Instance Selection Improves Cross-Lingual Model Training for Fine-Grained Sentiment Analysis

Roman Klinger[*‡]
[*]Institute for Natural Language Processing
University of Stuttgart
70569 Stuttgart, Germany

Philipp Cimiano[‡]
[‡]Semantic Computing Group, CIT-EC
Bielefeld University
33615 Bielefeld, Germany

roman.klinger@ims.uni-stuttgart.de
cimiano@cit-ec.uni-bielefeld.de

Abstract

Scarcity of annotated corpora for many languages is a bottleneck for training fine-grained sentiment analysis models that can tag aspects and subjective phrases. We propose to exploit statistical machine translation to alleviate the need for training data by projecting annotated data in a source language to a target language such that a supervised fine-grained sentiment analysis system can be trained. To avoid a negative influence of poor-quality translations, we propose a filtering approach based on machine translation quality estimation measures to select only high-quality sentence pairs for projection. We evaluate on the language pair German/English on a corpus of product reviews annotated for both languages and compare to in-target-language training. Projection without any filtering leads to 23 % F_1 in the task of detecting aspect phrases, compared to 41 % F_1 for in-target-language training. Our approach obtains up to 47 % F_1. Further, we show that the detection of subjective phrases is competitive to in-target-language training without filtering.

1 Introduction

An important task in fine-grained sentiment analysis and opinion mining is the extraction of mentioned aspects, evaluative subjective phrases and the relation between them. For instance, in the sentence

> "I really like the display but the battery seems weak to me."

the task is to detect evaluative (subjective) phrases (in this example "really like" and "seems weak") and aspects ("display" and "battery") as well as their relation (that "really like" refers to "display" and "seems weak" refers to "battery").

Annotating data for learning a model to extract such detailed information is a tedious and time-consuming task. Therefore, given the scarcity of such annotated corpora in most languages, it is interesting to generate models which can be applied on languages without manually created training data. In this paper, we perform annotation projection, which is one of the two main categories for cross-language model induction (next to direct model transfer (Agić et al., 2014)).

Figure 1 shows an example of a sentence together with its automatically derived translation (source language on top, target language on bottom) and the alignment between both. Such an alignment can be used to project annotations across languages, *e. g.*, from a source to target language, to produce data to train a system for the target language. As shown in the example, translation errors as well as alignment errors can occur. When using a projection-based approach, the performance of a system on the target language crucially depends on the quality of the translation and the alignment. In this paper we address two questions:

- What is the performance on the task when training data for the source language is projected into a target language, compared to an approach where training data for the target language is available?
- Can the performance be increased by selecting only high-quality translations and alignments?

Towards answering these questions, we present the following contributions:

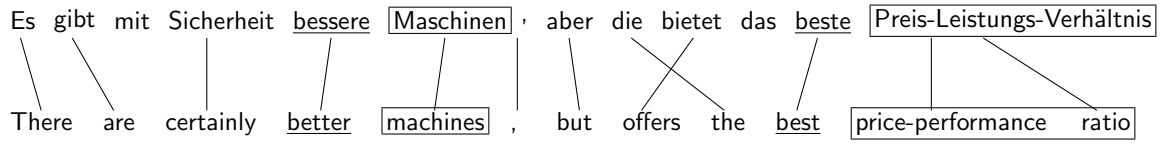

Figure 1: Example for the projection of an annotation from the source language to the target language. The translation has been generated with the Google translate API (https://cloud.google.com/translate/). The alignment is induced with FastAlign (Dyer et al., 2013).

- We propose to use a supervised approach to induce a fine-grained sentiment analysis model to predict aspect and subjective phrases on some target language, given training data in some source language. This approach relies on automatic translation of source training data and projection of annotations to the target language data.

- We present an instance selection method that only selects sentences with a certain translation quality. For this, we incorporate different measures of translation and alignment confidence. We show that such an instance selection method leads to increased performance compared to a system without instance selection for the prediction of aspects. Remarkably, for the prediction of aspects the performance is comparable to an upper baseline using manually annotated target language data for training (we refer to the latter setting as *in-target-language training*).

- In contrast, for the prediction of subjective phrases, we show that, while a competitive result compared to target language training can be observed when training with the projected training data, there is no beneficial effect of the filtering.

In the following, we describe our methodology in detail, including the description of the machine translation, annotation projection, and quality estimation methods (Section 2), and present the evaluation on manually annotated data (Section 3). Related work is discussed in Section 4. We conclude with Section 5 and mention promising future steps.

2 Methods

2.1 Supervised Model for Aspect and Subjective Phrase Detection

We use a supervised model induced from training data to detect aspect phrases, subjective (evaluative) phrases and their relations. The structure follows the proposed pipeline approach by Klinger and Cimiano (2013).[1] However, in contrast to their work, we focus on the detection of phrases only, and exploit the detection of relations only during inference, such that the detection of relations has an effect on the detection of phrases, but is not evaluated directly.

