

Investigating the Applicability of current Machine-Learning based Subjectivity Detection Algorithms on German Texts

Malik Atalla

Technische Universität Berlin
atalla@cs.tu-berlin.de,

Christian Scheel and Ernesto William De Luca and Sahin Albayrak

DAI-Labor, Technische Universität Berlin

{christian.scheel, ernesto.deluca, sahin.albayrak}@dai-labor.de

Abstract

In the field of subjectivity detection, algorithms automatically classify pieces of text into fact or opinion. Many different approaches have been successfully evaluated on English or Chinese texts. Nevertheless the assumption that these algorithms equally perform on all other languages cannot be verified yet. It is our intention to encourage more research in other languages, making a start with German. Therefore, this work introduces a German corpus for subjectivity detection on German news articles. We carry out this study in which we choose a number of state of the art subjectivity detection approaches and implement them. Finally we show and compare these algorithms' performances and give advice on how to use and extend the introduced dataset.

1 Introduction

The detection of subjective statements in natural language texts is necessary for the analysis of opinions and the extraction of facts for knowledge retrieval. The continuously increasing number of natural language texts on the Internet and the need for opinion detection and fact retrieval makes research on subjectivity detection more and more important.

The economic impact is rising just as much. The Internet has long turned into an open platform in which everybody can participate and contribute his or her share of opinions.

Subjectivity detection also affects upcoming fields of research like knowledge retrieval. Crawlers have to distinguish between objective and subjective texts in order to extract given facts only from the objective parts.

The other way round, in the field of opinion

analysis, many approaches are supposed to be applied on subjective texts. In *polarity classification* pieces of text are classified into complementary viewpoints. In this field of research facts are considered noise to the problem. So, finding the subjective parts beforehand can increase the accuracy of such a classifier (Pang and Lee, 2004). Subjectivity detection can support *question-answering systems*. The knowledge about the subjectivity of sentences and sections is important, especially for complex questions that cannot be answered with a single fact, but should rather treat different viewpoints on an issue. Also, subjectivity detection can be useful for *text summarization* which may want to list facts separately from different viewpoints.

In conclusion it can be stated that subjectivity detection is one of the most important pre-processing steps for many IR applications. Such pre-processing has to be language independent or at least the drawbacks for each language have to be known. Hence, the overall goal of this work is to investigate the differences in subjectivity detection in different languages, starting with German and English.

In this work we evaluate a number of machine learning based subjectivity detection approaches on German news texts and on the MPQA corpus, which is the current English standard corpus for subjectivity annotations. We focus on supervised learning approaches for sentence-wise binary classification between subjective and non-subjective sentences without polarity.

After giving an overview over the state of the art in subjectivity detection, we provide details about the MPQA corpus. Afterwards we introduce the *Subjectivity in News Corpus (SNC)*, which was created in the course of this work. The corpus, precisely the German part of the corpus, was annotated in such a way that it provides maximum compatibility with MPQA. Evaluation results on both corpora are compared to conclude if current

machine learning based subjectivity detection approaches are equally applicable on both languages. In the concluding part of this work we show which features and ideas are better fit to detect subjectivity in which language and give advice on how to handle subjectivity detection on German texts.

2 Related Work

The field of subjectivity detection can be roughly divided into lexical approaches and machine learning approaches. The lexical approaches are those that incorporate some sort of annotated dictionary. The machine learning approaches represent statements as feature vectors in order to learn to distinguish between subjective and objective statements. In this work, we decided to focus only on supervised learning approaches, which are reviewed in the remainder of this section. In the course of this work a new corpus was created. Therefore this section is completed by a description of corpora for subjectivity detection.

2.1 Subjectivity Detection

Yu and Hatzivassiloglou (2003) presented the first fully supervised machine learning approach in the field of subjectivity detection. As training data a set of Wall Street Journal (WSJ) articles with attached categories is used. In this work the use of a Naive Bayes classifier to distinguish subjective and objective texts has been proposed. These texts were represented by features like extracted unigrams, bigrams, trigrams and POS-tags.

