Extracting Opinion Targets in a Single- and Cross-Domain Setting with Conditional Random Fields

Niklas Jakob
Technische Universität Darmstadt
Hochschulstraße 10
64289 Darmstadt, Germany

Iryna Gurevych
Technische Universität Darmstadt
Hochschulstraße 10
64289 Darmstadt, Germany

http://www.ukp.tu-darmstadt.de/people

Abstract

In this paper, we focus on the opinion target extraction as part of the opinion mining task. We model the problem as an information extraction task, which we address based on Conditional Random Fields (CRF). As a baseline we employ the supervised algorithm by Zhuang et al. (2006), which represents the state-of-the-art on the employed data. We evaluate the algorithms comprehensively on datasets from four different domains annotated with individual opinion target instances on a sentence level. Furthermore, we investigate the performance of our CRF-based approach and the baseline in a single- and cross-domain opinion target extraction setting. Our CRF-based approach improves the performance by 0.077, 0.126, 0.071 and 0.178 regarding F-Measure in the single-domain extraction in the four domains. In the cross-domain setting our approach improves the performance by 0.409, 0.242, 0.294 and 0.343 regarding F-Measure over the baseline.

1 Introduction

The automatic extraction and analysis of opinions has been approached on several levels of granularity throughout the last years. As opinion mining is typically an enabling technology for another task, this overlaying system defines requirements regarding the level of granularity. Some tasks only require an analysis of the opinions on a document or sentence level, while others require an extraction and analysis on a term or phrase level. Amongst the tasks which require the finest level of granularity are: a) Opinion question answering - i.e. with questions regarding an entity as in “What do the people like / dislike about X?”. b) Recommender systems - i.e. if the system shall only recommend entities which have received good reviews regarding a certain aspect. c) Opinion summarization - i.e. if one wants to create an overview of all positive / negative opinions regarding aspect Y of entity X and cluster them accordingly. All of these tasks have in common that in order to fulfill them, the opinion mining system must be capable of identifying what the opinions in the individual sentences are about, hence extract the opinion targets.

Our goal in this work is to extract opinion targets from user-generated discourse, a discourse type which is quite frequently encountered today, due to the explosive growth of Web 2.0 community websites. Typical sentences which we encounter in this discourse type are shown in the following examples. The opinion targets which we aim to extract are underlined in the sentences, the corresponding opinion expressions are shown in italics.

(1) While none of the features are *earth-shattering*, eCircles does provide a *great* place to keep in touch.

(2) Hyundai’s *more-than-modest* refresh has largely addressed all the original car’s *weaknesses* while maintaining its price *competitiveness*.

The extraction of opinion targets can be considered as an instance of an information extraction (IE) task (Cowie and Lehnert, 1996). Conditional

Random Fields (CRF) (Lafferty et al., 2001) have been successfully applied to several IE tasks in the past (Peng and McCallum, 2006). A recurring problem, which arises when working with supervised approaches, concerns the domain portability. In the opinion mining context this question has been prominently investigated with respect to opinion polarity analysis (sentiment analysis) in previous research (Aue and Gamon, 2005; Blitzer et al., 2007). Terms as "unpredictable" can express a positive opinion when uttered about the storyline of a movie but a negative opinion when the handling of a car is described. Hence the effects of training and testing a machine learning algorithm for sentiment analysis on data from different domains have been analyzed in previous research. However to the best of our knowledge, these effects have not been investigated regarding the extraction of opinion targets.

The contribution of this paper is a CRF-based approach for opinion targets extraction which tackles the problem of domain portability. We first evaluate our approach in three different domains against a state-of-the art baseline system and then evaluate the performance of both systems in a cross-domain setting. We show that our CRF-based approach outperforms the baseline in both settings, and how the diffrrent combinations of features we introduce influence the results of our CRF-based approach. The remainder of this paper is structured as follows: In Section 2 we discuss the related work, and in Section 3 we describe our CRF-based approach. Section 4 comprises our experimental setup including the description of the dataset we employ in our experiments in Section 4.1 and the baseline system in Section 4.2. The results of our experiments and their discussion follow in Section 5. Finally we draw our conclusions in Section 6.

2 Related Work

In the following we will discuss the related work regarding opinion target extraction and domain adaptation in opinion mining. The discussion of the related work on opinion target extraction is separated in supervised and unsupervised approaches. We conclude with a discussion of the related work on domain adaptation in opinion mining.

