, which is also illustrated in Figure 6.5, (1) and (2) do not share any similar information, and have no similar propositions, it is close to impossible to say which one is more specific. Depending on the exact framing of asking for the specificity, however, it might still be possible to judge it. If, for instance, the question was which of the two sentences is more specific with regard to *service*, (1) would clearly be more specific, as (2) does not mention this topic. If, however, the question was which of the two sentences is more specific with regard to *landscaping*, (2) would clearly be more specific, as (1) does not mention this topic.

[1] We would like to stress that the explanation of the subsets is meant as an abstract illustration of our understanding of specificity. It does not automatically mean that infrequent professions (e.g. in the subset of PERSONNEL *hotel manager*) are more specific than frequent ones (such as *waitress* or *receptionist*). This remains an open question.

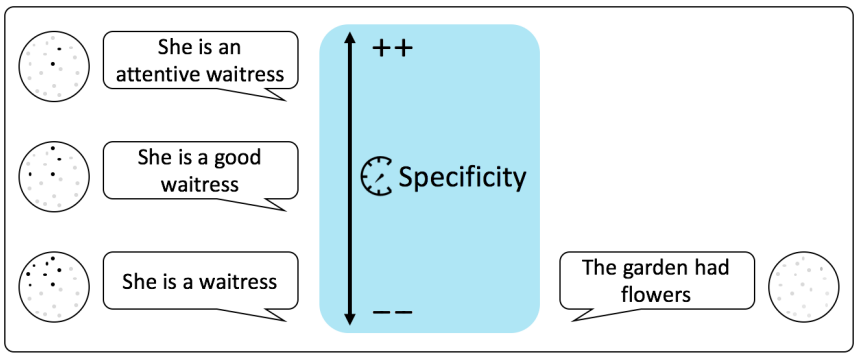


FIGURE 6.5: Statement pair from different semantic subsets

In that way, our definition of specificity is very similar to and as loose as the previous definitions.

Operationalization of Specificity Our operationalization of specificity is different from previous work, as will be discussed Section 6.1. So far, specificity has been regarded as both a binary (Louis and Nenkova, 2011, 2012) as well as a scaled phenomenon (Li et al., 2016; Swanson et al., 2015; Ko et al., 2019a) in more recent works. We will outline the development from binary to numeric annotations of specificity. As previously discussed in the Introduction, we believe that specificity, as well as other dimensions, are more reliably judged when statements are compared. Hence, after outlining the development from binary to numeric annotations of specificity performed on individual sentences, we will present how we measure the specificity of a statement in comparison to another statement in Gold and Zesch (2019); Kovatchev et al. (2020) and an unpublished study.

Automation of Specificity Determination Together with the concept of sentence specificity, Louis and Nenkova (2011) introduced an automatic solution. Depending on the operationalization of the task, the automation also makes a binary decision (Mathew and Katz, 2009; Louis and Nenkova, 2011; Li and Nenkova, 2015) or classifies on a nominal scale (Ko et al., 2019a,b). The automation will be discussed in more detail in Section 6.2.

Features Used in Automation There are linguistically informed and simple frequency features that correlate with specificity and thus can be used in its automatic determination. This will be further discussed in Section 6.3.

Specificity and Other Dimensions In Example 6.1, (1) entails (2), as "an attentive waitress" is also "a good waitress" (but not the other way around, as explained earlier). (1) is also more specific than (2)—possibly because (1) already contains the information entailed in (2). Hence, we believe that there is a relation between specificity and entailment. In Gold et al. (2019), we empirically study the relation between specificity and entailment, paraphrasing, contradiction, and semantic similarity. In this way, we can better understand the dimension of specificity and place it within the landscape of NLP. Furthermore, discovering relations

to other, well-researched phenomena, might be beneficial for specificity automation. For instance, we research whether the intuitive relation between textual entailment and specificity actually exists.

The results concerning specificity and its relations are discussed in Section 6.4.[2]

Decomposition of Specificity In Kovatchev et al. (2020), we decompose several dimensions, including specificity. In this study, we try to locate the part of a statement that makes it more specific than the other. In 80% of the pairs with specificity at textual level, our annotators were able to point at one or more particular elements that are responsible for the difference in specificity. In 97% of these cases, there was a further annotation that indicated the reason for the difference. In this way, we are able to get a more semantically informed understanding of the dimension. The results of of this study are presented in Section 6.5.[3]

Application of Specificity Knowing the specificity of a statement is helpful in any information extraction task, especially when extracting information from multiple sources with partially overlapping content,i.e. the statement with the most fitting specificity level could chosen from a set of statements with very similar information. Louis and Nenkova (2011) state that specificity might be an effective feature in many applications, such as prediction of writing quality, text generation, and information extraction. Ko et al. (2019a) annotated scaled specificity on Yelp and movie reviews, showing that specificity is applicable to this genre. Section 6.6 discusses how specificity is or could be used in summarization, argument extraction, and dialogue generation.

6.1 Operationalization of Specificity

Most work on specificity in linguistics is focused on noun phrases, meaning whether a noun references a unique entity in a given context (Reiter and Frank, 2010; Frawley, 2013). With regard to noun phrases, proper names are specific, whereas common nouns with an indefinite article are mostly not. In this work, however, we focus on sentence or propositional specificity—or as we call it, the specificity of statements.

The operationalization and the different scales we describe herein are focused on the annotation, not the automation. Specificity has been annotated and measured in three different ways that have developed chronologically—on a binary (or rather trinary) scale, on a nominal scale, and we introduced a comparative scale. In the following, we will discuss corpora that have been created using the different scales.

Table 6.1 shows the overview of the discussed operationalizations and the corpora based on these in particular. All corpora are available online, but do not provide a specific license for usage. As Inter-Annotator Agreement (IAA) is measured with different metrics, which is partly due to the different annotation schemes, we discuss it in the respective section, but do not to show it in the table. Furthermore, the table provides two exemplary sentences that

[2]The setting of the study is discussed in Chapter 5.1.
[3]The setting of the study is discussed in Chapter 5.2.

cannot be found in any of the discussed corpora. They solely serve the purpose to show the differences within the operationalization of specificity.

Corpus parameters				Exemplary sentences	
Authors	Size	Anno. type	Genre	(1) She is a **waitress**.	(2) Lucy is an **attentive waitress**.
Louis and Nenkova (2012)	894	Binary	News	General	Specific
Li et al. (2016)	543	Scalar	News	4	1
Ko et al. (2019a)	2,749	Scalar	Diverse	.75	.25
Gold et al. (2019)	1,040	Pairwise	Synthetic	(1) < (2)	

TABLE 6.1: Comparison of specificity corpora. Size is given in sentences or tweets.

6.1.1 Binary Scale

The first sentence specificity set, introduced by Louis and Nenkova (2011) is based on SPECIFICATION and INSTANTIATION relation annotations between sentences. Louis and Nenkova (2011, 2012) measured specificity on a binary, or rather trinary, scale (in the following we will refer to it as binary). They distinguished between GENERAL, SPECIFIC, and CAN'T TELL. Table 6.1 shows an exemplary annotation of two sentences. While (1) is GENERAL, as it works with a personal pronoun and a general label of the profession label, (2) gives a name and an adjective in addition to the profession label and thus makes it SPECIFIC. Louis and Nenkova (2011) used adjacent sentences from the Penn Discourse Treebank (PDTB) (Prasad et al., 2016) annotated with SPECIFICATION or INSTANTIATION and define the first sentence as being GENERAL and the subsequent one as SPECIFIC. The definitions of Specification or Instantiation describe the specificity of one sentence relative to the other.[4] However, as already mentioned, Louis and Nenkova (2011) do not define specificity as a relative relation towards another entity, but as a phenomenon of each individual sentence. Their definition (see Appendix A.3) is quite loose and not focused on references. The choice of sentences for the specificity annotation may be biased towards discourse relations and the references they contain (Louis and Nenkova, 2011). In short, their definition says that specific sentences contain details and can stand independently of other sentences. Additionally to the sentences taken from PDTB, Louis and Nenkova (2011) annotated a part of ACQUAINT (Graff, 2002)[5]. In a consecutive study, Louis and Nenkova (2012) annotated articles of the New York Times (NYT)-science news from the NYT corpus (Sandhaus, 2008). All of these are texts from the news genre. Louis and Nenkova (2011, 2012) used five workers on Amazon Mechanical Turk (AMT). The majority vote, which in 67% of the cases meant at least four annotators agreeing, was used as gold standard (Louis and Nenkova, 2011). Louis and Nenkova (2011, 2012) reported %-agreement as IAA. As they used more than two workers,

[4] Actually, the exemplary sentences given in Table 6.1 is an example of Instantiation.

[5] The AQUAINT Corpus of English News Text is a large collection of news texts (roughly 375 million words) prepared by the LDC for the AQUAINT Project, and planned to be used in official benchmark evaluations conducted by National Institute of Standards and Technology (NIST).

%-agreement is also not trivial to report. In summary we can say that there was only 1% of cases where no majority vote could be reached and, as previously stated, in 67% of the cases, there was either full agreement among the five annotators or one disagreement.

6.1.2 Numeric Scale

Li et al. (2016); Swanson et al. (2015) and Ko et al. (2019a) measured specificity on a numeric scale, arguing that the phenomenon of specificity is rather gradual than binary. Exemplary labels for both corpora are shown in Table 6.1.

Li et al. (2016) used a nominal scale from 0 (most specific: does not require any additional information to understand who or what is involved and what is the described event) to 6 (least specific) with three annotators.[6] They do not provide further information on the guidelines of the scaling annotation process.

As a continuation of the work by Louis and Nenkova (2011, 2012), Li et al. (2016) also worked on news texts, more specifically on articles from the NYT. They evaluated their annotation using Cronbach's α, which was .72.

Ko et al. (2019a) used similar instructions as Li et al. (2016) and re-scaled to real values from 0.0 - 1.0 (in 0.25 steps). In contrast to previous work, Ko et al. (2019a) used other and different genres as source data for specificity annotation—namely Twitter, Yelp and movie reviews. Their IAA, also measured with Cronbach's α, was between .68 and .70, which is only slightly lower than Li et al. (2016). The corpus referenced in Table 6.1 only shows the part that was manually annotated by them, which is their test corpus, whereas they used already available training corpora.

6.1.3 Comparative Scale

On the example of *sentiment intensity*, Kiritchenko and Mohammad (2017) showed that rating scales, in contrast to comparative ratings, are difficult to reproduce and also highly dependent on the annotator. Hence, in this thesis, we performed comparative ratings for specificity. The difference to the other operationalizations is shown in Table 6.1. While the other operationalizations label individual sentences, ours give comparative ratings. Although often discussed in theoretical settings or definitions of specificity, to our knowledge, we are the first to comparatively annotate specificity.

We did so in two ways, as illustrated in Table 6.2.

In our study that researched the links between different dimensions discussed in this thesis (see Section 3.1), we annotated all dimensions on pairs, meaning that we annotated which of two sentences was more specific (Gold et al., 2019).

In another, unpublished, study, we annotated specificity using best-worst scaling (BWS), which is a comparative rating score technique, where the annotator chooses the most specific and the least specific statement out of a list of statements.

[6] Additionally to the scaling, Li et al. (2016) asked the annotators 1) to mark underspecified parts of the sentence and 2) identify the cause of underspecification in the form of free text questions. This is relevant for Section 6.5 and will be repeated there, but remains a side note in this section.

Method	Simple Comparison	Best-Worst-Scaling
Input	2 statements	4 statements
Instruction	Is Statement 1 more specific than Statement 2?	Which statement is most specific? Which statement is least specific?
Output	Binary result for each tuple	One best and one worst for each quadruple

TABLE 6.2: Comparative methods of measuring specificity applied in this thesis

6.1.3.1 Comparing the specificity between two sentences

In Gold et al. (2019), we researched the links between specificity and other meaning relations, as well as semantic similarity. In order to be directly comparable to the other meaning relations, we chose the same annotation setup for all of them—each relation was annotated independently on a sentence pair by 10 annotators using crowdsourcing. The exact corpus creation procedure and the annotation of the other relations is described in more detail in Chapter 5.

Instructions for annotating specificity in a sentence pair Specificity, similar to textual entailment, was annotated directionally. This means that we asked the annotator, given a sentence pair, to state whether the first sentence is more specific than the second. For a better understanding of the specificity relation, we also added examples to the definition of specificity. The instructions given to the annotators can be found in Appendix A.1.2.1.

Gold label for specificity annotation Having 10 annotations per pair, we decided that the threshold for a gold label annotation is a majority vote for specificity. This means if at least six annotators annotated that the pair has a specificity relation, the gold label for this pair has a positive label for specificity.

Agreement on specificity in relations corpus For all nominal classification tasks we determined the majority vote and calculate the %-agreement between the annotators. Following the approach used with semantic similarity, we also calculate Cohen's κ between each annotator and the majority vote for their pairs. We report the average of all annotators. [7]

The overall %-agreement on this task was .80, which is slightly lower than on the other tasks (for the agreements of the other tasks, see Table 5.3), showing its complexity.

In our dataset, we added some control pairs for quality control. On these pairs, the %-agreement is .89, which is probably the upper boundary for specificity annotation. The averaged κ was .56.

[7] We are aware that κ does not fit the restrictions of our task very well (e.g. 1) there are more than two annotators and 2) it would even be difficult to make out two annotators who annotated the same pairs) and also that it is usually not averaged. However, we wanted to report a chance-corrected measure, which is non-trivial in a crowd-sourcing setting, where each pair is annotated by a different set of annotators.

6.1.3.2 Specificity using BWS

In an unpublished pilot study, we perform BWS to annotate specificity. BWS is a comparative scoring technique, which was shown to be more reliable that scoring techniques in many cases (Kiritchenko and Mohammad, 2017). In BWS, given a set of objects, the one fitting a particular characteristic best and the one fitting the characteristic worst have to be chosen. The methodology is explained in more detail in Section 1.2.3.

Source data We used sentences from the SemEval-2016 Task 5(Pontiki et al., 2016). These are sentences from restaurant reviews that contain sentiment annotations on given aspects. This being a pilot study, we used only the sentences that were marked to have a negative sentiment on AMBIANCE. These were 39 sentences which will be referenced as *statements* in the following.

Annotation setting In this study, we used AMT[8]. In Appendix A.1.4, we show the HIT instructions. We used four annotators per Human Intelligence Task (HIT). Each annotator is presented with a quadruple of statements. To get a reliable scoring after the annotation, not all permutations of statements need to be presented in quadruples. Typically the number of quadruples is $1.5 * the number of\ statements$.[9] Overall, we had 50 quadruples.[10] Presented with the quadruple, the annotator has to make two decisions in one HIT :

- which of the four statements is the most specific towards a given aspect
- which of the four statements is the least specific towards a given aspect

So the specificity is also attached to a given aspect. By limiting the specificity to one given aspect, we address the issue of several aspects of different specificity levels in one statement being mixed. In our study, we used the aspect AMBIANCE, as our source statements were taken from this aspect only.

Evaluation The resulting rating of the statements can be found in the Appendix (see Table A.4). The agreement is evaluated with Split-Half Reliability (SHR) as described in Section 1.2.3. Using SHR, the correlation was .93 and the Spearman-Brown coefficient was .96. This indicates that our annotation is reliable.

Conclusion In our pilot study, we show that specificity can be reliably annotated with BWS. However, we also found that in our study, specificity strongly correlates with sentence length—the correlation was .66 and the Spearman-Brown coefficient was .79. This is not surprising, as longer sentences potentially contain more information, which also correlates with specificity. Hence, when performing the real study, we plan to work with propositions instead of sentences.

[8] Details on crowdsourcing and its terminology can be found in Section 1.1.2.

[9] Details on the logic behind the creation of quadruples and the scoring process can be found in Section 1.2.3.

[10] We generated them using a script by Kiritchenko and Mohammad (2017) available on `https://www.saifmohammad.com/WebPages/BestWorst.html`.

6.1.4 Conclusion on Operationalization of Specificity

In this section, we have shown the development of specificity annotation from a binary to a numeric scale and ,as we believe, to an even more reliable scale—the comparative one. Prior to our studies, specificity has not been annotated comparatively. In our study described in Section 6.4, we used simple paired comparisons for the specificity annotation. The IAA was reliable, but could be improved by offering a focus of specificity to the annotator, as was done in our study described in Section 6.1.3.2. In this study, we used BWS and an aspect provided by the aspect-based sentiment analysis (ABSA)-corpus we used to annotate specificity. We were able to show that it can be reliably annotated using this operationalization. Hence, we were able to show that specificity can be reliably annotated using comparative methods.

6.2 Automation of Specificity Determination

Automatic sentence specificity classification has been introduced together with the task itself Louis and Nenkova (2011). In the following, different systems will be described. The systems can be divided in two different classification approaches: binary and numeric.

The data used for the annotation does not necessarily need to fit the training or the test set for the automation, as it can be transformed. Ko et al. (2019a) re-scaled a nominal values to a real values. After annotating on a comparative scale, the items can be ordered according to the strength of the annotation. This order can than be transformed to real values, so that automations for real values can also be applied. Accordingly, setting a threshold, real or nominal values can also be transformed to binary values.

The information on features used in the system was outsourced in Section 6.3, as it can be seen as a complex task on its own.

6.2.1 Binary Classification

In sentence specificity classification, binary classification distinguishes between SPECIFIC and GENERAL sentences.

Early Works Reiter and Frank (2010) presented an automatic approach to distinguish specific and generic noun phrases. However, the restriction to noun phrases does not generalize to our focus on statements. Mathew and Katz (2009) presented an automatic classification of generic and non-generic sentences. However, their interpretation of specificity is restricted to habitual interpretations of generic sentences, meaning a specific sentence refers to a single, specific event and a general sentence describes general facts.

First Sentential Specificity Classifier Louis and Nenkova (2011) not only introduced the task of sentential specificity, but also presented an automatic classifier. As previously described, they have a binary classification of specificity. They use a logistic regression classifier with sentence length, inverse document frequency (idf), count of numbers (identified using the part of speech), proper names, dollar signs, and plural nouns, syntax, polarity and

language models. In their analysis, they find that non-lexical features have the best performance. Louis and Nenkova (2011) evaluate their classifier using 10-fold Cross Validation (CV) using their own corpus.

SPECITELLER Building on the findings of Louis and Nenkova (2011), Li and Nenkova (2015) present SPECITELLER, a specificity annotation tool that significantly outperforms previous work. They experiment with a supervised and a semi-supervised approach. In their supervised approach, they use the same discourse-annotated source data as Louis and Nenkova (2011) to train a logistic classifier. They use shallow features, such as sentence surface features (number of words, Named Entitity (NE)s, capital letters, etc.) and dictionary features (using polarity metrics, and psychological metrics such as concreteness) and word representation features (such as word identity and neural network embeddings). We will discuss the relation the effectiveness of the features in Section 6.3 in more detail.

To train their complex approach, they use a semi-supervised approach via co-training using the same discourse-annotated source data as Louis and Nenkova (2011). Li and Nenkova's (2015) classification procedure consists of two steps: a supervised learning phase and a bootstrapping phase. In the first phase, two classifiers are trained separately on the data of the Instantiation discourse relation. One is trained on shallow features and the other one on word representation features, as described previously. In the bootstrapping phase, the classifiers take turns in creating examples for each other, adding the most probable examples to the training.

Both the simple and the complex approach are evaluated using the dataset by Louis and Nenkova (2012). In the complex approach, both shallow and word representation classifiers outperform Louis and Nenkova (2011), even without combining the two classifiers.

6.2.2 Numeric classification

According to Ko et al. (2019a), prior sentence specificity systems do not generalize well to other domains. Furthermore, they believe that real-valued classification of specificity is more helpful. Their

> "framework is an unsupervised domain adaptation system based on Self-Ensembling (Tarvainen and Valpola, 2017; French et al., 2018) that simultaneously reduces source prediction errors and generates feature representations that are robust against noise and across domains.[...] We further propose a posterior regularization technique Ganchev et al. (2010) that generally applies to the scenario where it is easy to get coarse-grained categories of labels, but fine-grained predictions are needed. Specifically, our regularization term seeks to move the distribution of the classifier posterior probabilities closer to that of a prespecified target distribution, which in our case is a specificity distribution derived from the source domain."
> (Ko et al., 2019a, p.2)

Figure 2 shows an abstraction of their system.

They evaluate their system on a re-rescaled version of Li et al.'s 2016 corpus, as well as their own corpus developed with the same annotation guidelines, as described in Section 6.1.2 using three metrics:

- the Spearman correlation between the labeled and predicted specificity values
- the pairwise Kendall's Tau correlation
- Mean Absolute Error (MAE)

According to their analysis, Speciteller does not generalize well to other domains and even performs worse then just using sentence length on two of three domains. The system presented by Ko et al. (2019a) performs best with adaption, distribution regularization using mean standard deviation or Kullback–Leibler divergence.

They further examine the usefulness of specificity in dialogue generation, which is discussed in Section 6.6.

6.2.3 Conclusion on Automation of Specificity

In this section, we describe the development from binary (Mathew and Katz, 2009; Reiter and Frank, 2010; Louis and Nenkova, 2011) to numeric Ko et al. (2019a) specificity determination systems. Ko et al. (2019a), presenting the most recent sentence specificity system, introduced a system that in contrast to prior ones generalizes domains vastly varying from the source domain and uses a real-valued classification of specificity.

6.3 Features Used in Automation

Specificity involves and is also related to many other linguistic phenomena. According to Ko et al. (2019b)

> ... past work in sentence specificity—the "quality of belonging or relating uniquely to a particular subject"— has shown that word frequency is only one aspect of specificity, and that specificity involves a wide range of phenomena including word usage, sentence structure (Louis and Nenkova, 2011; Li and Nenkova, 2015; Lugini and Litman, 2017) and discourse context (Dixon, 1987; Lassonde and O'Brien, 2009). Frequency-based specificity also does not exactly capture "the amount of information" as an information-theoretic concept.
> (Ko et al., 2019b, p.3456).

In this section, we describe studies that examine these links or examine which features are used to automatically calculate specificity, which shows that specificity is somehow linked to these phenomena. On the one hand, as already mentioned in the quote by Ko et al. (2019b), specificity seems to be closely related to frequency and information gain, as well as other frequency related features. Ko et al. (2019b) also state that insights of specificity studies showed that "sentence specificity encompasses multiple phenomena, including referring expressions, concreteness of concepts, gradable adjectives, subjectivity and syntactic structure." (Ko et al., 2019b, p.3457)

On the other hand, specificity is also related to other dimensions, namely textual entailment, paraphrases, and contradiction, as well as semantic similarity. Although these can also be used as features, we discuss them separately in Section 6.4.

Specificity is also related to other tasks, in which it can and is used to automatically compute them. These relations are discussed separately in Section 6.6.

6.3.1 Frequency-Based Features

There is a variety of frequency-based measures that are used to calculate specificity or examined with specificity. As specificity seems to be strongly linked with information gain (Ko et al., 2019b), there are also frequency-based features that are usually used to calculate it. Furthermore, we shortly describe features that purely capture frequencies of words, or of specific Part-of-Speech (POS). Frequency measures, such as idf, Normalized Inverse Response Frequency (NIRF), Normalized Inverse Word Frequency (NIWF), and Perplexity Per Word (PPW), indirectly capture the difference between function and content words, as function words are very frequent words in all contexts.

Word Frequencies Louis and Nenkova (2011) use the count of each word in the sentence as a feature. In their automation, this is the most effective feature. According to Louis and Nenkova (2011) this shows that there are strong lexical indicators for the distinction between specific and general sentences They state that in their study, discourse connectives such as "but", "also" and "however", and vague words such as "some" and "lot" are frequent and also top indicators for general sentences.

idf for a word w is defined as shown in Equation 6.1, where N is the number of documents in a large collection, and n is the number of documents that contain the word w.

$$idf_w = log\frac{N}{n} \tag{6.1}$$

According to Spärck Jones (1972), who introduced the term, "the specificity of a term is the number of documents to which it pertains" [p.13]. Louis and Nenkova (2011) used it as a feature for specificity classification. They used NYT articles of one year to compute idf.

NIRF and NIWF are introduced by Zhang et al. (2018) as specificity control variables to calculate specificity in the task of *neural response generation*. The assumption behind Normalized Inverse Response Frequency (NIRF) is the more often a response corresponds to an input in the corpus, the more general it is. Normalized Inverse Word Frequency (NIWF), on the other hand, is based on the assumption that the specificity or generality of a response corresponds to the sum of the specificity of the individual words it contains.

To calculate the Inverse Response Frequency (IRF), a collection of responses R is built. Y denotes a response and f_Y its frequency. IRF is calculated in the following way:

$$IRF_Y = log(1 + |R|)/f_Y \tag{6.2}$$

The IWF for a word y in response Y is calculated similarly:

$$IWF_Y = log(1 + |R|)/f_Y \tag{6.3}$$

As a response consists of several words, to calculate the value for response Y, the maximum specificity value of all words in the response is used:

$$IWF_Y = max_{y \in Y}(IWF_y) \tag{6.4}$$

For normalizing both IRF and IWF (to get IRF and IWF as a result), the metric m (IRF or IWF), the min-max normalization method by Jain, Anil and Nandakumar, Karthik and Ross, Arun (2005) is used:

$$Nm_Y = \frac{m_Y - min_{Y' \in R}(m_{Y'})}{max_{Y' \in R}(m_{Y'}) - min_{Y' \in R}(m_{Y'})} \tag{6.5}$$

NIRF was found not to correlate with specificity, as very general answers can also be infrequent in a dataset (Zhang et al., 2018). However, NIWF was found to be useful (Zhang et al., 2018). Hence, Ko et al. (2019b) also used NIRF to calculate specificity.

PPW according to Ko et al. (2019b) "is the exponentiation of the entropy, which estimates the expected number of bits required to encode the sentence" [p. 2460]. Ko et al. (2019b) train a neural language model on all gold responses and calculate cross-entropy of each sentence. In order to prevent PPW from over-fitting to long sentences, Ko et al. (2019b) normalize by sentence length.

Language Models namely simple uni-, bi, and trigram models from the NYT corpus were used by Louis and Nenkova (2011). According to Louis and Nenkova (2011), this was a helpful feature for sentence specificity classification. However, they do not elaborate on the individual models.

Sentence Length is mostly measured as the number of words in a sentence. Louis and Nenkova (2011) used the number of words and the number of nouns in a sentence as features for their automatic specificity classification. According to them, sentence length is the least indicative feature.

Ko et al. (2019b) used sentence length as a baseline for specificity classification. The baseline performed well on all domains and better than some complex methods on some domains.

Frequencies of Individual POS Louis and Nenkova (2011) used and analyzed frequencies of different POS, such as adjectives, adverbs, adjective phrases, adverbial phrases, verb phrases, prepositional phrases, numbers, NEs, plural nouns, etc.

In their analysis, they found that words (not POS) were the most predicative feature class. Hence, they analyzed those with the highest weights in their regression model. They found

that discourse connectives such as 'but', 'also', and 'however', and general words such as 'some' and 'lot' strongly correlate with general sentences. Pronouns and quantifiers such as 'a' and 'one', on the other hand, correlate with specific sentences. They also found a large amount of words that correlate with specificity but are domain-specific.

The most helpful feature after words were NEs together with numbers.

Content Density Li et al. (2016) found that content density, representing how factual, direct and succinct the content of a sentence is, positively correlates with specificity. They compared their specificity score annotations to manual content density score annotations by Yang and Nenkova (2014).

6.3.2 Measures Using External Knowledge Bases

Additionally to the word frequency measures described previously, there are also measures that calculate frequencies based on lexicons with semantic annotations.

Polarity is the distinction between positive and negative mentions in text. Chapter 7 discusses this phenomenon in more detail. According to Louis and Nenkova (2011), sentences with a strong opinion correlate with general sentences. Hence, they include the number of positive and negative words in a sentence as a feature using the General Inquirer (Stone et al., 1966) and the MPQA lexicon (Wilson et al., 2005). They also include a feature where these counts are normalized by sentence length. In the analysis of Louis and Nenkova (2011), polarity was a helpful feature for specificity classification.

Word Specificity has been researched prior to sentence specificity and was also used as a feature for sentence specificity detection by Louis and Nenkova (2011). For the word specificity measure they use the average, minimum, and maximum values of the hypernym relation paths of nouns and verbs in WordNet Miller (1995). Their assumption is that the longer the path to the root is, the more specific a word is. According to Louis and Nenkova (2011), word specificity was a helpful feature for specificity classification.

6.4 Specificity and its Links to the other Dimensions

In our study Gold et al. (2019), which is described in more detail in Chapter 5, we compare the specificity annotation described in Section 6.1.3.1 to the other dimensions using the Pearson correlation. With respect to specificity, we found that it does not correlate with other relations, showing that it is independent of those in our corpus. Specificity has a nearly equal overlap within all the other dimensions, meaning that it does not overlap with any other dimension in particular.[11] Furthermore, it also shows no clear trend on the similarity scale and no correlation with the difference in word length between the sentences. This indicates that

[11] Unlike Pearson correlation, the overlap is asymmetric—the fraction of specificity pairs that are also entailment pairs (UNI-DIRECTIONAL TEXTUAL ENTAILMENT in SPECIFICITY) is different from the fraction of entailment pairs that are also specificity (SPECIFICITY in UNI-DIRECTIONAL TEXTUAL ENTAILMENT).

specificity cannot be automatically predicted using the other dimensions and requires further study.

In the following, the intuitions behind the links are described. Moreover, we focus on interesting cases, which are complicated and unexpected, e.g. paraphrases that are not entailment or entailment pairs that do not differ in specificity. However, the full corpus also contains many conventional and non-controversial examples.

Specificity and Entailment Our intuition is that specificity is strongly related to entailment, as entailment implies a change of specificity. Intuitively, the entailing sentence is probably more specific than the entailed sentence.

1 She is an attentive waitress.
2 She is a waitress.

EXAMPLE 6.4: Statement pair showing intuitive relation behind specificity and entailment

In Example 6.4, (1) is more specific, as it gives more information on the waitress than (2). (1) entails (2), as if "She is an attentive waitress", "she" needs to be "a waitress" in the first place. Hence, the entailing statement could always be more specific, as it contains more information than the entailed statement.

In our study, however, we could not prove this intuition, as specificity had no correlation with entailment—uni-directional textual entailment had a -.09 correlation with specificity (cf. Table 5.6). There are 27% of uni-directional entailment relation pairs that are not in any specificity relation. One example of this is pair #8 in Table A.1:

1 You can find a good job if you only speak one language.
2 People who speak more than one language could only land pretty bad jobs.

EXAMPLE 6.5: Sentence pair #8 from Table A.1

Although the pair in Example 6.5 contains uni-directional entailment (backward entailment), none of the sentences was annotated as more specific than the other.

If we look at the other direction of the overlap, we can observe that in 62% of the cases involving difference in specificity, there is neither uni-directional nor bi-directional entailment. An example of such a relation pair is pair #9 in Table A.1:

1 All Christian priests need to study Persian, as the Bible is written in Ancient Greek.
2 Christian clergymen don't read the bible.

EXAMPLE 6.6: Sentence pair #9 from Table A.1

The two sentences are on the same topic and thus can be compared on their specificity. (1) is clearly more specific, as it gives information on what needs to be learned and where the

Bible was written, whereas (2) just gives information on what Christian clergymen do. These findings indicate that specificity is not tied to entailment.

Specificity and Paraphrases Our intuition is that paraphrases mostly have the same specificity, as they display approximately the same information. In Example 6.7, we show a paraphrase pair:

1 She is an attentive waitress.
2 She is an attentive waitperson.

EXAMPLE 6.7: Statement pair showing intuitive relation behind specificity and paraphrase

(1) and (2) have the same level of specificity. Although *waitress* is more specific than *waitperson*, the personal pronoun *she* already gives the information on the gender, which makes the sentences equivalent with regard to their specificity.

However, in our study we found that proportionally all other dimensions have the same overlap of sentence pairs with specificity, meaning that between paraphrase pairs there is as much specificity as in contradiction, entailment, or pairs without any of the other relations. More specifically, of the pairs annotated with paraphrase or entailment, 73% are also annotated with specificity. The high number of pairs that are in a paraphrase relation, but also have a difference in specificity is interesting, as our intuition was that paraphrases are on the same specificity level. One example of this is pair #7 in Table A.1 :

1 Speaking more than one language can be useful.
2 Languages are beneficial in life.

EXAMPLE 6.8: Sentence pair #7 from Table A.1

Although they are paraphrases (with 100% agreement), (1) is more specific, as it specifies the ability of speaking a language.

