Generalization Methods for In-Domain and Cross-Domain Opinion Holder Extraction

Michael Wiegand and **Dietrich Klakow**
Spoken Language Systems
Saarland University
D-66123 Saarbrücken, Germany
{Michael.Wiegand|Dietrich.Klakow}@lsv.uni-saarland.de

Abstract

In this paper, we compare three different generalization methods for in-domain and cross-domain opinion holder extraction being simple unsupervised word clustering, an induction method inspired by distant supervision and the usage of lexical resources. The generalization methods are incorporated into diverse classifiers. We show that generalization causes significant improvements and that the impact of improvement depends on the type of classifier and on how much training and test data differ from each other. We also address the less common case of opinion holders being realized in patient position and suggest approaches including a novel (linguistically-informed) extraction method how to detect those opinion holders without labeled training data as standard datasets contain too few instances of this type.

1 Introduction

Opinion holder extraction is one of the most important subtasks in sentiment analysis. The extraction of sources of opinions is an essential component for complex real-life applications, such as opinion question answering systems or opinion summarization systems (Stoyanov and Cardie, 2011). Common approaches designed to extract opinion holders are based on data-driven methods, in particular supervised learning.

In this paper, we examine the role of generalization for opinion holder extraction in both in-domain and cross-domain classification. Generalization may not only help to compensate the availability of labeled training data but also conciliate domain mismatches.

In order to illustrate this, compare for instance (1) and (2).

(1) Malaysia did not agree to such treatment of Al-Qaeda soldiers as they were prisoners-of-war and should be accorded treatment as provided for under the Geneva Convention.
(2) Japan wishes to build a $21 billion per year aerospace industry centered on commercial satellite development.

Though both sentences contain an opinion holder, the lexical items vary considerably. However, if the two sentences are compared on the basis of some higher level patterns, some similarities become obvious. In both cases the opinion holder is an entity denoting a person and this entity is an agent[1] of some predictive predicate (i.e. *agree* in (1) and *wishes* in (2)), more specifically, an expression that indicates that the agent utters a subjective statement. Generalization methods ideally capture these patterns, for instance, they may provide a domain-independent lexicon for those predicates. In some cases, even higher order features, such as certain syntactic constructions may vary throughout the different domains. In (1) and (2), the opinion holders are agents of a predictive predicate, whereas the opinion holder *her daughters* in (3) is a patient[2] of *embarrasses*.

(3) Mrs. Bennet does what she can to get Jane and Bingley together and embarrasses her daughters by doing so.

If only sentences, such as (1) and (2), occur in the training data, a classifier will not correctly extract the opinion holder in (3), unless it obtains additional knowledge as to which predicates take opinion holders as patients.

[1] By *agent* we always mean constituents being labeled as *A0* in PropBank (Kingsbury and Palmer, 2002).

[2] By *patient* we always mean constituents being labeled as *A1* in PropBank.

In this work, we will consider three different generalization methods being simple unsupervised word clustering, an induction method and the usage of lexical resources. We show that generalization causes significant improvements and that the impact of improvement depends on how much training and test data differ from each other. We also address the issue of opinion holders in patient position and present methods including a novel extraction method to detect these opinion holders without any labeled training data as standard datasets contain too few instances of them.

In the context of generalization it is also important to consider different classification methods as the incorporation of generalization may have a varying impact depending on how robust the classifier is by itself, i.e. how well it generalizes even with a standard feature set. We compare two state-of-the-art learning methods, conditional random fields and convolution kernels, and a rule-based method.

2 Data

As a labeled dataset we mainly use the MPQA 2.0 corpus (Wiebe et al., 2005). We adhere to the definition of opinion holders from previous work (Wiegand and Klakow, 2010; Wiegand and Klakow, 2011a; Wiegand and Klakow, 2011b), i.e. every source of a *private state* or a *subjective speech event* (Wiebe et al., 2005) is considered an opinion holder.

This corpus contains almost exclusively news texts. In order to divide it into different domains, we use the topic labels from (Stoyanov et al., 2004). By inspecting those topics, we found that many of them can grouped to a cluster of news items discussing human rights issues mostly in the context of combating global terrorism. This means that there is little point in considering every single topic as a distinct (sub)domain and, therefore, we consider this cluster as one single domain ETHICS.[3] For our cross-domain evaluation, we want to have another topic that is fairly different from this set of documents. By visual inspection, we found that the topic discussing issues regarding the International Space Station would suit our purpose. It is henceforth called SPACE.

