Multilingual Corpus Development for Opinion Mining

Julia Maria Schulz, Christa Womser-Hacker, Thomas Mandl

Department for Information Science and Natural Language Processing
University of Hildesheim, Germany
schulzju@uni-hildesheim.de, womser@uni-hildesheim.de, mandl@uni-hildesheim.de

Abstract

Opinion Mining is a discipline that has attracted some attention lately. Most of the research in this field has been done for English or Asian languages, due to the lack of resources in other languages. In this paper we describe our methodology for developing a manually annotated multilingual corpus with fine-grained opinion and target annotations. The languages represented in the corpus are English, German and Spanish. The tool for annotation and first results on the inter-annotator agreement for opinions and product features are presented.

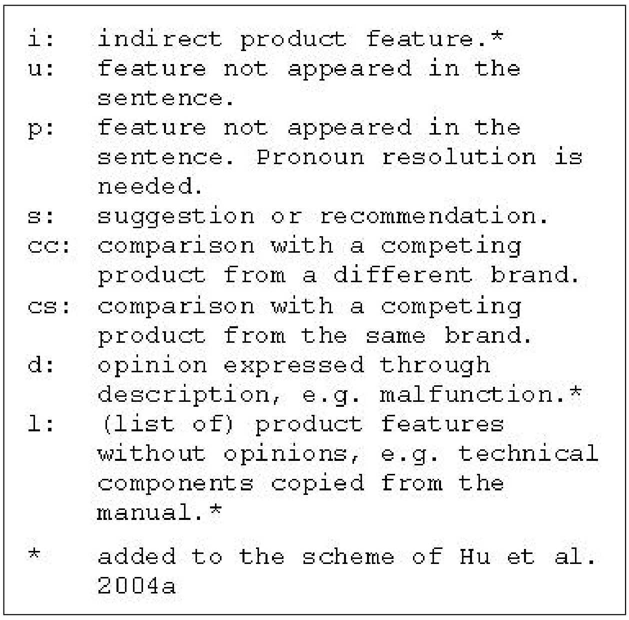
1. Introduction

Opinion Mining or Sentiment Analysis a recent discipline at the intersection of information retrieval, computational linguistics, and text mining is concerned with identifying and classifying opinions in unstructured texts, e.g. newspaper articles, forums, and product reviews. Research in this area produced a variety of different tasks and goals. Wiebe et al. (2004) classified text into subjective or objective, Wloka et al. (2007) analyse and classify emotions like joy, anger or grief, whereas Pang et al. (2002), and Turney and Littman (2003) focused on the classification of documents into positive or negative categories. This two-class problem can be extended rather easily by accessorily considering the strength of a given opinion (cf. Liu et al. (2005); Bautin et al. (2008); Subrahmanian and Reforgiato (2008)). Only few works have also considered the target of an opinion in addition (cf. Nasukawa and Yi (2003); Hu and Liu (2004); Popescu and Etzioni (2005)). Solely the latter have considered indirect opinion targets, like "size" in the sentence "This camera fits into every pocket.", in their research. Most of the past research has concentrated on English, due to the lack of resources in other languages. In this paper we describe an approach of building a manually annotated multilingual corpus, which can be used as a basis for fine-grained opinion analysis also considering direct and indirect opinion targets. An opinion target is the concept the opinion is referring to, which are products or product features in our context, e.g. picture quality in the following example: "The picture quality is great").

The rest of the paper is organized as follows. Section 2 looks at related work with respect to corpus building and annotation for opinion mining purposes. Section 3 describes the annotation process in more detail and in section 4 we present the results of the inter-annotator agreement. Section 5 gives some conclusions and in the last section we give an outlook on future work.

2. Related Work

There has been some effort in creating resources for opinion analysis tasks in recent years. Yu and Hatzivassiloglou (2003) manually annotated 400 sentences of the TREC news corpus with respect to facts and opinions and their polarity. 100 of these sentences were annotated by two annotators. Bethard et al. (2004) annotated a subset of the FrameNet and the PropBank corpus consisting of 5,139 sentences. In addition to subjectivity and objectivity annotations they also labeled the opinion holder, the topic of an opinion and the subjective part of a sentence. An even larger amount of sentences has been annotated by Wiebe et al. (2005), namely 10,657. In addition to the information annotated from Bethard et al. (2004) they captured also information about the intensity of an opinion and the nested holders of an opinion, e.g. the writer of the article and the person who actually said something. Stoyanov and Cardie (2008) enhanced the MPQA corpus[1] of Wiebe et al. (2005) by adding topic annotations to a subset of 150 of the 535 documents. The annotation scheme introduced by Wiebe et al. (2005) also contains opinion targets, but only a small subset of the corpus was annotated regarding the opinion topics.