Specificity and Semantic Similarity In our corpus, we found that there is no connection between the average semantic similarity score and the specificity gold label. The pairs with and without a specificity relation were equally distributed amongst all semantic similarity ratings. We must admit that given the corpus creation procedure, all sentence pairs have some kind of semantic similarity due to the shared topic they are on. So even if the semantic similarity is close to 0, the sentences are not 100% unrelated.[12]

Conclusion on Specificity and Other Dimensions In this study, we compare specificity with other dimensions in two ways: 1) the Pearson correlation and 2) overlap with other dimensions. In contrast to 1), 2) is not symmetric. We found that, against our intuition,

[12] As they are on a similar topic, there were no pairs with a similarity of 0.

specificity does not correlate with other relations, showing that it is independent of those in our corpus. Furthermore, paraphrase pairs contain as much overlap as uni-directional entailment pairs, i.e. paraphrases do not nessarily have the same specificity level as we assumed. Moreover, specificity also shows no clear trend on the similarity scale. In this study, we found no correlation with the difference in word length between the sentences, which differs from our finding in the unpublished study described in Section 6.1.3.2. This indicates that specificity cannot be automatically predicted using the other dimensions and requires further study.

6.5 Decomposing Specificity

Ko et al. (2019a) showed that specificity needs to be more linguistically and informational theoretically motivated to be more semantically plausible. We believe that decomposition may help with these issues. As specificity has not been studied extensively, it has also not been decomposed. To the best of our knowledge our work in Kovatchev et al. (2020) is the first work to do so.

Annotation Setting In Gold et al. (2019) (discussed in Section 5.1), we determine that in the therein created corpus there is no direct correlation between specificity and the other dimensions, including textual entailment. For that reason, we took a different approach to the decomposition of specificity and treat it separately from the other relations. We added one extra step in the annotation process focusing on the specificity relation (as opposed to the annotation of the other dimensions in the same study).

In Gold et al. (2019), we annotated for specificity at the textual level. That is, the crowd workers identified which of the two given sentences is more specific. In Example 6.9, the annotators would indicate that (2) is more specific than (1).

1 All children receive the same education.
2 The same education is received by all girls.

EXAMPLE 6.9: Statement pair with specificity

In Gold et al. (2019), we performed an additional annotation of the specificity and we identified the particular elements (words, phrases, clauses) in one sentence that were more specific than their counterpart. This is similar to the second step of the study performed by Li et al. (2016). After determining the specificity score in the first step, they identify which part of a sentence is underspecified. In our study, we have two sentences and determine which part is more specific than the other, whereas in the study by Li et al. (2016), the annotators are presented with one sentence only. However, due to the different settings in the annotation, further findings cannot be compared. In Example 6.9, we can identify that “girls” is more specific than “children”. The difference in the specificity of “girls” and “children” is the reason why (2) is annotated as more specific than (1). We called that SCOPE OF SPECIFICITY.

Results of Specificity Decomposition In 80% of the pairs with specificity at textual level, our annotators were able to point at one or more particular elements that are responsible for the difference in specificity. In 20% of the pairs, the specificity was not decomposable.

In our analysis on the nature of the specificity relation, we combined the annotation of SCOPE OF SPECIFICITY and the traditional annotation of linguistic and reason-based types discussed in the previous Chapter (see Section 5.2). In particular, we looked for overlap between the SCOPE OF SPECIFICITY and the scope of linguistic and reason-based types. Example 6.10 shows the two separate annotations side by side. In (1) and (2), we show the annotation of the linguistic and reason-based types: SAME POLARITY SUBSTITUTION (HABITUAL) of "children" and "girls", and DIATHESIS ALTERNATION of "receive" and "is received by". In (3) and (4), we show the annotation of the specificity: "children" and "girls". When we compare the two annotations we can observe that the SCOPE OF SPECIFICITY overlaps with the scope of SAME POLARITY SUBSTITUTION (HABITUAL).

1 All children *receive* the same education.
2 The same education *is received* by all girls.
3 All **children** receive the same education.
4 The same education is received by all **girls**.

EXAMPLE 6.10: SCOPE OF SPECIFICITY overlaps with the scope of SAME POLARITY SUBSTITUTION (HABITUAL)

We argue that when there is an overlap between the SCOPE OF SPECIFICITY and a linguistic or a reason-based type, it is the linguistic or reason-based phenomenon that is responsible for the difference in specificity. In Example 6.10, we can say that the substitution of "children" and "girls" is responsible for the difference in specificity.

ID	Type	Freq.
3	Derivational Changes	1
5	Same Pol. Sub. (habitual)	17
6	Same Pol. Sub. (contextual)	9
7	Same Pol. Sub. (named entity)	2
9	Opp. Pol. Sub (habitual)	2
11	Synthetic / Analytic sub.	9
14	Negation Switching	1
16	Anaphora	1
23	Addition / Deletion	50
27	Cause and Effect	7
28	Condition / Property	1
33	Transitivity	1
34	Other (General Inferences)	1

TABLE 6.3: Decomposition of specificity in percent

Table 6.3 shows the overlap between SCOPE OF SPECIFICITY and *atomic types*. In 97% of the cases where specificity was decomposable the more/less specific elements overlapped with an atomic type. In 50% of the cases the specificity was due to additional information (#23). The other frequent cases include *same polarity substitution* (#5, #6, and #7), *synthetic/analytic substitution* (#11), and *cause and effect* (#27) reasoning. While the overall tendencies are similar to the other meaning relations, specificity also has its unique characteristics. We found almost no specificity at the morphological level and the frequency of *Same polarity substitution* (#5, #6, and #7), while still high, was lower than that of paraphrasing and contradiction pairs. The relative frequency of *Synthetic/analytic substitution* (#11) was the highest of all relations and the reasoning types were almost as frequent as in entailment pairs, although the type distribution is different. We found no syntactic or discourse driven specificity changes.

6.6 Application of Specificity

According to Louis and Nenkova (2011), the automatic distinction of specificity would be beneficial in many applications, such as prediction of writing quality, text generation, and information extraction. In this section, we present studies and experiments in which the utility of specificity in other tasks was researched. Furthermore, we describe the application that serves as an example throughout the thesis—filtering for user-specific hotel reviews.

Summarization Louis and Nenkova (2011) used specificity classification to evaluate summarization. They used summaries from the Document Understanding Conference (DUC). 2005[13], which was the task to create either specific or general summaries. Louis and Nenkova (2011) used their automatic approach to predict the specificity of sentences taken from specific or general summaries. The automatically computed mean specificity value of individual summaries significantly varied between specific and general summaries. In this way, Louis and Nenkova (2011) showed that automatic specificity classification would be useful in summarization.

Argument Extraction Swanson et al. (2015) research whether specificity, amongst other factors, is helpful in argument extraction. They believe it is helpful, as a specific text mention may indicate a focused detailed argument and that specific arguments indicate good candidates for argument extraction. In their study, argument quality is crowdsourced to be used as gold standard. They analyze correlations between potential features and the training data. For two out of four topics (gun control and gay marriage) the specificity predicted by Speciteller (Li and Nenkova, 2015) was amongst the ten most correlated features. It was also found to be a helpful feature for these topics in the automatic classification of argument quality.

[13] http://duc.nist.gov/duc2005/

Dialogue Generation In a previous study, Ko et al. (2019a) were already able to show that specificity is useful in dialogue generation. In a more recent study, using a sequence-to-sequence dialogue model explicitly conditioning on specificity, Ko et al. (2019b) integrate linguistic, information-theoretic, and frequency-based specificity metrics. Their main framework is an attention-based sequence-to-sequence model with an Long short-term memory (LSTM) decoder and encoder with attention applied to the decoder. The encoder takes word embeddings as input.

To condition on an explicit specificity level, specificity is represented as a collection of real-valued metrics that can be estimated for each sentence independently. They learn embeddings of various specificity levels for each metric jointly with the model. The specificity metrics used by Ko et al. (2019b) are NIWF, PPW, and the system by Ko et al. (2019a), which is linguistically informed.[14]

Ko et al. (2019b) found that incorporating specificity to response generation in dialogue systems led to more interesting responses, with 6-10% improvement in informativeness and 3-7% improvement in topic relevance. However, their analysis also discovered that "30% of specific responses suffer from a range of problems from semantic incompatibility to flawed discourse" Ko et al. (2019b, 3457). These problems were then solved with plausibility re-ranking methods.

Filtering for User-Specific Reviews The application described in this thesis is a task that could be placed somewhere between summarization and argument extraction: On the hand we need to summarize information, on the other hand the best arguments mentioned in the reviews should be extracted. Hence, if it helps in these tasks, it is also potentially helpful in the filtering process. While Swanson et al. (2015) argue that a specific text mention may indicate a focused detailed argument and that specific arguments indicate good candidates for argument extraction, we believe that too specific statements might be unfavorable, e.g. there are reviews ranting about the size of keyholes in hotel rooms—which is very specific, but probably not helpful for other users. Hence, we believe that filtering on the right level of specificity is important. Ko et al. (2019a) showed the potential of specificity annotations by applying their system, which was trained on news texts, to vastly different domains on the example of Yelp, Twitter, and movie reviews. In this way, they showed that applying their system to the task of filtering for user-specific reviews is possible. The full outline of how the dimensions discussed in this thesis, including specificity, can be used for user-specific review filtering is presented in Section 8.2.5.

Figure 6.6 illustrates a concrete example for this step. Given several clusters of paraphrases, the specificity can be rated for each cluster, but also within the clusters. The level of specificity may vary according to the user's needs, as both too general and too specific statements should be excluded, e.g. in Figure 9 the statement "Comfy room" is too general, while "Keyhole too big" is too specific.

[14]The metrics are discussed in detail in Section 6.3.

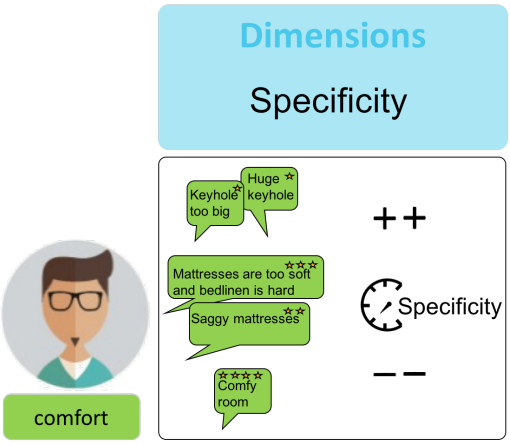


FIGURE 6.6: Illustration of specificity in exemplary user-specific filtering workflow

6.7 Conclusion on Specificity

In this chapter, we provided an overview of work performed on specificity, including both work by others as well as our own. In the following, we outline the most important conclusions on the aspects of specificity that were discussed in this chapter.

Understanding Specificity As specificity has not been widely researched, we tried to get a better understanding of this dimension. We tried to follow Ko et al. (2019a)'s claim that specificity needs to be more linguistically and information-theoretically motivated to be more semantically plausible by conducting three semantic annotations on this dimension. In one of these studies, we compare specificity to other semantic dimensions in order to get a better understanding of specificity itself (see Section 6.4). However, we find that specificity does not have any particular relation to other dimensions in our corpus. In another one of our studies, we decompose specificity and find that in most cases differences in the specificity level can be retraced to additional information in the more specific statement.

Operationalization of Specificity In our studies, we found that the best way to operationalize specificity is comparatively and with a given aspect. That is similar to that of sentiment, a dimension which is the focus of the next chapter.

Comparative Annotations In our three studies we perform comparative annotations, which has not been done before. In our study on relations between dimensions (see Section 6.4) and in our decomposition study (see Section 6.5), we used paired comparisons, while in our unpublished study we used BWS with one additional focus given. All three annotation studies have reliable IAA, showing that specificity can be annotated in a comparative way.

Specificity with a Given Aspect Furthermore, we believe that the setting of the unpublished study is more reliable, as it asked for specificity with a given aspect, similar to the task of ABSA. In Example 6.11, we show that given the aspect ATMOSPHERE, (1) is more

specific, as it provides more information on the atmosphere. Given the aspect of SERVICE, however, (2) is more specific.

1 The hotel room was small, but homely with a cozy lighting.
2 The quick and attentive room service brought a delicious breakfast to my comfy room.

EXAMPLE 6.11: Statement pair with different specificity aspects

Specificity on the Sentence Level However, our unpublished study using BWS as well as the decomposition study in Kovatchev et al. (2020) indicate that working on the sentence level for specificity is disadvantageous. In the BWS-study, we showed that sentence length strongly correlates with specificity. However, this finding could not be confirmed in our study in Gold et al. (2019). In Kovatchev et al. (2020), we showed that when decomposing sentence pairs with differing specificity, in most cases there is a segment in one of the sentences is not present in the other. The issue when working on the sentence level could be solved through working on the propositional level instead.

Automation of Specificity In our survey on the automation of specificity in Section 6.2, we describe a development from binary to numeric classification. Ko et al. (2019a), presenting the most recent sentence specificity system, introduced a system that in contrast to prior ones generalizes to other domains and uses a real-valued classification of specificity.

Chapter 7

Sentiment of Statements

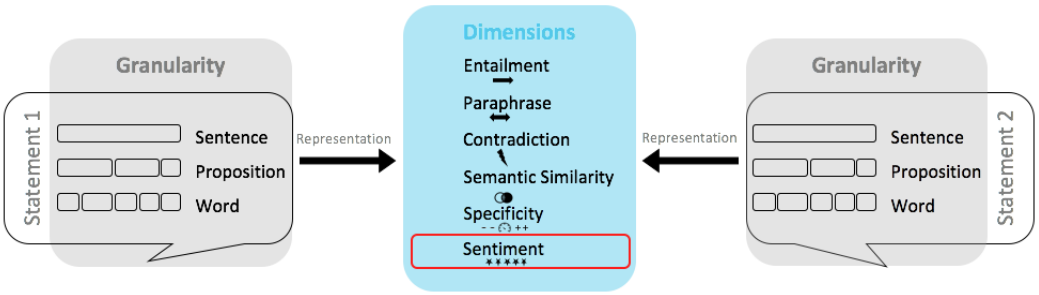


FIGURE 7.1: Illustration of sentiment amongst the other relation dimensions in this thesis

Sentiment is a sub-area of subjectivity analysis (Pang and Lee, 2008). According to Pang and Lee (2008)'s survey on sentiment analysis, interest in this topic began it's rise as early as with Carbonell (1979). Sentiment analysis can be used interchangeably with the term *opinion mining* (Pang and Lee, 2008) and denotes the task to systematically identify, extract, quantify, and analyze subjective information. Its goals include computers to recognize emotion in text (Pang and Lee, 2008).

Sentiment analysis is widely applied in genres including customer materials such as reviews and survey responses, but are also needed in the analysis of political text or social media texts. A current example of sentiment analysis being useful is the issue of hate speech—social media platforms are demanded to delete statement containing hateful and discriminating statements. In two of our studies, we annotated sentiment on individual statements Benikova et al. (2017); Gold et al. (2018), whereas we used best-worst scaling (BWS) in our most recent study Wojatzki et al. (2018a) on sentiment. These two studies are focus on an an extreme case of aspect-based sentiment—namely hate speech. Due to the importance and the coverage of this topic in this thesis, hate speech is discussed individually in the next chapter—Chapter 8.

According to Liu (2012) the task of aspect-based sentiment analysis (ABSA) consists of two subtasks:

- aspect extraction
- aspect sentiment classification

The first task is assigning an aspect to an utterance, mostly a sentence or a tweet. As in the ABSA shared tasks (Pontiki et al., 2014, 2015, 2016), these aspects are mostly predefined.

The second task is mostly a binary (POSITIVE, NEGATIVE) or trinary (POSITIVE, NEGATIVE, NEUTRAL) predefined set.

Mostly, sentiment is annotated towards one statement, as done in the ABSA tasks. An example of this is shown in Figure 7.2. Therein, three aspects are extracted: WIFI, COMFORT, and LOCATION. In the sentiment classification step, the first two are annotated as NEGATIVE, whereas the third is POSITIVE.

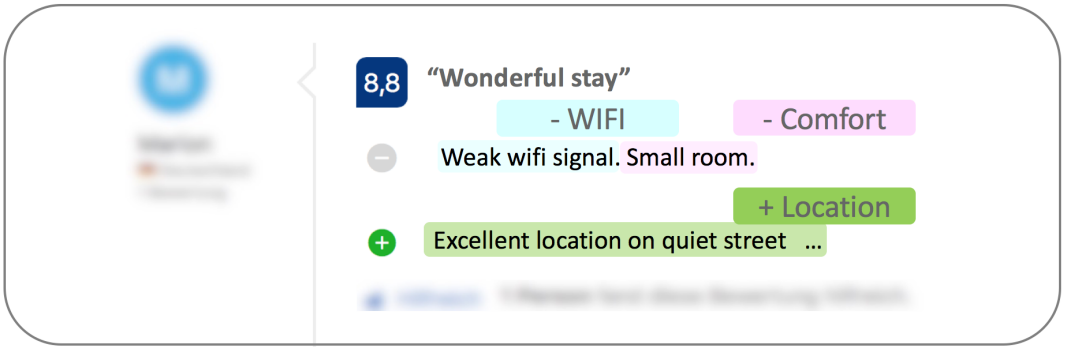


FIGURE 7.2: Example of aspect-based sentiment in hotel review

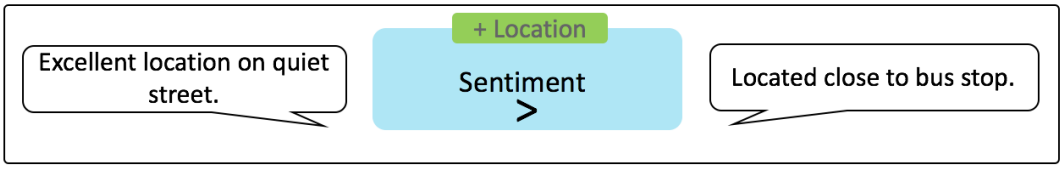


FIGURE 7.3: Example of a comparative aspect-based sentiment in hotel review

However, as previously stated, similar to the other dimensions, sentiment can also be annotated in a comparative way, as shown in Figure 7.1. In this case, several statements are compared regarding their sentiment towards a given aspect. A concrete example of a comparative sentiment annotation is shown in Figure 7.3. Given the aspect LOCATION, and a positive sentiment, the left statement would be annotated as more positive than the right one.

Aspect-Based Sentiment in Political Text In one of our works, we annotated and analyzed aspect-based sentiment within statements of the presidential debates of the 2016 US election (Gold et al., 2018). In this work, we are able to show numerically which topics are discussed in which way by each of the candidates and whether one of the candidates dominates the debates and how positive or negative they speak overall as well as about certain topics. We annotate the corpus according to two different schemata and analyze their differences. We show that the choice schema has a strong impact on the result of aspect-based sentiment analysis. Furthermore, we provide a corpus that can be used as a gold-standard for automatic aspect-based sentiment annotation of political debates. This study is presented in Section 7.1.

7.1 Aspect-Based Sentiment Analysis in Political Texts

In Gold et al. (2018), we present a corpus of political debates annotated with aspect-based sentiment and a corpus analysis. The source corpus consists of transcribed speeches taken from the two presidential debates of the 2016 US election. We conduct our study by first performing a manual annotation and analysis of the last presidential debates in the US, and then we show how this information can be extracted automatically.

Problem description Political debates are a fruitful source for ABSA, as the main goal of such a debate is the expression of sentiment towards certain aspects. Although there is much interest in the semantic annotation and analysis focusing in the presidential debate of the 2016 election, there is not much work available on aspect-based sentiment analysis in political debates.

Solution Idea In this study, we show that aspect-based sentiment annotations can help to obtain insights into aspects that are discussed in a political debate as well as the sentiment towards them. The corpus creation process is shown in Figure 7.4. In the herein presented corpus analysis, we are interested in how much they speak about different topics and whether they emphasize different topics, indicating different priorities. Furthermore, it is of interest how positive or negative they speak in general and if there are any peculiarities in the polarity with which they speak about a topic. The first and third of the three presidential debates between Hillary Clinton and Donald Trump were chosen for analysis. This gave us enough data and further enabled us to look for possible differences between the two debates. Furthermore, we trained a state-of-the-art classifier on the first debate and applied it to the third in order to show the applicability of the dataset for automatic aspect-based sentiment analysis.

Additionally, we research the impact of annotation schema for aspect-based sentiment on the resulting annotation, automation, and data analysis. Based on the assumption that annotation schema has a decisive effect on the outcome of the annotation, we annotated a part of the corpus using two different schemata for aspect-based sentiment and performed a comparative analysis of these.

Outcome The contributions of this study are

- a freely-available aspect-based sentiment annotated political debate corpus[1] (see Section 7.1.2)
- a comparison of two different aspect-based sentiment schemata (see Section 7.1.3)
- the analysis of the corpus (see Section 7.1.3)
- the discussion of the possibility to use this corpus for automatic training (see Section 7.1.4)

[1] `https://github.com/MeDarina/PoliticalABSA`

7.1.1 Related work of ABSA in political texts

In the political field, such analyses are used to track political views, detect consistency of political statements and actions, predict election results, or to determine the polarity of the blogoshpere. Semantic annotation and analysis is a current area of interest for the Natural Language Processing (NLP) community, many works focusing on the presidential debate of the 2016 election (Patwari et al., 2017; Gencheva et al., 2017; Nakov et al., 2018; Jaradat et al., 2018). However, there is not much work available on aspect-based sentiment analysis in political debates. Maynard and Funk (2011) extracted triples consisting of person, opinion and political party from pre-electional tweets. However, this kind of annotation is quite restrictive in the choice of data and possibly not applicable to debates between politicians. Balahur et al. (2009) investigated different approaches for binary sentiment and opinion classification on documents, on congressional floor debates. While this work is close to ours, Balahur et al. (2009) perform classification on whole documents, which is a coarse annotation. We, however, would like to extract as many sentiment mentions as possible in order to perform an extensive analysis. There are several corpora that extract stance, which can be shortly defined as aspect-based sentiment including implicit sentiment, from much-discussed political topics, such as death-penalty or same-sex marriage (Walker et al., 2012; Wojatzki and Zesch, 2016). To perform this kind of extraction, one needs much-discussed, controversial topics, we however want to capture the less discussed topics as well.

7.1.2 Presidential sentiment dataset

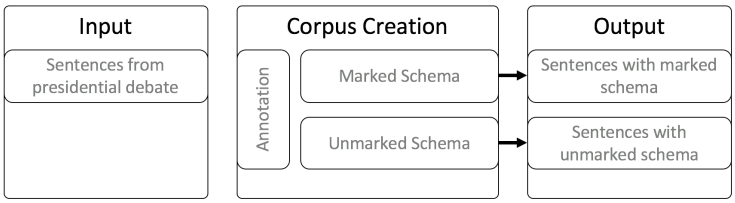


FIGURE 7.4: Corpus creation process of presidential sentiment dataset

Figure 7.4 shows the corpus creation process. Our input were transcripts of the presidential debates of 2016, consisting of over 2,000 sentences, which we annotated with aspect-based sentiment in a double-annotation process using two different schemata. We used a trinary sentiment annotation and the aspects AGENDA, UNITED STATES, GROUP, OPPOSITION, SELF, WOMEN, and OTHER. For both schemata, we performed a double-annotation with a subsequent curation using WebAnno (Yimam et al., 2013). All annotations were made considering the context of the election, the speaker (meaning whether it was spoken by Donald Trump or Hillary Clinton), and the context of the given sentence. Co-references outside the scope of the given sentence were not resolved, as we could not reliably provide this in an automated way, which is necessary for the automatic aspect-based sentiment classifier.

Source data and preprocessing As the basis for our dataset we used transcripts of the first and the third debate extracted from the website of the American Presidency Project[2]. After filtering for the parts spoken by the candidates, our source corpus consists of a total of 2,237 sentences (1,179 sentences in the first and 1,058 in the third debate). The data is preprocessed using the OpenNlpSegmenter provided by DKPro Core[3] (Eckart de Castilho and Gurevych, 2014). For the schema with noun and adjective markers, the data was further pre-annotated with nouns and adjectives using OpenNlpPosTagger.

Aspects We distinguish between seven pre-defined categories, which will be discussed in the following (we will not discuss the OTHER category). Their distribution in our dataset is shown in Table 7.1.

AGENDA refers to the speakers' political agenda. An exemplary excerpt from the debate containing this aspect is "I have a plan to fight ISIS", which also contains the aspect GROUP.

UNITED STATES refers mentions of the USA, including politics, economy, public figures, companies, etc. An exemplary excerpt from the debate containing this aspect is "Our country is suffering".

GROUP refers to any group other than the Americans, but also including Americans[4], e.g. ethnic minorities, countries and nations other than the US. An exemplary excerpt from the debate containing this aspect was named in AGENDA.

OPPOSITION refers to the other debater, including his or her agenda, biography, family, etc. An exemplary excerpt from the debate containing this aspect is "I call it trumped-up trickle-down".

SELF refers to the speaker, excluding his agenda, but including his beliefs, biography, family, etc. An exemplary excerpt from the debate containing this aspect is "I was secretary of state".

WOMEN refers to mentions of women and feminist topics, such as women rights, pay gap, and abortion. An exemplary excerpt from the debate containing this aspect is "Women's rights are human rights".

Annotation schemata To research the impact of annotation schema on sentiment analysis, we annotated the data using two schemata that share the same aspect and sentiment categories. Figure 7.5 shows an exemplary annotation of a sentence based on both schemata.

In the **unmarked schema** each aspect in a given sentence was annotated with its polarity. The exemplary sentence in Figure 7.5 reflects two aspects—US, which is annotated as NEUTRAL, and AGENDA, which is annotated as POSITIVE.

The **marked schema** is limited to aspect-based sentiment expressed through nouns and adjectives. In this way, the unitizing task of the aspect and sentiment markers is already given, which should further facilitate both the manual as well as their automatic detection. This excludes other occurrences of sentiment expressions that are not expressed using adjectives

[2] `http://www.presidency.ucsb.edu` (retrieved on June 14th, 2017)

[3] `https://dkpro.github.io/dkpro-core/`

[4] In the case of mentions such as American Christians or any hyphenated Americans, they are annotated as GROUP in the marked schema and as both UNITED STATE and GROUP in the unmarked schema.

	Marked Schema			Unmarked Schema [a].		
Aspect	Sum	κ	%	Sum	κ	%
Agenda	322	.59	7	140	.94	8
US	1127	.70	24	455	.84	27
Group	503	.86	11	168	.95	10
Opp.	362	.66	8	244	.96	14
Self	142	.59	3	315	.87	18
Women	93	.79	2	6	.99	0
Other	2194	.68	46	389	.79	23

TABLE 7.1: Distribution and inter-annotator agreement on individual classes across both annotation schemata in our corpus

[a]Note that the unmarked schema was annotated only for the first debate, whereas the marked schema was annotated for the first and the third

and nouns. However, we chose for this limitation as we believe that through it we gain a higher agreement of annotators and also automatic methods and thus a more reliable analysis. With the comparison to the unmarked schema, we are able to analyze whether and when this limitation is useful.

Our annotation schema consists of three layers: 1) Entity layer, 2) Aspect layer and 3) Sentiment layer. Figure 7.5 shows an exemplary excerpt of a sentence annotated with this schema.

In this way, the entity layer refers to nouns and adjectives, potentially expressing aspect and sentiment. In Figure 7.5, the entities are "new", "good", twice "jobs", "rising" and "incomes" . The aspect layer represents the aspect that a noun refers to. In Figure 7.5, in the marked schema, "jobs" and "incomes" refer to the aspect AGENDA. The sentiment layer represents the sentiment of an adjective expressed towards an aspect—"good", "new", and "rising" express a POSITIVE sentiment towards the aspect AGENDA.

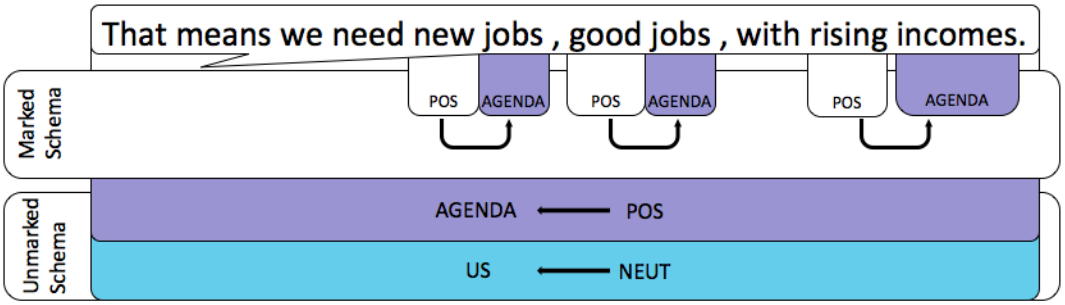


FIGURE 7.5: Example of aspect-based sentiment annotation schemata

Annotation process All data was double-annotated and consequently curated. The annotations followed a set of guidelines[5], which was improved iteratively. For the evaluation of each annotation we report Cohen's κ Cohen (1960). The different schemata were annotated consecutively to avoid bias.

[5]See either Appendix A.1.3.1 or `https://github.com/MeDarina/PoliticalABSA`

Aspect	κ Aspect	κ Sentiment
US	.84	.73
Group	.95	.88
Opp.	.96	.91
Self	.87	.81
Women	.99	.99
Other	.79	.73

TABLE 7.2: Inter-annotator agreement on individual classes of aspect and sentiment in the schema without markers

Only the first debate was annotated using the **unmarked schema**. As each sentence could be annotated with several labels, we report a binary κ for each class, which is presented in Table 7.2.

The agreement is the highest and nearly perfect for WOMEN, as it is very rare and thus the annotators mostly agree that it is not present.

In the **marked schema**, we manually annotated in three steps, each of which was followed by a curation. Each curated version of the previous step was used for the next step. We calculated κ for each of the steps individually (Table 7.3). The agreement of annotators and curation can be gathered from Table 7.4. The stree steps were the following:

1 The first step was to annotate the relations between adjectives and nouns on the **entity layer**. The aim for this step was to agree on which adjective referred to which noun. The inter-annotator agreement increased from the first to the third debate (Table 7.3). Agreement between annotators and curation was between $\kappa = .72$ and $\kappa = .85$ for the first debate and varied from $\kappa = .74$ to $\kappa = .87$ for the third debate.

2 The second step was a topical classification of the nouns on the **aspect layer**. As expected for such a high number of possible tags, the inter-annotator agreement was lower for this step. For the first debate it reached $\kappa = .71$, and slightly increased to $\kappa = .73$ for the third debate. The agreement between curation and annotation was between $\kappa = .75$ and the highest $\kappa = .93$. The agreement between curation did not get better overall, but became more stable—κ varying between .85 and .88.

3 The third step assigned a polarity to each of the curated relations on the **sentiment layer**. Agreement for this step was $\kappa = .66$ for the first debate, but dropped strongly to $\kappa = .50$ in the third debate. The agreement between annotators and curation varied from $\kappa = .78$ to $\kappa = .88$ for the first and $\kappa = .67$ and $\kappa = .79$ for the third debate.

Furthermore, Table 7.1 shows κ for each aspect individually for both debates together. AGENDA and SELF have the lowest agreement ($\kappa = 0.59$). The most disagreement for those classes is with the class OTHER, meaning that mostly one annotator saw the aspect and the other did not. This is mostly resolved through the curation, as it is not a classic disagreement, but rather a missing of the aspects.

	1st debate	3rd debate
Entity Layer	.62	.66
Aspect Layer	.71	.73
Sentiment Layer	.66	.50

TABLE 7.3: Inter-annotator agreement for both debates and all three annotation steps using κ

	Curated Version							
	Marked Schema						Unmarked Schema	
	Entity Layer		Sent. layer		Aspect Layer		Sent. Layer	Aspect Layer
Debate	1st	3rd	1st	3rd	1st	3rd	1st	1st
Annotator 1	.74	.87	.88	.79	.89	.89	.88 - .99	.91 - .99
Annotator 2	.72	.74	.78	.68	.81	.80	.84 - .99	.87 - .99

TABLE 7.4: Agreement using κ for each annotator and the curated versions

7.1.3 Corpus Analysis

First, we will report on the syntactic analysis, followed by a comparison of the polarity distribution for both speakers and the topics they choose to emphasize.

Comparison of the two schemata To make the schemata comparable, the annotation of the marked schema was slightly formatted:

- all aspects and their sentiments were attached to the full sentence, in this way deleting the marking
- if a sentence contained several sentiments towards one aspect, the neutral sentiment was dropped.[6]

This left each sentence with exactly one sentiment per aspect. The transformation for the marked schema making in comparable to the unmarked is shown in Figure 7.6. The example would have the aspect AGENDA with a positive sentiment and no other aspect.