2.1 Unsupervised Opinion Target Extraction

The first work on opinion target extraction was done on customer reviews of consumer electronics. Hu and Liu (2004) introduce the task of *feature based summarization*, which aims at creating an overview of the product features commented on in the reviews. Their approach relies on a statistical analysis of the review terms based on association mining. A dataset of customer reviews from five domains was annotated by the authors regarding mentioned product features with respective opinion polarities. The association mining based algorithm yields a precision of 0.72 and a recall of 0.80 in the extraction of a manually selected subset of product features. The same dataset of product reviews was used in the work of Yi et al. (2003). They present and evaluate a complete system for opinion extraction which is based on a statistical analysis based on the Likelihood Ratio Test for opinion target extraction. The Likelihood Ratio Test yields a precision of 0.97 and 1.00 in the task of opinion target (product feature) extraction, recall values are not reported.

Popescu and Etzioni (2005) present the OPINE system for opinion mining on product reviews. Their algorithm is based on an information extraction system, which uses the pointwise mutual information based on the hitcounts of a web-search engine as an input. They evaluate the opinion target extraction separately on the dataset by Hu and Liu (2004). OPINE's precision is on average 22% higher than the association mining based approach, while having an average 3% lower recall.

Bloom et al. (2007) manually create taxonomies of opinion targets for two datasets. With a handcrafted set of dependency tree paths their algorithm identifies related opinion expressions and targets. Due to the lack of a dataset annotated with opinion expressions and targets, they just evaluate the accuracy of several aspects of their algorithm by manually assessing an output sample. Their algorithm yields an accuracy of 0.75 in the identification of opinion targets.

Kim and Hovy (2006) aim at extracting opinion holders and opinion targets in newswire with semantic role labeling. They define a mapping of the semantic roles identified with FrameNet to the respective opinion elements. As a baseline, they implement an approach based on a dependency parser, which identifies the targets following the dependencies of opinion expressions. They measure the overlap between two human annotators and their algorithm as well as the baseline system. The algorithm based on semantic role labeling yields an F-Measure of 0.315 with annotator1 and 0.127 with annotator2, while the baseline yields an F-Measure of 0.107 and 0.109 regarding opinion target extraction

2.2 Supervised Opinion Target Extraction

Zhuang et al. (2006) present a supervised algorithm for the extraction of opinion expression - opinion target pairs. Their algorithm learns the opinion target candidates and a combination of dependency and part-of-speech paths connecting such pairs from an annotated dataset. They evaluate their system in a cross validation setup on a dataset of user-generated movie reviews and compare it to the results of the Hu and Liu (2004) system as a baseline. Thereby, the system by Zhuang et al. (2006) yields an F-Measure of 0.529 and outperforms the baseline which yields an F-Measure of 0.488 in the task of extracting opinion target - opinion expression pairs.

Kessler and Nicolov (2009) solely focus on identifying which opinion expression is linked to which opinion target in a sentence. They present a dataset of car and camera reviews in which opinion expressions and opinion targets are annotated. Starting with this information, they train a machine learning classifier for identifying related opinion expressions and targets. Their algorithm receives the opinion expression and opinion target annotations as input during runtime. The classifier is evaluated using the algorithm by Bloom et al. (2007) as a baseline. The support vector machine based approach by Kessler and Nicolov (2009) yields an F-Measure of 0.698, outperforming the baseline which yields an F-Measure of 0.445.

2.3 Domain Adaptation in Opinion Mining

The task of creating a supervised algorithm, which when trained on data from domain A, also performs well on data from another domain B, is a domain adaptation problem (Daumé III and Marcu, 2006; Jiang and Zhai, 2007). Aue and Gamon (2005) have investigated this challenge very early in the task of document level sentiment classification (positive / negative). They observe that increasing the amount of training data raises the classification accuracy, but only if the training data is from one source domain. Increasing the training data by mixing domains does not yield any consistent improvements. Blitzer et al. (2007) introduce an extension to a structural correspondence learning algorithm, which was specifically designed to address the task of domain adaptation. Their enhancement aims at identifying pivot features, which are stable across domains. In a series of experiments in document level sentiment classification they show that their extension outperforms the original structural correspondence learning approach. In their error analysis, the authors observe the best results were reached when the training - testing combinations were *Books - DVDs* or *Electronics - Kitchen appliances*. They conclude that the topical relatedness of the domains is an important factor. Furthermore they observe that training the algorithm on a smaller amount of data from a similar domain is more effective than increasing the amount of training data by mixing domains.

3 CRF-based Approach for Opinion Target Extraction

In the following we will describe the features we employ as input for our CRF-based approach. As the development data, we used 29 documents from the movies dataset, 23 documents from the web-services dataset and 15 documents from the cars & cameras datasets.