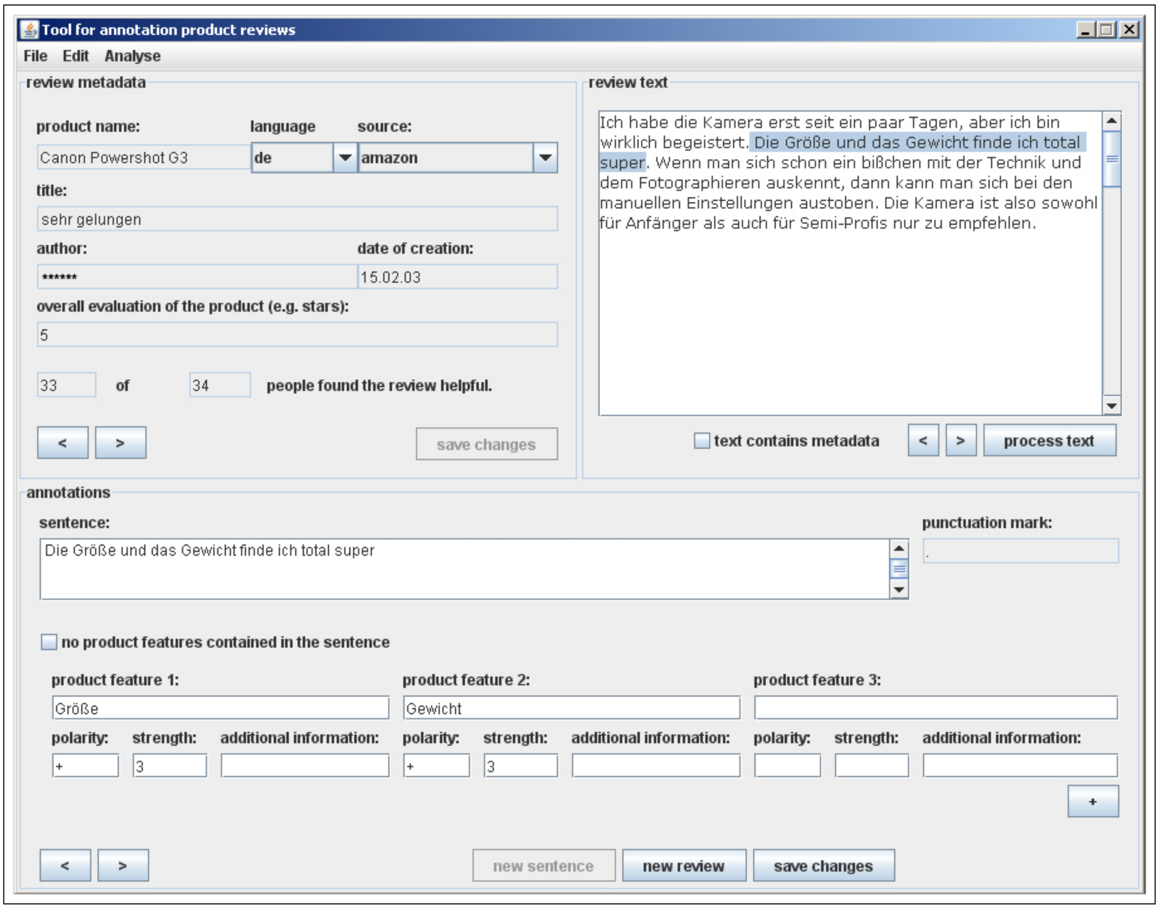


```
i:   indirect product feature.*
u:   feature not appeared in the
     sentence.
p:   feature not appeared in the
     sentence. Pronoun resolution is
     needed.
s:   suggestion or recommendation.
cc:  comparison with a competing
     product from a different brand.
cs:  comparison with a competing
     product from the same brand.
d:   opinion expressed through
     description, e.g. malfunction.*
l:   (list of) product features
     without opinions, e.g. technical
     components copied from the
     manual.*

*    added to the scheme of Hu et al.
     2004a
```

Figure 1: Additional information for annotated opinions.