The latter approach does not take the context of a sentence into account. A sentence is more likely to be subjective if the neighboring sentences are subjective as well. Pang and Lee try to tackle this issue with their minimum cut approach (Pang and Lee, 2004). They propose a similar machine learning approach, but additionally assign an association value for each pair of sentences, which is based on the distance of the two sentences in the text. This value encourages the classifier to assign the same label for sentences with a small distance. The structure of the article including the predictions and association values about the sentences are represented by a graph and the final classification is determined by a minimum-cut algorithm.

Wiebe et al. investigate another promising feature in (Wiebe et al., 2004), namely the one of "unique words". It is shown that "apparently, people are creative when they are being opinionated".

Note, that this work is a statistical study, where this idea was not carried on as an additional feature in a classifier. The study is based on a corpus of WSJ articles. It is argued that rare words are more likely to occur in opinionated pieces than in objective texts.

Another problem in subjectivity detection and opinion mining in general is the domain and context dependency of many words. It is true in many cases that a word can express an opinion in some context, but be perfectly neutral in another. In their entry for the 2006 Blog Trec, Yang et al. present an approach to this problem (Yang et al., 2006). They use two different sets of training data. One of them contains text about movies, the other about electronic devices. A classifier is trained with each of these data sets and only those features that were useful in both cases are extracted and used in the final classifier. Their rationale is to achieve a feature set that only contains domain-independent features.

Another approach at domain-dependency has been presented by Das and Bandyopadhyay (2009). Instead of discarding domain-dependent words, which could decrease recall, it is tried to determine the topic of a text and use it as a feature in their classifier. Therefore, an additional pre-processing step has been introduced, clustering the training data to determine all possible topics. It is claimed that this feature increases the performance on the MPQA corpus by 2.5%.

Another interesting feature has been proposed in the approach by Chen et al. (2007), where so-called long-distance-bigrams are introduced. Long-distance-bigrams are bigrams which are not consisting of neighboring terms, but of pairs of terms with a certain, fixed distance. 1-distance bigrams would be the same as regular bigrams, 2-distance bigrams have one term in between and so forth. They report a slightly better classification result using a feature set consisting of unigrams, bigrams and 2-distance bigrams, than by just using unigrams. This is interesting in the context of Pang et al. (2002) reporting that using only unigrams performs better than using a combination of unigram and n-gram features.

Banea et al. (2008) wondered if the large amount of NLP tools that already exist for English texts can be exploited for other languages and presented an approach based on machine translation. Different experiments are carried out with

English data and data that was automatically translated into Romanian. Machine translated data is either used as training or as test data. Encouraging results are achieved and it is claimed that machine translation is a viable alternative to creating resources and tools for subjectivity detection in other languages.

2.2 Subjectivity Annotated Corpora

For the English language there are currently two major corpora with subjectivity annotations. The first corpus is the movie data corpus presented in (Pang and Lee, 2004). It contains 5000 subjective and 5000 objective sentences, which have been automatically collected. The subjective sentences are extracted from movie reviews from *rottentomatoes.com* and the objective ones are from plot summaries from *imdb.com*. One drawback of this corpus is that it does not contain articles, but only a list of sentences without context.

The second corpus is the MPQA corpus¹, which is a 16000-sentence corpus made up of news articles which are tagged with a complex set of subjectivity annotations. The annotations not only mark subjective statements, but also their polarity, intensity, speaker and other things. Based on these fine-grained annotations the subjectivity of each sentence can be determined. The researchers consider a sentence to be subjective if it contains a *private state*, a subjective *speech event* or an *expressive subjective element*. Otherwise the sentence is objective (Wiebe, 2002). With the term *private state*, they refer to the definition by Quirk et al. (Quirk, 1985) which, according to them, includes mental or emotional states such as opinions, beliefs, thoughts, feelings, emotions, goals, evaluations and judgments.

A speech event refers to a speaking event, such as direct or indirect speech. Speech events are not automatically considered subjective. They can be objective if the credibility of the source is not in doubt and their content is portrayed as fact. The term *expressive subjective element* is based on a publication by Banfield (1982). It is to be used for statements that *"express private states, but do not explicitly state or describe the private state"*.

3 Settings

In this section we first introduce the Subjectivity in News Corpus (SNC), a set of corpora for subjectivity

detection. The German part of this corpus, namely *SNC.de*, which was created for this work, is based on German news. *SNC.de* marks the first corpus in a line of upcoming similar corpora for additional languages to be created in the near future. In the second part of this section, we present selected state of the art approaches. The evaluation results of these approaches will provide the baseline for future approaches.