Table 7.5 shows the binary κ between the marked and unmarked schema annotation. This means that the κ was calculated for each class individually due to the multi-label annotation of the unmarked schema, similarly to the IAA calculation of the unmarked schema. We only show the agreement on the aspect, as there was close to no agreement on their sentiment. Also, the agreement on the aspect annotation is very low, except for the label WOMEN, which is high due to its rarity. The agreement is not given for the class OTHER, as it has a κ of .03. The annotation according to the marked schema contained more annotations of this label. This may be due to the fact that according to our guidelines each noun had to be annotated with an aspect, although some did not represent any. In the unmarked version, each sentence had to be annotated with at least one label, too. As the information contained in a full sentence being potentially higher, the label was annotated less. SELF, also having a very

[6]There were three occasions in which there was both positive and negative sentiment towards one aspect in a single sentence. These sentences were excluded from the comparison.

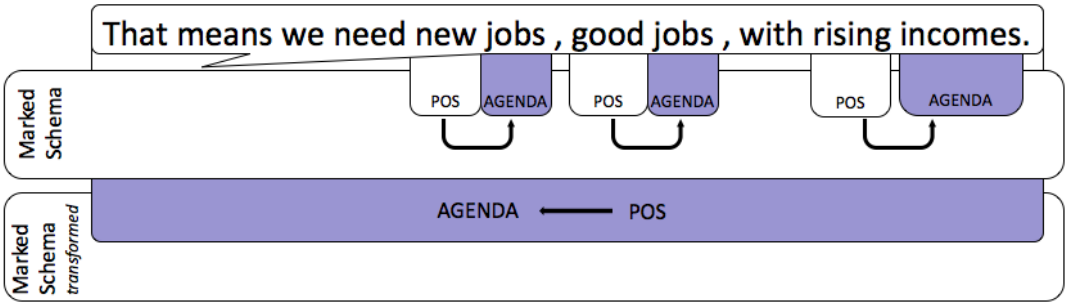


FIGURE 7.6: Example of transforming marked schema to unmarked schema

low inter-annotator agreement, was annotated much more often in the unmarked version (see Table 7.1), probably due to the use of first-person pronouns which were not annotated in the marked version.

Aspect	κ
Agenda	.60
United States	.46
Group	.57
Opposition	.44
Self	.19
Women	.86
Other	.03

TABLE 7.5: Binary κ between marked and unmarked annotation of aspect

When comparing the annotation of the two schemata, there is a great difference in the polarity distribution, which can be especially seen in the aspects AGENDA, US, and SELF. It could be explained with some parts, namely nouns and their adjectives, having a strong polarity, which is lost in the full sentence, e.g. in the case of AGENDA "My plan is to support our great school system.", while "great school system" is a positive point on ones agenda, the overall sentence is purely informative, meaning *neutral* towards AGENDA. Furthermore, the marked annotation denotes two debates, whereas the unmarked annotation denotes only the third debate, which is also an explanation for the big differences in the class distribution.

Analysis of polarities and aspects To perform the aspect-based sentiment analysis, we used the annotation as described in Section 7.1.2. Here, we compare the percentage amounts, if not mentioned otherwise.

Sentiment analysis The distribution of the polarity of these relations is shown in Table 7.6. While Clinton expresses more positive sentiment than negative sentiment in the marked schema, this is different in the unmarked schema. This may be due to her frequent use of several positive facts in a sentence, which in the sentence result in a rather neutral or even negative sentiment. An example for this is shown in Figure 7.5—while there are three positive annotations in the marked schema, there is a neutral annotation in the unmarked

	Marked				Unmarked	
	1st		3rd		1st	
	Clinton	Trump	Clinton	Trump	Clinton	Trump
pos	.32	.27	.28	.24	.17	.12
neut	.52	.41	.49	.38	.62	.59
neg	.17	.32	.23	.38	.21	.29

TABLE 7.6: Ratio of the polarities for both candidates and debates.

schema. Another possible explanation are her longer sentences overall[7], where she could list a lot of positive facts, which in the sentence sum up to only one positive mention, whereas her negative mentions may be fewer in one sentence.

In both schemata and debates, here is a clear predominance of negative relations with Trump. There is a decrease in the proportion of neutral relations from the first to the third debate for both candidates, indicating more polarized statements, as shown in Table 7.6.

Aspect analysis In this subsection, we discuss the distribution of individual aspects for each of the candidates. Table A.5 shows this distribution for each of the schemata. After OTHER, US is the most discussed aspect for both candidates in both schemata, which is understandable given the context of the debate. Both candidates discuss this topic with nearly the same frequency. This is also the case for OPPOSITION and OTHER, whereas the other aspects display differences in the frequency. These differences will be discussed in the following.

AGENDA Comparing the distribution of the aspects (see Table A.5), the biggest difference emerges with sentences referring to what a candidate intends to do once elected. It is also the second biggest difference in nouns referring to an aspect. While 9% of all nouns and 13.8% of sentences used by Hillary Clinton are classified as belonging to the AGENDA aspect, only 4.63% of nouns and 4.3% of sentences used by Donald Trump are, which is nearly half or one third as much. Irrespective of Donald Trump's overhead of negative polarities, adjectives referring to these nouns are positive in 80% of the cases. As shown in the table, only 25% of the sentences with this aspect are negative. This case is similar when comparing the percentages between the two schemata annotations for Hillary.

US Clinton expressed much less negative sentiment towards theUS aspect than Trump in both schemata, which reflects his criticism on the current situation, government and ruling party, while Clinton is positive on these sub-aspects.

GROUP Given existing prejudices accusing Donald Trump of racism, as indicated in some articles[8], the polarity of relations in reference to groups was of particular interest. For these nouns there was in fact a higher than average percentage of negative adjectives (40.74%) and sentences overall (30.6%) for him, whereas Clinton's sentiment was much less negative and

[7] In both debates, Clinton's sentences are on average four words longer.

[8] https://www.nytimes.com/interactive/2018/01/15/opinion/leonhardt-trump-racist.html
https://edition.cnn.com/2018/03/02/opinions/why-americans-think-trump-is-a-racist-louis/index.html

more neutral in both schemata. However, the fact that she also uses less positive adjectives and sentences than Donald Trump means that the prejudice could not be confirmed.

OPPOSITION Both candidates spoke similarly much and with similar sentiment on their opposition, namely more that 50% negatively, in both schemata.

SELF In the unmarked annotation, Trump speaks more about himself than Hillary, whereas in the marked annotation the percental amount is similar.

WOMEN There was not much talk on feminist issues, as suggested by some news articles[9]. Merely 70 nouns and six sentences of Hillary Clinton, and 23 nouns and no sentences of Donald Trump referred to women.

OTHER In the marked annotation schema, nearly half of the aspect annotations for both candidates are marked as OTHER, whereas for the unmarked it is much less. The difference is probably explainable with many individual nouns not referring to any of the aspects, but the overall sentence referring to at least one of them. However, in both schemata it is the most frequent label for both candidates, showing that there are still some aspects that are not covered by our schema, e.g. gun control or drug smuggling. This is a typical problem of pre-defined aspects and can be only partly solved by introducing new classes.

Comparison between debates The comparison between the first and the third debate can only be made on the marked annotation version. We summed up the changes in percentages of the noun classes between first and third debate for both candidates. This revealed a stable distribution for both candidates, the difference being nearly the same for each of the classes. Both candidates became nearly equally more negative and less neutral and positive in the third debate. Interestingly, the change towards the negative sentiment is the strongest in one aspect for both candidates: they both talk more negative about their opponent.

7.1.4 Automatic aspect-based sentiment annotation

In order to see whether the corpus can be used for training an aspect-based sentiment classifier, we trained an off-the-shelf system for both tasks, namely aspect recognition and sentiment recognition separately, using an Support Vector Machine (SVM) (Cortes and Vapnik, 1995), (LibSVM in DKProTC (Daxenberger et al., 2014)). For both tasks, we trained each aspect separately, as is usually done in ABSA-tasks. The data of the unmarked schema was processed in the same way as for the comparison of the schemata (see Section 7.1.3).

We evaluated our corpus by performing 10-fold Cross Validation (CV) on the first debate. We experimented with several feature sets—each feature individually as well as in combination with unigrams.

In the case of the marked schema, we tested the therein found best feature constellation on the third debate for aspect detection.

We experimented with n-gram features with $n \in \{1, 2, 3\}$, list features, and embeddings.

We used three lists that are usually used in the ABSA-tasks, namely the MPQA (Wiebe et al., 2005), the extended version of Bing Liu's dictionary (Hu and Liu, 2004), and the

[9] https://www.nytimes.com/2016/10/21/us/politics/hillary-clinton-women.html

Feature sets	Features	Marked schema						Unmarked schema					
		Aspects						Aspects					
		Agenda	US	Group	Opp	Self	Other	Agenda	US	Group	Opp	Self	Other
Majority Class Baseline		.92	.71	.87	.88	.94	.58	.88	.61	.88	.79	.73	.67
Individual Features	1gram	.93	.83	**.94**	**.93**	**.96**	**.79**	.92	.81	**.92**	.89	.91	**.79**
	2gram	.92	.78	.88	.90	.94	.66	**.93**	.75	.87	.85	.90	.72
	3gram	.93	.75	.88	.88	.95	.58	.92	.71	.86	.83	.87	.71
	list	.88	.71	.86	.79	.75	.71	.92	.75	.87	.88	**.94**	.70
	emb	.92	.78	.88	.87	.95	**.79**	.90	.78	.86	.80	.89	.74
1grams +	1+3 gram	.92	.79	.90	.92	.90	.76	**.93**	.80	.88	.89	.91	.77
	1gram+list	.92	.83	.92	**.93**	**.96**	**.79**	.91	.81	.89	.88	.90	.78
	1gram+emb	**.94**	**.84**	.93	**.93**	**.96**	**.79**	.92	**.82**	**.92**	**.90**	.91	**.79**
	1+2 gram	.91	.80	.91	.92	**.96**	.78	.93	.79	.89	.89	.90	.78

TABLE 7.7: F-scores for aspect models using CV on first debate

AFINN dictionary (Nielsen, 2011). These lists contain words and a corresponding sentiment that was mostly manually annotated, e.g. *good* has a *positive* sentiment in these lists.

To equip our classifier with semantic knowledge we used a feature derived from the Polyglot embeddings Al-Rfou et al. (2013).

There were too few occurrences of the label WOMEN to train a reliable model, hence we excluded the label from the training.

Furthermore, we only built a sentiment model for the unmarked schema, as the class distribution in this schema was too imbalanced and the occurrences of sentiment too sparse.[10]

Aspect extraction Table 7.7 shows the performance of several feature sets for aspects on the first debate for both schemata. Our performance measure is micro-F.

For both schemata, Table 7.7 shows that all models outperform the majority baseline. In the unmarked schema, the best-performing aspect, both in comparison with the majority baseline and with the other aspects, is SELF. The good performance may be explained by personal pronouns of the first person being a strong indicator for this class. Inter-annotator agreement is often regarded as an upper-bound for the performance of the classifier that is trained on this data. In the case of the unmarked schema, this bound is only reached for SELF.

Due to the imbalanced class distribution (see Table 7.1), the majority baseline is quite high for some classes, especially in the marked schema. Due to its higher majority class baseline, it is more difficult for the classifier to learn something meaningful from the data in the marked schema.

In the marked schema, it is not surprising that the aspects AGENDA and SELF, having a majority class baseline performance of >.9 are only slightly outperformed by some models. However, the models for the other aspects learn better. In the marked schema, the aspect model that classifies best when compared to the majority class baseline is OTHER. This is probably due to its more balanced class distribution and the fact that the performance of this model is mostly the poorest when compared to the other aspects.

[10]We experimented with models with the same features as described for the other classifiers, but these did not exceed the majority class baseline. Thus, we do not further report on this.

Aspects	Agenda	US	Group	Opp	Self	Other
Majority Class Baseline	.92	.73	.45	.90	.48	.59
1gram+emb	.92	.74	.45	.88	.49	.66

TABLE 7.8: Performance of best aspect model (1gram+emb) of first debate CV on third debate

	Features	Aspects					
		Agenda	US	Group	Opp	Self	Other
Majority Class Baseline		.87	.61	.86	.79	.73	.66
Individual features	1gram	.89	.71	**.88**	**.85**	.85	.72
	list	.88	.64	.86	.79	.75	.68
	emb	.89	.66	.86	.79	.83	.71
1-grams +	1gram+2gram	**.91**	.70	.86	.84	.85	.69
	1gram+3gram	**.91**	.71	.85	.84	**.86**	.71
	1gram+list	.89	.71	.86	.84	**.86**	**.74**
	1gram+emb	.90	**.72**	**.88**	.84	.85	.73

TABLE 7.9: Micro F-Scores for sentiment model

For most aspects, 1-grams models are amongst the best classifiers and are not highly outperformed by other models. Only in the case of SELF, the list-feature is .02 better than the 1-gram.

Table 7.8 shows the performance of the best model per aspect in the first debate and also how well it performed on the third debate for the marked schema. The performance of the best model is close to the majority class baseline, which shows that the features do not generalize well.

Aspect sentiment classification As shown by the majority baseline in Table 7.9 as well as the distribution in Table A.5, the class distribution for the sentiment is also uneven. However, all models outperform the majority baseline, even if not by far.

For the aspects AGENDA, SELF, and OPPOSITION, the classifier mostly distinguishes between neutral and one other sentiment—in the case of AGENDA and SELF positive and in the case of OPPOSITION negative, which clearly reflects the data as well as spirit of a presidential debate. In the case of SELF, the aspect can probably be learned better due to the use of pronouns, as explained in the previous subsection.

7.2 Summarization and Conclusion on Sentiment

We show that our manual aspect-based annotation of the presidential debates is reliable in the unmarked schema, but less so in the marked schema.

The marked schema had a worse annotator agreement, a more imbalanced class distribution and could only partly be used for automatic annotation. Overall, the marked schema was unfavorable and should not be used further. Through the use of all nouns and their adjectives we intended to improve the issue of unitizing when finding the marker for the aspects.

Additionally, we could show that Clinton talks about her agenda nearly thrice as much as Trump, while Trump talks a little more about the opposition than Clinton.

Overall, we could show that our dataset and the unmarked schema can be used perform aspect-based sentiment annotation and analysis in political debates in order to gain evidence on the discussed aspects and their sentiment. However, annotating aspect-based sentiment remains a challenge. Furthermore, we show that schema plays a big role in both manual and automatic aspect-based sentiment annotation.

Furthermore, we performed an off-the-shelf classification on the herein created dataset, which showed that the skewed class distributions represent a major obstacle for off-the-shelf methods.

We identified the uneven class distribution as one potential source for the difficulties in training. For the marked schema, we applied the aspect best model of the first debate to the third and found that the model is not well transferable. Probably some aspects, e.g. AGENDA and GROUP discussed different sub-aspects in the two debates, which may have lead to a decrease in the performance on the test set. Sentiment detection did not work for the marked schema, but seemed to work for the unmarked one. However, we did not have enough data to transform the findings of the CV to a test set.

Overall, in Gold et al. (2018), we showed that sentiment annotations are difficult, especially when performing them on sub-sentential level. In the next chapter, which will discuss an extreme form of negative sentiment—namely hate speech, we describe our study in Wojatzki et al. (2018a), where we perform a reliable comparative annotation on statement level. Although comparing these two studies is difficult due to their structural differences, the schema of Wojatzki et al. (2018a) seems more reliable and also cheaper in terms of time and money. We believe that comparative annotations are to be favored in sentiment tasks.

Chapter 8

Sentiment of Statements: The Case for Hate Speech

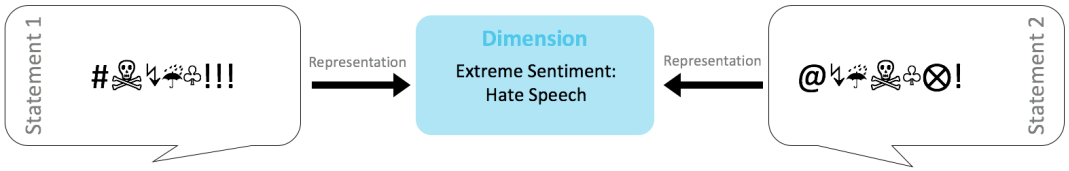


FIGURE 8.1: Illustration of operationalization of hate speech in this thesis

In the previous chapter, we discussed the superordinate topic of sentiment. We believe that hate is a very strong form of negative sentiment and that hate speech always has an *aspect* that is hated. In the spirit of the zeitgeist and the current debate on how to deal with the massive amount of hateful content in social media, we performed some studies on hate speech in statements. In earlier work on hate speech, hate speech has been framed as *abusive* or *hostile* messages or *flames* (Spertus, 1997). Other commonly used terms are *abusive language* (Waseem et al., 2017) or *offensive language* (Razavi et al., 2010), as well as sub-issues such as *cyberbullying* (Xu et al., 2012) or *trolling* (Mantilla, 2013). The effects of this unpleasant form of communication range from poisoning the atmosphere in social media to psychic or physical violence in the real world (Mantilla, 2013). To counteract the massive scale to which hate speech can occur in social media, automatic methods (pre-)identifying potentially hateful or threatening utterances are required. Its perception depends on linguistic, contextual, and social factors (Stefanowitsch, 2014). Hence, even for humans, the decision whether an utterance is hate speech or not is often difficult (Ross et al., 2016; Benikova et al., 2017). Research on hate speech focuses on an analysis of such utterances and methods for an automatic detection. We, however, focus on two factors influencing the perception of hate speech: implicitness/explicitness of a statement and group affiliation. For a better illustration of the herein presented results, all examples have been freely translated from German to English by the author. None of the examples reflects the opinion or political orientation of the author.

Implicitness and Explicitness as Influencing Factors of Hate Speech In Benikova et al. (2017), we analyze how the factors of implicitness and explicitness influence the phenomenon of hate speech. We create a corpus of parallel implicit and explicit statements annotated for

hate speech against refugees. Comparing the annotations, on average, it remains unclear which is more hateful, as the content is further moderated by content variables, e.g. explicit threats of violence are judged as more hateful than implicit ones, whereas implicit groundless suspicions are judged as more hateful than explicit ones. This study is presented in Section 8.1.

Group Affiliation as Influencing Factors of Hate Speech In Wojatzki et al. (2018a), we analyze how group affiliation affects the perception of hate speech. The examined group affiliation is gender. To do so, we create a corpus containing statements on topics concerning women such as gendered language, gender pay gap, or social roles. Hereafter, equally sized groups of women and men assess whether the statements are hate speech and whether they agree with them. We find that strong hate speech is judged as such regardless of gender affiliation. Furthermore, we find that there is a correlation between agreeing with a statement a judging it to be hate speech. Using this finding, we successfully use the agreement with a statement as a feature for the automatic training of hate speech judgments. This study is presented in Section 8.2.

8.1 Implicitness and Explicitness as Influencing Factors of Hate Speech

In Benikova et al. (2017), we examine the influence of implicitness and explicitness on the perception of hate speech. To do so, we used implicit 36 German tweets from the corpus of Ross et al. (2016) and paraphrase them to explicit versions. We perform a user study with 101 participants to obtain judgments on their perception as hate speech. Furthermore, we examine whether automatic hate speech detection is able to perceive a difference between implicitness and explicitness in hate speech.

Problem Description As mentioned earlier, the perception of hate speech depends on linguistic, contextual, and social factors (Stefanowitsch, 2014). In this study, we examine a specific dimension of this challenge—whether implicitness affects hate speech perception. Consider the following tweets:

Im. Alles recht ominäs mit dem Zugunglück. Wüsste gerne, ob die Lokführer Hassan, Ali oder Mohammed hießen #Fluechtlingskrise #Islamisierung

Ex. [...] Die Lokführer waren Muslime #Fluechtlingskrise #Islamisierung

(English translation)

Im. Everything was quite ominous with the train accident. Would like to know whether the train drivers were called Hassan, Ali, or Mohammed #RefugeeCrisis

Ex. [...] The train drivers were Muslims. #RefugeeCrisis

EXAMPLE 8.1: Implicit and explicit tweet with similar content

One could argue that that the first tweet in Example 8.1 is more offensive, since it evokes racist stereotypes by using allegedly prototypical Muslim first names as an implicit way of blaming Muslims in general. However, one could counter-argue that the second tweet is more offensive, as it explicitly accuses Muslims of being involved in a train accident. Additionally, the first tweet is hedged[1] by "Wüsste gerne" (*engl. "Would like to know whether"*), whereas it is implied that the second statement is rather factual. It remains unresolved whether implicit or explicit hate speech is perceived as more offensive and what the role of hedging is (Mamani Sanchez and Vogel, 2013).

In addition to the influence on the perception of hate speech, implicitness is a challenge for automatic hate speech detection. As most approaches rely on lists of abusive terms or phrases (Waseem and Hovy, 2016), hate speech often remains invisible if those explicit terms are missing. Or in terms of the above example, the classifier learns that it is hate speech to agitate against *Muslims*, but fails to learn the connection to *Hassan*.

Solution Idea To shed light on the influence of implicitness on the perception of hate speech, we construct a dataset in which we can experimentally control for implicitness. We select implicit hate speech instances from the German Hate Speech Twitter Corpus (Ross et al., 2016) and create explicit paraphrased counterparts. We then conduct a user study, wherein we ask participants to rate the offensiveness of either implicit or explicit tweets. We also show that a supervised classifier is unable to detect hate speech on both datasets. We hypothesize that there is a measurable difference in the perception of implicit and explicit statements in both human and automatic performance. However, we cannot estimate the direction of the difference, both alternatives seeming possible as shown in the previous discussion of the example.

Outcome We were able to show that there is a significant difference in the perception of explicit and implicit hate speech when comparing the direct paraphrases. However, on average, it is unclear which of them is more hateful, as the perception of hate speech seems to be moderated by content variables i.e. the implicit version is perceived as more hateful in insults, whereas the explicit version seems to be more insulting in threats. Furthermore, we were able to show that the phenomenon of implicitness and explicitness is invisible to automatic classification.

8.1.1 Theoretical Grounding

Our work is grounded in

- research on detecting hate speech
- the annotation and detection of implicit opinions
- paraphrasing

[1] As already mentioned in Chapter 4, "hedging is a textual construction that lessens the impact of an utterance. It is often expressed through modal verbs, adjectives, or adverbs, e.g. through "I believe that", "isn't it?", "I'm not an expert, but"." (see p.49)

Research on Detecting Hate Speech Hitherto, there has been no work on hate speech detection considering the issues posed by implicitness. Approaches based on n-grams or word lists (e.g. Sood et al. (2012); Chen et al. (2012)) are limited to detecting explicit insults or abusive language. Methods involving more semantics e.g. by incorporating Brown clusters (Waseem and Hovy, 2016; Warner and Hirschberg, 2012) are unlikely to cope with implicitness, as the necessary inferences go beyond word-relatedness.

Annotation and Detection of Implicit Opinions If we define hate speech as expressing a (very) negative opinion against a target, it is clear that there are relations to aspect-based sentiment analysis. However, sentiment analysis usually only models explicit expressions. For instance, the popular series of SemEval tasks on detecting aspect based sentiment, intentionally exclude implicit sentiment expressions and expressions requiring co-reference resolution in their annotation guidelines (Pontiki et al., 2014, 2015, 2016). Contrarily, the definition of stance, namely being in favor or against a target (i.e. a person, a group or any other controversial issue) explicitly allows to incorporate such inferences (for annotation guidelines see Mohammad et al. (2016) or Xu et al. (2016)). Thus, hate speech can also be considered as expressing a hateful stance towards a target. To clarify the difference between aspect-based sentiment and stance, we take the example in the introduction: The first tweet in Example 8.1 states a negative sentiment against "Hassan, Ali oder Mohammed" (*engl. "Hassan, Ali, or Mohammed"*). However, there is no explicit statement and thus no sentiment against MUSLIMS, which are, however, obviously targeted through the use of these prototypical names. As stance by definition also considers implicit postures against a target, it can be determined that in this example the target is MUSLIMS (which we integrate as shown in the explicit example). Consequently, we define explicit hate speech as expressing hateful sentiment and implicit hate speech as the instances which do not express hateful sentiment, but hateful stance. Therefore, this work relates to studies which use explicit opinion expressions to predict or rationalize stance (Boltužić and Šnajder, 2014; Hasan and Ng, 2014; Sobhani et al., 2015; Wojatzki and Zesch, 2016).

Paraphrasing The implicit and explicit versions of a tweet can be seen as paraphrases, i.e. units of texts containing semantically equivalent content (Madnani and Dorr, 2010). Paraphrases can be classified according to the source of difference between the two texts. Incorporating implicit stances is equivalent to the paraphrase class of *Ellipsis* (Kovatchev et al., 2020; Vila et al., 2014; Bhagat and Hovy, 2013) or the *Addition/Deletion* class (Kovatchev et al., 2020; Vila et al., 2014; Rus et al., 2014).

The modification of softening hedges indicating a clear stance into clearly stated statements corresponds to the classes of *modal-verb changes* (Kovatchev et al., 2020; Vila et al., 2014; Rus et al., 2014; Bhagat and Hovy, 2013), *Same-polarity substitutions* (Kovatchev et al., 2020; Vila et al., 2014), and *Synthetic/analytic substitutions* (Kovatchev et al., 2020), which are summarized as *Quantifiers* by Rus et al. (2014), and *General/Specific substitution* by Bhagat and Hovy (2013). To the best of our knowledge, paraphrasing techniques have not been used in the context of hate speech and its analysis.

8.1.2 Manufacturing Controllable Explicitness

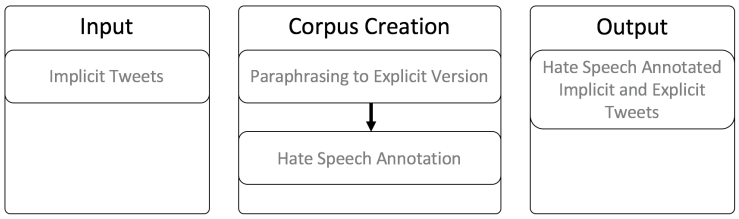


FIGURE 8.2: Corpus creation process of implicit and explicit hate speech

The corpus creation process is depicted in Figure 8.2. The basis of our data set is the German hate speech corpus (Ross et al., 2016) that contains about 500 German tweets annotated for expressing hate speech against refugees or not. We chose this corpus because it is freely available and addresses a current social problem, namely the debate on the so called *European refugee crisis*. To construct a data set in which we can control for implicitness, we perform the following consecutive steps:

- Restriction to tweets which contain hate speech, i.e. at least one annotator flagged a tweet as such
- Removal of tweets containing explicit hate speech markers
- Paraphrasing the remaining tweets to be explicit, so that we obtain a dataset which has both an implicit and an explicit version of each tweet

Indicators for Explicit Hate Speech We first identify tokens that are clear indicators for hate speech by retrieving words that are most strongly associated with hate speech. We hereby restrict ourselves to nouns, named entities, and hashtags, as we do not observe strong associations for other POS tags. We compute the collocation coefficient $Dice$ (Smadja et al., 1996) for each word and inspect the end of the spectrum associated with the hate speech class.

We observe the—by far—strongest association for the token "#rapefugee". Furthermore, we perceive strong association for cognates of "rape" such as "rapist" and "rapes". To further inspect the influence of these indicators, we compute the probability of their occurrence predicting whether a Tweet is hate speech or not. We find a probability of 65.8% for "#rapefugee" and of even 87.5% for the group of nouns related to "rape". When inspecting the tweets containing those explicit hate speech indicators, we observe that they are often considered as hate speech regardless of whether the rest of the tweet is protective of refugees. Because of this simple heuristic, we remove those tweets from our data set.

Paraphrasing To make the tweets explicit, we paraphrase them according to a set of rules[2], which correspond to previously mentioned paraphrase classes. We apply as many rules as possible to one tweet in order to make it as explicit as possible. As the corpus is concerned with the refugee crisis, we define ISLAM, MUSLIM, and REFUGEE as the targets of hate

[2] github.com/MeDarina/HateSpeechImplicit

speech. If a phrase does not explicitly contain them, we paraphrase it by adding this information as a new subject, object, or adjective or through co-reference resolution. An example for this rule is shown in [Ex.1] in Example 8.2.

Im. #Blutrache, #Zwangsbekehrung, #Scharia, #Kinderbräute, #Vielehe, #Genitalverstümmelung - kann nicht erkennen, was davon zu uns gehären soll.
Ex.1 [...] kann nicht erkennen, **wie der Islam** zu uns gehären soll.
Ex.2 [...] - **Es gehört nicht** zu uns.
Ex.3 [...] - **Der Islam** gehört nicht zu uns.

(English translation)
Im. #Vendetta, #ForcedConversion, #Sharia, #ChildBrides, #Polygamy, #GenitalMutilation - don't see how it belongs to us.
Ex.1 [...] - don't see how **Islam** belongs to us
Ex.2 [...] - It **doesn't** belongs to us.
Ex.3 [...] - **Islam doesn't** belongs to us.

EXAMPLE 8.2: Transforming implicit statement to explicit one using paraphrasing

If the message of the phrase is softened through hedges such as *modals* (e.g. "could", "should") and *epistemic modality with first person singular* (e.g. "I think", "in my opinion") these are either removed or reformulated to be more explicit. This reformulation is shown in [Ex.2]. However, as we apply as many rules as possible, the tweet would be paraphrased to its final version as shown in [Ex.3].

Rhetorical questions are paraphrased to affirmative phrases, as shown in Example 8.3.

Im. Gestern kamen die #Asylanten. Heute Nacht wurde versucht einzubrechen. #Zufall?
Ex. Gestern kamen die #Asylanten. [...] **#KeinZufall**

(English translation)
Im. Yesterday the refugees came. Today there's burglary. Coincidence?
Ex. Yesterday the refugees came. [...] **Not a coincidence!**

EXAMPLE 8.3: Transforming rhetorical questions to affirmative phrases

Furthermore, implicit generalizations are made explicit through the use of quantifiers, as shown in Example 8.4.

Im. 90% aller #Asylanten wollen nur deshalb nach D, weil sie nirgends auf der Welt mehr Geld bekommen! Nebenbei islamisieren. #Lanz
Ex. **Alle** #Asylanten wollen nur deshalb nach D, [...]

(English translation)
Im. 90% of all refugees want to come to Germany, only because nobody else will give them money! Islamize in passing. #Lanz
Ex. **All** refugees want to come to Germany, [...]

EXAMPLE 8.4: Transforming implicit generalizations through the use of quantifiers

The paraphrasing process was performed independently by two experts, who chose the same instances of implicit stance, but produced slightly differing paraphrases. The experts merged the two sets by choosing one of the two paraphrased versions after a discussion. The rules for incorporating the implicit stance were used more than thrice as often as the rules resolving softening hedges, showing the frequency of implicit stance in social media. The most frequently used rules were the incorporation of the stance target in the form of a noun phrase or an adjective. In 25% of the cases, two rules for paraphrasing were used, the rest was paraphrased using one rule only.

Automation To examine the influence of implicitness on automatic hate speech detection, we adapt and re-implement the hate speech systems of Waseem and Hovy (2016) and Warner and Hirschberg (2012) to German data. Thus, we rely on an Support Vector Machine (SVM) (Vapnik and Chervonenkis, 1981) equipped with type-token-ratio, emoticon ratio, character, token, and POS uni-, bi-, and trigams features. The system is implemented using the dkpro-tc framework (Daxenberger et al., 2014) and uses twitter specific tokenization (Gimpel et al., 2011) and the Stanford POS-tagger (Toutanova et al., 2003).

For our classification, we consider tweets as hate speech in which at least one annotator flagged it as such since we aim at training a high-recall classifier. The resulting class distribution is 33% HATE SPEECH and 67% NO HATE SPEECH. First, we establish baselines by calculating a majority class baseline and conducting a 10-fold Cross Validation (CV). We report macro-F_1 for all conducted experiments. While the majority class baseline results in a macro-F_1 of .4, we obtain a macro-F_1 of .65 for the CV.

To inspect the influence of implicitness, we conduct a train-test split with selected implicit tweets as test instances and the remaining tweets as train instances. We achieve a macro-F_1 of only .1, regardless whether we use the explicit or implicit version of the tweets. Although the performance is higher than the majority class baseline (which is 0, as all tweets are member of the HATE SPEECH class), the drop is dramatic compared to the CV. First, these results indicate that implicitness is a major problem in hate speech detection and thus should be addressed by future research. Second, as results are the same for the more explicit version, the classifier seems to be incapable of recognizing explicit paraphrases of implicit tweets. Although this was expected since we did not add hate speech indicating tokens during paraphrasing, this may be highly problematic as implicitness may alter human perception of hate speech.