The phrase detection follows the idea of semi-Markov conditional random fields (Sarawagi and Cohen, 2004; Yang and Cardie, 2012) and models phrases as spans over tokens as variables. Factor templates for spans of type *aspect* and *subjective* take into account token strings, prefixes, suffixes, the inclusion of digits, and part-of-speech tags, both as full string and as bigrams, for the spans and their vicinity. In addition, the length of the span is modeled by cumulative binning. The relation template indicates how close an aspect is to a subjective phrase based on token distance and on the length of the shortest path in the dependency tree. The edge names of the shortest path are also included as features. It is further checked if no other noun than the aspect is close to the subjective phrase.

Inference during training and testing is done via Markov Chain Monte Carlo (MCMC). In each sampling step (with options of adding a span, removing a span, adding an aspect as target to a subjective phrase), the respective factors lead to an associated model score. The model parameters are adapted based on sample rank (Wick et al., 2011) using an objective function which computes the fraction of correctly predicted tokens in a span. For details on the model configuration and its implementation in FACTORIE (McCallum et al., 2009), we refer to the description in the original paper (Klinger and Cimiano, 2013). The objective function to evaluate a span $\mathbf{r}$ during training is

$$f(\mathbf{r}) = \max_{\mathbf{g} \in \mathbf{s}} \frac{|\mathbf{r} \cap \mathbf{g}|}{|\mathbf{g}|} - \alpha \cdot |\mathbf{r} \backslash \mathbf{g}| \,,$$

[1] https://bitbucket.org/rklinger/jfsa

where $\mathbf{g}$ is the set of all gold spans, and $|\mathbf{r} \cap \mathbf{g}|$ is the number of tokens shared by gold and predicted span and $|\mathbf{r} \backslash \mathbf{g}|$ the number of predicted tokens which are not part of the gold span. The parameter α is set to 0.1 as in the original paper.[2]. The objective for the predictions in a whole sentence $\mathbf{s}$ containing spans is $f(\mathbf{s}) = \sum_{\mathbf{r} \in s} f(\mathbf{r})$.

This model does not take into account language-specific features and can therefore be trained for different languages. In the following, we explain our procedure for inducing a model for a target language for which no annotations are available.

2.2 Statistical Machine Translation and Annotation Projection

Annotating textual corpora with fine-grained sentiment information is a time-consuming and therefore costly process. In order to adapt a model to a new domain and to a new language, corresponding training data is needed. In order to circumvent the need for additional training data when addressing a new language, we project training data automatically from a source to a target language. As input to our approach we require a corpus annotated for some source language and a translation from the source to a target language. As the availability of a parallel training corpus cannot be assumed in general, we use statistical machine translation (SMT) methods, relying on phrase-based translation models that use large amounts of parallel data for training (Koehn, 2010).

While using an open-source system such as Moses[3] would have been an option, we note that the quality would be limited by whether the system can be trained on a representative corpus. A standard dataset that SMT systems are trained on is EuroParl (Koehn, 2005). EuroParl covers 21 languages and contains 1.920.209 sentences for the pair German/English. The corpus includes only 4 sentences with the term "toaster", 12 with "knives" (mostly in the context of violence), 6 with "dishwasher" (in the context of regulations) and 0 with "trash can". The terms "camera" and "display" are more frequent, with 208 and 1186 mentions, respectively, but they never occur together.[4] The corpus is thus not representative for product reviews as we consider in this paper.

Thus, we opt for using a closed translation system that is trained on larger amounts of data, that is Google Translate, through the available API[5]. The alignment is then computed as a post processing step relying on FastAlign (Dyer et al., 2013), a reparametrization of IBM Model 2 with a reduced set of parameters. It is trained in an unsupervised fashion via expectation maximization.

Projecting the annotations from the source to the target language works as follows: given an annotated sentence in the source language $s_1, \ldots, s_n$ and some translation of this sentence $t_1, \ldots, t_m$ into the target language, we induce an inductive mapping $a : [1 \ldots n] \rightarrow [1 \ldots m]$ using FastAlign. For a source language phrase $s_{i,j} = s_i, \ldots, s_j$ we refer by $a(s_{i,j})$ to the set of tokens that some token in $s_{i,j}$ has been aligned to, that is: $a(s_{i,j}) = \cup_{i \leq k \leq j} \{a(k)\}$. Note that the tokens in $a(s_{i,j})$ are not necessarily consecutive, therefore the annotation in the target language is defined as the minimal sequence including all tokens $t_k \in a(s_{i,j})$, *i. e.*, the most left and most right tokens define the span of the target annotation.

This procedure leads to the same number of span annotations in source and target language with the only exception that we exclude projected annotations for which $|n - m| > 10$.

2.3 Quality Estimation-based Instance Filtering

The performance of an approach relying on projection of training data from a source to a target language and using this automatically projected data to train a supervised model crucially depends on the quality of the translations and alignments. In order to reduce the impact of spurious translations, we filter out low-quality sentence pairs. To estimate this quality, we take three measures into consideration (following approaches described by Shah and Specia (2014), in addition to a manual assessment of the translation quality as an upper baseline):

1. The probability of the sentence in the source language given a language model build on unannotated text in the source language (measuring if the language to be translated is typical, referred to as *Source LM*).

2. The probability of the machine translated sentence given a language model built on unanno-

[2] Note that the learning is independent from the actual value for all $0 < \alpha < (\max_{g \in \text{Corpus}} |g|)^{-1}$.