[1] Available at www.cs.pitt.edu/mpqa/databaserelease/



review metadata

product name: Canon Powershot G3

language: de

source: amazon

title: sehr gelungen

author: ******

date of creation: 15.02.03

overall evaluation of the product (e.g. stars): 5

33 of 34 people found the review helpful.

review text

Ich habe die Kamera erst seit ein paar Tagen, aber ich bin wirklich begeistert. Die Größe und das Gewicht finde ich total super. Wenn man sich schon ein bißchen mit der Technik und dem Fotographieren auskennt, dann kann man sich bei den manuellen Einstellungen austoben. Die Kamera ist also sowohl für Anfänger als auch für Semi-Profis nur zu empfehlen.

□ text contains metadata

annotations

sentence: Die Größe und das Gewicht finde ich total super

punctuation mark: .

□ no product features contained in the sentence

product feature 1:			product feature 2:			product feature 3:		
Größe			Gewicht					
polarity:	strength:	additional information:	polarity:	strength:	additional information:	polarity:	strength:	additional information:
+	3		+	3				

Figure 2: Additional information for annotated opinions.

All the corpora previously mentioned consist of news documents, but as one of the major areas of application in opinion mining is the summarization of product reviews according to the expressed opinions, we like to focus on this text domain. To our knowledge in this domain there is only one corpus available, which includes information about the polarity and strength of an opinion as well as labeled opinion targets, which are product features (cf. Hu and Liu (2004); Ding et al. (2008))[2]. Information about the opinion holder is not as relevant in the domain of product reviews as it is in news documents, because the opinion holder is almost always the author of the review and is not included in the mentioned corpus. Since we want to build a multilingual corpus for opinion mining purposes, we started off with the English one annotated by Hu and Liu (2004) and extended by Ding et al. (2008) and added about 500 German and Spanish reviews for the same products as in the English corpus. A more detailed description of the annotation process is given in the next section.

3. The Annotation Process

3.1. Annotation Methodology

As described above, we used an existing English corpus as a basis for our multilingual corpus. To assure comparability between our corpus and the English one, we employed the same annotation scheme to our documents with slight enhancements.

For each sentence in a review, the mentioned product features (explicit and implicit ones) with their respective opinion polarity and strength on a scale from 0 to 3 are labeled manually by two annotators. We also capture additional information like comparison with other products from the same or a different brand, suggestions and recommendations etc. about the opinions and product features. We added three more categories of additional information to the original scheme of Hu and Liu (2004) (see Figure 1).

Furthermore, if a product feature does not appear in the sentence, where an opinion is expressed, or a pronoun is used instead, we also wanted to capture the cases where the product feature is mentioned implicitly, e.g. in a verb phrase or an adjective, separately. For example in the sentence "The camera is designed very well." the opinion clearly refers to the design of the camera, even though the noun "design" itself does not appear in the sentence. This kind of expressing an opinion is very frequent in the German corpus.

We also added a category for opinions, which were expressed indirectly, e.g. through a description of a malfunction of a product feature. In the sentence "The lens is visible in the optical viewfinder, when the lens is set to the wide angle" a negative sentiment towards the visibility of the lens in the optical viewfinder is expressed in the context of a product review, even though the sentence does not con-

[2] Available at www.cs.uic.edu/∼liub/FBS/sentiment-analysis.html.

