The USAGE review corpus for fine-grained, multi-lingual opinion analysis

Roman Klinger and Philipp Cimiano

Semantic Computing Group
Cognitive Interaction Technology – Center of Excellence (CIT-EC)
Bielefeld University
33615 Bielefeld, Germany
{rklinger,cimiano}@cit-ec.uni-bielefeld.de

Abstract

Opinion mining has received wide attention in recent years. Models for this task are typically trained or evaluated with a manually annotated dataset. However, fine-grained annotation of sentiments including information about aspects and their evaluation is very labour-intensive. The data available so far is limited. Contributing to this situation, this paper describes the Bielefeld University Sentiment Analysis Corpus for German and English (USAGE), which we offer freely to the community and which contains the annotation of product reviews from Amazon with both aspects and subjective phrases. It provides information on segments in the text which denote an aspect or a subjective evaluative phrase which refers to the aspect. Relations and coreferences are explicitly annotated. This dataset contains 622 English and 611 German reviews, allowing to investigate how to port sentiment analysis systems across languages and domains.
We describe the methodology how the corpus was created and provide statistics including inter-annotator agreement. We further provide figures for a baseline system and results for German and English as well as in a cross-domain setting. The results are encouraging in that they show that aspects and phrases can be extracted robustly without the need of tuning to a particular type of products.

Keywords: sentiment analysis, corpus, product reviews

1. Introduction

The task of analyzing sentiments and opinions of users about products, events, services etc. has generated wide interest not only in academia but also in industry due to its high commercial relevance. Approaches to develop sentiment analysis and opinion mining frameworks can be roughly divided into two categories. On the one hand, we find systems which rely on rules or dictionaries to extract evaluative phrases and the aspects they refer to. Such rule-based or dictionary-based methods typically exploit manually crafted or semi-automatically built resources like the subjectivity dictionary by Wilson et al. (2009) or the polarity dictionary by Ding et al. (2008).
On the other hand, there are approaches that exploit machine learning techniques to induce a sentiment extraction model from training data, either in a fully supervised or weakly supervised fashion. Fully supervised systems that train on manually annotated data are commonly used to extract aspects and subjective phrases (Klinger and Cimiano, 2013a; Klinger and Cimiano, 2013b; Li et al., 2010) or in order to classify the polarity or subjectivity of text (Täckström and McDonald, 2011; Sayeed et al., 2012; Shi and Li, 2011; Pang and Lee, 2004; Wiebe, 2000). In contrast to these fully supervised systems, Turney (2002) for instance proposed a system that is in this sense weakly supervised in that it relies on the two seed words "excellent" and "poor" and textual similarity to induce other "similar" adjectives that express a positive or negative sentiment, respectively. Completely unsupervised approaches have also been applied to the task (Titov and McDonald, 2008).
In most of the above mentioned cases, annotated data is needed, *e. g.*, to tune the parameters of a system in a supervised fashion or in order to evaluate the approach in question. However, creating annotated sentiment corpora is a labour-intense task, so that the availability and size of such datasets is limited so far.
With this paper, we provide the Bielefeld **U**niversity corpus for **S**entiment **A**nalysis in **G**erman and **E**nglish (USAGE), a resource based on Amazon product reviews for a variety of product classes, both in German and English. The annotation is fine-grained in the sense that not only course classes are assigned to sentences or whole reviews but word or token-based semantic information is provided as well. The corpus is freely and publicly available for future research.[1]

1.1. Previous Work

For sentiment analysis and opinion mining, several manually annotated corpora are available. An overview of the corpora mentioned in the following is given in Table 1. Examples include fine-grained annotations such as released by Hu and Liu (2004) and Ding et al. (2008), who have provided an annotated dataset consisting of Amazon reviews in which every sentence is annotated with an aspect and a polarity score. However, the actual offsets of phrases which denote the aspect or a subjective or evaluating phrase are not provided. The data set published in the context of the SemEval 2013 shared task provides annotations on Tweets (Nakov et al., 2013). These datasets focus on the task of extracting subjective phrases for given aspects and entities. Thus, aspects are pre-given and do not need to be extracted. The University review data set by Toprak et al. (2010) is annotated with opinion holders, targets, modifiers, anaphora as well as the relevance for a topic.
Restaurant reviews annotated on a sentence level with predefined aspects and polarities are made available by Ganu et al. (2009). Lakkaraju et al. (2011) have provided reviews for different product classes with predefined aspects and

[1] The corpus is available at http://dx.doi.org/10.4119/unibi/citec.2014.14. It will be further developed and future versions will be linked from that URL.

polarity annotations. The MPQA corpus consists of fine-grained annotations, focusing on debates and news articles (Ruppenhofer et al., 2008; Wiebe, 2000; Wiebe et al., 2005). The JDPA sentiment corpus consists of blog posts about cars and cameras and is annotated with a complex set of entities and relations, including aspects, subjective phrases, polarities, part-of relations, feature-of relations, opinion holders and others. The entities are provided on token level (Kessler et al., 2010). The Twitter data set by Spina et al. (2012) is annotated with offsets for aspect mentions (of given categories) and subjective phrases as well as overall subjectivity. Polarities are not given. Both corpora have been influential examples in the design of our annotation guidelines.