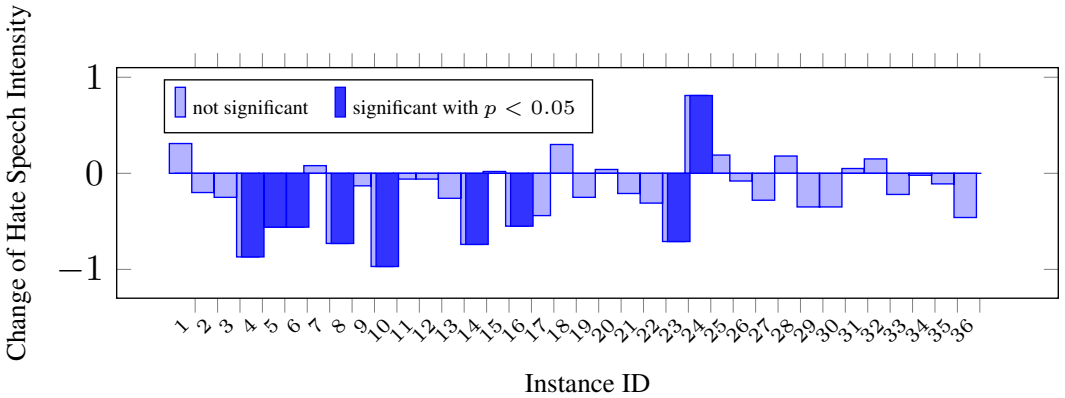


FIGURE 8.3: Change in hate speech intensity between implicit and explicit versions

8.1.3 User Study

After the above mentioned filtering, a set of 36 implicit tweets remained, which were paraphrased into an explicit version. To analyze the difference in their perception, we conducted an online survey and used a between-group design with implicitness as the experimental condition. The randomly assigned participants had to make a binary decision for each tweet on whether it is hate speech and rate its offensiveness on a six-point scale (a Likert scale from 1, denoting NOT OFFENSIVE AT ALL, to 6, denoting VERY OFFENSIVE), in accordance with Ross et al. (2016). The participants were shown the definition of *hate speech* of the European ministerial committee[3].

As understanding the content of the tweets is crucial, we filtered according to native knowledge of German, which resulted in 101 participants. They reported a mean age of 27.7 years, 53.4 39.6% had a university entrance qualification, 58.4% a university degree, and 1% had another education level. More than 90% stated that they identify as Germans which may question the representativeness of our study. Especially, the educational and ethnic background might be factors strongly influencing the perception of hate speech. 55 remained in the implicit condition and 46 in the explicit condition.

8.1.4 Results

First, we inspect how often the tweets are identified as hate speech. On average, we find that 31.6% of the tweets are rated as hate speech in the explicit ($M_{explicit} = 11.3$)[4] and 40.1% in the implicit condition ($M_{implicit} = 14.4$). Interestingly, we observe a high standard deviation ($SD_{explicit} = 11.3$ and $SD_{implicit} = 14.6$) for both conditions. These findings underline how difficult it is for humans to reliably detect hate speech and thus align with the findings of Ross et al. (2016). A χ^2 test shows that the answer to this question is not significantly differently distributed in the two conditions, ($\chi^2_{(22,N=57)} = 4.53, p < .05$).

Regarding intensity, encoded from 1-6, we do not find statistically significant differences between the explicit ($M = 3.9$, $SD = 0.94$) and the implicit ($M = 4.1$, $SD = 0.98$)

[3] http://www.egmr.org/minkom/ch/rec1997-20.pdf

[4] Statistical measures are reported according to the American Psychological Association (1994): M=Mean, SD = standard deviation, p = probability; N = number of participants/annotators

condition according to a t-test ($t(97.4) = 1.1, p > 0.05$). To further analyze this difference, we inspect the difference for each instance, which is visualized in Figure 8.3. All except one of the significantly differing instances are perceived as more hateful in the implicit version. For all cases, we observe that the implicit version is more global and less directed, which could be due to the fact that the vague and global formulation targets larger groups. Instances 6 and 10 contain rhetorical questions, which may be perceived as hidden or more accusative than the affirmative rather factual version. The one case in which the explicit form is more offensive is the only instance containing a threat of violence, which becomes more directed through making it explicit.

We also compute the change in the binary decisions between HATE SPEECH and NO HATE SPEECH on the level of individual instances using χ^2. Three of the eight significantly less offensive explicit instances on the scale are also significantly less often considered being hate speech in the binary decision. Similarly, instance 24, which is perceived significantly more offensive, is more frequently considered as hate speech. Thus, we conclude that there is a relationship between the offensiveness and the hate speech rating and that both are equally affected by implicitness. However, the direction of this relationship, is moderated by the contentual factors (e.g. the presence of a threat) which need further investigation.

8.1.5 Conclusion and Future Work on Implicitness and Explicitness of Hate Speech

In this study, we show that there are individual instances of explicit hate speech which are perceived significantly different compared to their implicit counterparts. However, on average, the polarity of this deviation remains unclear and seems to be moderated by content variables.

In all cases where the implicit version is perceived as more intensely hateful, the tweets were rather insulting than threatening. The perception change might be due to several reasons: the sly, potentially deceiving nature of implicitness might be perceived as more hateful, whereas the same content expressed clearly might be perceived as more honest and thus less hateful.

Furthermore, although implicitness has an influence on the human perception of hate speech, the phenomenon is invisible to automatic classifiers. This poses a severe problem for automatic hate speech detection, as it opens doors for more intense hate speech hiding behind the phenomenon of implicitness.

Since this study is based on 36 tweets, the generalizability of the findings may be limited. Thus, in future work a larger study with more data and more fine-grained distinctions between classes such as INSULTING and THREATENING CONTENT would give more insight in the correlation between implicitness and hate speech perception. Additionally it would be interesting to produce implicit paraphrases of explicitly expressed hate speech and see the effect. Furthermore, more diverse focus groups, such as representatives of diverse religions, origins, and educational backgrounds are required.

8.2 Group Affiliation as Influencing Factor of Hate Speech

In Wojatzki et al. (2018a), we examined he hypothesis of the whether being part of the hate targeted group or personally agreeing with a statement substantially effects hate speech perception on the example of gender affiliation. For instance, it is likely that females perceive the statement *women have a lower IQ than men* as more hateful than men. Furthermore, if anyone should strangely have the supposition that women really have a lower IQ than men, then this person will likely not attribute much hate to this statement.

Problem Description As already stated, hate speech is influenced by non-linguistic factors. In this study, we examine the influence of group affiliation. Furthermore, this study examines the possibility to annotate hate speech, or extreme negative sentiment, as a relation dimension.

Solution Idea To study these hypotheses, a dataset of 400 German statements targeting women was created and judged by both female and male subject according to how hateful these statements are.

The corpus creation process is shown in Figure 8.4. As a first step, we limit ourself to self-contained, explicit statements to which we will refer to as *statements*. Subsequently, we let 40 females and 40 males annotate the hatefullness of these statements. However, indicating the amount of hatefulness on a numerical scale is a hard task which is associated with inter-annotator inconsistencies (Ross et al., 2016; Benikova et al., 2017). Thus, we use the best-worst scaling (BWS) approach by Louviere (1991) for the latter, which results in more reliable scores for other opinion-related tasks (Kiritchenko and Mohammad, 2017). As already discussed in Section 1.2.3, the intuition underlying BWS is that although humans do not share a common absolute scale for a topic, they still tend to agree when picking the worst and best from a tuple of choices. We made all data publicly available.[5]

Outcome In the analysis, we found that extreme cases of hate speech are judged as such regardless of the subject's gender. For less extreme instances, however, there are clear differences e.g. when evaluating female quotas.

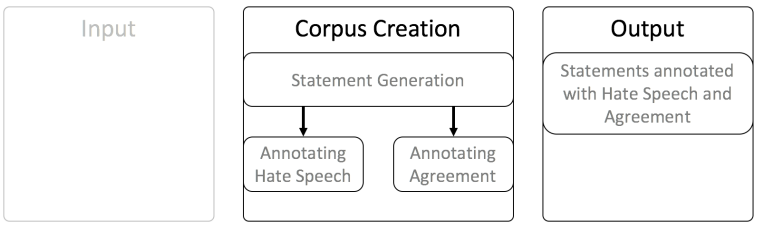


FIGURE 8.4: Corpus creation process of misogynist hate speech

[5] github.com/muchafel/femhate

8.2.1 Related Work

We now shed light on hate speech research and related methods as well as the various facets that make a formalization of hate speech difficult. Furthermore, we motivate our focus on misogyny in this study.

Misogyny as a Form of Hate Speech As noted by Mondal et al. (2017), hate speech is existent on many social media channels, resulting in many efforts to detect or eliminate hate speech and hate speech related phenomenons (Agarwal and Sureka, 2015; Bartlett et al., 2014; Gitari et al., 2015; Ting et al., 2013), often focusing on one specific form of hate speech, e.g. racism (Chaudhry, 2015; Waseem and Hovy, 2016). In this study, we decided to focus on another target group of hate speech—namely WOMEN.

Although there are some works on misogyny as a sub-form of hate speech (Mantilla, 2013; Bartlett et al., 2014; Cole, 2015), there is no dataset that serves as a gold standard for hate speech detection against women. The misogynist variant of hate speech was coined *Gendertrolling* by Mantilla (2013), which according to Mantilla (2013), is even more dangerous and destructive than regular trolling, often containing credible threats of physical and psychic violence. Bartlett et al. (2014) collected a corpus of tweets containing terms such as "rape" and "slut" in order to analyze their usage and origin. While this is a fruitful approach in analyzing misogynist behavior, it is also limited to tweets containing these terms.

Annotating Hate Speech In the construction of a hate speech corpus, there are basically two steps: 1) collection of potential hate speech 2) rating of these instances.

Most current studies rely on lists of offensive words and phrases for collecting potential hate speech (Mantilla, 2013; Njagi et al., 2015; Waseem and Hovy, 2016). However, such collection inevitably brings in biases due to the limited number of query terms. For example, if one collects tweets by searching for the term "bitch", it is not surprising that if there are hate speech annotations in this collection, it is strongly associated to this term.

As shown by Benikova et al. (2017), Waseem et al. (2017), and Ross et al. (2016), annotating hate speech using a numeric or binary scale on such data is a challenging task which is associated with low inter-annotator-agreement. Hypothesized reasons for these inconsistencies include differing thresholds from which a utterance should be classified as hateful, differing valuation of freedom of speech, and implicitness.

8.2.2 Dataset

Following the approach of Wojatzki et al. (2018b), we conduct the data collection in two steps (see Figure 8.4): In contrast to e.g. the corpus created in Section 7.1 or Section 8.1, we did not use external corpora as an input. Hence, in Figure 8.4, the *Input* label is missing. In the first step of the *Corpus Creation*, we asked subjects in a web survey to generate utterances about women to which they agree and disagree, including utterances they would not make in public (as they are highly controversial or provocative). This led to a new set of statements about women, related to women's rights, and their role in the society. In the second step,

	Number
Statements	400
Statement Judgments	32,000
BWS Judgments	4,800

TABLE 8.1: Overview on the collected dataset

we asked 40 female and 40 male subjects in a laboratory setting to indicate how hateful the statements are. For the latter we use a technique known as BWS (Louviere et al., 2015; Kiritchenko and Mohammad, 2016) (for general details on BWS see Section 1.2.3). The subjects received 15€ or subject hour certificates[6] as compensation for the participation. The experimental design was reviewed and approved by the ethics committee of our institution.[7] Table 8.1 gives an overview on the collected dataset.

Statement Generation To generate a large variety of different statements, we designed an online survey in which we directly asked participants to come up with statements that are relevant to our topic. To narrow the topic down for the subjects, we presented them with a list of (sub)-topics. The participants were explicitly instructed that these topics may be used as a source of inspiration for generating the statements but that they are not limited to them. These topics include: gendered language (e.g. *waitresses* vs. *wait staff*), legal differences between men and women (laws for divorce and custody), professional life (e.g. differences in salary, leadership positions, women in the army), social roles (e.g. 'typical women's interests', women and family, 'typical women's jobs'), biological differences, and gender identity. As we wanted to generate statements that differ in how controversial they are, we asked the subjects to provide us at least three statements with which they personally agree and three statements with which they disagree. On a voluntary basis, we also asked the subjects to generate at least three statements with which they personally agree, but which they would not express in public. In order to clarify the task, for each option we provided one example which takes a pro-woman stance and one example which takes the opposite position. In this phase of the data collection, we do not control for any possible bias, as we aim for collecting a diverse stimulus for the subsequent rating phase. However, due to the free generation of the statements, we are less prone to artifacts that occur in a key word based data collection (c.f. Section 8.2.1). Subjects were additionally instructed not to use expressions that indicate *subjectivity* (e.g. "I tend to think"), *co-reference* or *references to other statements*, and *hedged statements* (e.g. indicated by "maybe", "perhaps", or "possibly"). We removed statements which were duplicates, not self-contained and understandable without further context, or formulated in a way that a third person cannot agree or disagree with it.

Subjects We posted the link to our survey in various online forums to ensure a wide range of opinions including communities with a thematic connection to the topic (e.g. the German

[6] as needed by their study program

[7] Computer Science and Applied Cognitive Sciences at the Faculty of Engineering of the University of Duisburg-Essen (`uni-due.de/kognitionspsychologie/ethikkommission_eng`)

subreddit *from women for women* `r/Weibsvolk/`) or that are expected to have a critical attitude on the subject (e.g. the Facebook group *gender mich nicht voll* (*engl. don't gender me*)). Furthermore, we posted the link to topically unrelated communities such as the public Facebook group of the University of Duisburg-Essen to capture less extreme opinions.

We obtained 810 statements from 81 participants, which means that on average each subject generated 10 statements, although only a minimum of six was required. After clean up 627 statements remained of which we randomly subsampled 400 statements with which we will continue to work hereinafter.

Annotating Hate Speech We provide the subjects with a definition of hate speech following the definition made by the Council of Europe (McGonagle, 2013)[8]: "Hate speech is when people are attacked, devalued or when hate or violence is called for against them." We use BWS (cf. Section 1.2.3)—a comparative approach, in which each subject selects the most and least hateful statement from a 4-tuples of statements, which allows to rank the statements with considerably lower effort. We create 600 4-tuples using the script provided by Kiritchenko and Mohammad (2016). Each tuple is answered by four female and four male subjects. Given the comparative annotations we calculate a real value score of hatefulness for each statement using the formula by Orme (2009). The score ranges from -1 (least hate speech) to 1 (most hate speech).

Subjects As we hypothesize that identification with the target group is potentially an important factor in the perception of hate speech, we systematically control for this dimension in our target group. Therefore, we collect an equally sized sample of female and male subjects. In detail, we had 40 female and 40 male subjects in the second phase of our study.[9] In order to minimize the effect of other potentially influential variables such as age or education, we tried to obtain data from a otherwise rather homogeneous group by recruiting students of the University of Duisburg-Essen. The mean age of our subjects was 23.4 years (standard deviation: 4.3). 78% of our subjects were undergraduate students, 21% graduate students and 1% had a different educational level or did not provide this information.

8.2.3 Dataset Analysis

Now, we take a closer look at the generated scores and their distribution.

Hate Speech Scores In Figure 8.5, we show the histogram of the calculated hate speech score. We observe that the distribution of the scores resembles a normal distribution. The top- and bottom-ranked statements are particularly interesting to study the phenomenon of hate speech. The three statements ranked as most hateful are shown in Example 8.5:

[8] accessible at `https://no-hate-speech.de/en/knowledge/`

[9] We also provided an option for those who do not identify with either gender. This was never chosen in our survey.

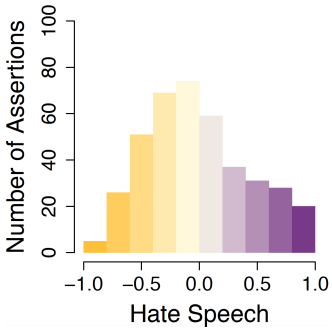


FIGURE 8.5: Distribution of hate speech score obtained using BWS. Scores are discretized into ten bins. Positive scores (gold) indicate a high degree of hatefulness and negative scores (violet) indicate a low degree of hatefulness. (Wojatzki et al., 2018a)

- Es gibt Frauen die Vergewaltigungen genießen. Diese Art von Frauen bilden den nicht-lesbischen Anteil von Gutmenschparteien wie den Grünen. *(hate speech score: 1.0)*
- Frauen gehören in die Küche. *(hate speech score: .98)*
- Frauen können nichts und sind zu dumm. *(hate speech score: .98)*

(English translation)

- There are women who enjoy being raped. This kind of women form the non-lesbian portion of the goody-goody parties such as the Greens. *(hate speech score: 1.0)*
- Women belong in the kitchen. *(hate speech score: .98)*
- Women are incapable and too stupid. *(hate speech score: .98)*

EXAMPLE 8.5: The three statements ranked most hateful overall

For people at large, these are hateful, unfounded, and invalid statements about women fulfilling the requirement of hate speech, by belittling violence against women as well as their abilities. For both the female and the male subgroup, these three statements are amongst the ten statements which were rated most hateful.

The three statements shown in Example 8.6 are the three least hateful ones, and among the ten least hateful statements for both gender subgroups:

- Frauen sind nicht schwach und sollten sich so auch nicht fühlen. *(hate speech score: -.84)*
- Männer und Frauen sind gleich intelligent. *(hate speech score: -.84)*
- Frauen sind stark und sie im Team zu haben macht jedes wirtschaftliche wie soziale Unternehmen stärker. *(hate speech score: -.89)*

(English translation)

- Women aren't weak and shouldn't feel that way. *(hate speech score: -.84)*
- Women and men are equally intelligent. *(hate speech score: -.84)*
- Women are strong and to have them in a team makes every economic as well as social company stronger. *(hate speech score: -.89)*

EXAMPLE 8.6: The three statements ranked least hateful overall

For most readers, these statements do not indicate any negative sentiment towards women. Overall, statements with a high hate speech score are easily perceived as being misogynistic, whereas low rated statements often do not display negative sentiment towards women, which both men and women agree upon.

The three statements with the biggest score difference between the female and male group are all concerned with socio-political topics—army requirements for women, gender pay gap, and structural disadvantages for men. All of these are topics that are usually discussed in political context and could be raised as such in official debates. However, there are also some statements that were rated as more hateful by women, which are not just political, but seem to be more personal: "Feministen sind männerhassende Frauen, die selber kein Glück im Leben gefunden haben." (*engl. Feminists are man-hating women, who found no happiness in life.*) (women hate speech score: .83; men hate speech score: .42), and "Gleichstellung bedeutet tatsächlich bloß "Bevorzugung der Frauen"." (*engl. Gender equality actually just means "favoring women".*) (women hate speech score: .67; men hate speech score: .25). Statements with a big score difference between men and women may contain belittlement of female activists or gender equality, but often do not display explicit hate towards women as a group.

Reliability of Hate-Speech Annotation The reliability of BWS is usually evaluated with Split-Half Reliability (SHR) (Louviere, 1993; Kiritchenko and Mohammad, 2016), as described in Section 1.2.3.

To avoid random effects, we repeat this procedure 100 times and compute the average correlation.[10] We compute the SHR for the whole group, females, and males. For the whole group we obtain a quite strong correlation coefficient of $r = .90$. The correlations of the female ($r = .82$) subjects and male ($r = .81$) subjects are significantly lower, however still substantial. Interestingly, the sexes do not differ in their consistency.

To examine the consistency of the scores of the two genders, we also compute the split-half reliability with one half being the group of males and one half being the group of females.

[10] As Pearson's r is defined in a probabilistic space it cannot be averaged directly. Therefore, we first z-transform the scores, average them and then transform them back into the original range of values.

This comparison results in a correlation coefficient of $r = .93$. This means that male and female subjects largely agree on the ranking of hate speech.

8.2.4 Conclusion & Future Work of Group Affiliation in Hate Speech

In this study, we present the FEMHATE dataset, which contains 400 statements that have been collected via crowdsourcing and that have subsequently been judged by 80 subjects (40 female and 40 male). We collected 4,800 judgments that indicate the strength of contained hate speech. The ratings were shown to be reliable.

Furthermore, we could show that the comparative ratings are relatively similar and robust throughout gender. Although there are cases of great disagreement, they are not cases of highly rated misogyny, neither by men nor women. In this way, we could provide evidence for the hypothesis that on both poles of the range of hate speech scores there is a high agreement between the male and female subjects. Hence, for cases of extreme misogyny, it is irrelevant whether men or women rate it.

8.2.5 Conclusion & Future Work on Hate Speech

In Benikova et al. (2017) and Wojatzki et al. (2018a), we annotated and analyzed hate speech, which we view as an intense form of negative sentiment, on statements. We showed that implicitness and group affiliation play a role in the perception of hate speech, which also influences its annotation.

In all our cases, the implicit version was perceived as more intensely hateful. Although women judged misogynist statements as more hateful than men on average, cases of strong misogyny were judged similarly by both. Furthermore, we found that there is a clear negative correlation between annotating a statement as hate speech and agreeing with it—annotators do not agree with statements which they regard as hate speech.

Findings of both studies show that when working on hate speech, more factors than the phenomenon itself need to be taken into account. The correlation of agreement with a judgment might be a good indicator for hate speech in further work.

Conclusion

In this thesis, we researched semantic representation dimensions for the application of information filtering of pieces of text which we defined as statements. Prior to our work, the disciplines of Computational Linguistics (CL) and Natural Language Processing (NLP) distinguished between dimensions that are operationalized as *single dimensions*—ones that are assigned to one single statement—and *inherent relation dimensions*—ones that are operationalized as relations between at least two statements. The main finding of this thesis is that all dimensions, including originally single ones, can be reliably annotated as relations.

Application Scenario Furthermore, throughout the thesis, we demonstrated how all the dimensions researched in this thesis can be used for an information filtering task.

As an example, we chose one kind of extractive summarization, namely user-specific review filtering:

- The filtering of *sentiment* for an aspect corresponding to the user's interest reflects the identification of relevant information.
- The bundling of redundant information to *paraphrase* clusters and the choice of the best statement out of these clusters using *specificity* corresponds to the choice of the most relevant information.

This example application is discussed in more detail in the Further Work section of this thesis.

Single Dimensions Operationalized as Relation Dimensions Overall, we find that although originally single dimensions can be reliably operationalized as relation dimensions, there are some restrictions to the operationalization process:

- The relation operationalization itself seems to have an influence.
- An additional focus or aspect also seems to improve the annotation process.

These restrictions are best seen on the example of specificity, a lesser researched dimension which we are the first to see as a relation. In the study described in Chapter 5, we use pairwise comparison between two statements to annotate specificity with rather moderate success. In a further study described in Chapter 6, we used best-worst scaling (BWS) in a crowdsourcing setting to operationalize specificity as a relation dimension. Additionally to the usual BWS setting, we provided a focus for the specificity. In this study (Chapter 6), we were able to reliably annotate specificity as a relation dimension. Similar to the second study on specificity, we successfully used BWS with an implicitly given aspect to annotate sentiment as a relation dimension in the study described in Section 8.2. Hence, we conclude that single dimensions can be reliably operationalized as relation dimensions with further

restrictions which is the cases of specificity and sentiment are the method of BWS and an additional focus, similar to the task of aspect-based sentiment analysis (ABSA).

Statement granularity The granularity of the statements on which the dimensions are annotated also has a significant influence on the operationalization of semantic representation dimensions. We conclude that propositions, although difficult to compute, are the granularity of choice for further research. In our study in Chapter 6, we found that sentences are too coarse. In Section 3.2, we conducted a survey showing the differences of various representation types. In this survey, we argued that predicate-argument structures fit our purpose best. We argue that although forfeiting some expressiveness, the most basic predicate-argument structure, namely *propositions* would be computable and human understandable. In Section 4.1, we perform a practical study of proposition extraction and find that complex sentences—those of greater interest—pose a big problem to automatic proposition extraction systems. Improving these systems seems like a difficult task, since their problems lie within the underlying dependency parsing, which has been worked on for decades. As a consequence, we could do one of the following:

- Accept the issues of automatic proposition extraction and focus on their upsides—we would probably still find more interesting dimensions than on the sentence level.
- Work on simple sentences only, as this would be more robust and accept the loss of interesting cases.

Decomposing Statements into different Granularities Further researching the different granularities and what is the best granularity for representing semantic dimensions, we performed two studies on the decomposition of dimensions. In Section 4.2, we decomposed the paraphrase annotation of statement pairs. It is the first work to analyze the compositionality of one of the discussed dimensions. In our study, about a third of complex sentences which are not paraphrases contain propositions which are paraphrases. This means that working on the sentence level one misses propositions of interest, which might be a big loss in applications using the relation dimensions. In Section 5.2, we presented a unifying typology that can be used for decomposing all the dimensions. It is the first successful step towards building a framework for studying and processing multiple meaning relations. Furthermore, we performed a case study proving the validity of our typology. Both studies (presented in Section 4.2 and Section 5.2) show that the decomposition of dimensions provides us with a deeper understanding of these. Moreover, in Section 5.2 we demonstrate that the linguistic and reasoning phenomena underlying the semantic dimensions are very similar and can be captured by a shared typology. A single framework for meaning relations can facilitate the analysis and comparison of the different relations and improve the transfer of knowledge between them.

Links between Relation Dimensions Apart from the operationalization of individual relation dimensions, we also researched if and how they are linked to each other. In Chapter 5 we describe a study which analyses the links between the relation dimensions researched in this

thesis. This is the first empirical study to do so. In this study, we present a new and successful methodology to create a manually annotated corpus containing all dimensions of interest. We provided empirical evidence that supports or rejects previously hypothesized connections in practical settings. The most important findings can be summarized as follows:

- There is a strong correlation between paraphrasing and entailment.
- Paraphrases and bi-directional entailment are not equivalent in practical settings.
  Most paraphrases include at least uni-directional entailment.
  There exist bi-directional entailments that are not paraphrases.
- The specificity dimension does not correlate strongly with the other relations and requires further study.
- Contradictions (in our dataset) are perceived as dissimilar.

In a small experiment using an Support Vector Machine (SVM) equipped with the dimensions excluding the one to be predicted, we were able to show that the relation dimensions can be used to automatically predict each other (for entailment, paraphrasing, and semantic similarity).

Overall, we have shown links between and operationalizations of relation dimensions. We found that using some restrictions, originally single dimensions can be operationalized as relation dimensions. Furthermore, we found that propositions are the best granularity level when working on semantic representation dimensions. Moreover, we found empirical evidence for links between the different relation dimensions.

In this thesis, we did not only provide new insights into semantic representation dimensions and develop operationalization methodologies enabling further research on the described relations, but we also created several new and available corpora, which facilitates new research including empirical analysis but also automatic methods. In the next chapter, we will discuss how our findings as well as our corpora can be used in further work.

Further Work

Herein, various ideas for future work using the findings of this thesis are discussed. First, we discuss how findings on all or several dimensions can be of benefit to the Natural Language Processing (NLP) and Computational Linguistics (CL) community. Then, we present how findings on individual dimensions can be used in future work.

Specificity Having performed several studies on specificity, we believe that another study accumulating this findings would be helpful for further automation. As already described in Chapter 6, specificity is most reliably annotatable in a comparative setting, given an aspect. Therefore, it is quite similar to aspect-based sentiment analysis (ABSA) at least formally. Hence, and also given the task of user-specific review filtering, an ABSA corpus with already annotated aspects could be used (ensuring that the aspect is present in the compared statement pair). In this way, we could also study the relation between sentiment and specificity. This might be helpful, as they, as mentioned above, seem at least formally similar. Furthermore, it would be of interest to study whether and which level of specificity is helpful to a potential user, which could be done on the ABSA corpus annotated with specificity and sentiment using best-worst scaling (BWS) with a given aspect.

Hate Speech The interaction between group affiliation and implicitness/explicitness would be interesting, as both phenomena seem to influence the judgment of hate speech in some way. Furthermore, the finding that agreement and disagreement with judgments correlates with hate speech judgments could be used in automatic hate speech detection in real world social media data using up-votes or down-votes of forum posts.

Working with Propositions Throughout this thesis we found that propositions are the type of representations needed for working with the described semantic dimensions. Although we found that all existing proposition extraction systems have issues with sentence complexity, we would choose the best-performing system—namely OpenIE—for future work, as it performed best on both simple and complex sentences.

Working upon Propositions As described in Section 3.2, in future work a representation that is more expressive than simple propositions could be achieved. Using propositions as a pre-processing step to get the predicates and their arguments. Then, sense clustering models could be performed and the clusters could be applied on the argument text. In future work, tasks such as:

- topic-model based frame labeling on the computed clusters

- pattern-based antonym detection in the clusters for enabling the operation of contradiction and improve the task of equivalence

could be tackled.

Using all Dimensions The following section provides a detailed description of how to use all dimensions discussed in this thesis for user-specific review filtering. In future work, this plan could be implemented.

Furthermore, the findings on the links between the different dimensions as discussed in Section 5.1 could be used in transfer learning, e.g. by accumulating corpora from various dimensions to train for another dimension.

Application: User-Specific Reviews

In the introduction and throughout this thesis, we already discussed this application. However, in this section we will outline it in more detail. We already stated that filtering for user-specific reviews is a kind of summarization. In Figure 6, we show which steps of the summarization workflow described in our work in Benikova et al. (2016) correspond to which steps in the filtering process. Furthermore, the figure shows how the dimensions described in this thesis can be used to perform these steps. In the following, we will outline how the discussed dimensions could be used in the filtering process and how this corresponds to the summarization workflow in more detail.

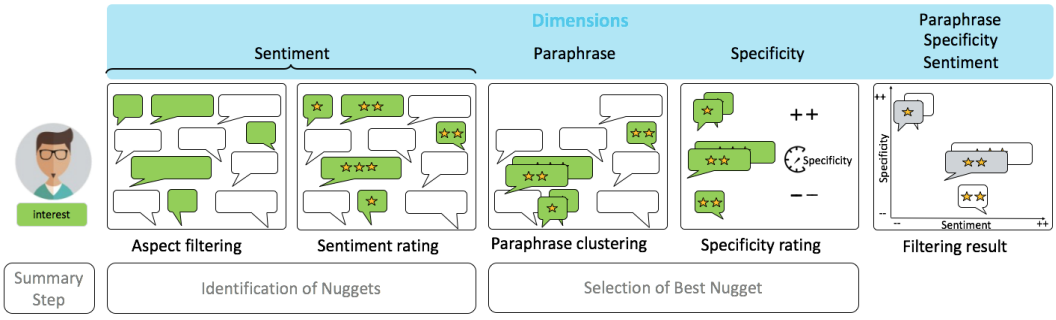


FIGURE 6: Illustration of exemplary user-specific filtering workflow using the dimensions in this thesis and its parallels to the summarization creation in Benikova et al. (2016)

Sentiment Dimension in the Review Filtering Process

In the first step, we reduce the amount of reviews by filtering for the aspect that is of interest to the user. For the user the sentiment of the filtered statements is also of interest. Hence, the dimension of aspect-based sentiment is described as *sentiment* in Figure 6, but also throughout this thesis, can be used for this filtering step. This step corresponds to the step of *Identification of Nuggets* in our summarization workflow described in Benikova et al. (2016). This step is simplified in the review filtering process, as the choice of nuggets is limited to statements containing the aspect of interest. Figure 7 shows a concrete example for this step.

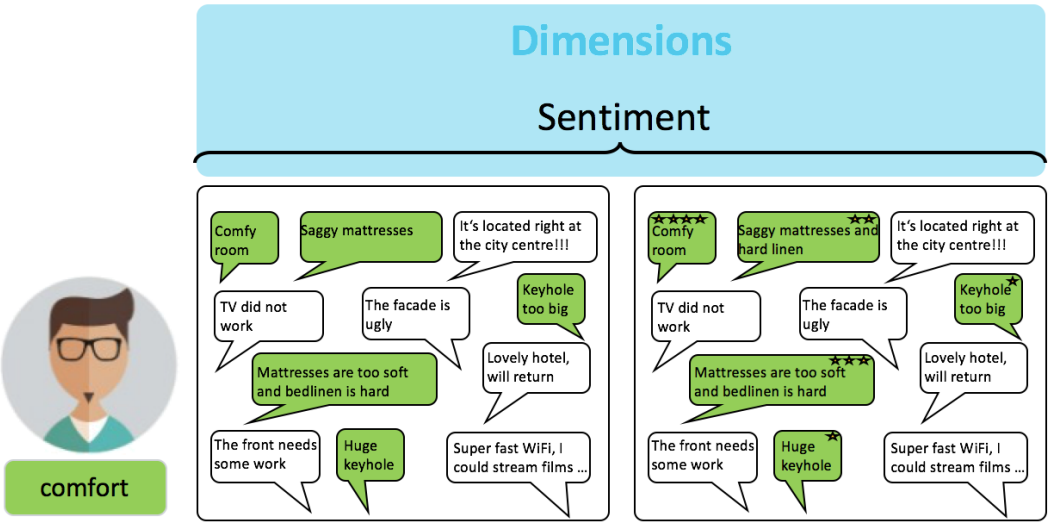


FIGURE 7: Illustration of sentiment in exemplary user-specific filtering workflow

Paraphrase and Entailment Dimension in the Review Filtering Process

In the *Paraphrase Clustering* step, the aspect-filtered reviews from the previous step are bundled into paraphrase clusters i.e. sets of statements with similar content. For reasons of illustration, in Figure 6 we show only the paraphrase dimension. However, the entailment dimension can be used for this step as well in order to create entailment clusters. In both cases, clusters containing redundant information would be created, which is the goal of this step. The clustering of redundant information is the first of two sub-steps for the *Selection of Best Nugget* in our summarization workflow. In the second step, the best nugget out of this cluster is chosen. Figure 8 shows a concrete example for this step, using the filtered content of the sentiment dimension.