Token
This feature represents the string of the current token as a feature. Even though this feature is rather obvious, it can have considerable impact on the target extraction performance. If the vocabulary of targets is rather compact for a certain domain (corresponding to a low target type / target ratio), the training data is likely to contain the majority of the target types, which should hence be a good indicator. We will refer to this feature as **tk** in our result tables.

POS
This feature represents the part-of-speech tag of the current token as identified by the Stanford POS Tagger[1]. It can provide some means of lexical disam-

[1] http://nlp.stanford.edu/software/tagger.shtml

biguation, e.g. indicate that the token “sounds” is a noun and not a verb in a certain context. At the same time, the CRF algorithm is provided with additional information to extract opinion targets which are multiword expressions, i.e. noun combinations. We will refer to this feature as **pos** in our result tables.

Short Dependency Path

Previous research has successfully employed paths in the dependency parse tree to link opinion expressions and the corresponding targets (Zhuang et al., 2006; Kessler and Nicolov, 2009). Both works identify direct dependency relations such as “amod” and “nsubj” as the most frequent and at the same time highly accurate connections between a target and an opinion expression. We hence label all tokens which have a direct dependency relation to an opinion expression in a sentence. The Stanford Parser[2] is employed for the constituent and dependency parsing. We will refer to this feature as **dLn** in our result tables.

Word Distance

From the work of Zhuang et al. (2006) we can infer that opinion expressions and their target(s) are not always connected via short paths in the dependency parse tree. Since we cannot capture such paths with the abovementioned feature we introduce another feature which acts as heuristic for identifying the target to a given opinion expression. Hu and Liu (2004) and Yi et al. (2003) have shown that (base) noun phrases are good candidates for opinion targets in the datasets of product reviews. We therefore label the token(s) in the closest noun phrase regarding word distance to each opinion expression in a sentence. We will refer to this feature as **wrdDist** in our result tables.

Opinion Sentence

With this feature, we simply label all tokens occurring in a sentence containing an opinion expression. This feature shall enable the CRF algorithm to distinguish between the occurence of a certain token in a sentence which contains an opinion vs. a sentence without an opinion. We will refer to this feature as **sSn** in our result tables.

Our goal is to extract individual instances of opinion targets from sentences which contain an opinion expression. This can be modeled as a sequence segmentation and labeling task. The CRF algorithm receives a sequence of tokens $t_1...t_n$ for which it has to predict a sequence of labels $l_1...l_n$. We represent the possible labels following the IOB scheme: *B-Target*, identifying the beginning of an opinion target, *I-Target* identifying the continuation of a target, and *O* for other (non-target) tokens. We model the sentences as a linear chain CRF, which is based on an undirected graph. In the graph, each node corresponds to a token in the sentence and edges connect the adjacent tokens as they appear in the sentence. In our experiments, we use the CRF implementation from the Mallet toolkit[3].

4 Experimental Setup

4.1 Datasets

In our experiments, we employ datasets from three different sources, which span four domains in total (see Table 1). All of them consist of reviews collected from Web 2.0 sites. The first dataset consists of reviews for 20 different movies collected from the Internet Movie Database. It was presented in Zhuang et al. (2006) and annotated regarding opinion target - opinion expression pairs. The second dataset consists of 234 reviews for two different web-services collected from epinions.com, as described in Toprak et al. (2010). The third dataset is an extended version of the data presented in Kessler and Nicolov (2009). The authors have provided us with additional documents, which have been annotated in the meantime. The version of the dataset used in our experiments consists of 179 blog postings regarding different digital cameras and 336 reviews of different cars. In the description of their annotation guidelines, Kessler and Nicolov (2009) refer to opinion targets as mentions. Mentions are all aspects of the review topic, which can be targets of expressed opinions. However, not only mentions which occur as opinion targets were originally annotated, but also mentions which occur in non-opinion sentences. In our experiments, we only use the mentions which occur as targets of opinion expressions.

[2] http://nlp.stanford.edu/software/lex-parser.shtml

[3] http://mallet.cs.umass.edu/

All three datasets contain annotations regarding the antecedents of anaphoric opinion targets. In our experimental setup, we do not require the algorithms to also correctly resolve the antecedent of an opinion target representy by a pronoun, as we are solely interested in evaluating the opinion target extraction not any anaphora resolution.