There is only a comparatively small number of corpora available in other languages. For instance, the only fine-grained corpus in German we are aware of is the manually annotated corpus with subjectivity and polarity annotation on sentence, phrase, and word level by Clematide et al. (2012). Another German resource is the Amazon review corpus by Boland et al. (2013), which is annotated on sentence level, whereas aspects are not annotated.

We are not aware of any dataset which supports the development of multi-lingual and cross-lingual sentiment analysis methods that are applicable to different languages or can be trained in one language and applied to another one. Further, we are neither aware of a large German corpus consisting of reviews that are annotated with fine-grained aspects, evaluative (subjective) expressions and the relation between both. The work presented in this paper aims to close this gap.

1.2. Motivation

We are especially interested in the automated analysis of product reviews. Such textual data is for instance collected on websites like Amazon[2], by shopping portals like Google[3] or Ciao[4]. In detail, we are investigating the following research questions:

- How can we detect mentions of aspects and the corresponding evaluating phrases with their polarity?
- How can a model trained on the domain of a specific product be adapted to another domain with limited supervision?
- Can we exploit multilingual features to train sentiment analysis systems to improve performance?
- Can we train a model on one language and transfer that model automatically to another language?

To the best of our knowledge, no dataset is currently available to investigate such research questions.

[2] `http://www.amazon.com/` or `http://www.amazon.de/`

[3] `http://shopping.google.com/` or `http://shopping.google.de/`

[4] `http://www.ciao.co.uk/` or `http://www.ciao.de/`

2. The Bielefeld University Sentiment Analysis corpus for German and English (USAGE)

We present the USAGE corpus, the Bielefeld University Sentiment Analysis corpus for German and English, consisting of annotations of Amazon reviews in German and English for 8 product categories. The corpus is annotated with aspects, subjective evaluating phrases, polarities and their relation.

2.1. Corpus selection

We used the search functionality of Amazon.com and Amazon.de[5] to retrieve lists of products for 8 classes of products. The search terms were "washing machine", "coffee machine", "trash can", "microwave", "vacuum cleaner", "dish washer", "toaster", and "cutlery" for English and "Waschmaschine", "Kaffeemaschine", "Mülleimer", "Mikrowelle", "Staubsauger", "Toaster", and "Besteck". For each search, we kept the top 60 results and downloaded up to 1000 reviews for each of the products for both English and German.

In order to provide the annotators with training material and to fine-tune the annotation guidelines provided to them, 5 sets of 16 English reviews (2 for each product) were selected. For the final corpus annotation, 800 English reviews and 800 German reviews were selected. Both annotators worked 10 hours a week for three months annotating as much reviews as possible within the given time.

2.2. Corpus annotation

The entity classes *aspect* and *evaluative (subjective) expression* are annotated in the corpus. Evaluative expressions are assigned a polarity (positive, negative, neutral) and a set of aspects they refer to. An aspect can be marked as "foreign" if a product or an aspect of a product is mentioned that is not an aspect of the main product discussed in the review. This is often the case in cross-product comparisons and mentions of envisioned or desired features of products. Co-references were to be annotated if the target is not in the same sentence as the evaluative expression.

The annotators were instructed to regard everything as an aspect that is part of a product or related to it and can influence the opinion about it, including the whole product itself. Evaluative phrases express an opinion. Negations are not separately annotated but are part of a phrase. Annotators were asked to avoid overlapping annotations if possible. The annotations should be as short as possible, as long as the meaning is understandable if only the annotations were given (without the sentence itself).