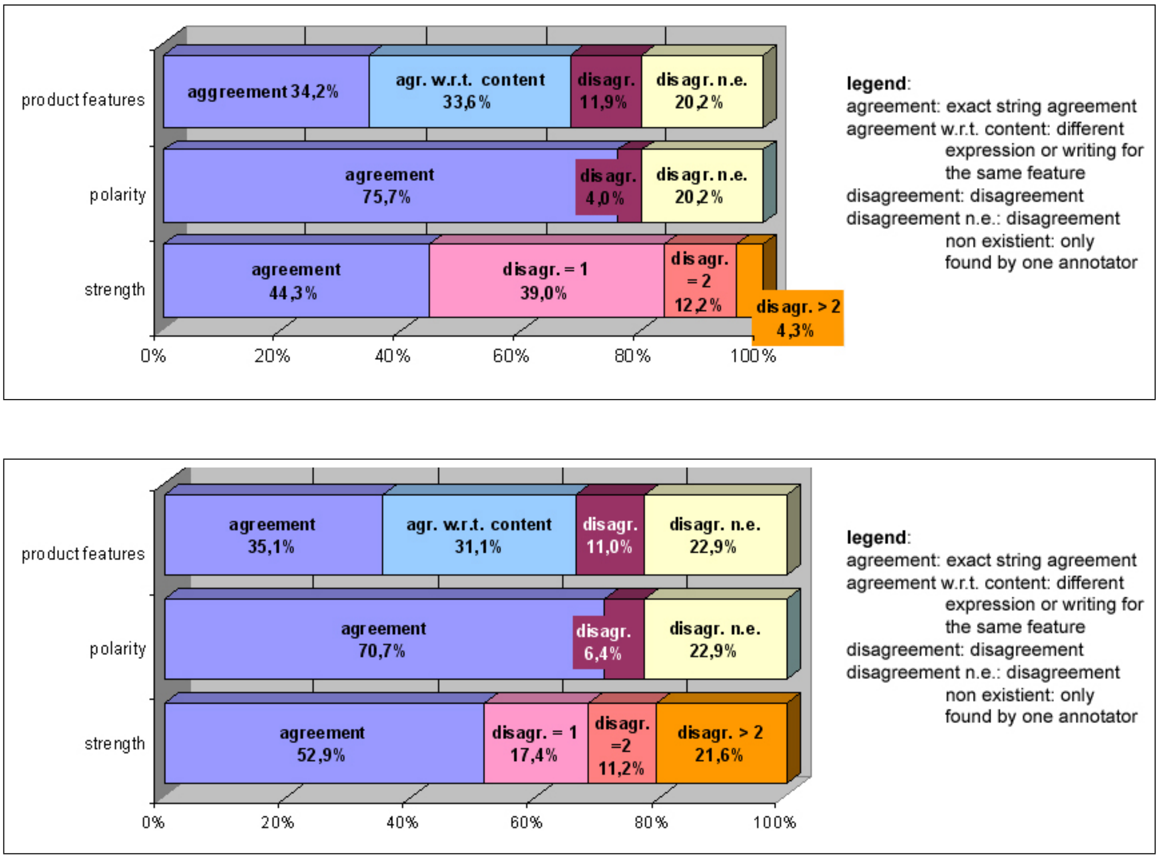


Figure 3: Results of the inter-annotator agreement (top: German, bottom: Spanish)

tain a word with an explicit negative sentiment orientation. The negative opinion can only be inferred from the context, the described position of the lens and the knowledge, that the lens should not be visible if you look through the optical viewfinder.

The last category we added is concerned with sentences, where a product feature is mentioned, but no opinion is expressed, like in “The player is shipped with an AC adaptor, firewire cable and USB cable”. This information can be used for a more detailed error analysis by identifying the modality of cases a system could not find.

In comparison to the English corpus we additionally incorporated the metadata of a review in our dataset. This information can later be used for further analysis, e.g. trend analysis. We also chose a more structured and flexible format to store our data than (Hu and Liu, 2004), who use a common text format. We employ an xml-format instead, which allows the use of different elements of the xml-file independently and permits more flexible test designs.

3.2. Tool for the Annotation Process

In order to facilitate the annotation process we developed a Java based tool with a graphical user interface (GUI) for the annotation process (see Figure 2). The GUI is divided into three parts: The upper left part presents the meta-information of the product review, namely the name of the product, the language of the review, the source of the review (e.g. amazon), the title, the author, the date of creation, the overall evaluation of the product, and the evaluation of the review by other users. The upper right part shows the entire text of the review, highlighting the sentence one is working on to give the annotator the possibility to see the context of a sentence. The third part on the lower half of the GUI is the core of the annotation tool. It shows the sentence to be annotated and holds the slots, where the product feature along with the respective polarity and strength of the opinion can be noted. It also contains a slot for each opinion/product feature pair to store additional information (see Figure 2).

3.3. Inter-annotator Agreement

The rating and annotating of opinions is a rather subjective task. In order to obtain a more reliable and objective perspective on the annotated opinions every review in the corpus was annotated by two persons. So far we have calculated an inter-annotator agreement for about 10 percent of the corpus, which means around 50 documents per language. The results can be seen in Figure 3. For the annotation of product features we calculated two kinds of agreement: an exact agreement, meaning an agreement on the string level, and an agreement with respect to the content, meaning the two annotators used different expressions or a different writing for a product feature. If we cumulate these two types of agreement we get an inter-annotator agreement of 68% for German and 66% for Spanish regarding the product features. About 20% of the product features in both languages were only annotated by one of the annotators. This also affects the agreement of the polarity and strength: for the German annotation we achieve an agreement of 76% (70% for Spanish) regarding the annotated polarity of an opinion. If we take the product features which are non existent in one of the annotators data out of the calculation for the agreement on the polarity and strength of an opinion, we get an inter-annotator agreement of 95% (91% for Spanish).