As shown in rows 4 and 5 of Table 1, the documents from the cars and the cameras datasets exhibit a much higher density of opinions per document. 53.5% of the sentences from the cars dataset contain an opinion and in the cameras dataset even 56.1% of the sentences contain an opinion, while in the movies and the web-services reviews just 22.1% and 22.4% of the sentences contain an opinion. Furthermore in the cars and the cameras datasets the lexical variability regarding the opinion targets is substantially larger than in the other two datasets: We calculate *target types* by counting the number of distinct opinion targets in a dataset. We divide this by the sum of all opinion target instances in the dataset. For the cars dataset this ratio is 0.440 and for the cameras dataset it is 0.433, while for the web-services dataset it is 0.306 and for the movies dataset only 0.122. In terms of reviews this means, that in the movie reviews the same movie aspects are repeatedly commented on, while in the cars and the cameras datasets many different aspects of these entities are discussed, which in turn each occur infrequently.

Table 1: Dataset Statistics

	movies	**web-services**	**cars**	**cameras**
Documents	1829	234	336	179
Sentences	24555	6091	10969	5261
Tokens / sentence	20.3	17.5	20.3	20.4
Sentences with target(s)	21.4%	22.4%	51.1%	54.0%
Sentences with opinion(s)	21.4%	22.4%	53.5%	56.1%
Targets	7045	1875	8451	4369
Target types	865	574	3722	1893
Tokens / target	1.21	1.35	1.29	1.42
Avg. targets / opinion sent.	1.33	1.37	1.51	1.53

4.2 Baseline System

In the task of opinion target extraction the supervised algorithm by Zhuang et al. (2006) represents the state-of-the-art on the movies dataset we also employ in our experiments. We therefore use it as a baseline. The algorithm learns two aspects from the labeled training data:

1. A set of opinion target candidates

2. A set of paths in a dependency tree which identify valid opinion target - opinion expression pairs

In our experiments, we learn the full set of opinion targets from the labeled training data in the first step. This is slightly different from the approach in (Zhuang et al., 2006), but we expect that this modification should be beneficial for the overall performance in terms of recall, as we do not remove any learned opinion targets from the candidate list. In the second step, the annotated sentences are parsed and a graph containing the words of a sentence is created, which are connected by the dependency relations between them. For each opinion target - opinion expression pair from the gold standard, the shortest path connecting them is extracted from the dependency graph. A path consists of the part-of-speech tags of the nodes and the dependency types of the edges. Example 3 shows a typical dependency path.

(3) NN - nsubj - NP - amod - JJ

During runtime, the algorithm identifies opinion targets from the candidate list in the training data. The opinion expressions are directly taken from the gold standard, as we focus on the opinion target extraction aspect in this work. The sentences are then parsed and if a valid path between a target and an opinion expression is found in the list of possible paths, then the pair is extracted. Since the dependency paths only identify pairs of single word target and opinion expression candidates, we employ a merging step. Extracted target candidates are merged into a multiword target if they are adjacent in a sentence. Thereby, the baseline system is also capable of extracting multiword opinion targets.

4.3 Metrics

We employ the following requirements in our evaluation of the opinion target extraction: An opinion target must be extracted with exactly the span boundaries as annotated in the gold standard. This is especially important regarding multiword targets. Extracted targets which partially overlap with the annotated gold standard are counted as errors. Hence a target extracted by the algorithm which does not exactly match the boundaries of a target in the gold standard is counted as a false positive (FP), e.g. if "battery life" is annotated as the target in the gold standard, only "battery" or "life" extracted as targets will be counted as FPs. Exact matches between the targets extracted by the algorithm and the gold standard are true positives (TP). We refer to the number of annotated targets in the gold standard as T_{GS}. Precision is calculated as $Precision = \frac{TP}{TP+FP}$, and recall is calculated as $Recall = \frac{TP}{T_{GS}}$. F-Measure is the harmonic mean of precision and recall.

5 Results and Discussion

We investigate the performance of the baseline and the CRF-based approach for opinion target extraction in a single- and cross-domain setting. The single-domain approach assumes that there is a set of training data available for the same domain as the domain the algorithm is being tested on. In this setup, we will both run the baseline and our CRF based system in a 10-fold cross-validation and report results macro averaged over all runs. In the cross-domain approach, we will investigate how the algorithm performs if given training data from domain A while being tested on another domain B. In this setting, we will train the algorithm on the entire dataset A, and test it on the entire dataset B, we hence report one micro averaged result set. In Subsection 5.1 we present the results of both the baseline system and our CRF-based approach in the single-domain setting, in Subsection 5.2 we present the results of the two systems in the cross-domain opinion target extraction.