The annotators worked on the corpus for 3 months for about 10 hours a week. The training phase took 20 days. After the training phase, the annotators were instructed to work on as many reviews as possible while trying to keep the number of German and English reviews comparable. Towards the end of annotation, the annotators were coordinated to work on

[5] `http://www.amazon.com/s/field-keywords=[searchterm]` and `http://www.amazon.de/s/field-keywords=[searchterm]`

	Hu(2004) Ding(2008)	Nakov(2013)	Toprak(2010)	Ganu(2009)	Lakkaraju(2011)	Wiebe(2005)	Kessler(2010)	Spina(2012)	Clematide(2012)	Boland(2013)
Text Source	Amazon	Twitter, SMS	Rateitall, eopinions	Restaurant reviews	Amazon	News	Blogs	Twitter	Web (German)	Amazon (German)
Aspects	Sentences no offsets	Yes Task A: offsets	Offsets	Predefined, Sentences	Predefined	Offsets	Offsets	Offsets	Layers/Tokens	No
Evaluation	Aspects	Aspects	Aspects Sentences	Sentences	Aspects	Aspects	Offsets	Aspects Sentences	Tokens/Phrases in layers	Sentences
Subj. phrases	No	No	Yes	No	No	Yes	Yes	Yes	Yes	No
Polarities	$[-3;3]$	Pos/Neg/ Neutr./Obj.	Polarity/ Subjectivity	Pos/Neg/ Neutr./Conflict	$[-2;2]$	Pos/Neg/ Both/None, Intensity	Prior. Polar., Negators, Intensifiers	No	Pos./Neg./ Neutral/ Intensifier/ Diminisher	Pos./Neg./ Neutr./Mix/ Conflict
Size	8897 sent. 14 products	39736 (Task A) 8184 (Task B)	474 reviews	$\approx$ 3400 sent. 652 reviews	2543 reviews.	1227 doc	335 posts	9238 posts	270 sent.	63067 sent.
Relations	Subj. for aspects	Phrases/Topics are evaluated	Yes	Subj. for aspects	Polarity for aspects	Yes	Yes	No	Yes	No
Format	Proprietary	TSV	MMax	XML	Proprietary	Gate DB	Knowtator, API	XML	TSV/XML	
Availability	Yes	Annotations	Yes	Yes	Yes	Yes	Annotations	Yes	Yes	No, In preparation

Table 1: Overview over characteristics of previously published corpora regarding aspects and evaluating subjective phrases. Some corpora contain additional annotated information. The availability field contains clickable links in the PDF version of this paper. "Fine" means that the actual phrase relation is annotated. The document number for Wiebe et al. (2005) is according to the number of entries in the doclist-files in the downloaded data.

the same reviews, such that the whole corpus is annotated twice.

Some examples are given below, with aspects marked in blue and subjective phrases marked in red:

- I had no problems with the return.
  - *return* is a target of *no problems*. *no problems* is positive.
- The washer itself is great, the included hose is junk.
  - *washer* is a target of *great*, *hose* is a target of *junk*. *great* is positive, *junk* is negative.
- It looks very neat, like a storage container, and using it is very simple and easy.
  - *looks* is a target of *very neat*, *using* is a target of *simple* and of *easy*.

3. Analysis

The training of the annotators and optimization of the guidelines has been conducted in four iterations. In order to estimate the inter-annotator agreement, Cohen's kappa was calculated (Cohen, 1960). In the first annotation round of 16 English reviews, the agreement between the annotators reached a κ-value of 0.524 (on token level). After discussion, the independent re-annotation of the same data lead to $\kappa = 0.608$. A further independent annotation round of new 16 reviews resulted in $\kappa = 0.62$, showing that the annotators converged in their understanding of the task. In the next step, the annotators were asked to annotate 16 reviews more in interaction with each other. In a subsequent independent annotation step involving 16 further reviews, an agreement of $\kappa = 0.66$ was reached, which can be regarded as a moderate agreement in comparison to agreement by chance. The agreement has been increased by several discussion and annotation rounds. The agreement in the full German corpus is 0.65 and in the English corpus 0.64.

Statistics of the German and the English full corpus as well as broken down by product domains are shown in Table 2. The German corpus consists of 611 annotated reviews describing 127 different products. The total number of annotated aspects is 6340 for Annotator 1 and 5055 for Annotator 2. There are 5086 (4881) subjective annotations in total, of which 3840 (3717) are positive and 1094 (1052) are negative. The number of subjective phrase-target relations is 4085 (4643). The most frequent ones are *'gut', 'sehr zufrieden', 'sehr gut', 'super', 'leicht', 'gute', 'schnell', 'sehr leise', 'einfach'*.

The English corpus consists of 622 annotated reviews describing 217 different products. The number of aspects is 8545 (6609) in total. There are 5321 (5518) subjective annotations from which 3426 (3600) are positive and 1799 (1792) are negative. The number of subjective phrase-target relations is 4481 (5180). The most frequent subjective phrases are *'recommend', 'best', 'nice', 'love', 'like', 'well', 'perfect', 'easy', 'good', 'love', 'great'*.