Rating the strength of an opinion can be considered the most challenging task in the annotation process. This can also be observed by looking at the inter-annotator agreement of the strength of an opinion. We achieved an agreement of about 44% for the German and 53% for the Spanish corpus. For about 33% of the German annotations (17% for Spanish) there is a difference of 1 regarding the strength of an opinion. That means for example, if one annotator thought an expression is very strong and assigned 3 to the opinion the other annotator assigned only the value 2 for the strength of an opinion. Due to the challenge of rating the strength of an opinion and the fact, that about 30% of the product features have only been annotated by one person, we think this is a quite good result, as this means only about 23% of the German annotations and 30% of the Spanish ones show a difference greater than 1.

4. Conclusions

We presented a methodology for building a multilingual corpus for opinion mining purposes, which does not exist in this size and granularity with respect to the annotation in the domain of product reviews so far.

5. Future work

In future work we want to further analyze the judgments of the two annotators and calculate an inter-annotator agreement for the entire German and English corpora. We also want to annotate a small subset of the corpus with a larger number of annotators to get an impression of the reliability of the annotations by only two annotators.

Aside from building the corpus we are working on a system to detect and classify opinions and their corresponding targets in a multilingual corpus. In future we plan to test the systems performance against our manually annotated corpus.

We intend to make the corpus available for research purposes, after clarifying copyright issues.

6. References

Mikhail Bautin, Lohit Vijayarenu, and Steven Skiena. 2008. International sentiment analysis for news and blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*.

Steven Bethard, Hong Yu, Ashley Thornton, Vasileios Hatzivassiloglou, and Dan Jurafsky. 2004. Automatic extraction of opinion propositions and their holders. In *AAAI Spring Symposium on Exploring Attitude and Affect in Text*, pages 22–24.

Xiaowen Ding, Bing Liu, and Philip S. Yu. 2008. A holistic lexicon-based approach to opinion mining. In *Proceedings of the international conference on Web search and web data mining*, pages 231–240. ACM, Palo Alto, California, USA.

Minqing Hu and Bing Liu. 2004. Mining opinion features in customer reviews. In L. Deborah Mcguinness, George Ferguson, L. Deborah Mcguinness, and George Ferguson, editors, *AAAI*, pages 755–760. AAAI Press / The MIT Press.

Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer analyzing and comparing opinions on the web. In *Special interest tracks and posters of the 14th international conference on World Wide Web*, pages 342–351, Chiba, Japan. ACM.

Tetsuya Nasukawa and Jeonghee Yi. 2003. Sentiment analysis: capturing favorability using natural language processing. In *K-CAP '03: Proceedings of the 2nd international conference on Knowledge capture*, pages 70–77, New York, NY, USA. ACM.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pages 79–86, Morristown, NJ, USA. Association for Computational Linguistics.

Ana-Maria Popescu and Oren Etzioni. 2005. Extracting product features and opinions from reviews. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 339–346. Association for Computational Linguistics, Vancouver, British Columbia, Canada.

Veselin Stoyanov and Claire Cardie. 2008. Annotating topics of opinions. In Choukri Bente Maegaard Joseph Mariani Jan Odjik Stelios Piperidis Daniel Tapias Khalid Nicoletta Calzolari, editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Venkatramana S. Subrahmanian and Diego Reforgiato. 2008. Ava: Adjective-verb-adverb combinations for sentiment analysis. *IEEE Intelligent Systems*, 23(4):43–50.

Peter D. Turney and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Trans. Inf. Syst.*, 21(4):315–346.

Janyce M. Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational Linguistics*, 30(3):277–308.

Janyce M. Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2):165–210.

Klaus Wloka, Johann Haller, and Silke Schirmer. 2007. Schlussbericht zum Forschungsprojekt: Technologie für ein internetbasiertes Marketingsystem zur automatischen qualitativen Bewertung von Meinungsäußerungen in Online-Quellen.

Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of EMNLP-03, 8th Conference on Empirical Methods in Natural Language Processing*, pages 129–136, Sapporo, JP.