An analysis of textual inference in German customer emails

Kathrin Eichler*, Aleksandra Gabryszak*, Günter Neumann†

*German Research Center for Artificial Intelligence (DFKI), Berlin
(kathrin.eichler|aleksandra.gabryszak@dfki.de)

†German Research Center for Artificial Intelligence (DFKI), Saarbrücken
(neumann@dfki.de)

Abstract

Human language allows us to express the same meaning in various ways. Recognizing that the meaning of one text can be inferred from the meaning of another can be of help in many natural language processing applications. One such application is the categorization of emails. In this paper, we describe the analysis of a real-world dataset of manually categorized customer emails written in the German language. We investigate the nature of textual inference in this data, laying the ground for developing an inference-based email categorization system. This is the first analysis of this kind on German data. We compare our results to previous analyses on English data and present major differences.

1 Introduction

A typical situation in customer support is that many customers send requests describing the same issue. Recognizing that two different customer emails refer to the same problem can help save resources, but can turn out to be a difficult task. Customer requests are usually written in the form of unstructured natural language text, i.e., when automatically processing them, we are faced with the issue of variability: Different speakers of a language express the same meanings using different linguistic forms. There are, in fact, cases where two sentences expressing the same meaning do not share a single word:

1. "Bild und Ton sind asynchron." [*Picture and sound are asynchronous.*]

2. "Die Tonspur stimmt nicht mit dem Film überein." [*The audio track does not match the video.*]

Detecting the semantic equivalence of sentences 1 and 2 requires several textual inference steps: At the lexical level, it requires mapping the word *picture* to *video* and *sound* to *audio track*. At the level of compositional semantics, it requires detecting the equivalence of the expressions *A and B are asynchronous* and *A does not match B*.

In this paper, we describe our analysis of a large set of manually categorized customer emails, laying the ground for developing an email categorization system based on textual inference. In our analysis, we compared each email text to the description of its associated category in order to investigate the nature of the inference steps involved. In particular, our analysis aims to give answers to the following questions: What text representation is appropriate for the email categorization task? What kind of inference steps are involved and how are they distributed in real-world data? Answering these questions will not only help us decide, which existing tools and resources to integrate in an inference-based email categorization system, but also, which non-existing tools may be needed in addition.

2 Related Work

The task of email categorization has been addressed by numerous people in the last decade. In the customer support domain, work to be mentioned includes Eichler (2005), Wicke (2010), and Eichler et al. (2012).

Approaching the task using textual inference relates to two tasks, for which active research is going on: Semantic Textual Similarity, which measures the degree of semantic equivalence (Agirre et al., 2012) of two texts, and Recognizing Textual Entailment (RTE), which is defined as recognizing, given a hypothesis H and a text T, whether the meaning of H can be inferred from (is *entailed* in) T (Dagan et al., 2005). The task of email categorization can be viewed as an RTE task, where T

refers to the email text and H refers to the category description. The goal then is to find out if the email text entails the category description, and if so, assign it to the respective category.

In connection with RTE, several groups have analyzed existing datasets in order to investigate the nature of textual inference. Bar-Haim (2010) introduces two levels of entailment, lexical and lexical-syntactic, and analyzes the contribution of each level and of individual inference mechanisms within each level over a sample from the first RTE Challenge test set (Dagan et al., 2005). He concludes that the main contributors are paraphrases and syntactic transformations.

Volokh and Neumann (2011) analyzed a subset of the RTE-7 (Bentivogli et al., 2011) development data to measure the complexity of the task. They divide the T/H pairs into three different classes, depending on the type of knowledge required to solve the problem: In class A, the relevant information is expressed with the same words in both T and H. In class B, the words used in T are synonyms to those used in H. In class C, recognizing entailment between H and T requires the use of logical inference and/or world knowledge. They conclude that for two thirds of the data a good word-level analysis is enough, whereas the remainder of the data contains diverse phenomena calling for a more sophisticated approach.

A detailed analysis of the linguistic phenomena involved in semantic inferences in the T-H pairs of the RTE-5 dataset was presented by (Cabrio and Magnini, 2013).

As the approaches described above, our analysis aims at measuring the contribution of inference mechanisms at different representation levels. However, we focus on a different type of text (customer request as compared to news) and a different language (German as compared to English). We thus expect our results to differ from the ones obtained in previous work.

3 Setup

3.1 Dataset

We analyzed a dataset consisting of a set of emails and a set of categories associated to these emails. The emails contain customer requests sent to the support center of a multimedia software company, and mainly concern the products offered by this company. Each email was manually assigned to one or more matching categories by a customer support agent (a domain expert). These categories, predefined by the data provider, represent previously identified problems reported by customers. All emails and category descriptions are written in German. As is common for this type of data, many emails contain spelling mistakes, grammatical errors or abbreviations, which make automatic text processing difficult. An anonymized[1] version of the dataset is available online[2]. Our data analysis was done on the original dataset. The data examples we use in the following, however, are taken from the anonymized dataset.