The average numbers of annotated aspect and subjective phrase mentions are comparable between the different domains and between the annotators. Annotator 1 tends to annotate more aspects than Annotator 2 (13.7 to 10.6 for English and 10.4 to 8.3 for German in the full corpora). The highest difference is between washing machines and cutlery with washing machines having the highest density of aspects and cutlery the lowest (9.9/7.5 versus 19.4/13 in English and 6.5/6.2 versus 17.4/10.7 in German). Examples for such differences are the inclusion of aspects by Annotator 1 like the product description itself ("the dishwasher") or aspects which are not directly connected to the product but clearly related to it ("hard water", "customer service", "dishes"). Obviously, these cases are hard to decide.

The differences in the average number of subjective phrases

		English									German							
		full	coffee machine	cutlery	microwave	toaster	trash can	vacuum cleaner	washing machine	dish washer	full	Kaffeemaschine	Besteck	Mikrowelle	Toaster	Mülleimer	Staubsauger	Waschmaschine
# reviews		622	75	49	100	100	100	51	49	98	611	108	72	100	4	99	140	88
# products		217	28	26	36	38	31	28	15	15	127	24	25	24	3	27	2	22
Aspects	num.	8545 6609	1102 904	484 366	1234 1055	1124 932	1015 824	896 676	950 638	1740 1214	6340 5055	925 817	468 447	895 803	55 44	973 713	1491 1289	1533 942
	avg. num.	13.7 10.6	14.7 12.1	9.9 7.5	12.3 10.6	11.2 9.3	10.2 8.2	17.6 13.3	19.4 13.0	17.6 12.4	10.4 8.3	8.6 7.6	6.5 6.2	9.0 8.0	13.8 11.0	9.8 7.2	10.1 9.2	17.4 10.7
	avg. length	9.5 8.7	9.4 9.1	9.5 9.1	10.1 9.3	9.2 8.4	9.1 8.3	9.0 8.5	9.3 8.6	9.8 8.5	10.9 10.1	10.4 10.2	9.8 8.8	10.6 10.0	11.0 9.0	10.0 9.5	10.7 10.0	12.4 11.3
Subjective Phrases	num.	5321 5517	678 742	355 357	815 869	761 783	748 786	512 545	476 471	976 964	5086 4881	781 808	540 526	723 736	41 41	740 730	1347 1329	914 711
	pos.	3426 3600	414 454	196 200	485 547	463 482	522 553	351 381	357 364	638 619	3840 3717	612 639	396 397	498 517	37 36	515 545	1029 1010	753 573
	neg.	1799 1792	258 268	153 148	311 308	276 281	215 222	152 151	109 98	325 316	1094 1052	145 152	134 123	195 193	4 4	192 170	288 283	136 127
	avg. num.	8.6 8.9	9.0 9.9	7.2 7.3	8.2 8.7	7.6 7.8	7.5 7.9	10.0 10.7	9.7 9.6	10.0 9.8	8.3 8.0	7.2 7.5	7.5 7.3	7.2 7.4	10.3 10.3	7.5 7.4	9.6 9.5	10.4 8.1
	avg. length	12.9 12.5	12.3 11.8	12.9 12.5	13.5 12.7	12.9 11.5	13.0 11.8	12.4 12.9	13.7 14.7	12.7 13.1	16.9 12.9	17.4 13.1	17.4 13.8	14.9 12.4	13.2 11.2	16.1 13.2	16.9 12.7	18.6 12.6
Asp.-Subj.	num.	4481 5180	601 723	290 296	704 829	664 777	609 741	413 514	349 384	851 916	4085 4643	662 774	376 473	631 703	35 36	554 686	1066 1286	761 685
	avg. num.	7.2 8.3	8.0 9.6	5.9 6.0	7.0 8.3	6.6 7.8	6.1 7.4	8.1 10.1	7.1 7.8	8.7 9.3	6.7 7.6	6.1 7.2	5.2 6.6	6.3 7.0	8.8 8.0	5.6 6.9	7.6 9.2	8.6 7.8
Coref.	num.	67 462	6 48	0 21	19 91	11 78	10 87	4 34	2 28	15 75	37 224	4 31	4 42	4 29	0 1	8 34	11 61	6 26
	avg. num.	0.1 0.7	0.1 0.6	0.0 0.4	0.2 0.9	0.1 0.8	0.1 0.9	0.1 0.7	0.0 0.6	0.2 0.8	0.06 0.36	0.04 0.29	0.06 0.58	0.04 0.29	0.00 0.25	0.01 0.34	0.08 0.44	0.01 0.30
Cohen's κ		0.64	0.67	0.67	0.65	0.66	0.66	0.66	0.59	0.59	0.65	0.68	0.66	0.73	0.72	0.64	0.68	0.53
F_1 Aspect		0.63	0.67	0.64	0.66	0.65	0.66	0.66	0.57	0.57	0.71	0.78	0.73	0.77	0.70	0.67	0.74	0.61
F_1 Subj.		0.54	0.58	0.53	0.53	0.54	0.59	0.54	0.48	0.52	0.55	0.54	0.54	0.62	0.53	0.53	0.56	0.51
F_1 Asp.-S.		0.38	0.42	0.42	0.40	0.42	0.42	0.36	0.35	0.36	0.42	0.45	0.43	0.48	0.43	0.38	0.42	0.39
F_1 Coref.		0.11	0.19	0.0	0.10	0.10	0.08	0.16	0.0	0.18	0.15	0.11	0.13	0.12	0.00	0.24	0.17	0.13