[3] www.statmt.org/moses/

[4] These example domains are taken from the USAGE corpus (Klinger and Cimiano, 2014), which is used in Section 3.

[5] https://cloud.google.com/translate/

	# reviews	
	en	de
coffee machine	75	108
cutlery	49	72
microwave	100	100
toaster	100	4
trash can	100	99
vacuum cleaner	51	140
washing machine	49	88
dish washer	98	0

Table 1: Frequencies of the corpus used in our experiments (Klinger and Cimiano, 2014).

tated text in the target language (measuring if the translation is typical, referred to as *Target LM*).

3. The likelihood that the alignment is correct, directly computed on the basis of the alignment probability (referred to as *Alignment*): $p(\mathbf{e} \mid \mathbf{f}) = \prod_{i=1^m} p(e_i \mid \mathbf{f}, m, n)$, where $\mathbf{e}$ and $\mathbf{f}$ are source and target sentences and m and n denote the sentence lengths (Dyer et al., 2013, Eq. 1f.).

For building the language models, we employ the toolkit SRILM (Stolcke, 2002; Stolcke et al., 2011). The likelihood for the alignment as well as the language model probability are normalized by the number of tokens in the sentence.

3 Experiments

3.1 Corpus and Setting

The proposed approach is evaluated on the language pair German/English in both directions (projecting German annotations into an automatically generated English corpus and testing on English annotations and vice versa). As a resource, we use the recently published USAGE corpus (Klinger and Cimiano, 2014), which consists of 622 English and 611 German product reviews from http://www.amazon.com/ and http://www.amazon.de/, respectively. The reviews are on coffee machines, cutlery sets, microwaves, toasters, trash cans, vacuum cleaners, washing machines, and dish washers. Frequencies of entries in the corpus are summarized in Table 1. Each review has been annotated by two annotators. We take into account the data generated by the first annotator in this work to avoid the design of an aggregation procedure. The corpus is unbalanced between the product classes. The average numbers of annotated aspects in each review in the German corpus (10.4) is smaller than in English (13.7). The average number of subjective phrases is more similar with 8.6 and 8.3, respectively. The total number of aspects is 8545 for English and 6340 in German, the number of subjective phrases is 5321 and 5086, respectively.

The experiments are performed in a leave-one-domain-out setting, *e. g.*, testing on coffee machine reviews is based on a model trained on all other products except coffee machines. This holds for the cross-language and the in-target-language training results and leads therefore to comparable settings. We use exact match precision, recall and F_1-measure for evaluation. However, it should be noted that partial matching scores are also commonly applied in fine-grained sentiment analysis due to the fact that boundaries of annotations can differ substantially between annotators. For simplicity, we limit ourselves to the more strict evaluation measure.

The language models are trained on 7,413,774 German and 9,650,633 English sentences sampled from Amazon product reviews concerning the product categories in the USAGE corpus. The FastAlign model is trained on the EuroParl corpus and the automatically translated USAGE corpus in both directions (German translated to English and English translated to German).

3.2 Results

We evaluate and compare the impact of the three automatic quality estimation methods and compare them to a manual sentence-based judgement for the language projection from German to English (testing on English). The manual judgement was performed by assigning values ranging from 0 (not understandable), over 1 and 2 (slightly understandable) to 8 (some flaws in translation), 9 (minor flaws in translation) to 10 (perfect translation).[6]

Figure 2 shows the results for all four methods (including manual quality assessment) from German to English for the product category of coffee machines compared to in-target-language training results. The x-axis corresponds to different values for the filtering threshold. Thus, when increasing the threshold, the number of sentences used for training decreases. For all quality estimation meth-

[6]This annotated data is available at http://www.romanklinger.de/translation-quality-review-corpus