Table 2: Single-Domain Extraction with Zhuang Baseline

Dataset	Precision	Recall	F-Measure
movies	0.663	0.592	0.625
web-services	0.624	0.394	0.483
cars	0.259	0.426	0.322
cameras	0.423	0.431	0.426

5.1 Single-Domain Results

5.1.1 Zhuang Baseline

As shown in Table 2, the state-of-the-art algorithm of Zhuang et al. (2006) performs best on the movie review dataset and worst on the cars dataset. The results on the movie dataset are higher than originally reported in (Zhuang et al., 2006) (Precision 0.483, Recall 0.585, F-Measure 0.529). We assume that this is due to two reasons: 1. In our task, the algorithm uses the opinion expression annotation from the gold standard. 2. We do not remove any learned opinion target candidates from the training data (See Section 4.2).

During training we observed that for each dataset the lists of possible dependency paths (see Example 3) contained several hundred entries, many of them only occurring once. We assume that the recall of the algorithm is limited by a large variety of possible dependency paths between opinion targets and opinion expressions, since the algorithm cannot link targets and opinion expressions in the testing data if there is no valid candidate dependency path. Furthermore, we observe that for the cars dataset the size of the dependency path candidate list (6642 entries) was approximately five times larger than the dependency graph candidate list for the web-services dataset (1237 entries), which has a comparable size regarding documents. At the same time, the list of target candidates of the cars dataset was approximately eight times larger than the target candidate list for the web-services dataset. We assume that a large number of both the target candidates as well as the dependency path candidates introduces many false positives during the target extraction, hence lowering the precision of the algorithm on the cars dataset considerably.

Table 3: Single-Domain Extraction with our CRF-based Approach

	movies			**web-services**			**cars**			**cameras**		
Features	**Prec**	**Rec**	**F-Me**	**Prec**	**Rec**	**F-Me**	**Prec**	**Rec**	**F-Me**	**Prec**	**Rec**	**F-Me**
tk, pos	0.639	0.133	0.220	0.500	0.051	0.093	0.438	0.110	0.175	0.300	0.085	0.127
tk, pos, wDs	0.542	0.181	0.271	0.451	0.272	0.339	0.570	0.354	0.436	0.549	0.375	0.446
tk, pos, dLn	0.777	0.481	0.595	0.634	0.380	0.475	0.603	0.372	0.460	0.569	0.376	0.453
tk, pos, sSn	0.673	0.637	0.653	0.604	0.397	0.476	0.453	0.180	0.257	0.398	0.172	0.238
tk, pos, dLn, wDs	**0.792**	0.481	0.598	0.620	0.354	0.450	0.603	0.389	0.473	0.596	0.425	0.496
tk, pos, sSn, wDs	0.662	0.656	0.659	0.664	0.461	0.544	0.564	0.370	0.446	0.544	0.381	0.447
tk, pos, sSn, dLn	0.791	0.477	0.594	0.654	0.501	0.568	0.598	0.384	0.467	0.586	0.391	0.468
tk, pos, sSn, dLn, wDs	0.749	**0.661**	**0.702**	**0.722**	**0.526**	**0.609**	**0.622**	**0.414**	**0.497**	0.614	**0.423**	**0.500**
pos, sSn, dLn, wDs	0.672	0.441	0.532	0.612	0.322	0.422	0.612	0.369	0.460	**0.674**	0.398	0.500

5.1.2 Our CRF-based Approach

Table 3 shows the results of the opinion target extraction using the CRF algorithm. Row 8 contains the results of the feature configuration, which yields the best performance regarding F-Measure across all datasets. We observe that our aproach outperforms the Zhuang baseline on all datasets. The gain in F-Measure is between 0.077 in the movies domain and 0.175 in the cars domain. Although the CRF-based approach clearly outperforms the baseline system on all four datasets, we also observe the same general trend regarding the individual results: The CRF yields the best results on the movies dataset and the worst results on the cars & cameras dataset.

As shown in the first row, the results when using just the token string and part-of-speech tags as features are very low, especially regarding recall. We observe that the higher the lexical variability of the opinion targets is in a dataset, the lower the results are. If we add the feature based on word distance (row 2), the recall is improved on all datasets, while the precision is slightly lowered on the movies and web-services datasets. The dependency path based feature performs better compared to the word distance heuristic as shown in row 3. The precision is considerably increased on all datasets, on the movies and cars & cameras datasets even reaching the overall highest value. At the same time, we observe an increase of recall on all datasets. The observation made in previous research that short paths in the dependency graph are a high precision indicator of related opinion expressions - opinion targets (Kessler and Nicolov, 2009) is confirmed on all datasets. Adding the information regarding opinion sentences to the basic features of the token string and the part-of-speech tag (row 4) yields the biggest improvements regarding F-Measure on the movies and web-services dataset (+0.433 / +0.383). On the cars & cameras dataset the recall is relatively low again. We assume that this is again due to the high lexical variability, so that the CRF algorithm will encounter many actual opinion targets in the testing data which have not occurred in the training data and will hence not be extracted.