3.1 Structure of the SNC Corpus

Although the introduced corpus will be a multi lingual corpus, the descriptions in this section focus on the German part (*SNC.de*). In order to be able to evaluate the approaches on German texts, we created an annotated, German corpus. The objective was to provide maximum compatibility with the MPQA corpus. So we abided by the annotation manual as closely as possible and also chose the topics for the texts similarly. Just like in the MPQA corpus, the articles in our corpus are ordered by topic. If we wanted to be completely consistent with MPQA we should also apply the same annotation set. The problem was that many of the very detailed annotations were of no relevance to this work. So we decided to only make binary, sentence-wise annotations, based on the definition of sentence subjectivity presented in (Riloff et al., 2003).

The corpus annotation was carried out by a single annotator using the graphical interface of the GATE² tool. So the entire corpus is saved in GATE's own xml serialization format. The annotator attached to each sentence one of the annotations "subjective" or "non-subjective". Suggestions for sentence splitting were provided by the user interface to speed up the annotation process. The annotated texts were saved as entire articles, in order to preserve the context of each sentence. The articles to annotate were randomly chosen from current world news in German.

A comparison of the SNC corpus and the MPQA corpus is given in Table 1.

3.2 Selected Approaches

In this part we list the approaches selected for evaluation and explain why they have been chosen.

Unique Words The feature of unique words, investigated by Wiebe et al. (2004), has never been tried out in a classifier. So, we will use a counter

¹Referring to version 2.0 in all explanations.

²<http://gate.ac.uk/>

Table 1: Text Statistics about the Corpora.
c: characters; *s*: sentences; *a*: article

	SNC	MPQA
Article Statistics		
<i>a</i> in the corpus	278	692
Avg. <i>a</i> length in <i>s</i>	24,6	22,8
Std.-dev. <i>a</i> lengths	14,3	27,1
Shortest <i>a</i> in <i>s</i>	3	2
Longest <i>a</i> in <i>s</i>	80	275
Sentence Statistics		
<i>s</i> in the corpus	6848	15802
Subjective <i>s</i>	3458	7675
Objective <i>s</i>	3390	8127
Avg. <i>s</i> length <i>c</i>	124,9	132,4
Std.-Dev. <i>s</i> lengths	67,0	80,1
Same annot. as neighbors	52,0%	59,1%
Avg. length subj. <i>s</i> in <i>c</i>	133,0	150,1
Avg. length obj. <i>s</i> in <i>c</i>	116,6	115,0
Word Statistics		
tokens in the corpus	141144	403116
words in the corpus	120128	342165
distinct word forms	18968	22736

for infrequent words in our classifier and evaluate if it contributes to the separation of the two classes. We can use the Leipzig Corpora Collection (LCC)³ and the BNC to figure out which words of each language classify as rare. We decided to define a rare word form, as one that is not one of the 600.000 most frequent word forms in its language.

POS-Trigrams Santini’s approach for genre detection (Santini, 2004) has not yet been picked up by researchers from subjectivity detection. The main idea is to use trigrams of POS-tags as features. We limit the number of features by taking the most frequent POS-trigrams and varying the maximum number.

Unigrams, Bigrams, Trigrams and POS-Tags

The machine learning approach by Yu and Hatzivassiloglou (2003) can be considered a standard for later approaches. Sentences are represented by a feature vector which is taken to train a model for separating objective from subjective sentences. A Naive Bayes classifier is used and the feature vector contains unigrams, bigrams, trigrams and POS-tags.

³<http://corpora.informatik.uni-leipzig.de/download.html>

Minimum-Cut Classifier Pang et al. presented the first idea to incorporate context into the classification decision (Pang and Lee, 2004). This classifier is not only based on the content of a sentence, but also on its neighboring sentences.

Long-Distance-Bigrams Chen et al. (2007) proposed the feature of long-distance-bigrams which is a novel idea and therefore worth investigating.

Machine-Translation of Training Data This work aims to investigate if a separate research effort is necessary for every language, or if the existing tools of the English language can be exploited for other languages with acceptable accuracy. Just like in the publication of Banea et al. (2008) we will automatically translate the MPQA corpus, in our case into German, and use it as training or test data. We denote the translation MPQA-G.