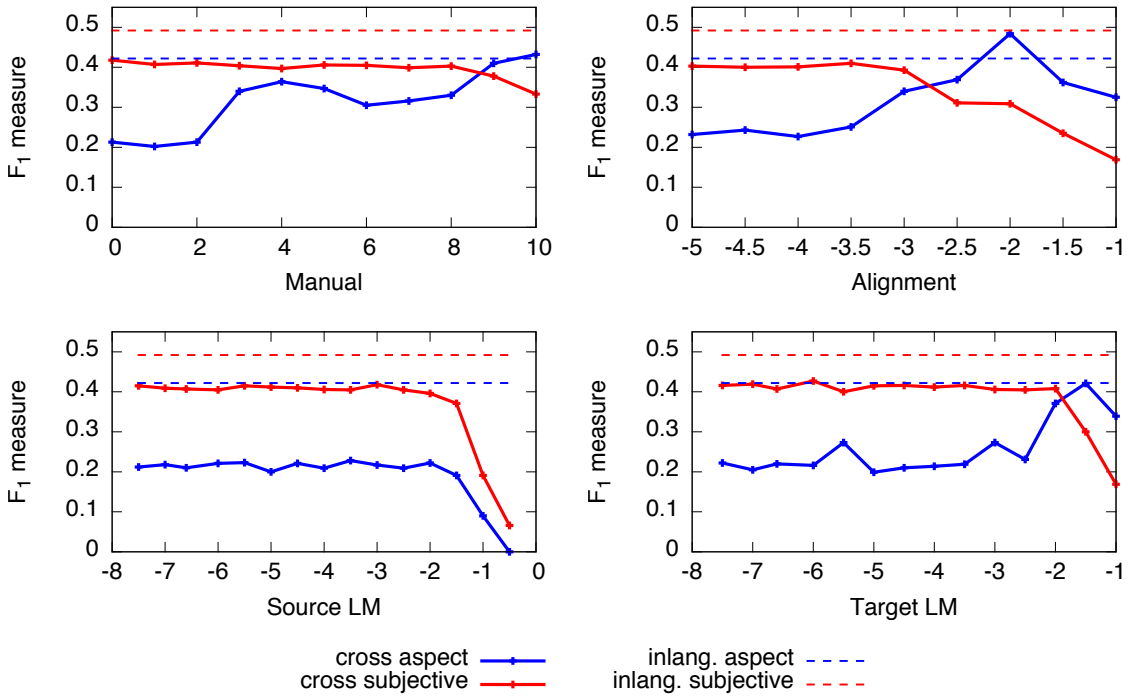

Figure 2: Complete results for the reviews for coffee machines for the projection direction German to English.

ods except for the language model for the source language, the performance on the target language increases significantly for the prediction of aspect phrases. The English in-target-language training performance represents an upper baseline, resulting from training the system on manually annotated data for the target language ($F_1 = 0.42$). Without any instance filtering, relying on all automatically produced translations to induce projected annotations for the target language, an F_1 measure of 0.21 is reached. With filtering based on the manually assigned translation quality estimation, a result of $F_1 = 0.43$ is reached. Using the alignment as quality score for filtering, the best result obtained is $F_1 = 0.48$. However, results start decreasing from this threshold value on, which is likely due to the fact that the positive effect of instance filtering is outweighed by the performance drop due to training with a smaller dataset. The filtering based on the target language model leads to $F_1 = 0.42$, while the source language model cannot filter the training instances such that the performance increases over the initial value.

Surprisingly, instance filtering has no impact on the detection of subjective phrases. Without any filtering, for the prediction of subjective phrases we get an F_1 of 0.42, which is close to the performance of in-target-language training of $F_1 = 0.49$. For the case of phrase detection, the difference between training with all data (21%) and in-target-language training (42%) is considerably higher. Decreasing the size of the training set by filtering only decreases the F_1 measure.

Figure 3 shows the macro-average results summarizing the different domains in precision, recall and F_1 measure. The thresholds for the filtering have been fixed to the best result of the product coffee machine for all products. The manual quality estimation as well as the alignment and target language model lead to comparable (or superior) results compared to target language training for aspect detection. This holds for nearly all product domains, only for trash cans and cutlery the performance achieved by filtering is slightly lower for the direction German-to-English. The initial performance for the whole data set is on average higher for the projection direction English to German; therefore the values for the source language model are comparably higher than for the other projection direction.

For the aspect detection, all filtering methods except using the source language model lead to an improvement over the baseline (without filtering) that is significant according to a Welch t-test ($\alpha =$