In our analysis, we manually compared the email texts to the descriptions of their associated categories in order to investigate the nature of the inference steps involved. In order to reduce the complexity of the task, we based our analysis on the subset of categories, for which the category text described a single problem (a single H, speaking in RTE terms). We also removed emails for which we were not able to relate the category description to the email text. However, we kept emails associated to several categories and analyzed all of the assignments. The reduced dataset we used for our analysis consists of 369 emails associated to 25 categories. The email lengths vary between 2 and 1246 tokens. Category descriptions usually consist of a single sentence or a phrase.

3.2 Task definition

The task of automatically assigning emails to matching categories can be viewed as an RTE task, where T refers to the email text and H refers to the category description. The goal then is to find out if the email text entails the category description, and if so, assign it to the respective category.

For the analysis of inference steps involved, we distinguish between two levels of inference: lexical semantics and compositional semantics. At the lexical level, we distinguish two different types of text representation: First, the bag-of-tokens representation, where both the email text and the category description are represented as the set of content word tokens contained in the respective text.

[1] The anonymization step was performed to eliminate references to the data provider and anonymize personal data about the customers. During this step, the data was transferred into a different product domain (online auction sales). However, the anonymized version is very similar to the original one in terms of language style (including spelling errors, anglicisms, abbreviations, and special characters).

[2] `http://www.excitement-project.eu/attachments/article/97/omq_public_email_data.zip`

Second, the bag-of-terms representation, where a "term" can consist of one or more content tokens occurring consecutively. At this level, following Bar Haim (2010), we assume that entailment holds between T (the email) and H (the category description) if every token (term) in H can be matched by a corresponding entailing token (term) in T.

At the level of compositional semantics, we represent each text as the set of complex expressions (combinations of terms linked syntactically and semantically) contained in it. At this level, we assume that entailment holds between T and H if every term in H is part of at least one complex expression that can be matched by a corresponding entailing expression in T.

The data analysis was carried out by two people separately (one of them an author of this paper), who analyzed each assignment of an email E to a category C based on predefined analysis guidelines. For each of the text representation types described above, the task of the annotators was to find, for each expression in the description of C, a semantically equivalent or entailing expression in E.[3] If such an expression was found, all involved inference steps were to be noted down in an annotation table. The predefined list of possible inference steps is explained in detail in the following.

4 Inference steps

4.1 Lexical semantics level

For each of the three different types of representation (token, term, complex expression), we distinguish various inference steps. At the lexical level, we distinguish among spelling, inflection, derivation, composition, lexical semantics at the token level and lexical semantics at the term level. This distinction was made based on the assumption that for each of these steps a different NLP tool or resource is required (e.g., a lemmatizer for inflection, a compound splitter for composition, a lexical-semantic net for lexical semantics). We also distinguish between token and term level lexical semantics, as, for term-level lexical semantics, we assume that a tool for detecting multi-token terms would be required.

[3] A preanalysis of the data revealed that in some cases, the entailment direction seemed to be flipped: Expressions in the category description entailed expressions in the email text, e.g. "Video" (*video*) → "Film" (*film*). In our analysis, we counted these as positive cases if the context suggested that both expressions were used to express the same idea. We consider this an interesting issue to be further investigated.

4.2 Compositional semantics level

At the level of compositional semantics, we consider inference steps involving complex expressions.[4] These steps go beyond the lexical level and would require the usage of at least a syntactic parser for detecting word dependencies and a tool for recognizing entailment between two complex expressions. At this level, we also record the frequency of three particular phenomena: particle verbs, negation, and light verb constructions, which we considered worth addressing separately.

Particle verbs are important when processing German because, unlike in English, they can occur both as one token or two, dependending on the syntactic construction, in which they are embedded (e.g., "aufnehmen" and "nehme [...] auf" [*(to) record*]. Recognizing the scope of negation can be required in cases where negation is expressed implicitly in one of the sentences, e.g., "A und B sind nicht synchron" [*A and B are not synchronous*] vs. "Es kommt zu Versetzung zwischen A und B" [*There is a misaligment between A and B*]. By *light verbs* we refer to verbs with little semantic content of their own, forming a linguistic unit with a noun or prepositional phrase, for which a single verb with a similar meaning exists, e.g., "Meldung kommt" [*message appears*] vs. "melden" [*notify*].