4 Experimental Results

Experiments were performed according to selected approaches of Section 3.2. For the experiments with SVMs we chose a linear SVM and used the implementation of Libsvm⁴. For the Naive Bayes classifier the weka⁵ implementation was used. The Minimum-Cut classifier was implemented by ourselves based on the description in the publication. For the creation and manipulation of graphs we used the JUNG⁶ API.

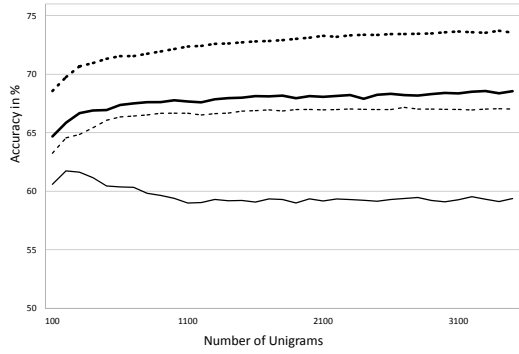
For the features of the **baseline classifier** we chose POS-tags and a limited number of the most frequent unigrams of the training corpus. It is a simple feature set which nevertheless performs strongly compared to other approaches. We carried out a number of experiments with a Naive Bayes classifier and an SVM and a variable number of unigrams as shown in Fig. 1a. It can be observed that the SVM on the English corpus is rising slightly more steeply than the SVM on the German corpus. This indicates that large numbers of unigrams are more useful for English texts than for German ones.

For the following experiments, the baseline classifier shall be the one using 1500 unigrams. This number seems a reasonable trade-off between computational cost and accuracy.

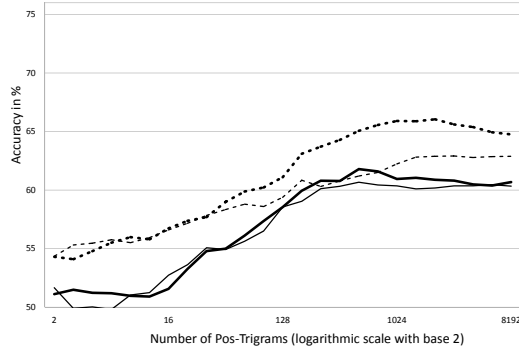
⁴www.csie.ntu.edu.tw/~cjlin/libsvm/

⁵www.cs.waikato.ac.nz/ml/weka/

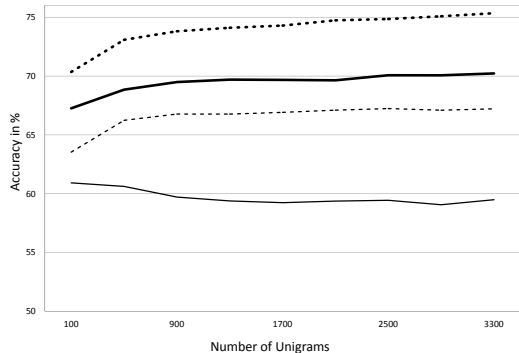
⁶<http://jung.sourceforge.net/>



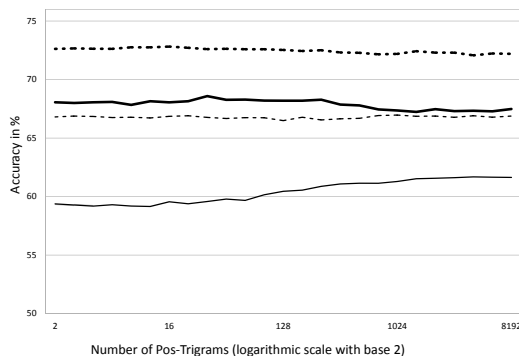
(a) baseline classifier



(b) POS-Trigrams (standalone)



(c) Pang's Minimum-Cut Classifier



(d) POS-Trigrams (merged feature vectors)

— SNC, SVM - - SNC, Naive Bayes ··· MPQA2, SVM - · - MPQA2, Naive Bayes

Figure 1: Comparing four approaches for different languages (a–d). Evaluations were performed on the English corpus MPQA and the German corpus SNC.