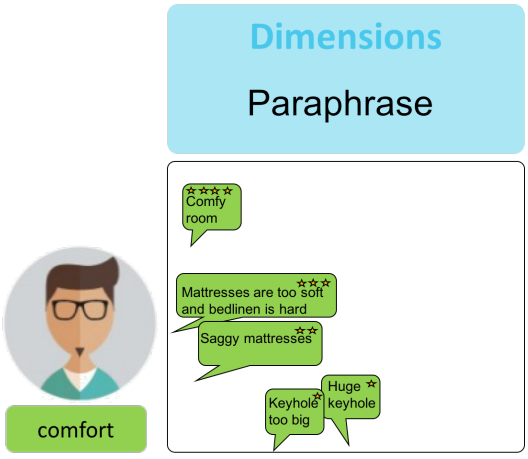


FIGURE 8: Illustration of paraphrase in exemplary user-specific filtering workflow

Specificity Dimension in the Review Filtering Process

The actual *Selection of Best Nugget* out of the *Paraphrase Cluster* (or *Entailment Cluster*) could be performed using the specificity dimension. Figure 9 shows a concrete example for

this step, using the paraphrase clusters created in the previous step. The level of specificity needs to be adjusted to the user's needs, as both too broad and too specific statements should be excluded. In Figure 9, the statement "Comfy room" is too broad, while "Keyhole too big" is too specific.

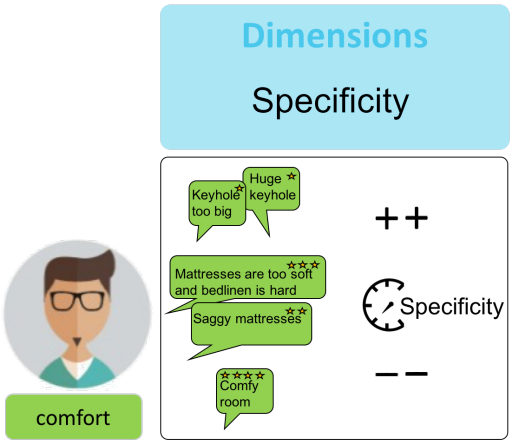


FIGURE 9: Illustration of specificity in exemplary user-specific filtering workflow

Filtering Result

In contrast to the summarization that we describe in Benikova et al. (2016), we are not interested in *coherent* extracts from reviews. Hence, the final step of *Formulation of Summaries* can be neglected in the review filtering process. A possible result that is displayed to the user is shown in the last step in Figure 6. The user is presented with some highlighted statements within reviews. Furthermore, the user can see the specificity and the sentiment of the reviews and choose based on these dimensions. The other dimensions may be displayed to the user as well. For instance, it might be of interest, how big the paraphrase or entailment cluster is, as this is a potential indicator of the validity of the presented statement. Contradiction relation might also be of interest, as they show conflicting information to the user.

Summary

Overall, the herein developed implementations of dimensions can be used for further research. For instance, using the findings on specificity and sentiment, the possibilities of other *single dimensions* being modeled as *relation dimensions* could be explored. Furthermore, all corpora can be used for automatic annotations of the annotated dimensions, especially for evaluation. These automatic annotations do not necessarily need to have the same implementation as described in this thesis, as they can be easily transformed to other formats, as has been discussed in the respective sections. In this way, the implementations, corpora, and findings of this thesis provide a vast variety of possibilities for future work, especially for the fields of NLP and CL.

Appendix A

Appendix

A.1 Guidelines Produced for Studies in this Thesis

A.1.1 Guidelines for Proposition Studies

A.1.1.1 Guidelines for Producing Reduced Sentences on AMT

Create phrases from sentences

We want to create a dataset of separate information pieces from one sentence.

- Create as many different and sensible minimal sentences from the content of the given sentence. You will be paid a bonus of 0.01 $ for each different correct phrase.
- A minimal sentence is a sentence that does not contain information that could be omitted. However, the sense of the information in the original sentence should remain the same, meaning that no necessary information should be omitted.
- Write each phrase in a new line.
- Use the meaningful words (mostly nouns, verbs, adjectives) from the sentence and add new words only if they are needed to understand the created phrase. So copy+paste of information pieces is encouraged.
- Do not reuse information from previous phrases.
- If you are not able to create at least two minimal sentences from the given sentence, copy the original sentence. (This is only a fall back option, please try creating new phrases)

Example

The sushi was fresh and authentic, but we had an unfriendly waiter, who refused to bring me the wine card.

- *The sushi was fresh.*
- *The sushi was authentic.*

The next two minimal sentences reveal information on the waiter:

- *We had an unfriendly waiter.*
- *The waiter refused to bring me the the wine card.*

However, putting the information in one sentence would be wrong, as they are two different pieces of information and one of them can be omitted. Thus, the following would be an incorrect minimal sentence:

- ~~*We had an unfriendly waiter, who refused to bring me the wine card.*~~

Although the following minimal sentence uses new words, these are not meaningful and the phrase is equivalent to "We had an unfriendly waiter" and is thus a correct minimal sentence:

- *The waiter was unfriendly.*

The following phrases would be incorrect, as they use meanigful words that were not used in the original sentence (namely "staff" and "food"):

- ~~*We had unfriendly staff.*~~
- ~~*The food was fresh.*~~

The following minimal sentences would be incorrect, as they reuse information from the previous correct minimal sentence above:

- ~~*The sushi was fresh and authentic.*~~
- ~~*The sushi was authentic, but we had an unfriendly waiter.*~~

${inst1}

A.1.1.2 Guidelines for Proposition Creation

Guidelines for Proposition Creation

Introduction

Propositions are semantic meaning representations of texts that are used in information extraction.
Here, several propositions may be extracted from a sentence.
The prototypic proposition consists of three elements, namely **subject**, **predicate**, and **object**:

Original sentence	Proposition 1
I ate an apple	I| ate| an apple

Example 1: Simple proposition extraction

A proposition needs to be as true on its own as it is in a sentence. All elements that are part of an information piece should be included.

Annotation

Tool instructions
This annotation is performed in Excel. Please write propositions in the column of the same row as the original sentence. Each proposition is to be written in a box of its own:

Original sentence 1	Proposition 1.1	Proposition 1.2
Original sentence 2	Proposition 2.1	Proposition 2.2

Proposition structure
The first element is always the subject and the second the predicate. Further element order is not restricted.
Here, propositions cannot only contain several objects, but also modifiers, e.g. temporal (e.g. tomorrow, now, in 1990 …) or local modifiers (e.g. here, in Paris, outside, …).
Subjects and objects that are semantically differing objects that are as true separately as they are together, are reformulated to several propositions:

I ate an apple and a pear	I| ate| an apple	I| ate| a pear

However, if the produced propositions would not be true, they cannot be made:

I ate an apple or a pear	~~I| ate| an apple~~	~~I| ate| a pear~~

Adjectives and adverb belong to the element they grammatically refer to:

I only ate an apple	I| only ate| an apple
I ate an apple only	I| ate| an apple| only
I ate only an apple	I| ate| only an apple

If a sentence is not split in several proposition, the proposition must contain each word of the sentence.

Features of propositions

Produce only propositions that are **asserted**, not implied or entailed in the original sentence:

I succeeded in eating the apple	I| succeeded in eating| the apple	~~I| ate| the apple~~

*The propositions should be as **minima**l as possible, but also be **complete**, meaning not omit any needed information:*

Subject

A subject mostly answers the question of Who? Or if it an inanimate object also What? In English, it mostly has the first position in a sentence. As an example see Example 1.
However, here, "there" can also act as a subject

There are apples	There| are| apples

If the **subjec**t is missing, it needs to be marked

Ate an apple	[subject missing]| ate| an apple

If a subject can be anaphorically inferred without doubt, insert the correct subject.

If you are hungry, eat an apple	[you]| eat| an apple| if you are hungry	you| are| hungry

Predicate

A predicate has to contain a verb. In English, it mostly has the second position in the sentence. As an example see Example 1.
If the **predicate** is not explicit, insert a fitting implicit verb (mostly an auxiliary verb e.g. to be, to have)

Sweet Apple!	apple| [is]| sweet
Sweet Pink Lady	Pink Lady| [is]| sweet
Green apple tree	apple tree| [is]| green

In the second case "Pink Lady" is a proper noun and is thus treated as an inseparable entity. In the third case, "apple tree" is a compound noun and is thus also treated as an inseparable entity.

If there is neither a clear subject nor a verb, try to use the phrase as a subject or object/modifier, depending on what makes more sense:

Apple!	Apple| is
Sweet	[subject missing]| is| sweet
tomorrow	[subject missing]| is| tomorrow

Sometimes the verb does not have the second position in the sentence. However, in the proposition, it must be on the second position:

After that| I ate the apple.	I| ate| the apple| after that

If there are **several verbs, that semantically describe the same action**, they are treated as one predicate. Mostly these are verbs that are written in one sequence, excluding to-infinitives and gerunds:

I would have eaten an apple	I| would have eaten| an apple

But:

I like eating apples	I| like| eating apples
I like to eat apples	I| like| to eat apples

I like eating apples	~~I| like eating| apples~~
I like to eat apples	~~I| like to eat| apples~~

Auxiliary and modal verbs are part of the predicate

I have eaten an apple	I| have eaten| an apple
I must eat an apple	I| must eat| an apple

This also applies to colloquial or rhetoric use of auxiliary verbs:

I did do eat this apple	I| did do eat| this apple

Constructions with infinitives are regarded as elements:

I tried to eat an apple	I| tried| to eat an apple
I have to eat an apple	I| have| to eat an apple

If there are **several verbs, that describe the different actions that are not both part of a main clause**, they are treated as two separate predicates.

Guess what| I ate an apple	[subject missing]| guess| what	I| ate| an apple
I ate an apple and went home	I| ate| an apple	I| went| home

Objects and Modifiers

The **object** is optional, meaning there can be propositions without an object:

I ate	I| ate

If there are **further objects**, or **modifiers**, they are also part of the proposition and are attached at the end:

I ate an apple with him yesterday	I| ate| an apple| with him| yesterday

There are prepositions that function as modifiers e.g. *there, here, in, out, after*, … . These are to be treated as independent objects:

I ate an apple there	I| ate| an apple| there
I was there for lunch	I| was| there| for lunch

There are also modifiers that contain further propositions

I ate an apple where I parked my car	I| ate| an apple| where I parked my car	I| parked| my car

Other phenomena
If there is a **negation**, it is attached as part of the element that it negates:

I did not eat an apple	I| did not eat| an apple
I ate no apple	I| ate| no apple

In case of unclearness of what element exactly is negated, it should be attached to the verb if it is "not" and to another element if it is "no".

Adverbs of frequency
Adverbs of frequency, such as always, finally, never, usually, etc., are independent elements:

I finally ate the apple	I| ate| the apple| finally
I almost never eat apples	I| eat| apples| almost never

Subordinate clauses
Difference between dependent and independent clause:
Dependent:

If you're sick, eat an apple	you| eat| apple| if you're sick	you| 're| sick

The subordinate clause is used as an element of the proposition in this case.
Also, in the case of dependent relative pronouns (no comma mostly), include the information in the proposition:

I ate an apple which was red	I| ate| an apple which was red	

~~Independent~~
~~In the case of independent relative pronouns, the information in the pronoun should left out in the first proposition, as shown in the first example:~~

~~I ate an apple| which was red~~	~~I| ate| an apple~~	~~the apple| was| red~~	
~~After I came home| I ate an apple~~	~~I| came| home~~	~~I| ate| an apple~~	

Independent
In the case of independent relative pronouns, treat it in the same way:

I ate an apple, which was red	I| ate| an apple, which was red	An apple| was| red	

After I came home, I ate an apple	I| ate| an apple| after I came home	I| came| home	

Orthographic mistakes / Grammar
Make proposition like you understand the text (with orthographic mistake)
Please do not correct orthographic mistakes!

The apples where sweet	The apples| where| sweet
	~~The apples| were| sweet~~

Connections with conjunctions
Let "with", "between" semantically together, if it belongs to one of the objects:

I ate the apple with honey	I| ate| the apple with honey

If it belongs to an independent object, it should be treated as such:

I ate the apple with you	I| ate| the apple| with you

Direct or indirect speech
Everything that is *said, told, answered, …* is treated as an object. If that object contains another sentence, make a proposition out of it.

I said that the apple was sweet	I| said| that the apple was sweet	the apple| was| sweet
I asked whether the apple was sweet	I| asked| whether the apple was sweet	the apple| was| sweet
I replied that the apple was sweet	I| replied| that the apple was sweet	the apple| was| sweet

Named Entities
Do not split named entities:

Snow White ate the apple	Snow White| ate| the apple	~~Snow| ate| the apple; White| ate| the apple~~

Prepositions
In case a preposition cannot semantically be attached to an object, it is part of the predicate

I asked for an apple.	I| asked| for an apple.
I found out about the apple	I| found out| about the apple

I ate a red apple and sneezed	I| ate| a red apple	I| sneezed	~~I| ate| a red apple| and sneezed~~	~~I| ate| an apple~~

The *sneezing* is a separate proposition. *Red* cannot be left out

Examples for difficult cases

Adjectives

I was hungry	I| was| hungry
It can be busy	It| can be| busy

Temporal adverbs

I am always hungry in the morning	I| am| always| hungry| in the morning	
The apple	It| can be| busy	

Indirect speech

I asked how tasty the apple from my parents' garden was with respect to its sweetness	I| asked| how tasty the apple from my parents' garden was with respect to its sweetness	the apple from my parents' garden| was| tasty with respect to its sweetness

Ungrammatical

How the apple tasted	how the apple| tasted	

Passive

The apple was eaten by me	the apple| was eaten| by me

Questions

If the question contains another verb\| it is treated as separate predicate

What did you do with the apple?	What| did| you do with the apple?	you| do| with the apple
How is that my fault?	how| is| that my fault?	

Imperative

Let's eat the apple!	[subject missing]| let eat| 's| the apple	
Do as I say!	[subject missing]| do| as I say	I| say

Participle used as an adjective

I am disappointed by the apples taste	I| am| disappointed by the apples taste
The apple shop feels crowded	the apple shop| feels| crowded
It can be disappointing	it| can be| disappointing

Dependence of elements / Subordinate clauses
If an element is dependent on one other element, be it a modifier or a subordinate clause, it must be placed with the element it is dependent on and not as an independent element

It was interesting that I ate an apple	It| was| interesting that I ate an apple	I| ate| an apple
When she finally came over, she said yes.	she| said| yes| when she finally came over	she| finally came| over

Independence of elements

I ate an apple here	I| ate| an apple| here	
I ate an apple while you talked about the pear.	I| ate| an apple| while you talked about the pear	you| talked| about the pear
The recommendations given by the waiter were amazing	The recommendations| given| by the waiter| were amazing	
He made me hungry	He| made| me hungry[1]	

Conjunctions

I ate neither apple nor pear.	I| ate| neither apple nor pear	
I ate no apple or pear	I| ate| no apple	I| ate| no pear

Make smb. do smth.

She made me eat the apple	She| made| me eat the apple
He made you feel hungry	He| made| you feel hungry

[1] "hungry" belongs to "me", it is not independent and does not belong to "make". If the sentence would be "He made me coffee", "coffee" would be independent of "me" and it would belong to coffee.

A.1.1.3 Guidelines for Paraphrase Annotation on Three Granularity Levels

Event Paraphrase Annotation Guidelines

State: 0.4; 25.03.17

Author: Darina Benikova, darina.benikova@uni-due.de

Table of Contents

Event Paraphrase Annotation Guidelines

Introduction

Events in Social Media are textual descriptions of actions and facts in the real world. The same real world events can be expressed in different ways. For instance, “Trump lives in Washington.” and “Donald Trump resides in the White House." are paraphrases of the same event.
Methodologies for detecting event paraphrases would be helpful in many tasks of natural language processing, such as summarization, information extraction, plagiarism detection, machine translation, and question answering. **Paraphrases describe differently worded pieces of text with the same content or in other words they can be described as bidirectional entailment**.
We believe that by detecting paraphrases on a more granular level than on the sentence level as it is currently performed, we detect more precise and accurate paraphrases.
However, previous works have mostly focused on paraphrases on the sentence level.
To measure and quantify the difference between paraphrases on different levels, namely the sentence and the event level, the annotation on both levels will be conducted as herein described.
You will be provided with tuples from existing sentence paraphrase corpora.
The three levels that you will be annotating on are depicted in Figure 1.
The two lower levels will be explained in more detail further on.
Shortly, it shall be stated that we consider events as so-called verb-argument structures: one main verb and all its arguments (subject, objects, …).

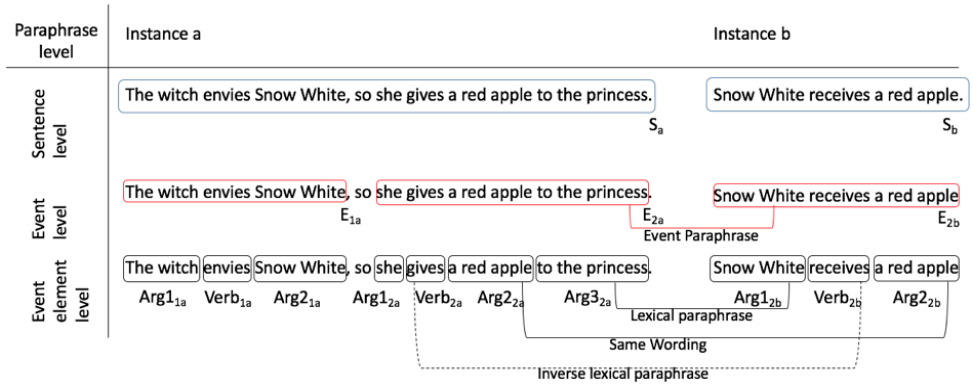


Figure 1 Example of paraphrases on sentence and event level

Event Paraphrase Annotation Guidelines

Definitions

Paraphrase

Here, a paraphrase will be defined as a piece of text with clear boundaries (Instance a) that has the same meaning as another piece of text (Instance b) with clear boundaries.
Also, we define paraphrase a bi-directional entailment, meaning that one text (Instance a) must result in the other (Instance 2) and the other way around. According to this definition, in Figure 1, the two instances

a) She gives a red apple to the princess
b) Snow White receives an apple

have the same meaning and also, a) results in b) and b) results in a).
However, the two instances

a) She gives a red apple to the princess
b) Snow White holds an apple

are not paraphrases. Although a) results in b), b) does not result in a).

Event

Here, an event consists of the main verb and its arguments. A main verb is sometimes also called *full verb* and is the opposite of an auxiliary verb. Arguments are those pieces of text that are grammatically dependent on the main verb, e.g. the subject and the objects.
In Figure 1, the first sentence, “The witch envies Snow White, so she gives a red apple to the princess.“, has two events, as it has two main verbs. The second event consists of the main verb, “gives” and the arguments “she”, “a red apple”, and “to the princess”.

Event Elements

Event Elements are the main verb and the arguments.
In case the main verb is a phrasal verb (consisting of more than one token), the whole span of the main verb is regarded as such, e.g. in the event

She asked him out

the verb would consist of both “asked” and “out”.

The arguments consist of the whole span of the argument, meaning of the head of the argument and everything that is dependent on it, e.g. in Figure 1, the argument “a red apple” consists of the head of the argument, namely “apple” and all its dependents, in this case the article “a” and the adjective “red”. Articles may consist of full clauses, e.g. “a red apple, that was given to her by her evil stepmother”.

Event Paraphrase Annotation Guidelines

Annotation Process

This section gives a short overview of the three sequential annotation steps. Each step will be discussed in a dedicated subsection.

First, you will be provided with two sentences. If you see any annotations (but your own), you are requested to turn them off.
In the first step of the annotation, you are requested to annotate whether the two presented sentences are paraphrases.

In the second step, you are requested to annotate the events, so switch the event structure annotations on and decide whether the span it is encompassing is a paraphrase of an event in the second sentence.
In the third step, you decide whether the individual event elements are paraphrases of event elements in the second sentence.
In this annotation study, we will be using the web-based annotation tool WebAnno. In case of questions concerning the tool, first consult
https://webanno.github.io/webanno/releases/2.3.0/docs/user-guide.html .

In each of the levels you will have to make an alignment label decision between three different kinds of paraphrase link:

- Same wording, meaning that the lemmas of the word(s) in the span have the same lemma(s)[1]
- Sure Paraphrasing, meaning that the wording differs, but you are sure that this is a paraphrase
- Unsure paraphrase, meaning that you are not entirely sure whether the words are paraphrases[2]

The last does not have to be explicitly annotated. The second one should only be chosen in cases in which you are unable to decide. As it is easiest to show this on the event element level, the examples for each of the paraphrase link kinds are presented in the corresponding section.

Paraphrases on sentence level

In the first step, you have to judge whether the two presented sentences are paraphrases. In order to be as unbiased as possible, please switch off the event annotation[3] and just make your decision based on the definition provided in the previous section.

[1] A lemma is the canonical form, dictionary form, or citation form of a set of words, e.g. the lemma of *giving*, *gave*, *gives*, and *given* is *give*.
[2] This label should only be used in case you are really unsure concerning the paraphrase, so please use it only in this occasion.
[3] The description of how to switch layers on and off can be found here:
https://webanno.github.io/webanno/releases/2.3.0/docs/user-guide.html#_layers

Event Paraphrase Annotation Guidelines

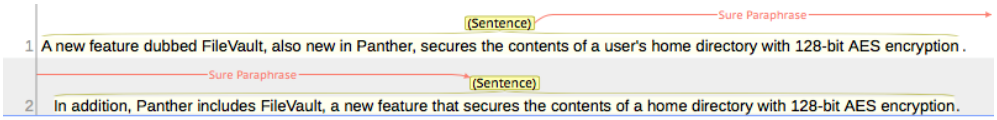


Figure 2Example for sentence paraphrase annotation

Annotation of event elements

The definition of event elements is described in the previous section. You will be presented with automatic pre-processing, the full verbs as well as their argument spans will be annotated in the first sentence.

First check the event elements marked in the first sentence.

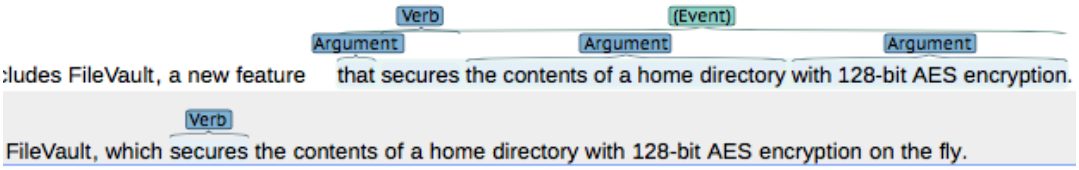
Then perform the following steps:

1) Find the event in the second sentence that corresponds best to the event in the first.

If it is difficult to find a matching event in the second sentence, first try to find an event that has similar arguments or a high lexical overlap. If this is also impossible, choose the first event in the sentence.

2) Annotate the verb of the event

When annotating the verb, please remember that we only consider main verbs, meaning that auxiliary and modal verbs, participles, and gerunds are not considered as verbs.



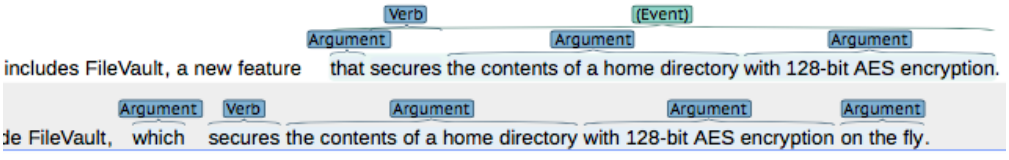
3) Find the heads of the arguments of the verb

The arguments of the verb are the subject and the objects on the one hand, but also temporal or local information, e.g. expressed through short phrases such as “last year” or “in Paris”, but also through clauses, such as “after seeing what had happened” or “where the statue was standing”.

Connectives such as e.g. “furthermore”, “in addition”, “but”, or “whether” are not considered as arguments.

4) Find all the elements that are dependent on the heads

Elements depending on the head might be adjectives, or clauses describing the head. Mark all the elements that semantically depend on the head of the argument as part of the argument.



In the example in Figure 5, the second argument is “by his wife’s conduct during their marriage”, as all the phrases refer to the argument head, “wife”. It could be assumed that “during their marriage” could be an argument of its own, but I does not refer to “saddened”.

Event Paraphrase Annotation Guidelines

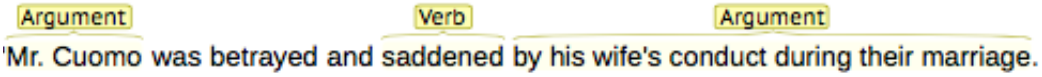


Figure 3 Example for argument span annotation

The example also shows that the span of the argument should also encompass the preposition referring to the verb.

If the main verb is used in connection with modals or negations, this should be annotated as shown in the following:

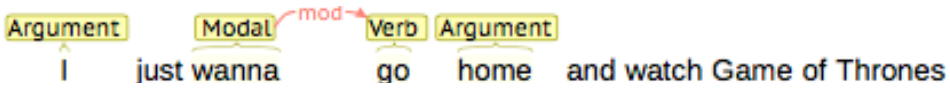


Figure 4 Example for modal annotation

The link between the modal and the verb is only for your understanding, you do not need to annotate this in your own annotation.
If the verb is linked to a modal, it should be annotated considering this information, e.g. in Figure 5, the verbs are not paraphrases due to the use of "could" in the first event.

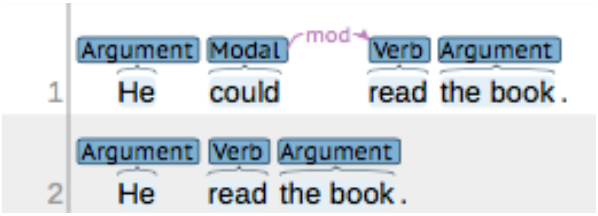


Figure 5 Example of events with modals

Phasal verbs cannot be annotated as one span if there would be also other words involved. In this case, they should be annotated followingly:

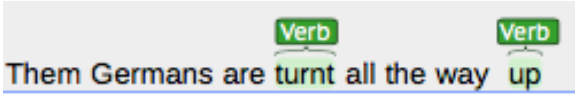


Figure 6 Annotation of phrasal verb

Paraphrases on the event level

Having annotated the event elements of the second sentence, mark the event span, as shown in Figure 7.

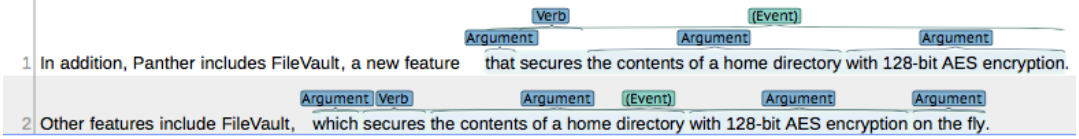


Figure 7 Example of event span annotation

Please consider that the event consists of the marked elements, so in case there is additional information in the span that is not marked through elements, it is not part of the event. For

Event Paraphrase Annotation Guidelines

instance, in Figure 8 the event is “That fire charred 469,000 acres”, not including the information that “[it] devastated timber on the reservation”.

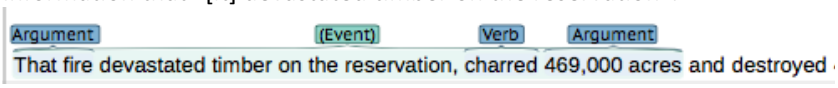


Figure 8 Span with information that is not part of the event

After having annotated the whole span of the event and align it with an event in the second sentence, if applicable.

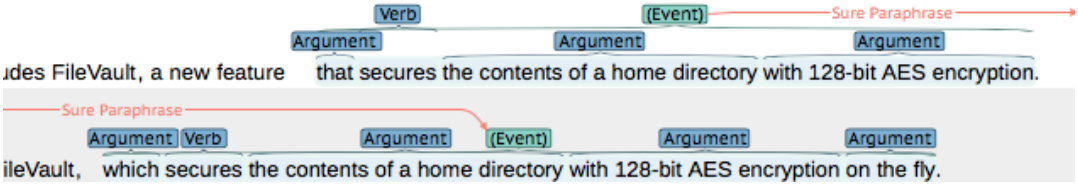


Figure 9 Example of event paraphrase alignment

Annotation of Element Paraphrases

Given the event element annotation in the two sentences, you should link those that are paraphrases of each other according to the classes mentioned previously. Figure 10 shows a full exemplary annotation of the event elements.

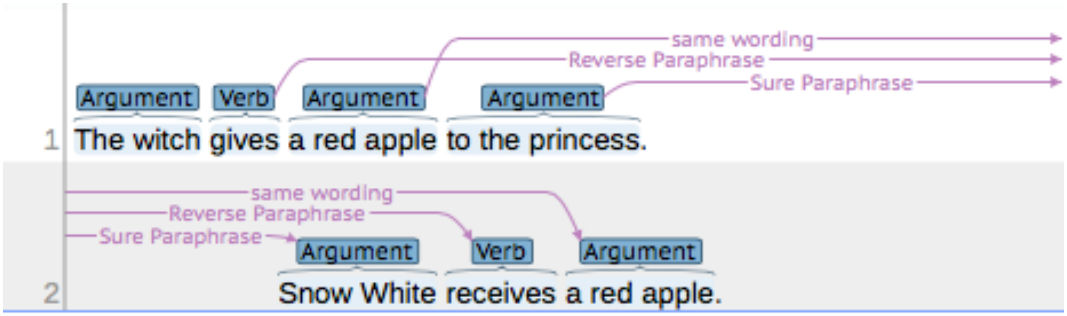


Figure 10 Example with all paraphrase link classes on the event element level

In the following, examples for the paraphrase classes “Same Wording” and “Sure Paraphrase” are shown:

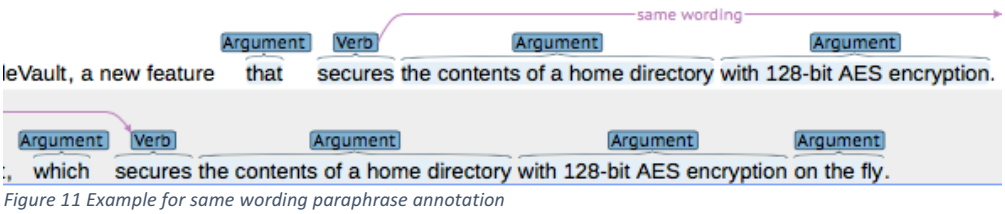


Figure 11 Example for same wording paraphrase annotation

Event Paraphrase Annotation Guidelines

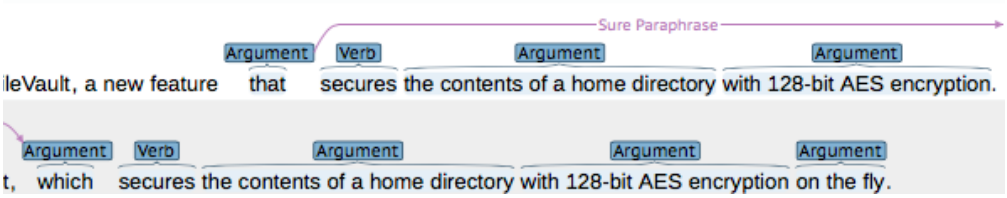


Figure 12 Example for sure paraphrase annotation

As the “Unsure paraphrase” class should only be used in cases you are not sure, we cannot give an example for that.

On the event paraphrase level, there is an additional paraphrase class for the verb element, the “Reverse paraphrase” and it should be used in cases where the same event is described with antonymic or opposite words, e.g.