[3]The cluster is the union of documents with the following MPQA-topic labels: *axisofevil*, *guantanamo*, *humanrights*, *mugabe* and *settlements*.

Domain	# Sentences	# Holders in sentence (average)
ETHICS	5700	0.79
SPACE	628	0.28
FICTION	614	1.49

Table 1: Statistics of the different domain corpora.

In addition to these two (sub)domains, we chose some text type that is not even news text in order to have a very distant domain. Therefore, we had to use some text not included in the MPQA corpus. Existing text collections containing product reviews (Kessler et al., 2010; Toprak et al., 2010), which are generally a popular resource for sentiment analysis, were not found suitable as they only contain few distinct opinion holders. We finally used a few summaries of fictional work (two Shakespeare plays and one novel by Jane Austen[4]) since their language is notably different from that of news texts and they contain a large number of different opinion holders (therefore opinion holder extraction is a meaningful task on this text type). These texts make up our third domain FICTION. We manually labeled it with opinion holder information by applying the annotation scheme of the MPQA corpus.

Table 1 lists the properties of the different domain corpora. Note that ETHICS is the largest domain. We consider it our primary (source) domain as it serves both as a training and (in-domain) test set. Due to their size, the other domains only serve as test sets (target domains).

For some of our generalization methods, we also need a large unlabeled corpus. We use the North American News Text Corpus (LDC95T21).

3 The Different Types of Generalization

3.1 Word Clustering (Clus)

The simplest generalization method that is considered in this paper is word clustering. By that, we understand the automatic grouping of words occurring in similar contexts. Such clusters are usually computed on a large unlabeled corpus. Unlike lexical features, features based on clusters are less sparse and have been proven to significantly improve data-driven classifiers in related tasks, such as named-entity recognition (Turian et

[4]available at: `www.absoluteshakespeare.com/guides/{othello|twelfth_night}/summary/{othello|twelfth_night}_summary.htm`
`www.wikisummaries.org/Pride_and_Prejudice`

I.	Madrid, Dresden, Bordeaux, Istanbul, Caracas, Manila, ...
II.	Toby, Betsy, Michele, Tim, Jean-Marie, Rory, Andrew, ...
III.	detest, resent, imply, liken, indicate, suggest, owe, expect, ...
IV.	disappointment, unease, nervousness, dismay, optimism, ...
V.	remark, baby, book, saint, manhole, maxim, coin, batter, ...

Table 2: Some automatically induced clusters.

ETHICS	SPACE	FICTION
1.47	2.70	11.59

Table 3: Percentage of opinion holders as patients.

al., 2010). Such a generalization is, in particular, attractive as it is cheaply produced. As a state-of-the-art clustering method, we consider Brown clustering (Brown et al., 1992) as implemented in the SRILM-toolkit (Stolcke, 2002). We induced 1000 clusters which is also the configuration used in (Turian et al., 2010).[5]

Table 2 illustrates a few of the clusters induced from our unlabeled dataset introduced in Section (§) 2. Some of these clusters represent location or person names (e.g. I. & II.). This exemplifies why clustering is effective for named-entity recognition. We also find clusters that intuitively seem to be meaningful for our task (e.g. III. & IV.) but, on the other hand, there are clusters that contain words that with the exception of their part of speech do not have anything in common (e.g. V.).

3.2 Manually Compiled Lexicons (Lex)

The major shortcoming of word clustering is that it lacks any task-specific knowledge. The opposite type of generalization is the usage of manually compiled lexicons comprising predicates that indicate the presence of opinion holders, such as *supported*, *worries* or *disappointed* in (4)-(6).

(4) I always *supported* this idea. holder:agent.
(5) This *worries* me. holder:patient
(6) He *disappointed* me. holder:patient

We follow Wiegand and Klakow (2011b) who found that those predicates can be best obtained by using a subset of Levin's verb classes (Levin, 1993) and the strong subjective expressions of the Subjectivity Lexicon (Wilson et al., 2005). For those predicates it is also important to consider in which argument position they usually take an opinion holder. Bethard et al. (2004) found the majority of holders are agents (4). A certain number of predicates, however, also have opinion holders in patient position, e.g. (5) and (6).