For every approach where it is applicable we carry out two types of experiments. The first type we would like to call the standalone experiment, in which we use exactly the feature vector described for the approach. The second experiment, the merged-feature-vector-experiment, is done by merging the feature set of our baseline classifier with that of the approach. The second experiment allows us to evaluate if an approach can improve a simple but effective classifier. This is important because some approaches may not be intended as full-blown classifiers, but rather as additional ideas to existing classifiers.

All experiments were carried out by 5-fold cross validation, except some of the experiments with machine-translated data, which is explained separately in the respective section. In all tables the column denotes the corpus used for the cross-validation and the row denotes the experiment's setting, i.e. feature set and classifier. Most experiments are compared to the results of the baseline classifier and to the corpus baseline. The latter is the percentage of the label that occurs more often in the corpus.

4.1 Unique Words

Table 2: Unique Words Feature Experiments.

UW: Unique Words; BUW: Base+Unique Words

	SNC	MPQA
<i>Corpus Baseline</i>	50,50	51,40
UW Training Set (NB)	49,46	51,24
UW Training Set (SVM)	50,76	51,10
UW Statistics (NB)	49,72	51,79
UW Statistics (SVM)	49,31	49,26
UW Tr.+Stats (NB)	48,01	51,65
UW Tr.+Stats (SVM)	53,79	49,81
<i>Baseline Classifier (NB)</i>	59,21	66,84
<i>Baseline Classifier (SVM)</i>	67,99	72,72
BUW Training (NB)	59,22	65,87
BUW Training (SVM)	67,76	72,20
BUW Statistics (NB)	59,18	66,80
BUW Statistics (SVM)	68,06	72,68
BUW Tr.+Stats (NB)	59,20	65,99
BUW Tr.+Stats (SVM)	67,91	72,68

For both experiment settings, the standalone setting and the merged-feature-vector setting, we carried out three variants with different features. In the first variant we used a counter for words, that are unique in the scope of the training data,

in the second one a counter for words unique according to statistics about the BNC or the LCC and thirdly a feature vector with both of the latter features (see Table 2).

The standalone setting of the unique words feature is not a serious attempt at a classifier. It contains only one or two attributes which is obviously not enough to separate the classes. But we can compare them to the corpus baselines to determine if the features contain any useful information.

Contrary to the claim made in (Wiebe et al., 2004), our experiments could not verify that unique words are a useful feature to detect subjectivity. Using only unique words results in accuracies very close to 50%, except for one setting, but the success for that setting could not be repeated when applying unique words additionally to baseline features. Since the feature does not seem to be useful for either language, no difference could be detected between them.

4.2 POS-Trigrams

When using only trigrams of POS-tags (Fig. 1b), most of the experiments stay far behind their respective baseline classifiers. The only classifier that reaches above the baseline classifier is NB on the German corpus, but only by a small margin.

The standalone experiments indicate that the POS-trigram approach is more useful for German texts than for English ones, when the baseline classifier is taken as comparison value. The best value from the German SVM is slightly closer to the baseline than the best value of the English SVM and for the NB classifiers it is even clearer. The German NB performs better than baseline and the English one significantly worse.

The chart about the experiments with the merged feature vectors (Fig. 1d) illustrates that all results are almost identical to the respective baseline result. Both the English classifiers and the German SVM perform exactly like the baseline classifier in all experiments. This indicates that the POS-trigrams do not contain much useful information that is not already contained in the baseline features. The only exception is again the Naive Bayes classifier applied on the German corpus, which considerably exceeds the baseline classifier’s accuracy. The diagram does not show a clear difference between the languages, but it does indicate that POS-trigrams are more useful for German with a classifier on the German cor-

pus being the only one above baseline.

4.3 Unigrams, Bigrams, Trigrams and POS-Tags

Table 3: Experiments with Unigrams, Bigrams, Trigrams and POS-Tags

	SNC	MPQA
<i>Baseline Classifier (NB)</i>	59,21	66,84
<i>Baseline Classifier (SVM)</i>	67,99	72,72
Naive Bayes	59,68	66,92
SVM	69,80	74,24

This experiment is carried out with feature vectors containing all unigrams, bigrams, trigrams and POS-tags that occur in the training data, which amounts to a very long vector. The setting with merged feature vectors is not applicable for this experiment because the feature set is a superset of the baseline classifier’s feature set.