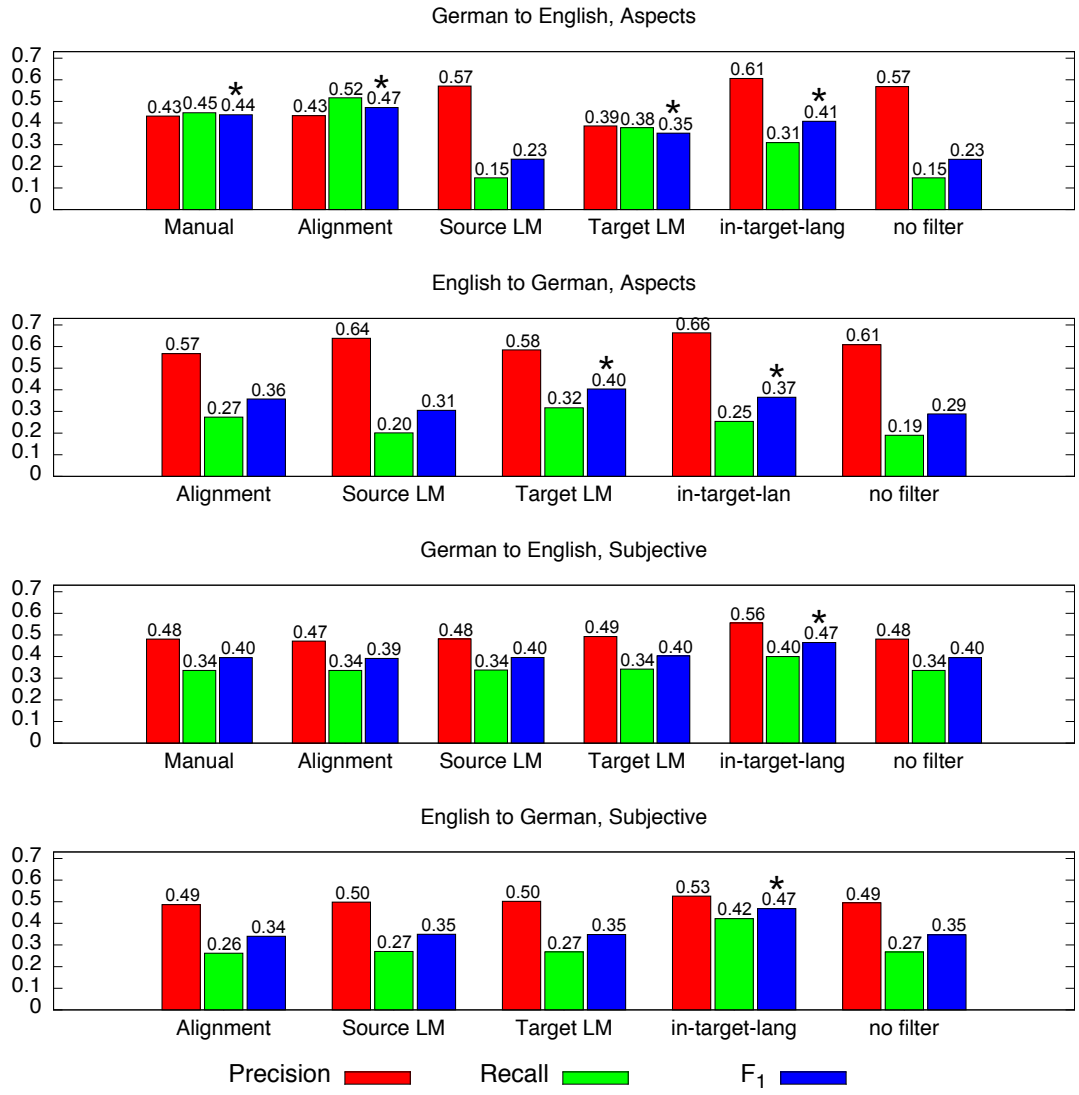

Figure 3: Macro-average results of cross-language training over the different domains showing precision, recall and F_1 measure. Significant differences of the F_1 results to the no-filter-baseline are marked with a star (Welch t-test comparing the different separate domain results, $p < 0.05$).

0.05). For English to German, in-target-language training and the target language model filtering provide improved results over the baseline that are significant. For subjective phrase detection, the in-language training is significantly better than the baseline.

It is notable that in all experiments the model's performance without filtering is mainly limited in recall, which drops the performance in F_1. Instance filtering therefore has mainly an effect on the recall.

3.3 Discussion

Our results show that an approach based on automatic training data projection across languages is feasible and provides competitive results compared to training on manually annotated target language data. We have in particular quantified the loss of performance of such an automatic approach compared to using a system trained on manually annotated data in the target language. We have shown that the performance of aspect detection of a system using all available projected training data yields a drop of $\approx 50\,\%$ in F_1-measure compared to a model trained using manually annotated data in the target language. The instance filtering approaches in which only the sentences with highest quality are selected to project training data to the target language using a threshold has a significant positive impact on the performance of a model trained on automatically projected training

data when predicting aspect phrases on the target language, increasing results from 23 % to 47 % on average. Our approach relying on instance filtering comes close to results produced by a system trained on manually annotated data for the target language on the task of predicting aspect phrases.

In contrast to these results for the aspect phrase recognition, the impact of filtering training instances is negligible for the detection of subjective phrases. The highest performance is achieved with the full set of training instances. Therefore, it may be concluded that for aspect name detection, high quality training is crucial. For subjective phrase detection, a greater training set is crucial. In contrast to aspect recognition, the drop in subjective phrase recognition performance is comparatively low when training on all instances.

Filtering translations by manually provided translation scores (as an upper baseline for the filtering) yields comparable results to using the alignment and the language model on the target language. Using the language model on the source language for filtering does not lead to any improvement. Predicting the quality of translation relying on the probability of the source sentence via a source language model therefore seems not to be a viable approach on the task in question. Using the target language model as a filter leads to the most consistent results and is therefore to be preferred over the source language model and the alignment score.

Including more presumably noisy instances by using a smaller filtering threshold leads to a decreased recall throughout all methods in aspect detection and to a lesser extent for subjective phrase detection. The precision is affected to a smaller degree. This can as well be observed in the number of predictions the models based on different thresholds generate: While the number of true positive aspects for the coffee machine subdomain is 1100, only 221 are predicted with a threshold of the manual quality assignment of 0. However, a treshold of 9 leads to 560 predictions and a threshold of 10 to 1291. This effect can be observed for subjective phrases as well. It increases from 465 to 827 while the gold number is 676. These observations hold for all filtering methods analogously.

4 Related Work

In-target-language training approaches for fine-grained sentiment analysis include those targeting the extraction of phrases or modelling it as text classification (Choi et al., 2010; Johansson and Moschitti, 2011; Yang and Cardie, 2012; Hu and Liu, 2004; Li et al., 2010; Popescu and Etzioni, 2005; Jakob and Gurevych, 2010b). Such models are typically trained or optimized on manually annotated data (Klinger and Cimiano, 2013; Yang and Cardie, 2012; Jakob and Gurevych, 2010a; Zhang et al., 2011). The necessary data, at least containing fine-grained annotations for aspects and subjective phrases instead of only an overall polarity score, are mainly available for the English language to a sufficient extent. Popular corpora used for training are for instance the J.D. Power and Associates Sentiment Corpora (Kessler et al., 2010) or the MPQA corpora (Wilson and Wiebe, 2005).