Wiegand and Klakow (2011b) found that many of those latter predicates are listed in one of Levin's verb classes called *amuse verbs*. While on the evaluation on the entire MPQA corpus, opinion holders in patient position are fairly rare (Wiegand and Klakow, 2011b), we may wonder whether the same applies to the individual domains that we consider in this work. Table 3 lists the proportion of those opinion holders (computed manually) based on a random sample of 100 opinion holder mentions from those corpora. The table shows indeed that on the domains from the MPQA corpus, i.e. ETHICS and SPACE, those opinion holders play a minor role but there is a notably higher proportion on the FICTION-domain.

3.3 Task-Specific Lexicon Induction (Induc)

3.3.1 Distant Supervision with Prototypical Opinion Holders

Lexical resources are potentially much more expressive than word clustering. This knowledge, however, is usually manually compiled, which makes this solution much more expensive. Wiegand and Klakow (2011a) present an intermediate solution for opinion holder extraction inspired by *distant supervision* (Mintz et al., 2009). The output of that method is also a lexicon of predicates but it is automatically extracted from a large unlabeled corpus. This is achieved by collecting predicates that frequently co-occur with prototypical opinion holders, i.e. common nouns such as *opponents* (7) or *critics* (8), if they are an agent of that predicate. The rationale behind this is that those nouns act very much like actual opinion holders and therefore can be seen as a proxy.

(7) Opponents *say* these arguments miss the point.
(8) Critics *argued* that the proposed limits were unconstitutional.

This method reduces the human effort to specifying a small set of such prototypes.

Following the best configuration reported in (Wiegand and Klakow, 2011a), we extract 250 verbs, 100 nouns and 100 adjectives from our unlabeled corpus (§2).

3.3.2 Extension for Opinion Holders in Patient Position

The downside of using prototypical opinion holders as a proxy for opinion holders is that it

[5] We also experimented with other sizes but they did not produce a better overall performance.

anguish*, astonish, astound, concern, convince, daze, delight, disenchant*, disappoint, displease, disgust, disillusion, dissatisfy, distress, embitter*, enamor*, engross, enrage, entangle*, excite, fatigue*, flatter, fluster, flummox*, frazzle*, hook*, humiliate, incapacitate*, incense, interest, irritate, obsess, outrage, perturb, petrify*, sadden, sedate*, shock, stun, tether*, trouble

Table 4: Examples of the automatically extracted verbs taking opinion holders as patients (*: not listed as *amuse verb*).

is limited to agentive opinion holders. Opinion holders in patient position, such as the ones taken by *amuse verbs* in (5) and (6), are not covered. Wiegand and Klakow (2011a) show that considering less restrictive contexts significantly drops classification performance. So the natural extension of looking for predicates having prototypical opinion holders in patient position is not effective. Sentences, such as (9), would mar the result.

(9) They criticized their opponents.

In (9) the prototypical opinion holder *opponents* (in the patient position) is not a true opinion holder.

Our novel method to extract those predicates rests on the observation that the past participle of those verbs, such as *shocked* in (10), is very often identical to some predicate adjective (11) having a similar if not identical meaning. For the predicate adjective, the opinion holder is, however, its subject/agent and not its patient.

(10) He had $shocked_{verb}$ me. holder:patient
(11) I was $shocked_{adj}$. holder:agent

Instead of extracting those verbs directly (10), we take the detour via their corresponding predicate adjectives (11). This means that we collect all those verbs (from our large unlabeled corpus (§2)) for which there is a predicate adjective that coincides with the past participle of the verb.

To increase the likelihood that our extracted predicates are meaningful for opinion holder extraction, we also need to check the semantic type in the relevant argument position, i.e. make sure that the agent of the predicate adjective (which would be the patient of the corresponding verb) is an entity likely to be an opinion holder. Our initial attempts with prototypical opinion holders were too restrictive, i.e. the number of prototypical opinion holders co-occurring with those adjectives was too small. Therefore, we widen the semantic type of this position from prototypical opinion holders to persons. This means that we allow personal pronouns (i.e. *I*, *you*, *he*, *she* and *we*) to appear in this position. We believe that this relaxation can be done in that particular case, as adjectives are much more likely to convey opinions a priori than verbs (Wiebe et al., 2004).

An intrinsic evaluation of the predicates that we thus extracted from our unlabeled corpus is difficult. The 250 most frequent verbs exhibiting this special property of coinciding with adjectives (this will be the list that we use in our experiments) contains 42% entries of the *amuse verbs* (§3.2). However, we also found many other potentially useful predicates on this list that are not listed as *amuse verbs* (Table 4). As *amuse verbs* cannot be considered a complete golden standard for all predicates taking opinion holders as patients, we will focus on a task-based evaluation of our automatically extracted list (§6).