Table 2: Statistics of the German and English full corpora as well as separated into different product domains. In cells with two numbers below each other, the first is for Annotator 1 and the second for Annotator 2.

is lower. However, differences between the two annotators can be observed for this class of segments as well: The average length (measured in characters) of annotated subjective phrases is higher than the lengths of aspect annotations. In addition, a difference in length between the two annotators can be observed, especially for the German subjective phrases.

Not every aspect or subjective phrase is actually in relation with a counterpart. The average number of aspect-subjective phrase relations is observed to be slightly lower than the number of aspects or subjective phrases. Annotator 2 tends to have more such relations, but the difference is only marginal. However, the annotation of coreferences differs a lot, with 67 such relations annotated by Annotator 1 and 462 by Annotator 2 for the English dataset. This difference is not based on a different understanding, but just by annotating more terms like "it" and "they". Annotator 1 annotated such terms only if a subjective phrase could not be linked to another aspect, while Annotator 2 annotated anaphora more frequently.

In order to be able to quantify the differences between the two annotators, the F_1 measure between them has been calculated. This serves as an upper bound for automatic extraction tools as well: If the agreement between two humans is lower than the agreement between a machine and a human, the result should be interpreted critically. This measure takes into account phrase boundaries and does not normalize over the probability of agreement, as Cohen's κ does. Note that the F_1 numbers in this table are all based on exact matches. Detection of aspects is generally better

		English									German							
		full	coffee machine	cutlery	microwave	toaster	trash can	vacuum cleaner	washing machine	dish washer	full	Kaffeemaschine	Besteck	Mikrowelle	Toaster	Mülleimer	Staubsauger	Waschmaschine
10-fold Cross-Validation	Aspect	0.56	0.50	0.53	0.50	0.47	0.55	0.49	0.52	0.50	0.63	0.68	0.55	0.61		0.58	0.64	0.56
		0.43	0.49	0.44	0.45	0.39	0.48	0.42	0.40	0.38	0.60	0.66	0.58	0.64		0.53	0.59	0.54
	Aspect Approx.	0.75	0.74	0.74	0.74	0.67	0.72	0.71	0.69	0.72	0.76	0.77	0.68	0.74		0.70	0.76	0.70
		0.58	0.65	0.65	0.62	0.52	0.61	0.58	0.50	0.49	0.69	0.73	0.66	0.69		0.62	0.70	0.63
	Subjective	0.48	0.41	0.38	0.39	0.42	0.41	0.41	0.31	0.38	0.48	0.35	0.38	0.44		0.37	0.43	0.42
		0.41	0.41	0.32	0.33	0.34	0.38	0.36	0.26	0.31	0.47	0.39	0.33	0.40		0.31	0.42	0.33
	Subjective Approx.	0.68	0.63	0.60	0.62	0.60	0.64	0.62	0.48	0.56	0.74	0.71	0.72	0.70		0.59	0.70	0.63
		0.60	0.62	0.55	0.55	0.53	0.57	0.56	0.50	0.49	0.68	0.67	0.62	0.63		0.54	0.64	0.46
	Asp-Subj	0.65	0.65	0.69	0.66	0.64	0.64	0.58	0.65	0.64	0.33	0.25	0.24	0.25		0.32	0.36	0.26
		0.64	0.65	0.66	0.68	0.67	0.68	0.61	0.57	0.63	0.42	0.29	0.31	0.32		0.43	0.41	0.40
	Asp-Subj Approx	0.68	0.66	0.72	0.68	0.67	0.67	0.60	0.66	0.68	0.46	0.33	0.30	0.33		0.41	0.47	0.37
		0.66	0.66	0.69	0.69	0.68	0.68	0.62	0.62	0.66	0.51	0.37	0.40	0.39		0.49	0.49	0.48
Cross-Domain	Aspect		0.50	0.37	0.50	0.50	0.45	0.39	0.50	0.47		0.53	0.36	0.48	0.42	0.43	0.39	0.43
			0.34	0.25	0.32	0.28	0.37	0.28	0.30	0.34		0.47	0.34	0.45	0.56	0.39	0.36	0.40
	Aspect Approx.		0.69	0.57	0.70	0.65	0.62	0.58	0.63	0.65		0.63	0.43	0.59	0.64	0.55	0.49	0.57
			0.56	0.34	0.46	0.40	0.52	0.42	0.40	0.45		0.55	0.39	0.50	0.59	0.46	0.45	0.47
	Subjective		0.50	0.46	0.49	0.49	0.50	0.48	0.45	0.45		0.46	0.48	0.52	0.43	0.44	0.46	0.42
			0.45	0.44	0.42	0.41	0.39	0.39	0.36	0.38		0.43	0.44	0.47	0.49	0.42	0.45	0.42
	Subjective Approx.		0.70	0.71	0.68	0.70	0.69	0.70	0.66	0.66		0.74	0.76	0.73	0.69	0.69	0.72	0.69
			0.63	0.64	0.60	0.61	0.59	0.60	0.60	0.57		0.64	0.69	0.65	0.64	0.67	0.66	0.60
	Asp-Subj		0.66	0.68	0.67	0.62	0.70	0.60	0.62	0.64		0.17	0.20	0.37	0.15	0.35	0.19	0.24
			0.63	0.65	0.66	0.67	0.67	0.61	0.59	0.61		0.46	0.33	0.43	0.38	0.31	0.36	0.35
	Asp-Subj Approx.		0.69	0.71	0.68	0.65	0.70	0.64	0.65	0.66		0.32	0.30	0.47	0.15	0.43	0.37	0.39
			0.67	0.66	0.67	0.68	0.69	0.65	0.63	0.65		0.54	0.42	0.52	0.50	0.43	0.50	0.43