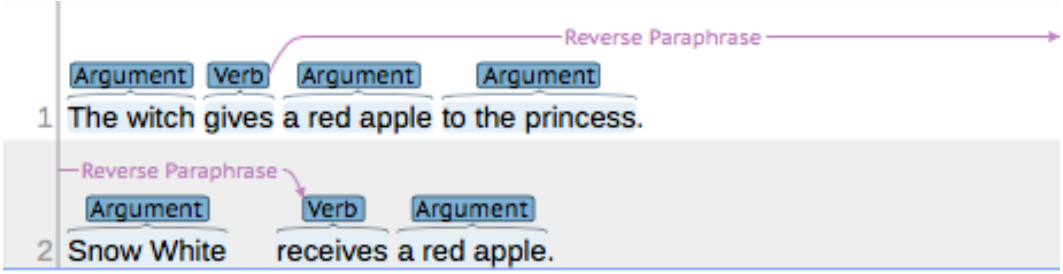


Figure 13 Example for reverse paraphrase

If there is no paraphrase, nothing is aligned or annotated.

Event Paraphrase Annotation Guidelines

Special cases

If one instance contains more or less information than the other, it may still be a paraphrase.
e.g.

They left the house.
They left the house on Monday.

If one instance contains information contradicting the other instance, they cannot be paraphrases.

They left the house on Wednesday.
They left the house on Monday.

If you have relative pronouns that are used instead of a full phrase, mark the pronoun as argument. In the first sentence there are co-references marked for your convenience, but you are not asked to do so.

The argument of interest is marked in bold in the following example.
e.g.

The pastry cook, who makes wonderful cakes, made a statement in the morning.

There are 2 events:

The pastry cook, [**who** makes wonderful cakes], made a statement in the morning.

[**The pastry cook, who makes wonderful cakes**, made a statement in the morning].

If you find orthographic or syntactic mistakes, please annotate as if they weren't there.

e.g.

Theyve done a great job.

"Theyve" would be an argument of done.

A.1.2 Guidelines for Studies of Relations between Semantic Dimensions

A.1.2.1 Guidelines for Study of Links between Relations

Guidelines for Sentence Generation

Guidelines for Sentence Generation

We want to research relationships between sentences. To do so, we need to generate a collection of sentences that are potentially related.

In the following tasks, you are given a reference sentence.

Task 1) Assuming that the information in the sentence is true, please write 2 sentences that are also true given the information in the sentence.
Example:

Given Sentence	True Sentence 1)	True Sentence 2)
Computer games make kids violent.	*Kids playing computer games are more violent than kids who don't.*	*Video games have an impact on the behavior of children.*
	Kids play computer games.	*Computer games affect human behavior.*

Task 2) Assuming that the information in the sentence is true, please write 2 sentences that cannot be true given the information in the sentence.

Given Sentence	False Sentence 1)	False Sentence 2)
Computer games make kids violent.	*Kids get kinder by playing video games.*	*Computer games do not trigger violent behavior.*
	Aggressive behavior comes solely from problems in the family.	*Children are not allowed to play computer games.*

When writing the sentences for both tasks, please try not to reuse the words from the given sentence.
Meaning please DO NOT produce sentences equivalent to the following:

Given Sentence	True Sentence 1)	True Sentence 2)
Computer games make kids violent.	*Computer games make children violent.*	*Kids are made violent by computer games.*

Especially, do not negate the given sentence when creating the false sentence.
Meaning please DO NOT produce sentences equivalent to the following:

Given Sentence	False Sentence 1)	False Sentence 2)
Computer games make kids violent.	Computer games don't make kids violent.	No computer game makes kids violent.

Task 1)

Assuming that the information in the sentence is true, please write 2 sentences that are also true given the information in the sentence.

Given Sentence	True Sentence 1)	True Sentence 2)
Getting a high educational degree is important for finding a good job, especially in big cities.		
In many countries, girls are less likely to get a good school education.		
Going to school socializes kids through constant interaction with others.		
One important part of modern education is technology, if not the most important.		
Modern assistants such as Cortana, Alexa, or Siri make our everyday life easier by giving quicker access to information.		
New technologies lead to asocial behavior by e.g. depriving us from face-to-face social interaction.		
Being able to use modern technologies is obligatory for finding a good job.		
Self-driving cars are safer than humans as they don't drink.		
Machines are good in strategic games such as chess and Go.		
Machines are good in communicating with people.		
Learning a second language is beneficial in life.		
Speaking more than one language helps in finding a good job.		
Christian clergymen learn Latin to read the bible.		

Task 2)

Assuming that the information in the sentence is true, please write 2 sentences that cannot be true given the information in the sentence.

Given Sentence	False Sentence 1)	False Sentence 2)
Getting a high educational degree is important for finding a good job, especially in big cities.		
In many countries, girls are less likely to get a good school education.		
Going to school socializes kids through constant interaction with others.		
One important part of modern education is technology, if not the most important.		
Modern assistants such as Cortana, Alexa, or Siri make our everyday life easier by giving quicker access to information.		
New technologies lead to asocial behavior by e.g. depriving us from face-to-face social interaction.		
Being able to use modern technologies is obligatory for finding a good job.		
Self-driving cars are safer than humans as they don't drink.		
Machines are good in strategic games such as chess and Go.		
Machines are good in communicating with people.		
Learning a second language is beneficial in life.		
Speaking more than one language helps in finding a good job.		
Christian clergymen learn Latin to read the bible.		

Guidelines for Entailment Annotation on AMT

Instructions

Background

We want to research causal relationships between sentences, which will help in information retrieval or summarization tasks. Thus, you are asked to determine whether given that the first sentence is true, the second sentence is also true.

Task

In this task, you are presented with **two sentences**. You are required to decide whether **if Sentence 1 is true, this also makes Sentence 2 true.**

In the case of pronouns (he, she, it, mine, his, our, ...) being used, you can assume they reference proper names, if your common sense does not suggest otherwise (e.g. "Linda" is a female name and can be referenced by "she, her, ...", but not "he, his, ...").

Examples for the option "Sentence 1 causes Sentence 2 to be true":

In that case, the first sentence causes the second sentence to be true, as assuming that John bought a car, it means that he has a car now.

- John bought a car from Mike.
- John has a car.

In that case, the first sentence causes the second sentence to be true, as the first sentence says that both boys and girls play games, it also contains the information that boys play games.

- Boys and girls play games.
- Boys play games.

Examples for the option "Sentence 1 does not cause Sentence 2 to be true":

If the second sentence makes the first sentence true (but the first doesn't make the second true), choose the option "Sentence 1 **does not cause** Sentence 2 to be true":

- John has a car.
- John bought a car from Mike.

If you cannot tell if the first sentence causes the second sentence to be true, choose the option "Sentence 1 **does not cause** Sentence 2 to be true":

- He works as a teacher in Peru.
- He is an English teacher.

Does Sentence 1 cause Sentence 2 to be true?

Sentence 1: ${inst1}

Sentence 2: ${inst2}

- ○ Yes, Sentence 1 **causes** Sentence 2 to be true.
- ○ No, Sentence 1 **does not cause** Sentence 2 to be true.

Guidelines for Paraphrase Annotation on AMT

Instructions

Background

We want to study the meaning relation between two texts. Thus you are asked to determine whether the two sentences mean (approximately) the same or not.

Task

In this task you are presented with **two sentences**. You are required to decide whether the two sentences **have approximately the same meaning or not**.

In the case of pronouns (he, she, it, mine, his, our, ...) being used, you can assume they reference proper names, if your common sense does not suggest otherwise (e.g. "Linda" is a female name and can be referenced by "she, her, ...", but not "he, his, ...").

Examples of the choce "approximately the same meaning":

- John goes to work every day with the metro.
- He takes the metro to work every day.

In the content of the task, we assume that "He" and "John" are the same person.

- Mary sold her Toyota to Jeanne.
- Jeanne bought her Toyota from Mary Smith.

In the content of the task, we assume that "Mary Smith" and "Mary" are the same person.

Examples of the choice of "not the same meaning":

- Mary sold her Toyota to Jeanne.
- Mary had a blue Toyota.

The two texts are related, but are not the same.

- John Smith takes the metro to work every day.
- John works from home every Tuesday.

The two texts are not closely related except for the person (John).

Do the sentences have approximately the same meaning?

Sentence 1: ${inst1}

Sentence 2: ${inst2}

- ◯ Yes, the sentences **have approximately the same meaning**.
- ◯ No, the sentences **do not have approximately the same meaning**.

Guidelines for Contradiction Annotation on AMT

Instructions

Background

We want to study the meaning relation between two texts. Thus you are asked to determine whether the two sentences contradict each other.

Task

In this task you are presented with **two sentences**. You are required to **decide whether the two sentences contradict each other**. Two contradicting sentences can't be true at the same time.

In the case of pronouns (he, she, it, mine, his, our, ...) being used, you can assume they reference proper names, if your common sense does not suggest otherwise (e.g. "Linda" is a female name and can be referenced by "she, her, ...", but not "he, his, ...").

Examples for the option "the sentences contradict each other":

- John bought a new house near the beach.
- John didn't buy the house near the beach.

The second sentence directly contradicts the first one - they can't both be true.

- Mary is on a vacation in Florida.
- Mary is at the office, working.

The two sentences can't be true at the same time - Mary is either on vacation in Florida, or at the office. She can't be in two places.

Examples for the option "the sentences do not contradict each other":

- Mary is on vacation in Florida.
- John is at the office.

John and Mary are two different persons. There is no contradiction. Both statements can be true.

Do the sentences contradict each other?

Sentence 1: ${inst1}

Sentence 2: ${inst2}

- ◯ Yes, the sentences **contradict** each other
- ◯ No, the sentences **do not contradict** each other

Guidelines for Similarity Annotation on AMT

Instructions

Background

We want to study the meaning relation between two texts. Thus you are asked to determine how similar two texts are.

Task

In this task you are presented with **two sentences**. You are required to **decide how similar the two sentences are on a scale** from 0 (completely dissimilar) to 5 (identical)

In the case of pronouns (he, she, it, mine, his, our, ...) being used, you can assume they reference proper names, if your common sense does not suggest otherwise (e.g. "Linda" is a female name and can be referenced by "she, her, ...", but not "he, his, ...").

Examples:

Similarity 0:

- John goes to work every day with the metro.
- The kids are playing baseball on the field.

The two texts are completely dissimilar.

Similarity 1-2:

- John goes to work every day with the metro.
- John sold his Toyota to Sam.

The two texts have some common elements, but are overall not very similar.

Similarity 3-4:

- Mary is writing the report on her Lenovo laptop.
- Mary has a Lenovo laptop.

The two texts have a lot in common, but also have differences.

Similarity 5:

- Mary was feeling blue.
- Mary was sad.

The two texts are (almost) identical.

How similar are the two sentences?

Sentence 1: ${inst1}

Sentence 2: ${inst2}

- 0
- 1
- 2
- 3
- 4
- 5

Guidelines for Specificity Annotation on AMT

Instructions

Background

We want to research whether displaying more specific sentences is helpful in information retrieval or summarization tasks. Thus, you are asked to determine whether the 1st sentence is more specific than the 2nd. The **specificity** of sentence is defined as a measure of how broad or specific its information level is.

Task

In this task, you are presented with **two sentences**. You are required to decide whether **the 1st sentence IS more specific than the 2nd**. If this is not the case, choose the option **the 1st sentence IS NOT more specific than the 2nd**.

Examples for the option "Sentence 1 IS more specific"

- I like cats.
- I like animals.

the 1st sentence IS more specific than the 2nd

As the 1st sentence gives the more specific information on which animal is liked, it is more specific. Hence, you have to choose the option that the 1st sentence is more specific.

- The cute cafe has great coffee.
- The cafe sells coffee.

the 1st sentence IS more specific than the 2nd

As the 1st sentence gives the more specific information on both the cafe and the coffee, it is more specific. Hence, you have to choose the option that the 1st sentence is more specific.

Examples for the option "Sentence 1 IS NOT more specific"

- I like animals.
- I like cats.

the 1st sentence IS NOT more specific than the 2nd

As the 2nd sentence gives the more specific information on which animal is liked, it is more specific. Hence, you have to choose the option that the 1st sentence is not more specific.

- I like dogs.
- I like cats.

the 1st sentence IS NOT more specific than the 2nd

Now, as in both cases the liked animal is mentioned, they have the same level of specificity. Hence, you have to choose the option that the 1st sentence is not more specific.

- I like black dogs.
- He saw a blind cat.

the 1st sentence IS NOT more specific than the 2nd

Now, as the information is very diverse, it is impossible to say which sentence is more specific. Hence, you have to choose the option that the 1st sentence is not more specific.

1. Is Sentence 1 more specific than Sentence 2?

Sentence 1: ${inst1}

Sentence 2: ${inst2}

- ○ Sentence 1 IS more specific
- ○ Sentence 1 IS NOT more specific

A.1.2.2 Guidelines for Extended Relations Typology

Annotation Guidelines ERT (Extended Relations Typology)

1. Presentation

This document sets out the guidelines for the annotation of atomic types using the Extended Typology for Relations. The task consists of annotating pairs of text that hold a textual semantic relation (paraphrasing, entailment, contradiction, similarity) with a textual label, and the atomic phenomena they contain. These guidelines have been used to annotate the ETRC corpus. For the purpose of the annotation, the WARP-Text annotation tool has been used. This document is divided as follows: Section 2 presents general considerations about the task and theoretical definitions. Section 3 presents the tagset definition and the ETR. Section 4 presents guidelines for annotating linguistic phenomena. Section 5 presents guidelines for annotating knowledge/reasoning phenomena.

2. The task

The task consists of
1) annotating the semantic relation between two texts at textual level
2) annotating the atomic relations between (parts of) the two texts.
At textual level, the two texts can have a relation of paraphrasing, entailment, or contradiction. The texts can also have difference in terms of their specificity. We also rate their overall similarity on a continuous scale. All texts that are sufficiently similar are also annotated for atomic types.

N.B.: Texts which are too dis-similar cannot be annotated for types. In that case the annotators should choose type "unrelated" with scope the whole texts. The two texts in 1a and 1b are substantially dis-similar and therefore they are marked as unrelated.

1a The country has no impact on girls ' school education.
1b In order to understand the bible it is important to learn French.

N.B.: It is possible that the two texts have some elements in common (i.e. they share participants, or they are on the same topic), but they do not hold any relation among paraphrasing, entailment, or contradiction. In that only the common elements are annotated. 2a and 2b show such example.

2a Humans drink.
2b Self-driving cars can be compared to humans.

If the two texts are deemed similar enough, the annotation continues with identifying atomic relations and their scope. We distinguish between two kinds of atomic relations – linguistic and reasoning. One of the objectives behind the annotation is to separate the linguistic capabilities needed to process the pair from the reasoning/knowledge capabilities.

The main driving principle behind the annotation of atomic relations is to make the two texts as similar as possible (hypothesis as similar to the text as possible in case of entailment). The goals behind this principle are twofold: 1) we make both the similarities and the key-differences as explicit as possible; 2) we identify the transformations a.k.a. the "steps" needed to obtain 1)

3. The tagset

The ERT (Extended Relations Typology) is based on EPT (Extended Paraphrase Typology). See EPT (Kovatchev et al. 2018) and Vila et al. (2014) typology for more details on the types. See also the list of examples. The full typology is listed in the table:

Category	Phenomena
Morphology based changes	Inflectional Changes
	Modal Verb Changes
	Derivational Changes
Lexicon based changes	Spelling Changes
	Same polarity substitution (habitual)
	Same polarity substitution (contextual)
	Same polarity substitution (named entities)
	Change of format
Lexical-syntactic based changes	Opposite polarity substitution (habitual)
	Opposite polarity substitution (contextual)
	Synthetic/analytic substitution
	Converse substitution
Syntax based changes	Diathesis alternation
	Negation switching
	Anaphora
	Ellipsis
	Coordination changes
	Subordination and nesting changes
Discourse based changes	Punctuation changes
	Direct/Indirect style alternations
	Sentence modality changes
	Syntax/Discourse structure changes
Other changes	Addition/Deletion
	Change of order
Extremes	Identity
	Unrelated
Reasoning	Cause and Effect
	Conditions and Properties
	Functionality and Mutual Exclusivity
	Transitivity
	Numerical Reasoning
	Named Entity Reasoning
	Temporal and Spatial Reasoning
	General Inference / Background Knowledge
	Specificity

ERT has several differences compared with EPT:
- We introduce a new type – "Anaphora" which was deemed necessary in several examples
- The "non-paraphrase" type has been changed to "unrelated", since the typology aims to annotate multiple relations (not just paraphrases).
- The "entailment" type has been removed (for the same reason as above)
- A new category "Reasoning" has been added. It includes all phenomena that have been listed as "Semantic/General Inference" in ETPC.
    - The reasoning category (and subtypes) account for the type of reasoning and knowledge that are needed to perform the inference
    - All reasoning types have direction, in order to handle entailment
- We relax the assumption on which types can have both positive and negative sense preserving. That is, we allow ALL phenomena to be both sense preserving and non-sense preserving.
- We introduce "Specificity" type as part of reasoning, in order to determine which parts of one sentence are more specific than their corresponding elements in the other one.

4. Annotating Linguistic Phenomena

There are two kinds of linguistic phenomena that we annotate: sense preserving and non-sense preserving. The objective behind the annotation of the linguistic phenomena is to make the semantic relation between (parts of) the two texts as explicit as possible.
In the case of **sense preserving phenomena**, the annotation indicates the kind of transformation that makes the parts of the texts **identical**.
In the case of **non-sense preserving phenomena**, the annotation indicates the kind of transformation that makes the parts of the texts **explicitly incompatible** or **contradictory**.

N.B.: We define sense preserving as context-dependent. We mark whether the linguistic transformation, within the context of the two given sentences, makes the two elements identical. 3a and 3b show an example. The linguistic phenomena that relates "students" and "children" is "same polarity substitution (contextual)" and it is sense preserving:

1) the elements fulfil approximately the same syntactic and semantic role in the two sentences. This determines that the involved phenomenon is one of the "same polarity substitution" phenomena.
2) the meaning relation between "students" and "children" is at least in part depending on the context (talking about "school"), therefore the phenomenon is "same polarity substitution (contextual)".
3) The two elements can be swapped between the sentences without (substantially) changing the meaning (ex.: "School makes students antisocial" – "School makes children antisocial"). This determines that the phenomena is "sense preserving".

3a School makes students antisocial.
3b School usually prevents children from socializing properly.

During this step of the annotation we assume linguistic knowledge (understanding of all transformations), dictionary knowledge (lexical relations between words), and world knowledge restricted to basic named entity properties (ex.: "Siri and Cortana" -> "virtual assistants"). During this step, the annotators should annotate all linguistic transformations that they can identify.

4.1 Annotating Sense Preserving Linguistic Phenomena

Sense preserving linguistic phenomena are linguistic transformations that can be applied to a (part of) one of the texts in order to make identical to the other text. The two texts in 4a and 4b show an example. There are two sense-preserving linguistic phenomena involved in the pair: Identity of "All receive the same education" and Same Polarity Substitution (habitual) of "kids" and "children".
N.B.: When the two sentences differ substantially, the easiest way to determine the sense-preserving of a phenomena is to try to apply the change (i.e. substitute the elements) in each sentence (see ex. 3a and 3b in the previous section)

4a All kids receive the same education.
4b All children receive the same education.

When annotating **sense preserving linguistic phenomena**, there are several important things:
- Annotate **all** possible phenomena **separately**. The aim is to annotate every token in both texts if possible.
- The scope of some phenomena can overlap. That means some tokens may be part of multiple scopes.
- When choosing the scope, we choose the **largest** scope possible.
- When annotating phenomena at morphological or lexical level and the phenomena is **sense preserving**, the affected units need not have similar syntactic or semantic role. See 5a and 5b.

5a Reading the Bible requires studying Latin.
5b The Bible is written in Latin.

4.2 Annotating Non-Sense Preserving Linguistic Phenomena

Annotating non-sense preserving linguistic phenomena has one main goal – to make explicit an important **incompatibility** or **contradiction** between the two texts. Like in ETPC non-sense preserving phenomena are involved in all pairs that are not paraphrases. The texts in 6 and 7 show some examples.
In 6a and 6b, the Same Polarity Substitution (habitual) is non-sense preserving as the relation between the substituted words ("children" and "girls") is hyponymy and not synonymy.
In 7a and 7b, the Same Polarity Substitution (habitual) is non-sense preserving as the two substituted words ("boys" and "girls") are substantially different.

N.B.: in both 3 and 4, the Identity of "All receive the same education" should be annotated regardless of the difference between the underlined words

6a All children receive the same education.
6b All girls receive the same education.

7a All boys receive the same education.
7b All girls receive the same education.

When annotating **non-sense preserving linguistic phenomena**, there are several important things:

- Annotate **all** possible phenomena **separately**. The aim is to annotate every token that is not already annotated as sense preserving.
- The scope of some phenomena can overlap. That means some tokens may be part of multiple scopes. Some tokens might even be part of sense preserving and non-sense preserving scopes at the same time.
- When choosing the scope, we choose the **smallest** scope possible. Unlike the sense-preserving, in this part of the annotation, the goal is to choose the most specific scope possible. For example, in 4a and 4b we could annotate "All boys" and "All girls", but in order to be as specific as possible, we choose to only annotate "boys" and "girls".
- When annotating phenomena at morphological or lexical level and the phenomena is **non-sense preserving**, the affected units **must** have similar syntactic or semantic role. Each token in the sentence is different from most of the other tokens (ex.: "boys" in 4a is different from "education" in 4b), so in order to annotate a difference, they must be comparable in terms of their role within the sentence.
- When there are multiple different elements (arguments of a same verb) that are different, each of them should be annotated separately.

N.B.: sometimes, it is unclear whether a phenomenon is sense preserving or non-sense preserving. Look at the annotation of the textual relation for additional information – if the texts are annotated as "paraphrases", they should NOT (in general) contain any non-sense preserving phenomena. On the contrary, if the pair is annotated as NOT-paraphrase, they should contain either non-sense preserving phenomena, addition/deletion, or reasoning phenomena.

4.3 Annotating Addition/Deletion

If parts of one of the texts are not presented in (or related to) the other text, they should be annotated as "Addition/Deletion". 5a and 5b show an example.

8a Clergymen never read the bible.
8b Christian clergymen learn Greek to read the bible.

When annotating addition/deletion, elements that belong to different syntactic units should be annotated separately. In example 8, there are two separate Addition/Deletion phenomena – "Christian" and "learn Greek".

4.4 Examples of Linguistic phenomena

4.4.1 Inflectional change – two words in the two texts with (approximately) the same syntactic or semantic role, which only differ in their inflection.

If they have (approximately) the same meaning, the phenomenon is sense-preserving: ``cities" and "city" in example 9 both have the meaning of "any big city", therefore the phenomena is sense preserving
9a It is harder to find a good job in big cities
9b It is harder to find a good job in a big city

If the meaning changes substantially, the phenomena is non-sense preserving: ``cities'' and ''city'' in example 10 refer to different entities.

10a He has a good job in a big city
10b He has a good job in big cities

4.4.2 Modal Verb Changes – changing the modal verb. Example 11 shows a sense preserving example of modal verb change, while example 12 shows a non-sense preserving use.

11a He could have a good job in the big city.
11b He might have a good job in the big city.

12a He could have a good job in the big city.
12b He should have a good job in the big city.

4.4.3 Derivational Changes – two words in the two texts with (approximately) the same syntactic or semantic role, which share a derivational relation

If they have (approximately) the same meaning, the phenomenon is sense-preserving:
13a Modern assistants such as Cortana , Alexa , or Siri speed up accessing to information .
13b Modern assistants such as Cortana , Alexa , or Siri speed up the access to information .

If the meaning changes substantially, the phenomena is non-sense preserving.

14a
14b

4.4.4 Spelling changes – two words that differ in their spelling. This includes different spelling variations ("color – colour"), contractions ("do not – don't") and abbreviations ("E.U. – European Union"). Spelling changes should always be sense-preserving.

4.4.5 Same Polarity Substitution (habitual) – two words or phrases in the two texts with (approximately) the same syntactic or semantic role, which have a clear, out-of-context semantic relation. The two words must be used in the text with (one of) their typical out-of-context meaning.

If the relation is synonymy, the SPS (habitual) is sense-preserving, as in example 15.
15a All kids receive the same education.
15b All children receive the same education.

If the relation is not synonymy, the SPS (habitual) is non-sense preserving, as in 16.
16a All kids receive the same education.
16b All girls receive the same education.

4.4.6 Same Polarity Substitution (contextual) – two words or phrases in the two texts with (approximately) the same syntactic or semantic role. The two words may or may not have a clear, out-of-context semantic relation. However at least part of the meaning of the words must depend on the context.

If the two words (or phrases) have approximately the same meaning in context, the phenomenon is sense preserving as in 17.

17a School makes students antisocial.
17b School prevents children from socializing.

Note that pronouns are also considered same polarity substitution (contextual). In the tasks of PI and RTE, we assume co-referentiality of entities unless the entities are incompatible and/or unless explicitly stated otherwise. Therefore example 18 is also SPS (contextual)

18a The man has a good job.
18b He has a good job.

If the meaning of the words differs substantially in the context, the phenomenon is non-sense preserving.
NB.: If two words or phrases have approximately the same syntactic and/or semantic roles but the two words differ substantially in meaning, they are annotated as non-sense preserving SPS(contextual) as in 19.

19a School makes students antisocial.
19b School makes teachers antisocial.

4.4.7 Same Polarity Substitution (Named Entity) – two words or phrases in the two texts with (approximately) the same syntactic or semantic role, which have a clear, out-of-context semantic relation to the same Named Entity. The two words (or phrases) could be different names of the same entity ("Stephen King" – "Richard Bachman") or a name and a property or a characteristic of the name ("Barack Obama" – "The 44th US President"). SPS (NE) is also used when substituting quantities.

When the two words or phrases have the same referent (or quantity), the phenomenon is sense preserving, as in 20.

20a Siri and Cortana can give you faster access to information.
20b Virtual assistants can give you faster access to information.

When the two words or phrases have different referents (or refer to different quantities), the phenomenon is non-sense preserving as in 21 and 22. In 21, while the two sentences are very similar in meaning, there is also a clear difference – not liking Cortana is not the same as not liking (all) virtual assistants.

21a Jane does not like Cortana
21b Jane does not like virtual assistants

22a Jane is born in 1995
22b Jane is born in 1996

4.4.8 Change of Format – two words that differ in their spelling. This includes changes such as "$"-"dollar", "%"-"percent". This phenomenon should always be sense-preserving.

4.4.9 Opposite polarity substitution (habitual) – two words or phrases in the two texts with (approximately) the same syntactic or semantic role, which have a clear, out-of-context

semantic relation of **contradictory/opposite meaning** (ex.: "good"- "bad"; "safe" – "dangerous").
N.B.: Polarity here is not meant in the sense of sentiment

When the two words or phrases have approximately the same meaning in the context of the sentences, the phenomenon is sense preserving. Usually in this case one of the words is **negated**. The negation, when presented, should also be part of the scope. In the example 23 the opposite polarity substitution (habitual) of "safe" and "not dangerous" includes the opposite words "safe" and "dangerous" and the negation. Note that the scope of negation could include more than just the word, like in 23c or 23d. The negation of the subject ("no") or the main verb ("n't") both have scope over the whole clause. Therefore, the phenomenon is OPS (habitual)

23a Autonomous cars are safe
23b Autonomous cars are not dangerous
23c No autonomous cars are dangerous
23d Autonomous cars are n't dangerous

When the two words or phrases have substantially different meaning in the context of the sentences, the phenomenon is non-sense preserving, like in 24.

24a Autonomous cars are safe
24b Autonomous cars are dangerous

4.4.10 Opposite polarity substitution (contextual) – two words or phrases in the two texts with (approximately) the same syntactic or semantic role. The two words may or may not have a clear, out-of-context semantic relation. However, in the context they have **contradictory / opposite meaning** and at least part of the meaning of the words must depend on the context.

When the two words or phrases have approximately the same meaning in the context of the sentences, the phenomenon is sense preserving. Usually in this case one of the words is **negated** like in 25.

25a Autonomous cars have n't passed all the tests
25b Autonomous cars are dangerous

When the two words or phrases have substantially different meaning in the context of the sentences, the phenomenon is non-sense preserving, like in 26.

26a Autonomous cars have passed all the tests
26b Autonomous cars are dangerous

4.4.11 Synthetic/analytic substitution – this is a meta phenomenon that includes various ways to represent similar meaning between words and/or phrases. The transformations involved in this phenomenon include **modifiers** ("technology" – "modern technology"), **genitive** ("Mexico's president" – "the president of Mexico"), **argument realization** ("Obama, the 44th

US President" – "Obama is the 44th US president") and other rewrite rules ("a sequence of ideas" – "ideas").

When the two words or phrases have approximately the same meaning in the context of the sentences, the phenomenon is sense preserving like in 27.

27a Modern education has no use for technology.
27b Education does not benefit from new technology.

When the two words or phrases have substantially different meaning in the context of the sentences, the phenomenon is non-sense preserving, like in 28.

28a Computers can play games
28b Computers can play strategic games

4.4.12 Converse substitution – referring to a (verbal) relation from the opposite points of view: "sell" – "buy", "is taller" – "is shorter". Converse substitution is often accompanied with a change of the syntactic roles, as seen in 29. John is the subject of "sell", while "Mary" is the subject of "buy".

When the relation between the entities is approximately the same (i.e. when the two points of view refer to the same relation), the phenomenon is sense preserving, as in 29 and 30.

29a John **sold** his car to Mary
29b Mary **bought** her car from John

30a John **is taller** than Mary
30b Mary **is shorter** than John

When the two points of view refer to different relations, the phenomenon is non-sense preserving, as in 31.

31a John **sold** his car to Mary
31b John **bought** his car from Mary

N.B.: In this phenomenon, we mark the scope (the whole clause that is affected by the change, can be the whole sentence or just part of it) and the key (the two verbs).

4.4.13 Diathesis alternation – referring to a change of the grammatic voice between the two sentences. Diathesis alternation is often accompanied with a change of the syntactic roles as in 32.

When the change of voice does not result in change of meaning, the phenomenon is sense preserving as in 32

32a John **buys** a car.
32b A car **is bought** by John.

When the change of voice also results in a substantial change of meaning, the phenomenon is non-sense preserving as in 33

33a The president **gave** a speech.
33b The president **was given** a speech.

N.B.: In this phenomenon, we mark the scope (the whole clause that is affected by the change, can be the whole sentence or just part of it) and the key (the two verbs).

4.4.14 Negation Switching – change in the manner in which the negation is expressed and/or change in the scope of negation.

N.B.: If there is negation in one of the sentences and that negation is expressed in another way (or missing) in the other sentence, negation switching must be annotated
N.B.: Often negation switching appears together with other phenomena (ex.: Opposite Polarity Substitution or Converse Substitution). **Both** phenomena have to be annotated independently.

If the meaning of the two texts is approximately the same, the phenomenon is sense-preserving like in 34 and 35.

34a Autonomous cars are **not** dangerous
34b **No** autonomous cars are dangerous

Negation can be more complex, involving lexical negation

35a Jane **did not want** to buy the house
35b Jane **refused** to buy the house

If the meaning of the two texts changes substantially as a result of the negation switching, the phenomenon is non-sense preserving like in 36 and 37.

36a Autonomous cars are **not** dangerous
36b Autonomous cars are dangerous

37a Jane did **n't** see John
37b **Nobody** saw John

N.B.: In this phenomenon, we mark the scope (the whole clause that is affected by the change, can be the whole sentence or just part of it) and the key (the negation markers, and in the case of lexical negation – the verbs).

4.4.15

5. Annotating Reasoning Phenomena

Reasoning phenomena account for relations that cannot be expressed and processed using only linguistic knowledge. Like the linguistic phenomena, the reasoning phenomena can be sense-preserving or non-sense preserving. Our goals with the annotation of reasoning phenomena are twofold: 1) we want to make a precise and explicit annotation of the units involved in the inference; 2) we want to determine the kind of reasoning and background knowledge required. 6a and 6b show an example of an "existential" reasoning – "speaking X" entails "X exists". 7a and 7b show an example of "causal" reasoning – "X is written in Y (language)" entails "reading X requires Y (language)" (not sure – should we only annotate "reading requires" or "reading the bible requires latin" or "reading requires latin").

6a Speaking more than one language is imperative today.
6b There is more than one language.

7a Reading the Bible requires studying Latin.
7b The Bible is written in Latin.