Table 3 shows that the SVM performs clearly better than the baseline classifier for both corpora, unlike the Naive Bayes classifier which does not show any improvement. The approach is based on a huge amount of different features. The SVM is able to handle this number of features better than the Naive Bayes classifier. Since the features also contain a large amount of redundant data, the Naive Bayes classifier does not perform as well.

For both languages the classifiers achieve about the same distance from the baseline classifiers, namely roughly 2%. So it appears that there is no language dependence for this feature set, but in fact it is much harder to achieve an improvement of 2% when starting from a higher baseline. So, the fact that the distances to the baselines are equal might actually be an indication that the approach is more useful for the English corpus.

4.4 Minimum-Cut Classifier

Fig. 1c shows the results of the experiments with our native implementation of the minimum-cut classifier. The feature set we used is the same as in the baseline classifier. The parameter c , which determines how much influence the context of a sentence has on its classification, was set to the value determined in a parameter optimization. The values are different for SVM (2^{-3}) and Naive Bayes ($2^{-1.9}$). It can be seen that when the SVM is used as base-predictor by the minimum-cut classifier, the accuracy is well above the one of the baseline

classifier, but when Naive Bayes is used, the accuracy increased only minimally.

With respect to the difference between the two languages, there is certainly none for the Naive Bayes classifiers. But for the SVMs it seems that the classifier for the German corpus achieves a bigger distance to its baseline classifier than the classifier for the English corpus. The average distances to the baseline classifiers are 1.85% and 1.64% respectively. This might signify that the approach is better fit for German texts, but the reason for the difference might also be that the English baseline classifier is much better in the first place and there is not as much room for improvement as for the German baseline classifier.

4.5 Long-Distance Bigrams

Table 4: Long-Distance Bigrams Experiments; (DB: Distance Bigrams)

	SNC	MPQA
<i>Corpus Baseline</i>	<i>50,50</i>	<i>51,40</i>
2-DB (NB)	55,55	60,57
2-DB (SVM)	60,29	66,74
2+3-DB (NB)	56,16	61,11
2+3-DB (SVM)	60,63	66,51
2+3+4-DB (NB)	56,20	61,22
2+3+4-DB (SVM)	60,41	66,08
<i>Baseline Classifier (NB)</i>	<i>59,21</i>	<i>66,84</i>
<i>Baseline Classifier (SVM)</i>	<i>67,99</i>	<i>72,72</i>
Base+2-DB (NB)	59,67	65,91
Base+2-DB (SVM)	67,59	72,78
Base+2+3-DB (NB)	61,06	66,11
Base+2+3-DB (SVM)	67,29	72,34
Base+2+3+4-DB (NB)	61,63	66,11
Base+2+3+4-DB (SVM)	66,65	72,05

Table 4 shows the results of using only long-distance-bigrams as features. The types of experiments we carried out are similar to those in described in (Chen et al., 2007). First we used only 2-distance-bigrams as features. Then we extended the feature set by adding 3- and 4-distance-bigrams.

All SVMs perform much worse than the baseline classifiers and their distance to that value is about the same for all settings and corpora. Naive Bayes also performs worse than baseline, but a difference between the English and German corpus can be observed. For the German corpus the distances to the baseline classifier are between 3% and 4%, whereas for the English corpora the dis-

tance is between 5% and 6%. This observation is affirmed when we apply long-distance-bigrams together with the baseline features.

4.6 Corpus Translation

Table 5: Machine-Translated Data Experiments.

	MPQA-G	SNC CV
Baseline Class. (NB)	57,70	59,21
Baseline Class. (SVM)	63,43	67,99
Base+Most Freq. (NB)	59,80	60,05
Base+Most Freq. (SVM)	61,34	64,22
POS-Tri 2048 (NB)	57,83	60,18
POS-Tri 2048 (SVM)	58,56	60,88
Uni+Bi+Tri+POS (NB)	56,31	59,68
Uni+Bi+Tri+POS (SVM)	63,52	69,80
Baseline (MinCut-NB)	57,71	59,24
Baseline (MinCut-SVM)	63,41	69,68

Banea et al. proposed machine translation as a way of saving the effort to create NLP tools in languages other than English. Our experiments with machine translated data are shown in Table 5. The middle column shows the results for using our translation of the MPQA corpus (MPQA-G) as training data and SNC as test data. The right column gives the upper bounds of accuracy that can be achieved, which we determined by cross validation on the test data.