Table 3: F_1 measures serving as baselines for different experiments on the USAGE corpus. "10-fold cross-validation" refers to a cross-validation experiment on the full corpora or the product class specific subsets. "Cross-Domain" refers to a cross-domain experiment in which the model is trained on all data of the respective language except for the product class indicated in the table. This 'left-out' product class is used for testing, the results of which are included in the table.

compared to the detection of subjective phrases. German aspect detection has higher measures than for English (with 0.63 over 0.71 for the whole corpus), while there is no such big difference for subjective phrases (0.54 for English and 0.55 for German). The detection of relations yields comparable results for both languages (0.38 and 0.42). The results for coreferences are very low as the difference in annotation frequency between the two annotators already hints at. In order to exploit the coreference data, a deeper analysis of the annotation differences between the two annotators would be required.

4. Prediction Baseline

To provide a strong baseline for future systems to be developed based on the USAGE corpus, we perform experiments based on our previously published approach on aspect and subjective phrase-oriented fine-grained sentiment analysis (Klinger and Cimiano, 2013a; Klinger and Cimiano, 2013b). This method is based on an undirected probabilistic model with Markov Chain Monte Carlo inference which can perform prediction of aspects, subjective phrases and their relation in a joint manner or in a pipeline setting.

In more detail, spans of aspects and subjective phrases are represented similarly to a semi-Markov conditional random field (Sarawagi and Cohen, 2005). Each span variable can have a list of other spans to be related with. In the case of aspects, this can be used to model coreferences. In the case of subjective phrases, a reference to the target of the phrase is kept. In addition, each subjective phrase can be positive, negative, or neutral.

In the pipeline setting, a classifier estimating if an aspect and a subjective phrase are in relation is trained. We report the results under the assumption of perfect knowledge about aspect and subjective phrases, estimating the difficulty and performance for relation extraction in isolation. In our previous work, we detected a higher performance for aspect detection in the joint inference setting and a higher result for subjective phrase detection in the pipeline setting. We report the best results over both learning settings (joint and pipeline), as a productive system would obviously use a hy-

brid approach combining the inferences of both the joint and the pipeline model. However, the configuration is the same as reported by Klinger and Cimiano (2013b) for English. An adaptation of the system to other languages would demand for inclusion of a language-specific dependency parser, which is still future work. Thus, the German sentiment analysis system does not make use of features computed on the basis of dependency parse information.
The experiments performed are the following, each for German and for English separately:

1. **10-fold cross-validation on the full corpus:** including all product categories (denoted as 'full' in Table 2). Cross-validation is performed on the document level such that no characteristics of one text are shared between the respective training and validation sets.

2. **10-fold cross-validation for each product category:** *i. e.*, coffee machine, cutlery, microwave, toaster, trash can, vacuum cleaner, washing machine and dish washer for English, and Kaffeemaschine, Besteck, Mikrowelle, Mülleimer, Staubsauger and Waschmaschine for German. Toaster is not taken into account for German due to the small number of reviews, not being suitable for a cross-validation setting.[6] The aim of these experiments is to yield a class-specific baseline and in order to understand whether the difficulty of the task differs across product types.