When annotating **reasoning phenomena**, there are several important things:

- Annotate **all** possible phenomena **separately**. The aim is to annotate every token that is not already annotated as linguistic or addition-deletion.
- The scope of some phenomena can overlap. That means some tokens may be part of multiple scopes.
- When choosing the scope, we choose the **smallest** scope possible. Unlike the sense-preserving, in this part of the annotation, the goal is to choose the most specific scope possible. For example, in 6a and 6b we could annotate "Speaking more than one language" and "There is more than one language", but in order to be as specific as possible, we choose to only annotate "Speaking" and "There is".
- When choosing the scope, if possible, try to annotate whole linguistic units without breaking them. For example in 7a and 7b, we could only annotate "Reading requires" and "is written in"
- Like in the linguistic phenomena – the **sense preserving reasoning phenomena** need not relate units that have similar syntactic or semantic role; however, the **non-sense preserving reasoning phenomena** must relate units that have similar syntactic or semantic role.

A.1.3 Guidelines for Sentiment Studies

A.1.3.1 Guidelines for Sentiment Annotation on Political Speeches

Guidelines 1.2

Goal:

+ Sprachliche Nomentypen sind für unsere Aufgabenstellungen irrelevant.

Doppelannotation:
+ much/very/lot
+ Zahlen

Adjektiv - Nomen Relation:

+ Pfeilrichtung: Vom **Nomen** auf das **Adjektiv**.
+ Keine Annotation über Satzgrenzen.
+ Nomen, die aus mehreren Wörtern bestehen sollen auf das „Kernnomen“ reduziert werden, damit die Übereinstimmung nicht gefährdet wird.
+ Pos Neg Neut der Relation bestimmen,
  dann gucken ob negiert,
  dann gucken ob Irónie.
  + *eliminate underpaid jobs, underpaid negativ.*

Bsp.:

University of Hampstead	-> Nur University markieren,
Federal Court	-> Nur Court markieren,

+ mehrere Adjektive können sich auf ein Nomen beziehen

Bsp.: „..our entire adult life.“ Pfeil vom Nomen life auf A entire und A adult

+Adjektiv kann sich auf mehrere Nomen beziehen, z.B bei Sätzen mit Komma zwischen den Nomen

Bsp.: advanced manufacturing, innovation and technology;
Relation von N manufacturing, N innovation und N technology auf A advanced

Adjektive klassifizieren (positiv, negativ, neutral)

+ Anhand der Nomen, auf die sie sich beziehen.
+ Kontextabhängig (Bsp. Ironie) „so wie der Sprecher das meint“

Nomen kategorisieren, anhand der folgenden Kategorien: (1.2)

+ alle Nomen werden annotiert, nicht nur die mit Relation auf ein Adjektiv
+ zusammengesetztes Nomen : **„Nomen-Bezugsregel“** => Nomen A bezieht sich auf Nomen B, also bekommt Nomen A selben Tag wie Nomen B
+ Rangfolge der Tags beachten

1. **WMN** Bezugnahme auf Frauen
2. **GRP** Nationalitäten, „andere Länder“, Religion
3. **OPP** Referenz auf Instanz der Opposition, andere Partei
4. **AG** Bezugnahme auf zukünftiges Vorhaben während der Präsidentschaft,
5. **AM** Amerika: Wirtschaft, Politik, Gesellschaft.
6. **SELF** Beschreibung der eigenen Person (persönlich, privat), Beschreibung eigene Partei
7. **OTHER** Sonstiges das nicht anhand der anderen Kategorien eingeordnet werden kann, Satz Teile ohne Relevanz Bsp. „…,end of story.“

+ Hinweisworte für Tags:
„our“ —> **AM** außer es ist die Rede von Hillary + Trump

„your“ —> **OPP** / **AM**
„going to“ —> **AG**
„my“ —> **SELF**

Bei Prüfung Doppelannotation:

- **Phrasen** raus gelöscht („Take a Look…“ usw.)
- Wenn das Wort als Ganzes im Wörterbuch existiert Bsp. „federal judge“ „ general election“ werden beide Wörter als N getaggt

A.1.3.2 Guidelines for Producing Explicit Hate Speech

Guidelines for making Implicit Hatespeech Explicit

If the given Tweet contains an implicit stance towards Islam, Muslims, or refugees, please make it explicit by paraphrasing each sentence using one the following rules:

1) Built it in as an additional argument, meaning subject or object (and adjust the sentence to it):

Refugees are criminals. Return to where they came from!

$$\#Refugees \xrightarrow{transform\ to}$$

Refugees are criminals. **Refugees must** return to where they came from!

2) Replace co-references of the implicit stance.

Refugees are criminals. They must return to where they came from!

$$\#Refugees \xrightarrow{transform\ to}$$

Refugees are criminals. **Refugees must** return to where they came from!

3) Built the stance in as a noun-adjective conversion specifying one of the arguments, meaning the subject or the object.

Other countries don't have issues with Muslims. Merkel's curse!

$$\#Muslims \xrightarrow{transform\ to}$$

Other countries don't have issues with Muslims. Merkel's **Muslim** curse!

If the message of the Tweet is softened through the use of modals, quantifiers, or specifications, make it more explicit by paraphrasing it in the following way:

4) Delete the specifying phrase

Criminal refugees aren't punished for their crimes.

$$\xrightarrow{transform\ to}$$

Refugees aren't punished for their crimes.

5) Replace the specifying phrase

Many refugees are criminal.

$$\xrightarrow{transform\ to}$$

All refugees are criminal.

6) Deleting the softening phrase

I **believe** refugees are criminals.

$$\xrightarrow{transform\ to}$$

Refugees are criminals.

7) Changing the softening phrase

They **should** be sent back.

$$\xrightarrow{transform\ to}$$

They **must** be sent back.

A.1.4 Guidelines for Specificity using BWS

1. Which of the following sentences is most specific regarding the ambiance of a restaurant?

- This place is not inviting and the food is totally weird.
- The atmosphere was pretty nice but had a bit lacking, which it tries to make up for with a crazy scheme of mirrors.
- The food is decent at best, and the ambience, well, it's a matter of opinion, some may consider it to be a sweet thing, I thought it was just annoying.
- Despite the confusing mirrors this will likely be my go-to for modern Japanese food for the foreseeable future.

2. Which of the following sentences is least specific regarding the ambiance of a restaurant?

- This place is not inviting and the food is totally weird.
- The atmosphere was pretty nice but had a bit lacking, which it tries to make up for with a crazy scheme of mirrors.
- The food is decent at best, and the ambience, well, it's a matter of opinion, some may consider it to be a sweet thing, I thought it was just annoying.
- Despite the confusing mirrors this will likely be my go-to for modern Japanese food for the foreseeable future.

A.2 Statistics, Illustrations, Typologies and Examples from Studies in this Thesis

A.2.1 Relations between Semantic Dimensions

#	Sentence 1	Sentence 2	PP	FTE	BTE	Cont	FSpec	BSpec	Sim
1	The importance of technology in modern education is overrated.	Technology is not mandatory to improve education	✓	✓	✓				2.8
2	Machines cannot interact with humans.	No machine can communicate with a person.	✓	✓	✓				4.9
3	The modern assistants make finding data slower.	Today's information flow is greatly facilitated by digital assistants.				✓		✓	1.9
4	The bible is in Hebrew.	Bible is not in Latin.		✓			✓		2.7
5	All around the world, girls have higher chance of getting a good school education.	Girls get a good school education everywhere.	✓					✓	4.7
6	Reading the Bible requires studying Latin.	The Bible is written in Latin.		✓	✓			✓	3.6
7	Speaking more than one language can be useful.	Languages are beneficial in life.	✓	✓	✓			✓	4.4
8	You can find a good job if you only speak one language.	People who speak more than one language could only land pretty bad jobs.			✓				2.3
9	All Christian priests need to study Persian, as the Bible is written in Ancient Greek.	Christian clergymen don't read the bible.						✓	0.9

TABLE A.1: Annotations of sentence pairs on all meaning relations taken from our corpus studying the links between relation dimensions (continued)

#	Sentence 1	Sentence 2	PP	FTE	BTE	Cont	FSpec	BSpec	Sim
10	School makes students antisocial.	School usually prevents children from socializing properly.	✓	✓	✓			✓	3.9

TABLE A.1: Annotations of sentence pairs on all meaning relations taken from our corpus studying the links between relation dimensions (end)

ID	**Type**
Morphology-based changes	
1	Inflectional changes
2	Modal verb changes
3	Derivational changes
Lexicon-based changes	
4	Spelling changes
5	Same polarity substitution (habitual)
6	Same polarity substitution (contextual)
7	Same polarity sub. (named entity)
8	Change of format
Lexico-syntactic based changes	
9	Opposite polarity sub. (habitual)
10	Opposite polarity sub. (contextual)
11	Synthetic/analytic substitution
12	Converse substitution
Syntax-based changes	
13	Diathesis alternation
14	Negation switching
15	Ellipsis
16	Anaphora
17	Coordination changes
18	Subordination and nesting changes
Discourse-based changes	
18	Punctuation changes
20	Direct/indirect style alternations
21	Sentence modality changes
22	Syntax/discourse structure changes
Other changes	
23	Addition/Deletion
24	Change of order
Extremes	
25	Identity
26	Unrelated
Reason-based changes	
27	Cause and Effect

TABLE A.2: The SHARel (continued)

ID	Type
28	Conditions and Properties
29	Functionality and Mutual Exclusivity
30	NE Reasoning
31	Numerical Reasoning
32	Temporal and Spatial Reasoning
33	Transitivity
34	Other (General Inference)

TABLE A.2: The SHARel (end)

ID	Type	Paraphrasing	Entailment	Contradiction	ETPC
Morphology-based changes					
1	Inflectional changes	4 %	4 %	1.9 %	2.78 %
2	Modal verb changes	0.25 %	1 %	0	0.83 %
3	Derivational changes	2 %	0	0.6 %	0.85 %
Lexicon-based changes					
4	Spelling changes	0.25 %	0.4 %	0	2.91 %
5	Same polarity substitution (habitual)	25.2 %	17 %	26 %	8.68 %
6	Same polarity substitution (contextual)	9.7 %	6.3 %	5.5 %	11.66 %
7	Same polarity sub. (named entity)	0.7 %	0.4 %	1.2 %	5.08 %
8	Change of format	0.7 %	0.9 %	0	1.1 %
Lexico-syntactic based changes					
9	Opposite polarity sub. (habitual)	2.7 %	3.5 %	7.5%	0.07 %
10	Opposite polarity sub. (contextual)	0.5 %	0.9 %	1.2 %	0.02 %
11	Synthetic/analytic substitution	6.7 %	6.8 %	3.7 %	3.80 %
12	Converse substitution	2.5 %	3.2 %	3.1 %	0.20 %
Syntax-based changes					
13	Diathesis alternation	1.5 %	2.2%	1.9 %	0.73 %
14	Negation switching	4 %	4 %	11.2 %	0.09 %
15	Ellipsis	0	0	0	0.30 %
16	Anaphora	1.7 %	2.7 %	0.6 %	0
17	Coordination changes	0	0	0	0.22 %
18	Subordination and nesting changes	0.25 %	0	0	2.14 %
Discourse-based changes					
18	Punctuation changes	0	0	0	3.77 %
20	Direct/indirect style alternations	0	0	0	0.30 %
21	Sentence modality changes	0	0	0	0
22	Syntax/discourse structure changes	0	0	0	1.39 %
Other changes					
23	Addition/Deletion	16.25 %	16.4 %	16.2 %	25.94 %
24	Change of order	0.5 %	0.9 %	0.6 %	3.89 %
Extremes					
25	Identity	12.5 %	14.5 %	11.8 %	17.5 %

TABLE A.3: Type Frequency distribution of SHARel Typology in Paraphrasing, Entailment, and Contradiction in our corpus and ETPC for comparison (continued)

ID	Type	Paraphrasing	Entailment	Contradiction	ETPC
26	Unrelated	0	0	0	3.81 %
	Reasoning				
27	Cause and Effect	4.7 %	5.4 %	5 %	n/a
28	Conditions and Properties	2 %	6 %	0.6 %	n/a
29	Functionality and Mutual Exclusivity	0	0.4 %	0	n/a
30	NE Reasoning	0	0	0	n/a
31	Numerical Reasoning	0	0	0	n/a
32	Temporal and Spatial Reasoning	0	0	0	n/a
33	Transitivity	0.25 %	0.9 %	0	n/a
34	Other (General Inference)	0.5 %	0.4 %	0.6 %	1.53 %

TABLE A.3: Type Frequency distribution of SHARel Typology in Paraphrasing, Entailment, and Contradiction in our corpus and ETPC for comparison (end)

A.2.2 Specificity

Rank	Sent
1	Traditional French decour was pleasant though the hall was rather noisy - the restaurant was full and we had to raise our voices to be able to maintain a conversation.
2	The place is small and intimate and you may feel a little crowded, but the service is excellent and it's great for friends out, a romantic date, or a special occassion.
3	Although the tables may be closely situated, the candle-light, food-quality and service over-compensate.
4	This place has totally weird decor, stairs going up with mirrored walls - I am surprised how no one yet broke their head or fall off the stairs - mirrored walls make you dizzy and delusional...
5	The music playing was very hip, 20-30 something pop music, but the subwoofer to the sound system was located under my seat, which became annoying midway through dinner.
6	The atmosphere was pretty nice but had a bit lacking, which it tries to make up for with a crazy scheme of mirrors.
7	The food was great and tasty, but the sitting space was too small, I don't like being cramp in a corner.
8	The wait here is long for dim sum, but if you don't like sharing tables or if the typical raucous dim sum atmosphere is not your gig, this is a sleek (for Chinatown) alternative.
9	Despite the confusing mirrors this will likely be my go-to for modern Japanese food for the foreseeable future.
10	The tables are crammed way too close, the menu is typical of any Italian restaurant, and the wine list is simply overpriced.
11	The seats are uncomfortable if you are sitting against the wall on wooden benches.
12	The space kind of feels like an Alice in Wonderland setting, without it trying to be that.
13	The atmosphere is noisy and the waiters are literally walking around doing things as fast as they can.
14	The staff has been nice, but they seemed really stressed and the unisex bathroom needs to be cleaned more often.
15	The decor is night tho...but they REALLY need to clean that vent in the ceiling...its quite un-appetizing, and kills your effort to make this place look sleek and modern.

TABLE A.4: Rating Result of best-worst scaling (BWS) for Specificity (continued)

Rank	Sent
16	Calling the place Hampton Chutney Co. does warn you that these folks offer more style than substance, but in this unattractive room with unhelpful clerks there was a dearth of the former too.
17	Admittedly some nights inside the restaurant were rather warm, but the open kitchen is part of the charm.
18	The food is decent at best, and the ambience, well, it's a matter of opinion, some may consider it to be a sweet thing, I thought it was just annoying.
19	I have it a 4 instead of 5 because of the price (just chicken tikka masala - no bread of rice - is $25), which I would expect at a upscale Indian restaurant but this place doesn't have an upscale feel.
20	It's a rather cramped and busy restaurant and it closes early.
21	If you like your music blasted and the system isnt that great and if you want to pay at least 100 dollar bottle minimun then you'll love it here.
22	Prices too high for this cramped and unappealing resturant.
23	The place is small and cramped but the food is fantastic.
24	The waiter was a bit unfriendly and the feel of the restaurant was crowded.
25	Service ok but unfriendly,filthy bathroom.
26	Decor needs to be upgraded but the food is amazing!
27	First of all, this place is *not* romantic, as claimed by Citysearch's editorial review.
28	Excellent food, although the interior could use some help.
29	oh speaking of bathroom , the mens bathroom was disgusting.
30	The only thing that strikes you is the decor?(not very pleasant).
31	This place is not inviting and the food is totally weird.
32	Even though the place is not beautiful, the food speaks for itself.
33	Mazing interior.
34	Service was also horrible and the ambience is not that great.
35	Well, this place is so Ghetto its not even funny.
36	Zero ambiance to boot.

TABLE A.4: Rating Result of BWS for Specificity (end)

A.2.3 Sentiment

	Marked				Unmarked			
	Clinton		Trump		Clinton		Trump	
Aspect	Sum	%	Sum	%	Sum	%	Sum	%
AGENDA	210	9,00	112	4,63	96	13.8	44	4.3
no sent	122		79					
w. sent	99		35					
pos	69	70,00	28	80,00	14	14.6	11	25,0
neut	24	24,00	4	11,43	82	85.4	33	75,0
neg	6	6,00	3	8,57	0	0,0	0	0,0
US	580	24,97	547	22,60	207	29.8	248	24.3
no sent	475		441					
w. sent	121		115					
pos	30	24,79	32	27,83	66	31.9	21	8.5

TABLE A.5: Distribution of the aspects and the sentiments within aspects for the whole dataset (continued)

	Marked				Unmarked			
	Clinton		Trump		Clinton		Trump	
Aspect	Sum	%	Sum	%	Sum	%	Sum	%
neut	79	65,29	48	41,74	110	53.1	102	41.1
neg	12	9,92	35	30,43	31	15,0	125	50.4
GROUP	223	9,60	280	11,57	70	10.1	98	9.6
no sent	177		235					
w. sent	54		54					
pos	3	5,56	5	9,26	5	7.1	8	8.2
neut	44	81,48	27	50,00	48	68.6	60	61.2
neg	7	12,96	22	40,74	17	24.3	30	30.6
OPP.	171	7,36	191	7,89	93	13.4	151	14.8
no sent	143		162					
w. sent	28		32					
pos	1	3,57	4	12,50	3	3.2	7	4.6
neut	8	28,57	6	18,75	26	28,0	67	44.4
neg	19	67,86	22	68,75	64	68.8	77	51,0
SELF	66	2,84	76	3,14	98	14.1	217	21.2
no sent	53		46					
w. sent	13		31					
pos	6	46,15	25	80,65	20	20.4	59	27.2
neut	6	46,15	6	19,35	76	77.6	158	72.8
neg	1	7,69	0	0,00	2	2,0	0	0,0
WOMEN	70	3,01	23	0,95	6	0.9	0	0,0
no sent	64		21					
w. sent	8		2					
pos	1	12,50	1	50,00	1	16.7	0	0
neut	4	50,00	1	50,00	4	66.7	0	0
neg	3	37,50	0	0,00	1	16.7	0	0
OTHER	1003	43,18	1191	49,21	125	18,0	264	25.8
no sent	759		818					
w. sent	269		413					
pos	67	24,91	80	19,37	10	8,0	16	6.1
neut	132	49,07	176	42,62	85	68,0	182	68.9
neg	70	26,02	157	38,01	30	24,0	66	25,0

TABLE A.5: Distribution of the aspects and the sentiments within aspects for the whole dataset (end)

A.3 Materials by Others

Specificity Guidelines by Others

Sentences could vary in how much detail they contain. One distinction we might make is whether a sentence is general or specific. General sentences are broad statements made about a topic. Specific sentences contain details and can be used to support or explain the general sentences further. In other words, general sentences create expectations in the minds of a reader who would definitely need evidence or examples from the

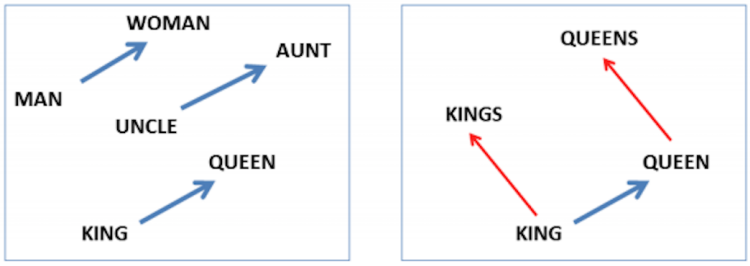


FIGURE 1: Exemplary word embedding representation (Mikolov et al., 2013b)

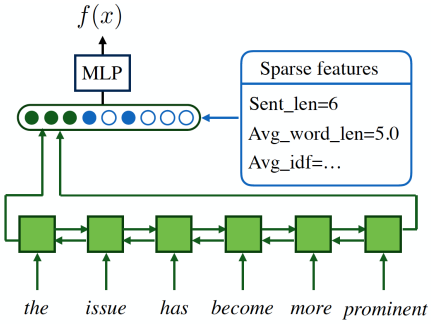


Figure 1: Base model for sentence specificity prediction. The sentence x is encoded with a BiLSTM combined with sparse features, and fed to a MLP to predict specificity $f(x)$.

FIGURE 2: Illustration of system by Ko et al. (2019a)

> author. Specific sentences can stand by themselves. For example, one can think of the first sentence of an article or a paragraph as a general sentence compared to one which appears in the middle. In this task, use your intuition to rate the given sentence as general or specific.

(Louis and Nenkova, 2011, p.609).

Bibliography

Swati Agarwal and Ashish Sureka. Using KNN and SVM based one-class Classifier for Detecting Online Radicalization on Twitter. In *International Conference on Distributed Computing and Internet Technology*, pages 431–442, Bhubaneswar, India, 2015.

Rodrigo Agerri, Itziar Aldabe, Egoitz Laparra, German Rigau, Antske Fokkens, Paul Huijgen, Ruben Izquierdo, Marieke Van Erp, Piek Vossen, Anne-lyse Minard, and Bernardo Magnini. Multilingual Event Detection using the NewsReader Pipelines. In *Proceedings of LREC*, volume 17, pages 42–46, Reykjavík, Iceland, 2014.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. * SEM 2013 shared task: Semantic textual similarity. In *Proceedings of *SEM/CoNLL*, volume 1, pages 32–43, Atlanta, GA, USA, 2013.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. Semeval-2014 Task 10: Multilingual semantic textual similarity. In *Proceedings of SemEval*, pages 81–91, Dublin, Ireland, 2014.

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Polyglot: Distributed Word Representations for Multilingual NLP. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria, 2013.

Erin L. Allwein, Robert E. Schapire, and Yoram Singer. Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers. *Journal of Machine Learning Research*, 1:113–141, 2000.

Salha Alzahrani and Naomie Salim. Fuzzy Semantic-Based String Similarity for Extrinsic Plagiarism Detection. In *CLEF Labs and Workshops, Notebook Papers*, volume 1176, pages 1–8, Padua, Italy, 2010.

American Psychological Association. *Publication manual of the American Psychological Association*. American Psychological Association Washington, 1994.

Ion Androutsopoulos and Prodromos Malakasiotis. A Survey of Paraphrasing and Textual Entailment Methods. *Journal of Artificial Intelligence Research*, 38:135–187, 2010.

Gabor Angeli, Melvin J. Johnson Premkumar, and Christopher D. Manning. Leveraging Linguistic Structure For Open Domain Information Extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, volume 1, pages 344–354, Beijing, China, 2015.

Ron Artstein and Massimo Poesio. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596, 2008.

Collin F. Baker and Josef Ruppenhofer. FrameNet's frames vs. Levin's verb classes. In *Proceedings of the 28th annual meeting of the Berkeley Linguistics Society*, pages 27–38, Berkeley, CA, USA, 2002.

Collin F. Baker, Charles J. Fillmore, and Beau Cronin. The Structure of the Framenet Database. *International Journal of Lexicography*, 16(3):281–296, 2003.

Alexandra Balahur, Zornitsa Kozareva, and Andrés Montoyo. Determining the Polarity and Source of Opinions Expressed in Political Debates. In *Proceedings of CICLing*, pages 468–480, Mexico City, Mexico, 2009.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. Abstract Meaning Representation for Sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria, 2013.

Daniel Bär, Torsten Zesch, and Iryna Gurevych. Text Reuse Detection using a Composition of Text Similarity Measures. *Proceedings of COLING 2012*, pages 167–184, 2012.

Jamie Bartlett, Jeremy Reffin, Noelle Rumball, and Sarah Williamson. Anti-social media. *Demos*, pages 1–51, 2014.

Hans Baumgartner and Jan-Benedict E.M. Steenkamp. Response styles in marketing research: A cross-national investigation. *Journal of marketing research*, 38(2):143–156, 2001.

Asa Ben-Hur, David Horn, Hava T. Siegelmann, and Vladimir Vapnik. Support Vector Clustering. *Journal of machine learning research*, 2(Dec):125–137, 2001.

Darina Benikova and Torsten Zesch. Bridging the gap between computable and expressive event representations in Social Media. In *Proceedings of the Workshop on Uphill Battles in Language Processing: Scaling Early Achievements to Robust Methods*, pages 6–10, 2016.

Darina Benikova and Torsten Zesch. Same Same, but different: Compositionality of Paraphrase Granularity Levels. In *Proceedings of RANLP*, pages 90–96, Varna, Bulgaria, 2017.

Darina Benikova, Margot Mieskes, Christian M. Meyer, and Iryna Gurevych. Bridging the gap between extractive and abstractive summaries: Creation and evaluation of coherent extracts from heterogeneous sources. In *Proceedings of COLING 2016*, pages 1039–1050, 2016.

Darina Benikova, Michael Wojatzki, and Torsten Zesch. What does this imply? Examining the impact of implicitness on the perception of hate speech. In *Proceedings of GSCL*, pages 171–179, Berlin, Germany, 2017.

E. M. Bennett, R. Alpert, and A.C. Goldstein. Communications Through Limited-Response Questioning. *Public Opinion Quarterly*, 18(3):303–308, 1954.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. The Seventh PASCAL Recognizing Textual Entailment Challenge. In *Proceedings of TAC*, 2011.

Rahul Bhagat and Eduard Hovy. What Is a Paraphrase? *Computational Linguistics*, 39(3):463–472, 2013.

Chris Biemann, Kalina Bontcheva, Richard Eckart de Castilho, Iryna Gurevych, and Seid Muhie Yimam. Collaborative Web-based Tools for Multi-layer Text Annotation In: Handbook of Linguistic Annotation. In *Handbook of Linguistic Annotation*, chapter 7, pages 229–256. Springer, 2017.

Steven Bird and Mark Liberman. A Formal Framework for Linguistic Annotation. *Speech communication*, 33(1-2):23–60, 2001.

Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.

Filip Boltužić and Jan Šnajder. Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58, Baltimore, MD, USA, 2014.

Kalina Bontcheva, Ian Roberts, Leon Derczynski, and Samantha Alexander-Eames. The GATE crowdsourcing plugin: Crowdsourcing annotated corpora made easy. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 97–100, 2014.

Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.

Wauter Bosma and Chris Callison-Burch. Paraphrase Substitution for Recognizing Textual Entailment. In *Proceedings of the Workshop of CLEF*, pages 502–509, Évora, Portugal, 2006.

Gabor Bowman, Samuel R .and Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of EMNLP*, pages 632–642, Lisbon, Portugal, 2015.

Elena Cabrio and Bernardo Magnini. Decomposing Semantic Inferences. *Linguistics Issues in Language Technology - LiLT. Special Issues on the Semantics of Entailment*, 9(1), August 2013.

Jaime Guillermo Carbonell. *Subjective Understanding, Computer Models of Belief Systems*. PhD thesis, Yale University, New Haven, CT, USA, 1979.

Jean Carletta, Stefan Evert, Ulrich Heid, Jonathan Kilgour, Judy Robertson, and Holger Voormann. The NITE XML toolkit: flexible annotation for multimodal language data. *Behavior Research Methods, Instruments, & Computers*, 35(3):353–363, 2003.

Julio J. Castillo and Marina E. Cardenas. Using sentence semantic similarity based on WordNet in recognizing textual entailment. In *Ibero-American Conference on Artificial Intelligence*, pages 366–375, Bahía Blanca, Argentina, 2010.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. *Proceedings of SemEval*, pages 1–14, 2017.

Irfan Chaudhry. # Hashtagging hate: Using Twitter to track racism online. *First Monday*, 20(2), 2015.

Wei-Te Chen and Will Styler. Anafora: A Web-based General Purpose Annotation Tool. In *Proceedings of NAACL*, pages 14–19, Atlanta, GA, USA, 2013.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. In *Proceedings of PASSAT and SocialCom*, pages 71–80, Amsterdam, Netherlands, 2012.

Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

Trevor Cohn, Chris Callison-Burch, and Mirella Lapata. Constructing Corpora for the Development and Evaluation of Paraphrase Systems. *Computational Linguistics*, 34(4):597–614, 2008.

Kirsti K. Cole. It's Like She's Eager to be Verbally Abused: Twitter, Trolls, and Gendering Disciplinary Rhetoric. *Feminist Media Studies*, 15(2):356–358, 2015.

Bonaventura Coppola, Aldo Gangemi, Alfio Gliozzo, Davide Picca, and Valentina Presutti. Frame detection over the Semantic Web. In *Proceedings of ESWC*, pages 126–142, Hersonissos, Greece, 2009.

Corinna Cortes and Vladimir Vapnik. Support-Vector Networks. *Machine learning*, 20(3):273–297, 1995.

D. Alan Cruse. The Pragmatics of Lexical Specificity. *Journal of linguistics*, 13(2):153–164, 1977.

Ido Dagan and Oren Glickman. Probabilistic textual entailment: Generic applied modeling of language variability. In *PASCAL workshop on Learning Methods for Text Understanding and Mining*, pages 26–29, Grenoble, France, 2004.

Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL recognising textual entailment challenge. In *Proceedings of MLCW*, pages 177–190, Southampton, UK, 2005.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering*, 15(4):i–xvii, 2009.

Dipanjan Das. Statistical Models for Frame-Semantic Parsing. In *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore*, pages 26–29, Baltimore, MD, USA, 2014.

Dipanjan Das, Nathan Schneider, Desai Chen, and Noah A. Smith. SEMAFOR 1.0: A probabilistic frame-semantic parser. Technical report, Carnegie Mellon University, 2010. Technical Report CMU-LTI-10-001.

Herbert Aron David. *The Method of Paired Comparisons*. Charles Griffin, 1963.

Mark Davies and Joseph L Fleiss. Measuring Agreement for Multinomial Data. *Biometrics*, 38(4): 1047–1051, 1982.

Randall Davis, Howard Shrobe, and Peter Szolovits. What Is a Knowledge Representation? *AI magazine*, 14(1):17–33, 1993.

Johannes Daxenberger, Oliver Ferschke, Iryna Gurevych, and Torsten Zesch. DKPro TC: A Java-based Framework for Supervised Learning Experiments on Textual Data. In *Proceedings of ACL*, pages 61–66, Baltimore, MD, USA, 2014.

David Day, Janet Hitzeman, Michael L. Wick, Keith Crouch, and Massimo Poesio. A Corpus for Cross-Document Co-reference. In *Proceedings of LREC*, pages 2996–2999, Marrakech, Morocco, 2008.

Robert De Beaugrande and Wolfgang U. Dressler. *Introduction to Text Linguistics*. Routledge, 1981.

Luciano Del Corro and Rainer Gemulla. ClausIE: Clause-Based Open Information Extraction. In *Proceedings of WWW*, pages 355–366, Rio de Janeiro, Brazil, 2013.

Peter Dixon. The processing of organizational and component step information in written directions. *Journal of memory and language*, 26(1):24–35, 1987.

George R. Doddington, Alexis Mitchell, Mark A. Przybocki, Lance A. Ramshaw, Stephanie Strassel, and Ralph M. Weischedel. The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation. In *Proceedings of LREC*, pages 837–840, Lisbon, Portugal, 2004.

Bill Dolan, Chris Quirk, and Chris Brockett. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of COLING*, page 350–356, Geneva, Switzerland, 2004.

William B. Dolan and Chris Brockett. Automatically Constructing a Corpus of Sentential Paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing*, pages 9–16, Jeju Island, South Korea, 2005.

Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of KDD*, pages 601–610, New York, NY, USA, 2014.

Harris Drucker, Chris J. C. Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. Support Vector Regression Machines. *Advances in neural information processing systems*, 9:155–161, 1997.

Richard Eckart de Castilho and Iryna Gurevych. A broad-coverage collection of portable NLP components for building shareable analysis pipelines. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT*, pages 1–11, Dublin, Ireland, 2014.

Mürvet Enç. The Semantics of Specificity. *Linguistic inquiry*, 22(1):1–25, 1991.

Anthony Fader, Stephen Soderland, and Oren Etzioni. Identifying Relations for Open Information Extraction. In *Proceedings of EMNLP*, pages 1535–1545, Edinburgh, Scotland, UK, 2011.

Donka F. Farkas. Specificity Distinctions. *Journal of semantics*, 19(3):213–243, 2002.

Samuel Fernando and Mark Stevenson. A Semantic Similarity Approach to Paraphrase Detection. In *Proceedings of CLUK*, pages 45–52, 2008.

Jessica Ficler and Yoav Goldberg. A Neural Network for Coordination Boundary Prediction. In *Proceedings of EMNLP*, pages 23–32, Austin, TX, USA, 2016.

Charles J. Fillmore. Frame Semantics and the Nature of Language. *Annals of the New York Academy of Sciences*, 280(1):20–33, 1976.

Charles J. Fillmore, Christopher R. Johnson, and Miriam R.L. Petruck. Background to Framenet. *International journal of lexicography*, 16(3):235–250, 2003.

Peter Flach. *Machine learning: the art and science of algorithms that make sense of data*. Cambridge University Press, 2012.

Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76 (5):378–382, 1971.

William Frawley. *Linguistic Semantics*. Routledge, 2013.

Geoffrey French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. In *Proceedings of ICLR*, 2018.

Kuzman Ganchev, Joao Graça, Jennifer Gillenwater, and Ben Taskar. Posterior regularization for structured latent variable models. *The Journal of Machine Learning Research*, 11:2001–2049, 2010.

Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. Opinosis: A Graph Based Approach to Abstractive Summarization of Highly Redundant Opinions. In *Proceedings of COLING*, pages 340–348, Beijing, China, 2010.

Kiril Gashteovski, Rainer Gemulla, and Luciano Del Corro. MinIE: Minimizing Facts in Open Information Extraction. In *Proceedings of EMNLP*, pages 2630–2640, Copenhagen, Denmark, 2017.

Pepa Gencheva, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. A Context-Aware Approach for Detecting Worth-Checking Claims in Political Debates. In *Proceedings of RANLP 2017*, pages 267–276, Varna, Bulgaria, 2017.

Danilo Giampiccolo, Hoa Trang Dang, Bernardo Magnini, Ido Dagan, Elena Cabrio, and Bill Dolan. The Fourth PASCAL Recognizing Textual Entailment Challenge. In *TAC*, 2008.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 42–47, Portland, USA, 2011.

Christian Girardi, Manuela Speranza, Rachele Sprugnoli, and Sara Tonelli. CROMER: a Tool for Cross-Document Event and Entity Coreference. In *LREC*, pages 3204–3208, Reykjavík, Iceland, 2014.

Rachayita Giri, Yosha Porwal, Vaibhavi Shukla, Palak Chadha, and Rishabh Kaushal. Approaches for information retrieval in legal documents. In *Proceedings of IC3*, pages 1–6, Noida, India, 2017.

Njagi D. Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. A Lexicon-based Approach for Hate Speech Detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10 (4):215–230, 2015.

Ana-Maria Giuglea and Alessandro Moschitti. Semantic Role Labeling via FrameNet, VerbNet and PropBank. In *Proceedings of ACL*, pages 929–936, Sydney, Australia, 2006.

Amir Globerson, Gal Chechik, Fernando Pereira, and Naftali Tishby. Euclidean Embedding of Co-occurrence Data. *Journal of Machine Learning Research*, 8(76):2265–2295, 2007.

Max Glockner, Vered Shwartz, and Yoav Goldberg. Breaking NLI Systems with Sentences that Require Simple Lexical Inferences. In *Proceedings of ACL*, pages 650–655, Melbourne, Australia, 2018.

Darina Gold and Torsten Zesch. Divide and Extract – Disentangling Clause Splitting and Proposition Extraction. In *Proceedings of RANLP*, pages 399–408, Varna, Bulgaria, 2019.

Darina Gold, Marie Bexte, and Torsten Zesch. Corpus of Aspect-based Sentiment in Political Debates. In *KONVENS*, pages 89–99, 2018.

Darina Gold, Venelin Kovatchev, and Torsten Zesch. Annotating and analyzing the interactions between meaning relations. In *Proceedings of LAW*, pages 26–36, Florence, Italy, 2019.

David Graff. The AQUAINT Corpus of English News Text. Technical report, Linguistic Data Consortium, Philadelphia, PA, USA, 2002. Technical Report LDC2002T31.

Harold Guetzkow. Unitizing and categorizing problems in coding qualitative data. *Journal of Clinical Psychology*, 6(1):47–58, 1950.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of NAACL*, pages 107–112, New Orleans, Louisiana, 2018.

Viviana Haase, Maria Spychalska, and Markus Werning. Investigating the Comprehension of Negated Sentences Employing World Knowledge: An Event-Related Potential Study. *Frontiers in psychology*, 10:2184, 2019.

Patrick Hanks. Corpus Pattern Analysis. In *Proceedings of Euralex*, pages 87–97, Lorient, France, 2004.

Sanda Harabagiu and Andrew Hickl. Methods for Using Textual Entailment in Open-Domain Question Answering. In *Proceedings of ACL*, pages 905–912, 2006.

Kazi Saidul Hasan and Vincent Ng. Why are You Taking this Stance? Identifying and Classifying Reasons in Ideological Debates. In *Proceedings of the EMNLP*, pages 751–762, Doha, Qatar, 2014.

Graeme Hirst. Paraphrasing Paraphrased. In *Keynote Address for IWP: Paraphrase Acquisition and Applications*, Sapporo, Japan, 2003.

Eduard Hovy, Andrew Philpot, and Marina Rey. Events are Not Simple : Identity , Non-Identity , and Quasi-Identity. In *Proceedings of the The 1st Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 21–28, Atlanta, GA, USA, 2013.

Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, 2004.

Nancy Ide and James Pustejovsky. *Handbook of Linguistic Annotation*. Springer, 2017.

Jain, Anil and Nandakumar, Karthik and Ross, Arun. Score normalization in multimodal biometric systems. *Pattern recognition*, 38(12):2270–2285, 2005.

Nathalie Japkowicz and Mohak Shah. *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press, 2011.

Israa Jaradat, Pepa Gencheva, Alberto Barrón-Cedeño, Lluís Màrquez, and Preslav Nakov. ClaimRank: Detecting Check-Worthy Claims in Arabic and English. *arXiv preprint arXiv:1804.07587*, 2018.

Cordeiro Joao, Dias Gaël, and Brazdil Pavel. New Functions for Unsupervised Asymmetrical Paraphrase Detection. *Journal of Software*, 2(4):12–23, 2007.

David A. Jurgens. Embracing ambiguity: A comparison of annotation methodologies for crowdsourcing word sense labels. In *Proceedings of NAACL*, pages 556–562, Atlanta, GA, USA, 2013.

David A. Jurgens, Peter D. Turney, Saif M. Mohammad, and Keith J. Holyoak. SemEval-2012 Task 2: Measuring degrees of relational similarity. In *Proceedings of *SEM/CoNLL*, pages 356–364, Montréal, QC, Canada, 2012.

Tushar Khot, Ashish Sabharwal, and Peter Clark. Answering Complex Questions Using Open Information Extraction. In *Proceedings of ACL*, volume 2, pages 311–316, Vancouver, BC, Canada, 2017.

Paul Kingsbury and Martha Palmer. PropBank: the Next Level of TreeBank. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories*, pages 105–116, Vaxjo, Sweden, 2003.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. Extending VerbNet with novel verb classes. In *Proceedings of LREC*, pages 1027–1032, Genoa, Italy, 2006.

Svetlana Kiritchenko and Saif M. Mohammad. Capturing Reliable Fine-Grained Sentiment Associations by Crowdsourcing and Best–Worst Scaling. In *Proceedings of NAACL*, pages 811–817, San Diego, CA, USA, 2016.

Svetlana Kiritchenko and Saif M. Mohammad. Best-Worst Scaling More Reliable than Rating Scales: A Case Study on Sentiment Intensity Annotation. In *Proceedings of ACL*, pages 465–470, Vancouver, BC, Canada, 2017.

Fabian Kneißl. *Crowdsourcing for linguistic field research and e-learning*. PhD thesis, Ludwig-Maximilians-Universität München, 2014.

Wei-Jen Ko, Greg Durrett, and Junyi Jessy Li. Domain Agnostic Real-Valued Specificity Prediction. In *Proceedings of AAAI*, volume 33, pages 6610–6617, Honolulu, HI, USA, 2019a.

Wei-Jen Ko, Greg Durrett, and Junyi Jessy Li. Linguistically-Informed Specificity and Semantic Plausibility for Dialogue Generation. In *Proceedings of NAACL*, pages 3456–3466, Minneapolis, MN, USA, 2019b.

Venelin Kovatchev, M. Antònia Martí, and Maria Salamó. ETPC - A Paraphrase Identification Corpus Annotated with Extended Paraphrase Typology and Negation. In *Proceedings of LREC*, pages 1384–1392, Miyazaki, Japan, 2018a.

Venelin Kovatchev, M. Antònia Martí, and Maria Salamó. WARP-Text: a Web-Based Tool for Annotating Relationships between Pairs of Texts. In *Proceedings of COLING*, pages 132–136, Santa Fe, New Mexico, 2018b.

Venelin Kovatchev, Darina Gold, and Torsten Zesch. RELATIONS-Workshop on meaning relations between phrases and sentences. In *RELATIONS-Workshop on meaning relations between phrases and sentences*, 2019a.

Venelin Kovatchev, M. Antònia Martí, Maria Salamó, and Javier Beltrán. Qualitative Evaluation of Paraphrase Identification Systems. In *Proceedings of RANLP*, pages 568–577, Varna, Bulgaria, 2019b.

Venelin Kovatchev, Darina Gold, M. Antònia Martí, Maria Salamó, and Torsten Zesch. Decomposing and Comparing Meaning Relations: Paraphrasing, Textual Entailment, Contradiction, and Specificity. In *Proceedings of LREC*, pages 5782–5791, 2020.

Manfred Krifka. Nominal Reference , Temporal Constitution and Quantification in Event Semantics. *Semantics and Contextual Expression*, pages 75–115, 1989.

Klaus Krippendorff. *Content analysis: An introduction to Its methodology*. Sage, 1 edition, 1980.

Klaus Krippendorff. On the Reliability of Unitizing Continuous Data. *Sociological Methodology*, 25: 47–76, 1995.

Klaus Krippendorff. *Content analysis: An introduction to Its methodology*. Sage, 4 edition, 2018.

Karla A Lassonde and Edward J O'Brien. Contextual Specificity in the Activation of Predictive Inferences. *Discourse Processes*, 46(5):426–438, 2009.

Rémi Lebret and Ronan Collobert. Word emdeddings through hellinger PCA. *arXiv preprint arXiv:1312.5542*, 2013.

Adrian Leemann, Marie-José Kolly, Ross Purves, David Britain, and Elvira Glaser. Crowdsourcing Language Change with Smartphone Applications. *PLoS ONE*, 11(1):1–25, 2016.

Beth Levin. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, 1993.

Omer Levy and Yoav Goldberg. Linguistic Regularities in Sparse and Explicit Word Representations. In *Proceedings of CoNLL*, pages 171–180, Baltimore, MD, USA, 2014a.

Omer Levy and Yoav Goldberg. Neural Word Embedding as Implicit Matrix Factorization. In *Proccedings of NIPS*, pages 2177–2185, Montréal, QC, Canada, 2014b.

Hao Li and Heng Ji. Cross-genre Event Extraction with Knowledge Enrichment. In *Proceedings of NAACL*, pages 1158–1162, San Diego, CA, USA, 2016.

Hao Li, Xiang Li, Heng Ji, and Yuval Marton. Domain-Independent Novel Event Discovery and Semi-Automatic Event Annotation. In *Proceedings of PACLIC*, pages 233–242, Sendai, Japan, 2010.

Junyi J. Li, Bridget O'Daniel, Yi Wu, Wenli Zhao, and Ani Nenkova. Improving the Annotation of Sentence Specificity. In *LREC*, pages 3921–3927, Portorož, Slovenia, 2016.

Junyi Jessy Li and Ani Nenkova. Fast and Accurate Prediction of Sentence Specificity. In *Proceedings of AAAI*, page 2281–2287, Austin, TX, USA, 2015.

Rensis Likert. A Technique for the Measurement of Attitudes. *Archives of Psychology*, 22(140):5–55, 1932.

Dekang Lin. An Information-Theoretic Definition of Similarity. In *Proceedings of ICML*, pages 296–304, 1998.

Bing Liu. Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167, 2012.

Elena Lloret, Oscar Ferrández, Rafael Munoz, and Manuel Palomar. A Text Summarization Approach under the Influence of Textual Entailment. In *In Proceedings of NLPCS*, pages 22–31, Barcelona, Spain, 2008.

Anna Lobanova, Tom Van der Kleij, and Jennifer Spenader. Defining Antonymy: A Corpus-based Study of Opposites by Lexico-syntactic Patterns. *International Journal of Lexicography*, 23(1): 19–53, 2010.

Peter LoBue and Alexander Yates. Types of Common-sense Knowledge Needed for Recognizing Textual Entailment. In *Proceedings of ACL*, pages 329–334, Stroudsburg, PA, USA, 2011.

Iñigo Lopez-Gazpio, Montse Maritxalar, Aitor Gonzalez-Agirre, German Rigau, Larraitz Uria, and Eneko Agirre. Interpretable Semantic Textual Similarity: Finding and explaining differences between sentences. *Knowledge-Based Systems*, 119:186–199, 2017.

Alexander Löser, Sebastian Arnold, and Tillmann Fiehn. The GoOlap Fact Retrieval Framework. In *European Business Intelligence Summer School*, pages 84–97, Paris, France, 2011.

Annie Louis and Ani Nenkova. Automatic identification of general and specific sentences by leveraging discourse annotations. In *Proceedings of IJCNLP*, pages 605–613, Chiang Mai, Thailand, 2011.

Annie Louis and Ani Nenkova. A corpus of general and specific sentences from news. In *Proceedings of LREC*, pages 1818–1821, Instanbul, Turkey, 2012.

Jordan J. Louviere. Best-worst scaling: A model for the largest difference judgments, 1991.

Jordan J. Louviere. The best-worst or maximum difference measurement model: Applications to behavioral research in marketing. In *The American Marketing Association's Behavioral Research Conference*, Phoenix, Arizona, 1993.

Jordan J. Louviere and Gary J. Gaeth. Decomposing the determinants of retail facility choice using the method of hierarchical information integration-a supermarket illustration. *Journal of Retailing*, 63(1):25–48, 1987.

Jordan J. Louviere, Terry N. Flynn, and A. A. J. Marley. *Best-Worst Scaling: Theory, Methods and Applications*. Cambridge University Press, 2015.

Luca Lugini and Diane Litman. Predicting Specificity in Classroom Discussion. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–61, 2017.

Nitin Madnani and Bonnie J. Dorr. Generating Phrasal and Sentential Paraphrases: A Survey of Data-Driven Methods. *Computational Linguistics*, 36(3):341–387, 2010.

Liliana Mamani Sanchez and Carl Vogel. IMHO: An Exploratory Study of Hedging in Web Forums. In *Proceedings of SIGDIAL*, pages 309–313, Metz, France, 2013.

Karla Mantilla. Gendertrolling: Misogyny adapts to new media. *Feminist Studies*, 39(2):563–570, 2013.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of LREC*, pages 216–223, Reykjavík, Iceland, 2014.

Thomas Mathew and Graham Katz. Supervised Categorization for Habitual versus Episodic Sentences. *Sixth Midwest Computational Linguistics Colloquium*, pages 2–3, 2009.

Mausam, Michael Schmitz, Robert Bart, Stephen Soderland, and Oren Etzioni. Open Language Learning for Information Extraction. In *Proceedings of EMNLP-CoNLL*, pages 523–534, Jeju Island, Korea, 2012.

Diana Maynard and Adam Funk. Automatic Detection of Political Opinions in Tweets. In *Proceedings of ESWC*, pages 88–99, Heraklion, Greece, 2011.

Philipp Mayring. Qualitative Inhaltsanalyse. In *Handbuch qualitative Forschung in der Psychologie*, pages 601–613. Springer, 2010.

Tarlach McGonagle. The Council of Europe against online hate speech: Conundrums and challenges. *Expert paper, doc. no*, 1900(2013):005, 2013.

Gabor Melli, Zhongmin Shi, Yang Wang, Yudong Liu, Anoop Sarkar, and Fred Popowich. Description of SQUASH, the SFU Question Answering Summary Handler for the DUC-2006 Summarization Task. In *Proceedings of DUC*, 2006.

Christian M. Meyer, Darina Benikova, Margot Mieskes, and Iryna Gurevych. MDSWriter: Annotation tool for creating high-quality multi-document summarization corpora. *Proceedings of ACL*, pages 97–102, 2016.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of NIPS*, pages 3111–3119, Lake Tahoe, NV, USA, 2013a.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of NAACL*, pages 746–751, Atlanta, GA, USA, 2013b.

George A. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11): 39–41, 1995.

Ashutosh Modi, Ivan Titov, and Alexandre Klementiev. Unsupervised Induction of Frame-Semantic Representations. In *Proceedings of the NAACL-HLT Workshop on the Induction of Linguistic Structure*, pages 1–7, Montréal, QC, Canada, 2012.

Saif M. Mohammad and Felipe Bravo-Marquez. Emotion Intensities in Tweets. *arXiv preprint arXiv:1708.03696*, 2017.

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. SemEval-2016 Task 6: Detecting Stance in Tweets. In *Proceedings of SemEval*, pages 31–41, San Diego, USA, 2016.

Mainack Mondal, Leandro Araújo Silva, and Fabrício Benevenuto. A Measurement Study of Hate Speech in Social Media. In *Proceedings of ACMHT*, pages 85–94, New York, NY, USA, 2017.

Thomas S. Morton and Jeremy LaCivita. WordFreak: An Open Tool for Linguistic Annotation. In *Proceedings of NAACL*, pages 17–18, Edmonton, AB, Canada, 2003.

Éva Mújdricza-Maydt, Silvana Hartmann, Iryna Gurevych, and Anette Frank. Combining Semantic Annotation of Word Sense & Semantic Roles: A Novel Annotation Scheme for VerbNet Roles on German Language Data. In *Proceedings of LREC*, pages 3031–3038, Portorož, Slovenia, 2016.

Andreas Müller. *Introduction to Machine Learning with Python*. O'Reilly Media, Inc., 1st edition edition, 2016.

Christoph Müller and Michael Strube. Multi-level annotation of linguistic data with MMAX2. *Corpus technology and language pedagogy: New resources, new tools, new methods*, 3:197–214, 2006.

Robert Munro, Steven Bethard, Victor Kuperman, Vicky T. Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen, and Harry Tily. Crowdsourcing and language studies: the new generation of linguistic data. In *Proceedings of NAACL Workshop on Creating Speech and Language Data With Amazon's Mechanical Turk*, pages 122–130, 2010.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. Stress Test Evaluation for Natural Language Inference. In *Proceedings of COLING*, pages 2340–2353, Santa Fe, NM, USA, 2018.

Preslav Nakov, Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Wajdi Zaghouani, Pepa Atanasova, Spas Kyuchukov, and Giovanni Da San Martino. Overview of the CLEF-2018 CheckThat! Lab on Automatic Identification and Verification of Political Claims. In *Proceedings of the Ninth International Conference of the CLEF Association: Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Lecture Notes in Computer Science, pages 372–387, Avignon, France, 2018.

Arvind R. Neelakantan. *Knowledge Representation and Reasoning with Deep Neural Networks*. PhD thesis, University of Massachusetts Amherst, 2017.

Finn Å. Nielsen. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of CEUR Workshop*, number 718, pages 93–98, 2011.

Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. A Survey on Open Information Extraction. In *Proceedings of COLING*, pages 3866–3878, Santa Fe, NM, USA, 2018.

Dennis Njagi, Zhang Zuping, Damien Hanyurwimfura, and Jun Long. A Lexicon-based Approach for Hate Speech Detection. In *International Journal of Multimedia and Ubiquitous Engineering*, volume 10, pages 215–230, 2015.

Mick O'Donnell. Demonstration of the UAM CorpusTool for text and image annotation. In *Proceedings of ACL*, pages 13–16, Columbus, OH, USA, 2008.

Bryan Orme. Maxdiff analysis: Simple counting, individual-level logit, and HB. *Sawtooth Software*, 2009.

Sebastian Padó, Michel Galley, Dan Jurafsky, and Christopher D. Manning. Textual Entailment Features for Machine Translation Evaluation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 37–41, 2009.

Alexis Palmer and Caroline Sporleder. Evaluating FrameNet-style semantic parsing: the role of coverage gaps in FrameNet. In *Proceedings of COLING*, pages 928–936, Uppsala, Sweden, 2010.

Alexander Panchenko. Best of Both Worlds: Making Word Sense Embeddings Interpretable. In *Proceedings of LREC*, pages 2649–2655, Portorož, Slovenia, 2016.

Bo Pang and Lillian Lee. Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135, 2008.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA, USA, 2002.

Ayush Patwari, Dan Goldwasser, and Saurabh Bagchi. TATHYA: A multi-classifier system for detecting check-worthy statements in political debates. In *Proceedings of CIKM*, pages 2259–2262, Singapore, 2017.

Fabio Petroni, Luciano Del Corro, and Rainer Gemulla. CORE: Context-Aware Open Relation Extraction with Factorization Machines. In *Proceedings of EMNLP*, pages 1763–1773, Lisbon, Portugal, 2015.

John C. Platt, Nello Cristianini, and John Shawe-Taylor. Large Margin DAGs for Multiclass Classification. In *MIT Press*, volume 12, pages 547–553, 2000.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. Semeval-2014 Task 4: Aspect based sentiment analysis. In *Proceedings of SemEval 2014*, pages 27–35, Dublin, Ireland, 2014.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. SemEval-2015 Task 12: Aspect Based Sentiment Analysis. In *Proceedings of SemEval 2015*, pages 486–495, Denver, CO, USA, 2015.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphee De Clercq, Veronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Núria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In *Proceedings of SemEval 2016*, pages 19–30, San Diego, CA, USA, 2016.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K. Joshi, and Bonnie L. Webber. The Penn Discourse TreeBank 2.0. In *In Proceedings of LREC*, pages 2961–2968, Portorož, Slovenia, 2016.

James Pustejovsky. The syntax of event structure. *Cognition*, 41(1-3):47–81, 1991.

Randolph Quirk. *A Grammar of Contemporary English*. Longman, London, 11. impression edition, 1985.

Jaroslav Ramík. Pairwise Comparison Matrices in Decision-Making. In *Pairwise Comparisons Method*, pages 17–65. Springer, 2020.

Amir H. Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. Offensive Language Detection Using Multi-level Classification. In *Proceedings of the 23rd Canadian Conference on Advances in Artificial Intelligence*, pages 16–27, Ottawa, ON, Canada, 2010.

Nils Reiter and Anette Frank. Identifying Generic Noun Phrases. In *Proceedings of ACL*, pages 40–49, Uppsala, Sweden, 2010.

Philip Resnik. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proccedings of IJCAI*, pages 448–453, Montréal, QC, Canada, 1995.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M Marlin. Relation Extraction with Matrix Factorization and Universal Schemas. In *Proceedings of NAACL*, pages 74–84, Atlanta, GA, USA, 2013.

Martin Riedl. *Unsupervised Methods for Learning and Using Semantics of Natural Language*. PhD thesis, TU Darmstadt, 2016.

Alan Ritter, Oren Etzioni, and Sam Clark. Open Domain Event Extraction from Twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1104–1112, Beijing, China, 2012.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, pages 6–9, 2016.

Michael Roth and Anette Frank. Aligning Predicate Argument Structures in Monolingual Comparable Texts: A New Corpus for a New Task. In *Proceedings of *SEM/CoNLL*, pages 218–227, Montréal, QC, Canada, 2012.

Eugen Ruppert, Manuel Kaufmann, Martin Riedl, and Chris Biemann. JOBIMVIZ: A Web-based Visualization for Graph-based Distributional Semantic Models. In *Proceedings of ACL*, pages 103–108, 2015.

Vasile Rus, Rajendra Banjade, and Mihai C Lintean. On Paraphrase Identification Corpora. In *Proceedings of LREC*, pages 2422–2429, Reykjavík, Iceland, 2014.

Swarnadeep Saha and Mausam. Open Information Extraction from Conjunctive Sentences. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2288–2299, 2018.

Mark Sammons, V.G. Vinod Vydiswaran, and Dan Roth. "Ask not what Textual Entailment can do for You...". In *Proceedings of ACL*, pages 1199–1208, 2010.

Evan Sandhaus. The New York Times Annotated Corpus. *Linguistic Data Consortium, Philadelphia*, 6(12), 2008.

Rudolf Schneider, Tom Oberhauser, Tobias Klatt, Felix A. Gers, and Alexander Löser. Analysing Errors of Open Information Extraction Systems. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 11–18, Copenhagen, Denmark, 2017.

Karin Kipper Schuler. *VerbNet: A broad-coverage, comprehensive verb lexicon*. PhD thesis, University of Pennsylavania, 2005.

Howard Schuman and Stanley Presser. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context (Quantitative Studies in Social Relation)*. Sage, 1996.

William A. Scott. Reliability of Content Analysis: The Case of Nominal Scale Coding. *Public Opinion Quarterly*, 19(3):321, 1955.

Vered Shwartz, Gabriel Stanovsky, and Ido Dagan. Acquiring Predicate Paraphrases from News Tweets. In *Proceedings of *SEM*, pages 155–160, Vancouver, BC, Canada, 2017.

Frank Smadja, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. Translating collocations for bilingual lexicons: A statistical approach. *Computational linguistics*, 22(1):1–38, 1996.

Parinaz Sobhani, Diana Inkpen, and Stan Matwin. From Argumentation Mining to Stance Classification. In *Proceedings of NAACL*, pages 67–77, Denver, CO, USA, 2015.

Richard Socher, Eric H. Huang, Jeffrey Pennin, Christopher D. Manning, and Andrew Y. Ng. Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection. In *Proceedings of NIPS*, pages 801–809, Granada, Spain, 2011.

Sara Owsley Sood, Judd Antin, and Elizabeth F. Churchill. (using crowdsourcing to improve profanity detection.). In *AAAI Spring Symposium: Wisdom of the Crowd*, volume 12, pages 69–74, Palo Alto, CA, USA, 2012.

Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.

Ellen Spertus. Smokey: Automatic Recognition of Hostile Messages. In *Proceedings of AAAI/IAAI*, pages 1058–1065, Providence, RI, USA, 1997.

Gabriel Stanovsky and Ido Dagan. Creating a Large Benchmark for Ipen Information Extraction. In *Proceedings of EMNLP*, pages 2300–2305, 2016.

Gabriel Stanovsky, Ido Dagan, and Mausam. Open IE as an intermediate structure for semantic tasks. In *Proceedings of ACL*, volume 2, pages 303–308, Beijing, China, 2015.

Gabriel Stanovsky, Jessica Ficler, Ido Dagan, and Yoav Goldberg. Getting more out of syntax with PROPS. *arXiv preprint arXiv:1603.01648*, 2016.

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. Supervised Open Information Extraction. In *Proceedings of NAACL*, pages 885–895, New Orleans, LA, USA, 2018.

Anatol Stefanowitsch. Was ist überhaupt Hate-Speech? *Geh sterben. Umgang mit Hate-Speech und Kommentaren im Internet*, pages 11–13, 2014.

Philip J. Stone, Dexter C. Dunphy, and Marshall S. Smith. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press, 1966.

Maria Sukhareva, Judith Eckle-Kohler, Ivan Habernal, and Iryna Gurevych. Crowdsourcing a Large Dataset of Domain-Specific Context-Sensitive Semantic Verb Relations. In *Proceedings of LREC*, pages 2131–.2137, Portorož, Slovenia, 2016.

Reid Swanson, Brian Ecker, and Marilyn Walker. Argument Mining: Extracting Arguments from Online Dialogue. In *Proceedings of SIGDIAL*, pages 217–226, Prague, Czech Republic, 2015.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proceedings of NIPS*, pages 1195–1204, 2017.

Louis L Thurstone. A law of comparative judgment. *Psychological review*, 34(4):273–286, 1927.

Ye Tian and Richard Breheny. Dynamic Pragmatic View of Negation Processing. In *Negation and polarity: Experimental perspectives*, pages 21–43. Springer, 2016.

I-Hsien Ting, Hsing-Miao Chi, Jyun-Sing Wu, and Shyue-Liang Wang. An Approach for Hate Groups Detection in Facebook. In *The 3rd International Workshop on Intelligent Data Analysis and Management*, pages 101–106, 2013.

Howard E. Tinsley and David J. Weiss. Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology*, 22(4):358–376, 1975.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of NAACL*, pages 173–180, Sapporo, Japan, 2003.

Vladimir N. Vapnik. *Statistical learning theory*. Wiley, 1998.

Vladimir N. Vapnik and Alexei Y. Chervonenkis. Necessary and sufficient conditions for the uniform convergence of means to their expectations. *Theory of Probability & Its Applications*, 26(3):532–553, 1981.

Paola Velardi, Maria T. Pazienza, and Michela Fasolo. How to Encode Semantic Knowledge: A Method for Meaning Representation and Computer-Aided Acquisition. *Computational Linguistics*, 17(2):153–170, 1991.

Marta Vila, M. Antònia Martí, and Horacio Rodríguez. Is This a Paraphrase? What Kind? Paraphrase Boundaries and Typology. *Open Journal of Modern Linguistics*, 4(01):205–2018, 2014.

Marilyn A. Walker, Pranav Anand, Rob Abbott, Jean E. Fox Tree, Craig Martell, and Joseph King. That is your evidence?: Classifying stance in online political debate. *Decision Support Systems*, 53 (4):719–729, 2012.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal Adversarial Triggers for Attacking and Analyzing NLP. In *Proceedings of EMNLP-IJCNLP*, pages 2153–2162, Hong Kong, China, 2019.

William Warner and Julia Hirschberg. Detecting Hate Speech on the World Wide Web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Stroudsburg, PA, USA, 2012.

Zeerak Waseem and Dirk Hovy. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California, 2016.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, 2017.

Jason Weston and Chris Watkins. Multi-class support vector machines. Technical report, Royal Holloway, University of London, 1998. Technical Report CSD-TR-98-04.

Aaron S. White, Drew Reisinger, Keisuke Sakaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. Universal Decompositional Semantics on Universal Dependencies. In *Proceedings of EMNLP*, pages 1713–1723, Austin, TX, USA, 2016.

Lyndon White, Roberto Togneri, Wei Liu, and Mohammed Bennamoun. How Well Sentence Embeddings Capture Meaning. In *Proceedings of the 20th Australasian Document Computing Symposium*, page 9, Parramatta, Australia, 2015.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating Expressions of Opinions and Emotions in Language. *Language resources and evaluation*, 39(2-3):165–210, 2005.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Proceedings of HLT/EMNLP*, page 347–354, Vancouver, BC, Canada, 2005.

Michael Wojatzki and Torsten Zesch. Stance-based Argument Mining - Modeling Implicit Argumentation Using Stance. In *Proceedings of the KONVENS*, pages 313–322, Bochum, Germany, 2016.

Michael Wojatzki, Tobias Horsmann, Darina Gold, and Torsten Zesch. Do Women Perceive Hate Differently: Examining the Relationship Between Hate Speech, Gender, and Agreement Judgments. In *Proceedings of KONVENS*, pages 110–120, Vienna, Austria, 2018a.

Michael Wojatzki, Saif M. Mohammad, Torsten Zesch, and Svetlana Kiritchenko. Quantifying Qualitative Data for Understanding Controversial Issues. In *Proceedings of LREC*, pages 1405 – 1418, Miyazaki, Japan, 2018b.

Zhibiao Wu and Martha Palmer. Verb semantics and lexical selection. In *Proceedings of ACL*, pages 133–138, Las Cruces, NM, USA, 1994.

Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. Learning from Bullying Traces in Social Media. In *Proceedings of NAACL HLT*, pages 656–666, Stroudsburg, PA, USA, 2012.

Ruifeng Xu, Yu Zhou, Dongyin Wu, Lin Gui, Jiachen Du, and Yun Xue. Overview of NLPCC Shared Task 4: Stance Detection in Chinese Microblogs. In *Proceedings of ICCPOL*, pages 907–916, Kunming, China, 2016.

Wei Xu, Alan Ritter, Chris Callison-Burch, William B. Dolan, and Yangfeng Ji. Extracting lexically divergent paraphrases from Twitter. In *Transactions of the Association for Computational Linguistics*, volume 2, pages 435–448, Baltimore, MD, USA, 2014.

Ronald R. Yager. Default knowledge and measures of specificity. *Information Science*, 61(1-2):1–44, 1992.

Yinfei Yang and Ani Nenkova. Detecting Information-Dense Texts in Multiple News Domains. In *Proceedings of AAAI*, pages 1650–1656, Québec City, QC, Canada, 2014.

Seid M. Yimam, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations. In *Proceedings of ACL*, pages 1–6, Baltimore, MA, USA, 2013.

Ken-ichi Yokote, Shohei Tanaka, and Mitsuru Ishizuka. Effects of Using Simple Semantic Similarity on Textual Entailment Recognition. In *TAC*, Gaithersburg, MD, USA, 2011.

Lotfi A. Zadeh. PRUF—a meaning representation language for natural languages. *International Journal of Man-Machine Studies*, 10(4):395–460, 1978.

Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, Jun Xu, and Xueqi Cheng. Learning to Control the Specificity in Neural Response Generation. In *Proceedings of ACL*, pages 1108–1117, Melbourne, Australia, 2018.

DOI: 10.17185/duepublico/74633
URN: urn:nbn:de:hbz:464-20210805-151946-9