Non-English resources are scarce. Examples are a YouTube corpus consisting of English and Italian comments (Uryupina et al., 2014), a not publicly available German Amazon review corpus of 270 sentences (Boland et al., 2013), in addition to the USAGE corpus (Klinger and Cimiano, 2014) we have used in this work, consisting of German and English reviews. The (non-fine-grained annotated) Spanish TASS corpus consists of Twitter messages (Saralegi and Vicente, 2012). The "Multilingual Subjectivity Analysis Gold Standard Data Set" focuses on subjectivity in the news domain (Balahur and Steinberger, 2009). A Chinese corpus annotated at the aspect and subjective phrase level is described by Zhao et al. (2014).

There has not been too much work on approaches to transfer a model either directly or via annotation projection in the area of sentiment analysis. One example is based on sentence level annotations which are automatically translated to yield a resource in another language. This approach has been proven to work well across several languages (Banea et al., 2010; Mihalcea et al., 2007; Balahur and Turchi, 2014). Recent work approached multilingual opinion mining on the above-mentioned multi-lingual Youtube corpus with tree kernels predicting the polarity of a comment and whether it concerns the product or the video in which the product is featured. (Severyn et al., 2015). Brooke et al. (2009) compare dictionary and classification transfer from English to Spanish in a similar classification setting.

While cross-lingual annotation projection has been investigated in the context of polarity computation, we are only aware of two approaches exploiting cross-lingual annotation projection on

the task of identifying aspects specifically with an evaluation on manually annotated data in more than one language. The CLOpinionMiner (Zhou et al., 2015) uses an English data set which is transfered to Chinese. Models are further improved by co-training. Xu et al. (2013) perform self-training based on a projected corpus from English to Chinese to detect opinion holders. Due to the lack of existing manually annotated resources, to our knowledge no cross-language projection approach for fine-grained annotation at the level of aspect and subjective phrases has been proposed before.

The projection of annotated data sets has been investigated in a variety of applications. Early work includes an approach to the projection of part-of-speech tags and noun phrases (Yarowsky et al., 2001; Yarowsky and Ngai, 2001) and parsing information (Hwa et al., 2005) on a parallel corpus. Especially in syntactic and semantic parsing, heuristics to remove or correct spuriously projected annotations have been developed (Padó and Lapata, 2009; Agić et al., 2014). It is typical for these approaches to be applied on existing parallel corpora (one counter example is the work by Basili et al. (2009) who perform postprocessing of machine translated resources to improve the annotation for training semantic role labeling models). In cases in which no such parallel resources are available containing pertinent annotations, models can be transfered after training. Early work includes a cross-lingual parser adaption (Zeman and Resnik, 2008). A recent example is the projection of a metaphor detection model using a bilingual dictionary (Tsvetkov et al., 2014). A combination of model transfer and annotation projection for dependency parsing has been proposed by Kozhevnikov and Titov (2014).

To improve quality of the overall corpus of projected annotations, the selection of data points for dependency parsing has been studied (Søgaard, 2011). Similarly, Axelrod et al. (2011) improve the average quality of machine translation systems by selection of promising training examples and show that such a selection approach has a positive impact. Related to the latter, a generic instance weighting scheme has been proposed for domain adaptation (Jiang and Zhai, 2007).

Other work has attempted to exploit information available in multiple languages to induce a model for a language for which sufficient training data is not available. For instance, universal tag sets take advantage of annotations that are aligned across languages (Snyder et al., 2008). Delexicalization allows for applying a model to other languages (McDonald et al., 2011).

Focusing on cross-lingual sentiment analysis, joint training of classification models on multiple languages shows an improvement over separated models. Balahur and Turchi (2014) analyzed the impact of using different machine translation approaches in such settings. Differences in sentiment expressions have been analyzed between English and Dutch (Bal et al., 2011). Co-training with non-annotated corpora has been shown to yield good results for Chinese (Wan, 2009). Ghorbel (2012) analyzed the impact of automatic translation on sentiment analysis.

Finally, SentiWordNet has been used for multilingual sentiment analysis (Denecke, 2008). Building dictionaries for languages with scarce resources can be supported by bootstrapping approaches (Banea et al., 2008).

Estimating the quality of machine translation can be understood as a ranking problem and thus be modeled as regression or classification. An important research focus is on investigating the impact of different features on predicting translation quality. For instance, sentence length, the output probability, number of unknown words of a target language as well as parsing-based features have been used (Avramidis et al., 2011). The alignment context can also be taken into account (Bach et al., 2011). An overview on confidence measures for machine translation is for instance provided by Ueffing et al. (2003). The impact of different features has been analyzed by Shah et al. (2013). A complete system and framework for quality estimation (including a list of possible features) is QuEst (Specia et al., 2013).

For an overview of other cross-lingual applications and methods, we refer to Bikel and Zitouni (2012).