3. **Cross-domain testing:** training on the reviews from all but one product class and test on the hold-out product class. These experiments are performed for each product category. The goal is to get insights about how easy a model trained on certain products can be transferred to a new product domain. It therefore allows for estimating if newly annotated corpora are actually needed when developing an opinion mining system for a specific product class.

The results of these experiments are summarized in Table 3. We report the F_1 measures with exact match between prediction and annotation and approximate (partial) match which regards an annotation which overlaps in at least one token with the gold standard annotation as a true positive. We take into account aspects, subjective phrases, and relations between both.
The results for 10-fold cross-validation are comparable to the figures published earlier (Klinger and Cimiano, 2013a; Klinger and Cimiano, 2013b). The recognition of aspects yields higher F_1 measures than subjective phrase recognition. Approximate measures are especially higher for subjective phrases as these are typically longer than aspects. While the performance of relation detection is similar for English, the values for German are generally much lower. Note that no dependency parser has been used for German and the set of informative features is therefore very limited.

[6] One might propose to perform the cross-validation on the segment level or sentence level instead of full review level. However, such approach is known to be overly optimistic (Pyysalo et al., 2008).

The results for the cross-domain transfer experiments are especially interesting. We observe a drop in performance when compared to 10-fold cross-validation, *e. g.*, for Annotator 1, cutlery's aspects drop from 0.53 to 0.37. Most aspect performance rates drop in English and German but some remain stable. In contrast, for subjective phrase detection, F_1 measures increase in the cross-domain setting for all sub-domains. These results license the conclusion that there is a fraction of shared vocabulary between the domains that is used in similar contexts and grammatical structures.

5. Availability and File Formats

The corpus is made available via document object identifier `10.4119/unibi/citec.2014.11` and therefore accessible via `http://dx.doi.org/10.4119/unibi/citec.2014.14` in a tabular separated file format which will be explained in the following. The annotation has been performed in Knowtator (Ogren, 2006), which is a plugin for the ontology building environment Protégé. The original files can be provided on request.
The corpus consists of a set of file quintuples, each quintuple being a `.txt` file providing necessary information to be able to retrieve the reviews from Amazon, two files with the extension `.csv` storing the offsets and attributes of aspects and subjective phrases for each annotator, and two `.rel` files with the information about relations between phrases for each annotator, respectively.
In detail, the `.txt` needs to be the input for a crawling workflow which is also provided. The output of that workflow will be another `.txt` file consisting of an ID and the review title and text. The exact guidelines are available online. Note that we do not publish the Amazon reviews but only the (stand-off) annotations.
The `.csv` files consist of a column indicating whether the phrase represents an aspect or a subjective phrase, the ID to denote the correct entry in the `.txt` file, left and right offset, the string representation and an ID uniquely identifying this phrase. In addition, subjective phrases can have an unknown, positive, negative, or neutral polarity and aspects can have the label *'foreign'*, each in a separate column.
The `.rel` file stores target-subjective phrase relations and coreference relations. It specifies the kind of the relation, provides the `.txt`-ID and the two participating phrase IDs. In addition, the textual representations of the phrases are repeated, which simplifies error detection and statistical evaluations.
A more detailed explanation is available on the download web site.

6. Summary, Conclusion and Future Research Opportunities

The corpus presented in this papers is, to the best of our knowledge, the largest manually annotated resource for fine-grained sentiment analysis with annotations of aspects, subjective evaluating phrases, their polarities and relations between them in two languages (German and English).
We are sure that this dataset will motivate and enable an array of novel research questions to be investigated and foster the development of sentiment analysis methods which work on

multiple languages (multilingual mode), approaches which exploit multilingual features in one model (joint model), or methods that allow one to train a sentiment analysis system in one language and apply it to another language (cross-lingual transfer mode). In addition, the selection of reviews from different product categories will enable research in the areas of domain adaption for such fine-grained annotations.

Acknowledgments

Roman Klinger has been funded by the "It's OWL" project ("Intelligent Technical Systems Ostwestfalen-Lippe", `http://www.its-owl.de/`), a leading-edge cluster of the German Ministry of Education and Research. We thank Frederike Strunz and Luci Fillinger for discussions and annotation and Robin Schiewer for implementation of data crawling and management. This work was conducted using the Protégé resource, which is supported by grant GM10331601 from the National Institute of General Medical Sciences of the United States National Institutes of Health.

7. References

Katarina Boland, Andias Wira-Alam, and Reinhard Messerschmidt. 2013. *Creating an Annotated Corpus for Sentiment Analysis of German Product Reviews*, volume 2013/05. GESIS Institute.