As shown in row 5, if we combine the dependency graph based feature with the word distance heuristic, the results regarding F-Measure are consistently higher than the results of these features in isolation (rows 2 - 4) on all datasets. We conclude that these two features are complementary, as they apparently indicate different kinds of opinion targets which are then correctly extracted by the CRF. If we combine each of the opinion expression related features with the label which identifies opinion sentences in general (rows 6 & 7), we observe that this feature is also complementary to the others. On all datasets the results regarding F-Measure are consistently higher compared to the features in isolation (rows 2 - 4). Row 8 shows the results of all features in combination. Again, we observe the complementarity of the features, as the results of this feature combination are the best regarding F-Measure across all datasets.

In row 9 of the results, we exclude the token string as a feature. In comparison to the full feature combination of row 8 we observe a significant decrease of F-Measure on the movies and the web-services dataset. On the cars dataset we only observe a slight decrease of recall. Interestingly on the cameras dataset we even observe a slight increase of precision which compensates a slight decrease of recall, in turn resulting in stable F-Measure of 0.500 as in the full feature set of row 8.

We have run some additional experiments in which we did not rely on the annotated opinion expressions, but employed a general pupose subjectivity lexicon[4]. Already in the single-domain extraction, we observed that the results declined substantially (e.g. web-services F-Measure: 0.243, movies F-Measure: 0.309, cars F-Measure: 0.192 and cameras F-Measure: 0.198).

We performed a quantitative error analysis on the results of the CRF-based approach in the single-domain setting. In doing so, we focused on misclassifications of B-Target and I-Target instances, as the recall is consistently lower than the precision across all datasets. We observe that most of the recall errors result from one-word opinion targets or the beginning of opinion targets (B-Targets) being missclassified as non-targets (movies 83%, web-services 73%, cars 68%, cameras 64%). For the majority of these missclassifications neither the *dLn* nor the *wDs* features were present (movies 82%, web-services 56%, cars 64%, cameras 61%). We assume that our features cannot capture the structure of more complex sentences very well. Our results indicate that the *dLn* and *wDs* features are complementary, but apparently there are quite a few cases in which the opinion target is neither directly related to the opinion expression in the dependency graph nor close to it in the sentence. One of these sentences, in this case from a camera review, in shown in Example 4.

(4) A lens cap and a strap may not sound very important, but it makes a *huge difference* in the speed and usability of the camera.

In this sentence, the *dLn* and *wDs* features both labeled “speed” which was incorrectly extracted as the target of the opinion. None of the actual targets “lens cap”, “strap” and “camera” have a short dependency path to the opinion expression and “speed” is simply the closest noun to it. Note that although both “speed” and “usability” are attributes of a camera, the opinion in this sentence is about the “lens cap” and “strap”, hence only these attributes are annotated as targets.

5.2 Cross-Domain Results

5.2.1 Zhuang Baseline

Table 4 shows the results of the opinion target extraction with the state-of-the-art system in the cross-domain setting. We observe that the results on all domain combinations are very low. A quantitative error analysis has revealed that there is hardly any overlap in the opinion target candidates between domains, as reflected by the low recall in all configurations. The vocabularies of the opinion targets are too different, hence the performance of the algorithm by Zhuang et al. (2006) is so low. The overlap regarding the dependency paths between domains was however higher. Especially identical short paths could be found across domains which at the same time typically occured quite often. For future work it might be interesting to investigate how the algorithm by Zhuang et al. (2006) performs in the cross-domain setting if the target candidate learning is performed differently, e.g. with a statistical approach as outlined in Section 2.1.

5.2.2 CRF-based Approach

The results of the cross-domain target extraction with the CRF-based algorithm are shown in Table 5. Due to the increase of system configurations introduced by the training - testing data combinations, we had to limit results of the feature combinations reported in the Table. The feature combination *pos, sSn, wDs, dLn* yielded the best results regarding F-Measure. Hence, we report its result as the basic feature set. When comparing the results of the best performing feature / training data combination of our CRF-based approach with the baseline, we observe that our approach outperforms the baseline on all four domains. The gain in F-Measure is 0.409 in the movies domain, 0.242 in the web-services domain, 0.294 in the cars domain and 0.343 in the cameras domain.