5 Conclusion and Future Work

We have presented an approach that alleviates the need of training data for a target language when adapting a fine-grained sentiment analysis system to a new language. Our approach relies on training data available for a source language and on automatic machine translation, in particular statistical methods, to project training data to the target language, thus creating a training corpus on which

a supervised sentiment analysis approach can be trained on. We have in particular shown that our results are competitive to training with manually annotated data in the target language, both for the prediction of aspect phrases as well as subjective phrases. We have further shown that performance for aspect detection can be almost doubled by estimating the quality of translations and selecting only the translations with highest quality for training. Such an effect cannot be observed in the prediction of subjective phrases, which nevertheless delivers results comparable to training using target language data using all automatically projected training data. Predicting translation quality by both the alignment probability and the target language model probability have been shown to deliver good results, while an approach exploiting source language model probability does not perform well.

Our hypothesis for the failure of translation filtering for the prediction of subjective phrases is that translation quality for subjective phrases is generally higher as their coverage in standard parallel corpora is reasonable and they are often domain-independent. A further possible explanation is that subjective phrases have a more complex structure (for instance, their average length is 2.38 tokens in English and 2.57 tokens in German, while the aspect length is 1.6 and 1.3, respectively). Therefore, translation as well as filtering might be more challenging. These hypotheses should be verified and investigated further in future work.

Further work should also be devoted to the investigation of other quality estimation procedures, in particular combinations of those investigated in this paper. Preliminary experiments have shown that the correlation between the filters incorporated in this paper is low. Thus, their combination could indeed have an additional impact. Similarly, the projection quality can be affected by the translation itself and by the alignment. These two aspects should be analyzed separately.

In addition, instead of Boolean filtering (using an instance or not), weighting the impact of the instance in the learning procedure might be beneficial as lower-quality instances can still be taken into account, although with a lower impact proportional to their corresponding score or probability.

In addition to the presented approach of projecting annotations, a comparison to directly transferring a trained model across languages would allow for a deeper understanding of the processes involved. Finally, it is an important and promising step to apply the presented methods on other languages.

Acknowledgments

We thank Nils Reiter, John McCrae, and Matthias Hartung for fruitful discussions. Thanks to Lucia Specia for her comments on quality estimation methods in our work. We thank Chris Dyer for his help with FastAlign. Thanks to the anonymous reviewers for their comments. This research was partially supported 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 and by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG).

References

Željko Agić, Jörg Tiedemann, Danijela Merkler, Simon Krek, Kaja Dobrovoljc, and Sara Moze. 2014. Cross-lingual dependency parsing of related languages with rich morphosyntactic tagsets. In *Workshop on Language Technology for Closely Related Languages and Language Variants, EMNLP*.

Eleftherios Avramidis, Maja Popović, David Vilar, and Aljoscha Burchardt. 2011. Evaluate with confidence estimation: Machine ranking of translation outputs using grammatical features. In *Workshop on Statistical Machine Translation, ACL*.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *EMNLP*.

Nguyen Bach, Fei Huang, and Yaser Al-Onaizan. 2011. Goodness: A method for measuring machine translation confidence. In *ACL-HLT*.

Daniella Bal, Malissa Bal, Arthur van Bunningen, Alexander Hogenboom, Frederik Hogenboom, and Flavius Frasincar. 2011. Sentiment analysis with a multilingual pipeline. In *Web Information System Engineering WISE 2011*.

Alexandra Balahur and Ralf Steinberger. 2009. Rethinking sentiment analysis in the news: from theory to practice and back. In *Proceeding of Workshop on Opinion Mining and Sentiment Analysis (WOMSA)*.

Alexandra Balahur and Marco Turchi. 2014. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech & Language*, 28(1).

Carmen Banea, Rada Mihalcea, and Janyce Wiebe. 2008. A bootstrapping method for building subjectivity lexicons for languages with scarce resources. In *LREC*.

Carmen Banea, Rada Mihalcea, and Janyce Wiebe. 2010. Multilingual subjectivity: Are more languages better? In *COLING*.

Roberto Basili, Diego De Cao, Danilo Croce, Bonaventura Coppola, and Alessandro Moschitti. 2009. Cross-language frame semantics transfer in bilingual corpora. In *CICLING*, volume 5449 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg.

Daniel Bikel and Imed Zitouni, editors. 2012. *Multilingual Natural Language Processing Applications: From Theory to Practice*. IBM Press, 1st edition.

Katarina Boland, Andias Wira-Alam, and Reinhard Messerschmidt. 2013. Creating an annotated corpus for sentiment analysis of german product reviews. Technical Report 2013/05, GESIS Institute.

Julian Brooke, Milan Tofiloski, and Maite Taboada. 2009. Cross-linguistic sentiment analysis: From english to spanish. In *RANLP*, Borovets, Bulgaria, September. Association for Computational Linguistics.

Yoonjung Choi, Seongchan Kim, and Sung-Hyon Myaeng. 2010. Detecting Opinions and their Opinion Targets in NTCIR-8. In *NTCIR8*.

Kerstin Denecke. 2008. Using sentiwordnet for multilingual sentiment analysis. In *ICDEW*.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *NAACL*.

Hatem Ghorbel. 2012. Experiments in cross-lingual sentiment analysis in discussion forums. In Karl Aberer, Andreas Flache, Wander Jager, Ling Liu, Jie Tang, and Christophe Guret, editors, *Social Informatics*, volume 7710 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *ACM SIGKDD*.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Nat. Lang. Eng.*, 11(3), September.