4 Data-driven Methods

In the following, we present the two supervised classifiers we use in our experiments. Both classifiers incorporate the same levels of representations, including the same generalization methods.

4.1 Conditional Random Fields (CRF)

The supervised classifier most frequently used for information extraction tasks, in general, are conditional random fields (CRF) (Lafferty et al., 2001). Using CRF, the task of opinion holder extraction is framed as a tagging problem in which given a sequence of observations $x = x_1 x_2 \dots x_n$ (words in a sentence) a sequence of output tags $y = y_1 y_2 \dots y_n$ indicating the boundaries of opinion holders is computed by modeling the conditional probability $P(x|y)$.

The features we use (Table 5) are mostly inspired by Choi et al. (2005) and by the ones used for plain support vector machines (SVMs) in (Wiegand and Klakow, 2010). They are organized into groups. The basic group *Plain* does not contain any generalization method. Each other group is dedicated to one specific generalization method that we want to examine (*Clus*, *Induc* and *Lex*). Apart from considering generalization features indicating the presence of generalization types, we also consider those types in conjunction with semantic roles. As already indicated above, semantic roles are especially important for the detection of opinion holders. Unfortunately, the cor-

Group	Features
Plain	Token features: unigrams and bigrams POS/chunk/named-entity features: unigrams, bigrams and trigrams Constituency tree path to nearest predicate Nearest predicate Semantic role to predicate+lexical form of predicate
Clus	Cluster features: unigrams, bigrams and trigrams Semantic role to predicate+cluster-id of predicate Cluster-id of nearest predicate
Induc	Is there predicate from induced lexicon within window of 5 tokens? Semantic role to predicate, if predicate is contained in induced lexicon Is nearest predicate contained in induced lexicon?
Lex	Is there predicate from manually compiled lexicons within window of 5 tokens? Semantic role to predicate, if predicate is contained in manually compiled lexicons Is nearest predicate contained in manually compiled lexicons?

Table 5: Feature set for CRF.

responding feature from the *Plain* feature group that also includes the lexical form of the predicate is most likely a sparse feature. For the opinion holder *me* in (10), for example, it would correspond to *A1_shock*. Therefore, we introduce for each generalization method an additional feature replacing the sparse lexical item by a generalization label, i.e. *Clus*: *A1_CLUSTER-35265*, *Induc*: *A1_INDUC-PRED* and *Lex*: *A1_LEX-PRED*.[6]

For this learning method, we use CRF++.[7] We choose a configuration that provides good performance on our source domain (i.e. ETHICS).[8] For semantic role labeling we use SWIRL[9], for chunk parsing CASS (Abney, 1991) and for constituency parsing Stanford Parser (Klein and Manning, 2003). Named-entity information is provided by Stanford Tagger (Finkel et al., 2005).

4.2 Convolution Kernels (CK)

Convolution kernels (CK) are special kernel functions. A kernel function $K : X \times X \rightarrow \mathbb{R}$ computes the similarity of two data instances x_i and x_j ($x_i \wedge x_j \in X$). It is mostly used in SVMs that estimate a hyperplane to separate data instances from different classes $H(\vec{x}) = \vec{w} \cdot \vec{x} + b = 0$ where $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$ (Joachims, 1999). In convolution kernels, the structures to be compared within the kernel function are not vectors comprising manually designed features but the underlying discrete structures, such as syntactic parse trees or part-of-speech sequences. Since they are directly provided to the learning algorithm, a classifier can be built without taking the effort of implementing an explicit feature extraction.

We take the best configuration from (Wiegand and Klakow, 2010) that comprises a combination of three different tree kernels being two tree kernels based on constituency parse trees (one with *predicate* and another with *semantic scope*) and a tree kernel encoding predicate-argument structures based on semantic role information. These representations are illustrated in Figure 1. The resulting kernels are combined by plain summation.