Simon Clematide, Stefan Gindl, Manfred Klenner, Stefanos Petrakis, Robert Remus, Josef Ruppenhofer, Ulli Waltinger, and Michael Wiegand. 2012. MLSA – A Multi-layered Reference Corpus for German Sentiment Analysis. In *Proceedings of the 8th International Conference on Language Ressources and Evaluation (LREC)*, pages 3551–3556, Marrakesh, Marocco, May. European Language Resources Association (ELRA).

Jakob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.

Xiaowen Ding, Bing Liu, and Philip S. Yu. 2008. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 231–240, New York, NY, USA. ACM.

Gayatree Ganu, Noemie Elhadad, and Amélie Marian. 2009. Beyond the stars: Improving rating predictions using review text content. In *Twelfth International Workshop on the Web and Databases (WebDB 2009)*, Providence, Rhode Island, USA, June.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, New York, NY, USA. ACM.

Jason S. Kessler, Miriam Eckert, Lyndsie Clark, and Nicolas Nicolov. 2010. The 2010 ICWSM JDPA Sentiment Corpus for the Automotive Domain. In *4th International AAAI Conference on Weblogs and Social Media Data Workshop Challenge (ICWSM-DWC 2010)*.

Roman Klinger and Philipp Cimiano. 2013a. Bi-directional inter-dependencies of subjective expressions and targets and their value for a joint model. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 848–854, Sofia, Bulgaria, August. Association for Computational Linguistics.

Roman Klinger and Philipp Cimiano. 2013b. Joint and pipeline probabilistic models for fine-grained sentiment analysis: Extracting aspects, subjective phrases and their relations. In *IEEE International Conference on Data Mining Workshops (ICDMW)*, Dallas, TX, USA.

Himabindu Lakkaraju, Chiranjib Bhattacharyya, Indrajit Bhattacharya, and Srujana Merugu. 2011. Exploiting coherence for the simultaneous discovery of latent facets and associated sentiments. In *SIAM International Conference on Data Mining*, pages 498–509. SIAM / Omnipress.

Fangtao Li, Minlie Huang, and Xiaoyan Zhu. 2010. Sentiment analysis with global topics and local dependency. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, pages 1371–1376, Atlanta, Georgia, USA.

Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. 2013. Semeval-2013 task 2: Sentiment analysis in twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Philip V. Ogren. 2006. Knowtator: a protégé plug-in for annotated corpus construction. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 273–275, Morristown, NJ, USA. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics, Main Volume*, pages 271–278, Barcelona, Spain, July.

Sampo Pyysalo, Rune Sætre, Jun'ichi Tsujii, and Tapio Salakoski. 2008. Why biomedical relation extraction results are incomparable and what to do about it. In Tapio Salakoski, Dietrich Rebholz-Schuhmann, and Sampo Pyysalo, editors, *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine*, pages 149–152. Turku Centre for Computer Science (TUCS).

Josef Ruppenhofer, Swapna Somasundaran, and Janyce Wiebe. 2008. Finding the sources and targets of subjective expressions. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Sunita Sarawagi and William W. Cohen. 2005. Semimarkov conditional random fields for information extraction. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 1185–1192. MIT Press, Cambridge, MA.

Asad Sayeed, Jordan Boyd-Graber, Bryan Rusk, and Amy

Weinberg. 2012. Grammatical structures for word-level sentiment detection. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 667–676, Montréal, Canada, June. Association for Computational Linguistics.

Han-Xiao Shi and Xiao-Jun Li. 2011. A sentiment analysis model for hotel reviews based on supervised learning. In *Proceedings of the International Conference on Machine Learning and Cybernetics*, Guilin, July.

Damiano Spina, Edgar Meij, Maarten de Rijke, Andrei Oghina, Minh Thuong Bui, and Mathias Breuss. 2012. Identifying entity aspects in microblog posts. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pages 1089–1090, New York, NY, USA. ACM.

Oscar Täckström and Ryan McDonald. 2011. Semi-supervised latent variable models for sentence-level sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 569–574, Portland, Oregon, USA, June. Association for Computational Linguistics.

Ivan Titov and Ryan McDonald. 2008. A Joint Model of Text and Aspect Ratings for Sentiment Summarization. In *Proceedings of ACL-08: HLT*, pages 308–316, Columbus, Ohio, June. Association for Computational Linguistics.

Cigdem Toprak, Niklas Jakob, and Iryna Gurevych. 2010. Sentence and expression level annotation of opinions in user-generated discourse. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 575–584, Uppsala, Sweden, July. Association for Computational Linguistics.

Peter D. Turney. 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424, Philadelphia, USA, July.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210.

Janyce Wiebe. 2000. Learning subjective adjectives from corpora. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 735–740. AAAI Press.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2009. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399–433, September.