Niklas Jakob and Iryna Gurevych. 2010a. Extracting opinion targets in a single- and cross-domain setting with conditional random fields. In *EMNLP*.

Niklas Jakob and Iryna Gurevych. 2010b. Using anaphora resolution to improve opinion target identification in movie reviews. In *ACL*.

Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in nlp. In *ACL*.

Richard Johansson and Alessandro Moschitti. 2011. Extracting opinion expressions and their polarities: exploration of pipelines and joint models. In *ACL-HLT*.

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

Roman Klinger and Philipp Cimiano. 2013. Bidirectional inter-dependencies of subjective expressions and targets and their value for a joint model. In *ACL*.

Roman Klinger and Philipp Cimiano. 2014. The usage review corpus for fine grained multi lingual opinion analysis. In *LREC*.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*. AAMT.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition.

Mikhail Kozhevnikov and Ivan Titov. 2014. Cross-lingual model transfer using feature representation projection. In *ACL*.

Fangtao Li, Minlie Huang, and Xiaoyan Zhu. 2010. Sentiment analysis with global topics and local dependency. In *AAAI*.

Andrew McCallum, Karl Schultz, and Sameer Singh. 2009. FACTORIE: Probabilistic programming via imperatively defined factor graphs. In *NIPS*.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *EMNLP*.

Rada Mihalcea, Carmen Banea, and Janyce Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *ACL*.

Sebastian Padó and Mirella Lapata. 2009. Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research (JAIR)*, 36.

Ana-Maria Popescu and Oren Etzioni. 2005. Extracting product features and opinions from reviews. In *HLT-EMNLP*.

Xabier Saralegi and Iñaki San Vicente. 2012. Tass: Detecting sentiments in spanish tweets. In *Workshop on Sentiment Analisis at SEPLN (TASS)*. SEPLN, 09/2012.

Sunita Sarawagi and William W. Cohen. 2004. Semimarkov conditional random fields for information extraction. In *NIPS*.

Aliaksei Severyn, , Alessandro Moschitti, Olga Uryupina, Barbara Plank, and Katja Filippova. 2015. Multi-lingual opinion mining on youtube. *Information Processing and Management*. in press.

Kashif Shah and Lucia Specia. 2014. Quality estimation for translation selection. In *Conference of the European Association for Machine Translation*, EAMT, Dubrovnik, Croatia.

Kashif Shah, Trevor Cohn, and Lucia Specia. 2013. An investigation on the effectiveness of features for translation quality estimation. In *Machine Translation Summit XIV*.

Benjamin Snyder, Tahira Naseem, Jacob Eisenstein, and Regina Barzilay. 2008. Unsupervised multilingual learning for POS tagging. In *EMNLP*.

Anders Søgaard. 2011. Data point selection for cross-language adaptation of dependency parsers. In *ACL-HLT*.

Lucia Specia, Kashif Shah, Jose G.C. de Souza, and Trevor Cohn. 2013. Quest - a translation quality estimation framework. In *ACL*.

Andreas Stolcke, Jing Zheng, Wen Wang, and Victor Abrash. 2011. SRILM at sixteen: Update and outlook. In *Proceedings IEEE Automatic Speech Recognition and Understanding Workshop*.

Andreas Stolcke. 2002. Srilm-an extensible language modeling toolkit. In *Proceedings International Conference on Spoken Language Processing*.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *ACL*.

Nicola Ueffing, Klaus Macherey, and Hermann Ney. 2003. Confidence measures for statistical machine translation. In *In Proc. MT Summit IX*.

Olga Uryupina, Barbara Plank, Aliaksei Severyn, Agata Rotondi, and Alessandro Moschitti. 2014. Sentube: A corpus for sentiment analysis on youtube social media. In *LREC*.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In Keh-Yih Su, Jian Su, and Janyce Wiebe, editors, *ACL/IJCNLP*.

M. Wick, K. Rohanimanesh, K. Bellare, A. Culotta, and A. McCallum. 2011. SampleRank: Training factor graphs with atomic gradients. In *ICML*.

Theresa Wilson and Janyce Wiebe. 2005. Annotating attributions and private states. In *CorpusAnno, ACL*.

Ruifeng Xu, Lin Gui, Jun Xu, Qin Lu, and Kam-Fai Wong. 2013. Cross lingual opinion holder extraction based on multi-kernel svms and transfer learning. *World Wide Web Journal*.

Bishan Yang and Claire Cardie. 2012. Extracting opinion expressions with semi-markov conditional random fields. In *EMNLP-CoNLL*.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual pos taggers and np bracketers via robust projection across aligned corpora. In *NAACL*.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *HLT*.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *IJCNLP-08 Workshop on NLP for Less Privileged Languages*.

Qi Zhang, Yuanbin Wu, Yan Wu, and Xuanjing Huang. 2011. Opinion mining with sentiment graph. In *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01*.

Y. Zhao, B. Qin, and T. Liu. 2014. Creating a fine-grained corpus for chinese sentiment analysis. *Intelligent Systems, IEEE*, PP(99).

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2015. Clopinionminer: Opinion target extraction in a cross-language scenario. *IEEE/ACM Transactions on Audio, Speech & Language Processing*, 23(4).