In order to integrate our generalization methods into the convolution kernels, the input structures, i.e. the linguistic tree structures, have to be augmented. For that we just add additional nodes whose labels correspond to the respective generalization types (i.e. *Clus*: CLUSTER-*ID*, *Induc*: INDUC-PRED and *Lex*: LEX-PRED). The nodes are added in such a way that they (directly) dominate the leaf node for which they provide a generalization.[10] If several generalization methods are used and several of them apply for the same lexical unit, then the (vertical) order of the generalization nodes is LEX-PRED $\succ$ INDUC-PRED $\succ$ CLUSTER-ID.[11] Figure 2 illustrates the predicate argument structure from Figure 1 augmented with INDUC-PRED and CLUSTER-IDs.

For this learning method, we use the SVMLight-TK toolkit.[12] Again, we tune the parameters to our source domain (ETHICS).[13]

5 Rule-based Classifiers (RB)

Finally, we also consider rule-based classifiers (RB). The main difference towards CRF and CK is that it is an unsupervised approach not requiring training data. We re-use the framework by Wiegand and Klakow (2011b). The candidate set are all noun phrases in a test set. A candidate is classified as an opinion holder if all of the following

[6] Predicates in patient position are given the same generalization label as the predicates in agent position. Specially marking them did not result in a notable improvement.

[7] `http://crfpp.sourceforge.net`

[8] The soft margin parameter $-c$ is set to 1.0 and all features occurring less than 3 times are removed.

[9] `http://www.surdeanu.name/mihai/swirl`

[10] Note that even for the configuration *Plain* the trees are already augmented with named-entity information.

[11] We chose this order as it roughly corresponds to the specificity of those generalization types.

[12] `disi.unitn.it/moschitti`

[13] The cost parameter $-j$ (Morik et al., 1999) was set to 5.

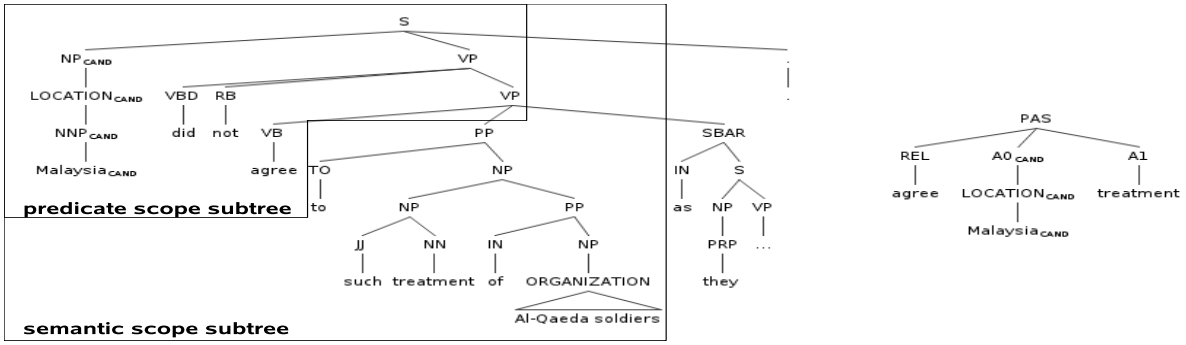

Figure 1: The different structures (*left*: constituency trees, *right*: predicate argument structure) derived from Sentence (1) for the opinion holder candidate *Malaysia* used as input for convolution kernels (CK).

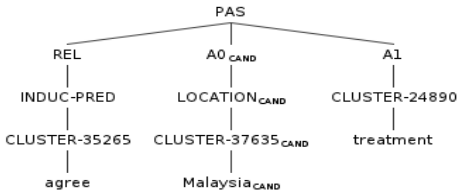

Figure 2: Predicate argument structure augmented with generalization nodes.

conditions hold:

- The candidate denotes a person or group of persons.
- There is a predictive predicate in the same sentence.
- The candidate has a pre-specified semantic role in the event that the predictive predicate evokes (*default: agent-role*).

The set of predicates is obtained from a given lexicon. For predicates that take opinion holders as patients, the default agent-role is overruled.

We consider several classifiers that differ in the lexicon they use. RB-Lex uses the combination of the manually compiled lexicons presented in §3.2. RB-Induc uses the predicates that have been automatically extracted from a large unlabeled corpus using the methods presented in §3.3. RB-Induc+Lex considers the union of those lexicons. In order to examine the impact of modeling opinion holders in patient position, we also introduce two versions of each lexicon. AG just considers predicates in agentive position while AG+PT also considers predicates that take opinion holders as patients. For example, RB-Induc$_{AG+PT}$ is a classifier that uses automatically extracted predicates in order to detect opinion holders in both agent and patient argument position, i.e. RB-Induc$_{AG+PT}$ also covers our novel extraction method for patients (§3.3.2).