Effects of Features

Interestingly with the best performing feature combination from the single-domain extraction, the results regarding recall in the cross-domain extraction are very low[5]. This is due to the fact that the CRF attributed a relatively large weight to the token string

[4] http://www.cs.pitt.edu/mpqa/

[5] Not shown in any result table due to limited space.

Table 4: Cross-Domain Extraction with Zhuang Baseline

Training	Testing	Precision	Recall	F-Measure
web-services	movies	**0.194**	0.032	0.055
cars	movies	0.032	0.034	0.033
cameras	movies	0.155	0.084	**0.109**
cars + cameras	movies	0.071	0.104	0.084
web-services + cars + cameras	movies	0.070	0.103	0.083
movies	web-services	**0.311**	0.073	**0.118**
cars	web-services	0.086	0.091	0.089
cameras	web-services	0.164	0.081	0.108
cars + cameras	web-services	0.086	**0.104**	0.094
movies + cars + cameras	web-services	0.074	0.100	0.080
movies	cars	0.182	0.014	0.026
web-services	cars	0.218	0.028	0.049
cameras	cars	**0.250**	0.121	0.163
cameras + web-services	cars	0.247	**0.131**	**0.171**
movies + web-services	cars	0.246	0.045	0.076
movies	cameras	0.108	0.012	0.022
web-services	cameras	**0.268**	0.048	0.082
cars	cameras	0.125	**0.160**	**0.140**
cars + web-services	cameras	0.119	0.157	0.136
movies + web-services	cameras	0.245	0.063	0.100

feature. As we also observed in the analysis of the baseline results, the overlap of the opinion target vocabularies between domains is low, which resulted in a very small number of targets extracted by the CRF. As shown in Table 5 the results are promising regarding F-Measure if we just leave the token feature out of the configuration.

Effects of Training Data

When analyzing the results of the different training - testing domain configurations we observe the following: In isolation training data from the cameras domain consistently yields the best results regarding F-Measure when the algorithm is run on the datasets from the other three domains. This is particularly interesting since the cameras dataset is the smallest of the four (see Table 1). We investigated whether the CRF algorithm was overfitting to the training datasets by reducing their size to the size of the cameras dataset. However, the reduction of the training data sizes never improved the extraction results regarding F-Measure for the movies, web-serviecs and cars datasets. The good results when training on the cameras dataset are in line with our observations from Section 5.1.2. We noticed that on the cameras dataset the results regarding F-Measure remained stable if the token feature is not used in the training.

In isolation, training only on the cars data yields the second highest results on the movies and web-services datasets and the highest results regarding F-Measure on the cameras data. However, the results of the cars + cameras training data combination indicate that the cameras data does not contribute any additional information during the learning, since the results on both the movies and the web-services datasets are lower than when training only on the cameras data.

Our results also confirm the insights gained by Blitzer et al. (2007), who observed that in cross-domain polarity analysis adding more training data is not always beneficial. Apparently even the smallest training dataset (cameras) contain enough feature instances to learn a model which performs well on the testing data.

We observe that the results of the cross-domain extraction regarding F-Measure come relatively close to the results of the single-domain setting, especially if the token string feature is removed there (see Table 3 row 9). On the cars and the cameras dataset the cross-domain results are even closer to the single-domain results. The features we employ seem to scale well across domains and compensate the difference between training and testing data and the lack of information regarding the target vocabu-

Table 5: Cross-Domain Extraction with our CRF-based Approach

		Testing					
		web-services			**movies**		
		Pre	**Rec**	**F-Me**	**Pre**	**Rec**	**F-Me**
Training	web-services	-	-	-	0.560	0.339	0.422
	movies	**0.565**	0.219	0.316	-	-	-
	cars	0.538	0.248	0.340	0.642	0.382	0.479
	cameras	0.529	0.256	0.345	0.642	0.408	0.499
	movies + cars	0.554	0.249	0.344	-	-	-
	movies + cameras	0.530	**0.273**	**0.360**	-	-	-
	movies + cars + cameras	0.562	0.250	0.346	-	-	-
	cars + cameras	0.538	0.254	0.345	0.641	0.395	0.489
	web-services + cars	-	-	-	**0.651**	0.396	0.492
	web-services + cameras	-	-	-	0.642	**0.435**	**0.518**
	web-services + cars + cameras	-	-	-	0.639	0.405	0.496
		cars			**cameras**		
		Pre	**Rec**	**F-Me**	**Pre**	**Rec**	**F-Me**
	web-services	0.391	0.277	0.324	0.505	0.330	0.399
	movies	0.512	0.307	0.384	0.550	0.303	0.391
	cars	-	-	-	0.665	0.369	0.475
	cameras	**0.589**	0.384	**0.465**	-	-	-
	cameras + movies	0.567	**0.394**	**0.465**	-	-	-
	cameras + web-services	0.572	0.381	0.457	-	-	-
	movies + web-services	0.489	0.327	0.392	0.553	0.339	0.421
	movies + cars	-	-	-	0.634	0.376	0.472
	web-services + cars	-	-	-	**0.678**	0.376	**0.483**
	web-services + movies + cars	-	-	-	0.635	**0.378**	0.474
	movies + web-services + cameras	0.549	0.381	0.450	-	-	-