For example, for the text pair "Das Brennen bricht ab mit der Meldung X" [*Burning breaks with message X*] and "Beim Brennen kommt die Fehlermeldung X" [*When burning, error message X appears*], the word "Meldung" [*message*] was recorded as inference at the token level because it can be derived from "Fehlermeldung" [*error message*] using decomposition. The verb "bricht ab" [*break*] was considered inference at the level of compositional semantics because there is no lexical-semantic relation to the verb "kommt" [*appears*]. The verb can thus only be matched by considering the complete expression.

4.3 Possible effects on precision

The focus of the analysis described so far was on ways to improve recall in an email categorization system: We count the inference steps required to increase the amount of mappable information (similar to query expansion in information retrieval). However, the figures do not show the impact of these mappings on precision, i.e.,

[4] Additional lexical inference steps required at this level are not recorded.

whether an inference step we take would negatively affect the precision of the system. Taking a more precision-oriented view at the problem, we also counted the number of cases for which a more complex representation could be "helpful" (albeit not necessary). For example, inferring the negated expression "Programm kann die DVD nicht abspielen" [*Program cannot play the DVD*] from "Programm kann die DVD nicht laden"' [*Program does not load the DVD*] is possible at the lexical level, assuming that "abspielen" [*(to) play*] entails "laden" [*(to) load*]. However, knowing that both verbal expressions are negated is expected to be beneficial to precision, in order to avoid wrongly inferring a negated from a non-negated expression.

5 Results

5.1 Interannotator agreement

Our analysis was done by two people separately, which allowed us to measure the reliability of the annotation for the different inference steps. The kappa coefficient (Cohen, 1960) for spelling, inflection, derivation and composition ranged between 0.46 and 0.67, i.e., moderate to substantial agreement according to the scale proposed by Landis and Koch (1977). For lexical semantics, the value is only fair (0.38). An analysis showed that the identification of a lexical semantic relation is often not straightforward, and may require a good knowledge of the domain. For example, the verbs "aufrufen" [*call*] and "importieren" [*import*], which would usually not be considered to be semantically related, may in fact be used to describe the same action in the computer domain, referring to files. Also for the more complex inference steps, we measured only fair agreement, due to the number of positive and negative cases being very skewed. For the "helpful" cases, the values ranged between 0.73 and 0.79 (substantial agreement).

5.2 Distribution of inference steps

Table 1 summarizes the distribution of inference steps identified in our data for each text representation type, ordered by their frequency of occurrence.[5] For multi-token terms, particle verbs, and negation, the number of "helpful" cases is given in brackets.

Our results show that the most important inference step at the lexical level is lexical semantics. At the lexical level, we found 157 different word mappings. Only 26 of them correspond to a relation in GermaNet (Hamp and Feldweg, 1997), version 7.0. 48 of the involved words had no GermaNet entry at all, due to the word being an anglicism (e.g., "Error" instead of "Fehler"), a non-lexicalized compound (e.g., "Bildschirmbereich" [*screen area*]) or a highly domain- or application-specific word (for only 37.5% of the words missing in GermaNet, we found an entry in Wikipedia). In 72 cases, both words had a GermaNet entry, but no relation existed, usually because the relation was too domain-specific.

For more than 30% of the words (as compared to 10.1% in Bar-Haim's (2010) analysis on English), a morphological transformation is required, which can be explained by the high complexity of German morphology as compared to the morphology of English. Spelling mistakes or differences, which are not considered in other analyses, are also found in a considerable number of words, the reason being that customer emails are less well-formed than, for example, news texts.

The significance of multi-token terms was surprisingly high for German, where word combinations are usually expressed in the form of compounds (i.e., a single token). In our data, multi-token terms were usually compounds consisting of at least one anglicism (e.g., "USB Anschluss" [*USB port*]). This suggests that texts written in a domain language with a high proportion of English loan words may be more difficult to process than general language texts, as multi-token terms have to be recognized.

At the level of compositional semantics, it should be noted that, in many cases, recognizing the entailment relation between two expressions requires world or domain knowledge. Several of the mappings involved particle verbs or light verbs. Detecting negation scope is expected to be important in a precision-oriented system.

5.3 Comparing text representations

We also had a look at the amount of information left unmapped at each level. For the lexical level, we determined for how many of the content tokens (terms) occurring in the category descriptions, no matching expression was found in the associated emails. For the level of compositional semantics, we looked at each term left unmapped at the lexical level and tried to map a complex expression in

[5] Based on the steps agreed on after a consolidation phase.