The output of clustering will exclusively be evaluated in the context of learning-based methods, since there is no straightforward way of incorporating this output into a rule-based classifier.

	Features				
	Induc		Lex		Induc+Lex
Domains	**AG**	**AG+PT**	**AG**	**AG+PT**	**AG+PT**
ETHICS	50.77	50.99	52.22	52.27	**53.07**
SPACE	45.81	46.55	47.60	**48.47**	45.20
FICTION	46.59	49.97	54.84	59.35	**63.11**

Table 6: F-score of the different rule-based classifiers.

6 Experiments

CK and RB have an instance space that is different from the one of CRF. While CRF produces a prediction for every word token in a sentence, CK and RB only produce a prediction for every noun phrase. For evaluation, we project the predictions from RB and CK to word token level in order to ensure comparability. We evaluate the sequential output with precision, recall and F-score as defined in (Johansson and Moschitti, 2010; Johansson and Moschitti, 2011).

6.1 Rule-based Classifier

Table 6 shows the cross-domain performance of the different rule-based classifiers. RB-Lex performs better than RB-Induc. In comparison to the domains ETHICS and SPACE the difference is larger on FICTION. Presumably, this is due to the fact that the predicates in *Induc* are extracted from a news corpus (§2). Thus, *Induc* may slightly suffer from a domain mismatch. A combination of the two classifiers, i.e. RB-Lex+Induc, results in a notable improvement in the FICTION-domain. The approaches that also detect opinion holders as patients (AG+PT) including our novel approach (§3.3.2) are effective. A notable improvement can

		Training Size (%)				
Features	**Alg.**	**5**	**10**	**20**	**50**	**100**
Plain	CRF	32.14	35.24	41.03	51.05	55.13
	CK	42.15	46.34	51.14	56.39	59.52
+Clus	CRF	33.06	37.11	43.47	52.05	56.18
	CK	42.02	45.86	51.11	56.59	59.77
+Induc	CRF	37.28	42.31	46.54	54.27	56.71
	CK	46.26	49.35	53.26	57.28	60.42
+Lex	CRF	40.69	43.91	48.43	55.37	58.46
	CK	46.45	50.59	53.93	58.63	61.50
+Clus+Induc	CRF	37.27	42.19	47.35	54.95	57.14
	CK	45.14	48.20	52.39	57.37	59.97
+Clus+Lex	CRF	40.52	44.29	49.32	55.44	58.80
	CK	45.89	49.35	53.56	58.74	61.43
+Lex+Induc	CRF	**42.23**	**45.92**	49.96	55.61	58.40
	CK	**47.46**	**51.44**	**54.80**	58.74	**61.58**
All	CRF	41.56	45.75	**50.39**	**56.24**	**59.08**
	CK	46.18	50.10	54.04	**58.92**	61.44

Table 7: F-score of in-domain (ETHICS) learning-based classifiers.

only be measured on the FICTION-domain since this is the only domain with a significant proportion of those opinion holders (Table 3).

6.2 In-Domain Evaluation of Learning-based Methods

Table 7 shows the performance of the learning-based methods CRF and CK on an in-domain evaluation (ETHICS-domain) using different amounts of labeled training data. We carry out a 5-fold cross-validation and use $n\%$ of the training data in the training folds. The table shows that CK is more robust than CRF. The fewer training data are used the more important generalization becomes. CRF benefits much more from generalization than CK. Interestingly, the CRF configuration with the best generalization is usually as good as plain CK. This proves the effectiveness of CK. In principle, *Lex* is the strongest generalization method while *Clus* is by far the weakest. For *Clus*, systematic improvements towards no generalization (even though they are minor) can only be observed with CRF. As far as combinations are concerned, either *Lex+Induc* or *All* performs best. This in-domain evaluation proves that opinion holder extraction is different from named-entity recognition. Simple unsupervised generalization, such as word clustering, is not effective and popular sequential classifiers are less robust than margin-based tree-kernels.

Table 8 complements Table 7 in that it compares the learning-based methods with the best rule-based classifier and also displays precision and recall. RB achieves a high recall, whereas the learning-based methods always excel RB in precision.[14] Applying generalization to the learning-based methods results in an improvement of both recall and precision if few training data are used. The impact on precision decreases, however, the more training data are added. There is always a significant increase in recall but learning-based methods may not reach the level of RB even though they use the same resources. This is a side-effect of preserving a much higher precision. It also explains why learning-based methods with generalization may have a lower F-score than RB.