It can be seen in many of the settings that the translation approach comes rather close to its upper bound. For many settings the difference is only 2%. We have to acknowledge though, that exactly for those settings where the upper bound is high, the distance to the upper bound is also pretty large. There are three settings with 68% and almost 70% accuracy in CV, but using the translated corpus for training achieves only 63,5% in all of these settings. That means the highest accuracy of the translation approach is significantly lower than the highest cross validation on the test data.

5 Conclusion

While searching for the best machine learning approach for subjectivity detection on multi-lingual texts, we have observed several differences concerning the quality of subjectivity detection in different languages. These differences depend on the chosen features for the individual machine learning approach, but we have also seen that the differences along the languages are very subtle. Most approaches do not show a clear preference for one

specific language. Also, the differences are difficult to interpret because the results of the baseline classifiers for each language are very far apart from each other and the variety caused by different classifiers is much bigger than the language dependency.

The evaluated approaches performed better on English texts than on German. Whenever an approach improved according to the English baseline, this approach also improved according to the German baseline.

Focusing on the chosen features, we have seen that using large numbers of unigrams is more useful for English, compared to using only POS-tags. Since there is no objective comparison value, it remains unclear if this means that POS-tags are less useful or unigrams more useful for English. We have furthermore observed that POS-trigrams are more useful for German. These two aspects indicate that German subjectivity is more grammaticalized as opposed to English subjectivity which is more based in lexis.

Another feature that seems to be more useful for German are the long-distance-bigrams.

The efficacy of the minimum-cut approach strongly depends on the distribution of class labels in the articles. If many sentences are tagged with the same labels as their neighbors, the approach will be very useful, otherwise it will not. In the evaluation we found that the approach seems to work slightly better for MPQA than for SGN. The statistics about the corpora confirm this, indicating that MPQA has more consecutive sentences with equal annotations (see Table 1).

Another important observation we made is that machine-translation of training data is not a viable alternative to manually creating it. The results only came close to their comparison values for approaches that did not perform so well in the first place. The effectiveness of the approach depends of course on the quality of the translation. So it can be expected that it becomes more useful in the future as machine-translation improves. On the other hand, the quality of translations between English and German is quite high compared to other language pairs.

Summarizing we can state that there is no subjectivity detection approach which is more suitable for German texts than for English texts.

The dataset was published at <http://130.149.154.91/corpus/snc/SNC.de.zip>.

References

- Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. 2008. Multilingual subjectivity analysis using machine translation. In *Proceedings of the EMNLP '08*, pages 127–135, Morristown, NJ, USA. Association for Computational Linguistics.
- Ann Banfield. 1982. *Unspeakable sentence*. Routledge.
- Bo Chen, Hui He, and Jun Guo. 2007. Language feature mining for document subjectivity analysis. In *ISDPE '07: First International Symposium on Data, Privacy, and E-Commerce*, pages 62–67, Washington, DC, USA. IEEE Computer Society.
- Amitava Das and Sivaji Bandyopadhyay. 2009. Subjectivity detection in english and bengali: A crf-based approach. In *Proceedings of ICON 2009*, Hyderabad.
- Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proc. of the ACL*, pages 271–278.
- Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the EMNLP '02*, pages 79–86.
- Randolph Quirk. 1985. *A comprehensive grammar of the English language*. Longman.
- Ellen Riloff, Janyce Wiebe, and Theresa Wilson. 2003. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 25–32, Morristown, NJ, USA. Association for Computational Linguistics.
- M. Santini. 2004. A shallow approach to syntactic feature extraction for genre classification. In *Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics*.
- Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational Linguistics*, 30(3):277–308, January.
- Janyce Wiebe. 2002. Instructions for annotating opinions in newspaper articles. Technical report, University of Pittsburgh.
- Hui Yang, Luo Si, and Jamie Callan. 2006. Knowledge transfer and opinion detection in the TREC2006 blog track. In *Proceedings of TREC*.
- Hong Yu and Vasileios Hatzivassiloglou. 2003. Separating facts from opinions and identifying the polarity of opinion sentences. In *Proc. of EMNLP '03*, pages 129–136.