Type of inference	Data example	Total (Share)
Lexical semantics (Token)	"Anfang" [*start*] → "Beginn" [*beginning*]	310 (20.2%)
Inflection	"startet" [*starts*] → "starten" [*start*]	206 (13.4%)
Derivation	"Import" [*import*] → "importieren" [*(to) import*]	164 (10.7%)
Composition	"Fehlermeldung" [*error message*] → "Meldung" [*message*]	158 (10.3%)
Spelling	"Dateine" → "Dateien" [*files*]	47 (3.1%)
Lexical semantics (Term)	"MPEG Datei" [*MPEG file*] → "Video" [*video*]	60 (4.1%) [+124 (8.6%)]
Particle verbs	"spielt [...] ab" [*play*] → "abspielen" [*play*]	26 (1.8%) [+34 (2.4%)]
Light verbs	"Meldung kommt" [*message appears*] → "melden" [*notify*]	17 (1.2%)
Negation	"Brennergerät kann nicht gefunden werden" [*Burning device cannot be found*] → "Es wird kein Brenner gefunden" [*No burner is found*]	8 (0.6%) [+121 (8.4%)]
Other complex expressions	"Das Brennen bricht ab mit der Meldung X" [*Burning breaks with message X*] → "Beim Brennen kommt die Fehlermeldung X" [*Burning yields error message X*]	83 (5.7%)

Table 1: Distribution of inference steps in the dataset.

which the term occurred. If for none of these expressions a matching expression was found in the email, the term was counted as non-mappable at this level.

Representation	Non-mappable	Share
Tokens	428 / 1538	27.8%
Terms	365 /1446	25.2%
Complex expressions	229 / 1446	15.8%

The above table shows that the majority of the required inference relates to the lexical level. Choosing a representation that allows us to map more complex expressions, increases the amount of mappable terms by almost 10%. However, even with this more complex representation, a considerable amount of terms (15.8%) cannot be mapped at all because the email text does not contain all information specified in the category description.

6 Conclusions

In our analysis, we examined the inference steps required to determine that the text of a category description can be inferred from the text of a particular email associated to this category. We identified major inference phenomena and determined their distribution in a German real-world dataset. Our analysis supports previous results for English data in that a large portion of the required inference relates to the lexical level. Choosing a representation that allows us to map more complex expressions significantly increases the amount of mappable expressions, but some expressions simply cannot be mapped because the categorization was done relying on partial information in the email.

Our results extend previous results by investigating inference steps specific to the German language (such as morphology, composition, and particle verbs). Some outcomes are unexpected for the German language, such as the high share of multi-token terms. Our analysis also stresses the importance of inference steps relying on domain-specific resources, i.e., for this type of data, the development of tools and resources to support inference in highly specialized domains is crucial.

We are currently using the results of our analysis to build an email categorization system that integrates linguistic resources and tools to expand the linguistic expressions in an incoming email with entailed expressions. This will allow us to measure the performance of such a system, in particular with respect to the effect on precision.

Acknowledgements

This work was partially supported by the EXCITEMENT project (EU grant FP7 ICT-287923) and the German Federal Ministry of Education and Research (Software Campus grant 01—S12050). We would like to thank OMQ GmbH for providing the dataset, Britta Zeller and Jonas Placzek for the data anonymization, and Stefania Racioppa for her help in the annotation phase.

References

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics (SemEval 2012)*, pages 385–393, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

Roy Bar-Haim. 2010. *Semantic Inference at the Lexical-Syntactic Level*. Ph.D. thesis, Department of Computer Science, Bar Ilan University, Ramat Gan, Israel.

Luisa Bentivogli, Peter Clark, Ido Dagan, Hoa T. Dang, and Danilo Giampiccolo. 2011. The Seventh PASCAL Recognizing Textual Entailment Challenge. In *Proceedings of TAC*.

Elena Cabrio and Bernardo Magnini. 2013. Decomposing Semantic Inferences. *Linguistics Issues in Language Technology - LiLT. Special Issues on the Semantics of Entailment*, 9(1), August.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*.

Kathrin Eichler, Matthias Meisdrock, and Sven Schmeier. 2012. Search and Topic Detection in Customer Requests - Optimizing a Customer Support System. *KI*, 26(4):419–422.

Kathrin Eichler. 2005. *Automatic classification of Swedish email messages*. Bachelor thesis, Eberhard-Karls-Universität, Tübingen, Germany.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a Lexical-Semantic Net for German. In *In Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15.

J. R. Landis and G. G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174, March.

Alexander Volokh and Günter Neumann. 2011. Using MT-Based Metrics for RTE. In *Proceedings of the 4th Text Analysis Conference (TAC 2011)*, Gaithersburg, Maryland, USA, November. National Institute of Standards and Technology.

Janine Wicke. 2010. *Automated Email Classification using Semantic Relationships*. Master thesis, KTH Royal Institute of Technology, Stockholm, Sweden.