6.3 Out-of-Domain Evaluation of Learning-based Methods

Table 9 presents the results of out-of-domain classifiers. The complete ETHICS-dataset is used for training. Some properties are similar to the previous experiments: CK always outperforms CRF. RB provides a high recall whereas the learning-based methods maintain a higher precision. Similar to the in-domain setting using few labeled training data, the incorporation of generalization increases both precision and recall. Moreover, a combination of generalization methods is better than just using one method on average, although *Lex* is again a fairly robust individual generalization method. Generalization is more effective in this setting than on the in-domain evaluation using all training data, in particular for CK, since the training and test data are much more different from each other and suitable generalization methods partly close that gap.

There is a notable difference in precision between the SPACE- and FICTION-domain (and also the source domain ETHICS (Table 8)). We strongly assume that this is due to the distribution of opinion holders in those datasets (Table 1). The FICTION-domain contains much more opinion holders, therefore the chance that a predicted opinion holder is correct is much higher.

With regard to recall, a similar level of performance as in the ETHICS-domain can only be achieved in the SPACE-domain, i.e. CK achieves a recall of 60%. In the FICTION-domain, however, the recall is much lower (best recall of CK is below 47%). This is no surprise as the SPACE-domain is more similar to the source domain than

[14]The reason for RB having a high recall is extensively discussed in (Wiegand and Klakow, 2011b).

the FICTION-domain since ETHICS and SPACE are news texts. FICTION contains more out-of-domain language. Therefore, RB (which exclusively uses domain-independent knowledge) outperforms both learning-based methods including the ones incorporating generalization. Similar results have been observed for rule-based classifiers from other tasks in cross-domain sentiment analysis, such as subjectivity detection and polarity classification. High-level information as it is encoded in a rule-based classifier generalizes better than learning-based methods (Andreevskaia and Bergler, 2008; Lambov et al., 2009).

We set up another experiment exclusively for the FICTION-domain in which we combine the output of our best learning-based method, i.e. CK, with the prediction of a rule-based classifier. The combined classifier will predict an opinion holder, if either classifier predicts one. The motivation for this is the following: The FICTION-domain is the only domain to have a significant proportion of opinion holders appearing as patients. We want to know how much of them can be recognized with the best out-of-domain classifier using training data with only very few instances of this type and what benefit the addition of using various RBs which have a clearer notion of these constructions brings about. Moreover, we already observed that the learning-based methods have a bias towards preserving a high precision and this may have as a consequence that the generalization features incorporated into CK will not receive sufficiently large weights. Unlike the SPACE-domain where a sufficiently high recall is already achieved with CK (presumably due to its stronger similarity towards the source domain) the FICTION-domain may be more severely affected by this bias and evidence from RB may compensate for this.

Table 10 shows the performance of those combined classifiers. For all generalization types considered, there is, indeed, an improvement by adding information from RB resulting in a large boost in recall. Already the application of our induction approach *Induc* results in an increase of more than 8% points compared to plain CK. The table also shows that there is always some improvement if RB considers opinion holders as patients (AG+PT). This can be considered as some evidence that (given the available data we use) opinion holders in patient position can only be effectively extracted with the help of RBs. It is also further evidence that our novel approach to extract those predicates (§3.3.2) is effective.

		CRF			***CK***		
Size	**Feat.**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
10	Plain	52.17	26.61	35.24	58.26	38.47	46.34
	All	62.85	35.96	45.75	63.18	41.50	50.10
50	Plain	59.85	44.50	51.05	59.60	53.50	56.39
	All	62.99	50.80	56.24	61.91	56.20	58.92
100	Plain	64.14	48.33	55.13	62.38	56.91	59.52
	All	**64.75**	54.32	**59.08**	**63.81**	59.24	**61.44**
	RB	47.38	**60.32**	53.07	47.38	**60.32**	53.07

Table 8: Comparison of best RB with learning-based approaches on in-domain classification.