lary.

6 Conclusions

In this paper, we have shown how a CRF-based approach for opinion target extraction performs in a single- and cross-domain setting. We have presented a comparative evaluation of our approach on datasets from four different domains. In the single-domain setting, our CRF-based approach outperforms a supervised baseline on all four datasets. Our error analysis indicates that additional features, which can capture opinions in more complex sentences, are required to improve the performance of the opinion target extraction. Our CRF-based approach also yields promising results in the cross-domain setting. The features we employ scale well across domains, given that the opinion target vocabularies are substantially different. For future work, we might investigate how machine learning algorithms, which are specifically designed for the problem of domain adaptation (Blitzer et al., 2007; Jiang and Zhai, 2007), perform in comparison to our approach. Since three of the features we employed in our CRF-based approach are based on the respective opinion expressions, it is to investigate how to mitigate the possible negative effects introduced by errors in the opinion expression identification if they are not annotated in the gold standard. We observe similar challenges as Choi et al. (2005) regarding the analysis of complex sentences. Although our data is user-generated from Web 2.0 communities, a manual inspection has shown that the documents were of relatively high textual quality. It is to investigate to which extent the approaches taken in the analysis of newswire, such as identifying targets with coreference resolution, can also be applied to our task on user-generated discourse.

Acknowledgments

The project was funded by means of the German Federal Ministry of Economy and Technology under the promotional reference “01MQ07012”. The authors take the responsibility for the contents. This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806.

References

Anthony Aue and Michael Gamon. 2005. Customizing sentiment classifiers to new domains: A case study. In *Proceedings of the 5th International Conference on Recent Advances in Natural Language Processing*, Borovets, Bulgaria, September.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, Prague, Czech Republic, June.

Kenneth Bloom, Navendu Garg, and Shlomo Argamon. 2007. Extracting appraisal expressions. In *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, pages 308–315, Rochester, New York, USA, April.

Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. Identifying sources of opinions with conditional random fields and extraction patterns. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Vancouver, Canada, October.

James R. Cowie and Wendy G. Lehnert. 1996. Information extraction. *Communications of the ACM*, 39(1):80–91.

Hal Daumé III and Daniel Marcu. 2006. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research (JAIR)*, 26:101–126.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177, Seattle, Washington, USA, August.

Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in nlp. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 264–271, Prague, Czech Republic, June. Association for Computational Linguistics.

Jason Kessler and Nicolas Nicolov. 2009. Targeting sentiment expressions through supervised ranking of linguistic configurations. In *Proceedings of the Third International AAAI Conference on Weblogs and Social Media*, pages 90–97, San Jose, California, USA, May.

Soo-Min Kim and Eduard Hovy. 2006. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the ACL Workshop on Sentiment and Subjectivity in Text*, pages 1–8, Sydney, Australia, July.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289, Williamstown, MA, USA, June.

Fuchun Peng and Andrew McCallum. 2006. Information extraction from research papers using conditional random fields. *Information Processing and Management*, 42(4):963–979, July.

Ana-Maria Popescu and Oren Etzioni. 2005. Extracting product features and opinions from reviews. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 339–346, Vancouver, Canada, October.

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.

Jeonghee Yi, Tetsuya Nasukawa, Razvan Bunescu, and Wayne Niblack. 2003. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Proceedings of the 3rd IEEE International Conference on Data Mining*, pages 427–434, Melbourne, Florida, USA, December.

Li Zhuang, Feng Jing, and Xiao-Yan Zhu. 2006. Movie review mining and summarization. In *Proceedings of the ACM 15th Conference on Information and Knowledge Management*, pages 43–50, Arlington, Virginia, USA, November.