Algorithms	**Generalization**	**Prec**	**Rec**	**F**
CK (Plain)		66.90	41.48	51.21
CK	Induc	67.06	45.15	53.97
CK+RB$_{AG}$	Induc	60.22	54.52	57.23
CK+RB$_{AG+PT}$	Induc	61.09	58.14	59.58
CK	Lex	69.45	46.65	55.81
CK+RB$_{AG}$	Lex	67.36	59.02	62.91
CK+RB$_{AG+PT}$	Lex	68.25	63.28	65.67
CK	Induc+Lex	**69.73**	46.17	55.55
CK+RB$_{AG}$	Induc+Lex	61.41	65.56	63.42
CK+RB$_{AG+PT}$	Induc+Lex	62.26	**70.56**	**66.15**

Table 10: Combination of out-of-domain CK and rule-based classifiers on FICTION (i.e. distant domain).

The combined approach in Table 10 not only outperforms CK (discussed above) but also RB (Table 6). We manually inspected the output of the classifiers to find also cases in which CK detect opinion holders that RB misses. CK has the advantage that it is not only bound to the relationship between candidate holder and predicate. It learns further heuristics, e.g. that sentence-initial mentions of persons are likely opinion holders. In (12), for example, this heuristics fires while RB overlooks this instance as *to give someone a share of advice* is not part of the lexicon.

(12) She later *gives* Charlotte her *share of advice* on running a household.

7 Related Work

The research on opinion holder extraction has been focusing on applying different data-driven approaches. Choi et al. (2005) and Choi et al. (2006) explore conditional random fields, Wiegand and Klakow (2010) examine different combinations of convolution kernels, while Johansson and Moschitti (2010) present a re-ranking approach modeling complex relations between multiple opinions in a sentence. A comparison of

	SPACE (similar target domain)						*FICTION (distant target domain)*					
	CRF			**CK**			**CRF**			**CK**		
Features	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
Plain	47.32	48.62	47.96	45.89	57.07	50.87	68.58	28.96	40.73	66.90	41.48	51.21
+Clus	49.00	48.62	48.81	49.23	57.64	53.10	71.85	32.21	44.48	67.54	41.21	51.19
+Induc	42.92	49.15	45.82	46.66	**60.45**	52.67	71.59	34.77	46.80	67.06	45.15	53.97
+Lex	49.65	49.07	49.36	49.60	59.88	54.26	71.91	35.83	47.83	69.45	46.65	55.81
+Clus+Induc	46.61	48.78	47.67	48.65	58.20	53.00	71.32	35.88	47.74	67.46	42.17	51.90
+Lex+Induc	48.75	50.87	49.78	49.92	58.76	53.98	**74.02**	37.37	49.67	69.73	46.17	55.55
+Clus+Lex	49.72	50.87	50.29	**53.70**	59.32	**56.37**	73.41	37.15	49.33	70.59	43.98	54.20
All	**49.87**	51.03	**50.44**	51.68	58.76	54.99	72.00	37.44	49.26	**70.61**	44.83	54.84
best RB	41.72	**57.80**	48.47	41.72	57.80	48.47	63.26	**62.96**	**63.11**	63.26	**62.96**	**63.11**

Table 9: Comparison of best RB with learning-based approaches on out-of-domain classification.

those methods has not yet been attempted. In this work, we compare the popular state-of-the-art learning algorithms conditional random fields and convolution kernels for the first time. All these data-driven methods have been evaluated on the MPQA corpus. Some generalization methods are incorporated but unlike this paper they are neither systematically compared nor combined. The role of resources that provide the knowledge of argument positions of opinion holders is not covered in any of these works. This kind of knowledge should be directly learnt from the labeled training data. In this work, we found, however, that the distribution of argument positions of opinion holders varies throughout the different domains and, therefore, cannot be learnt from any arbitrary out-of-domain training set.

Bethard et al. (2004) and Kim and Hovy (2006) explore the usefulness of semantic roles provided by FrameNet (Fillmore et al., 2003). Bethard et al. (2004) use this resource to acquire labeled training data while in (Kim and Hovy, 2006) FrameNet is used within a rule-based classifier mapping frame-elements of frames to opinion holders. Bethard et al. (2004) only evaluate on an artificial dataset (i.e. a subset of sentences from FrameNet and PropBank (Kingsbury and Palmer, 2002)). The only realistic test set on which Kim and Hovy (2006) evaluate their approach are news texts. Their method is compared against a simple rule-based baseline and, unlike this work, not against a robust data-driven algorithm.

(Wiegand and Klakow, 2011b) is similar to (Kim and Hovy, 2006) in that a rule-based approach is used relying on the relationship towards predictive predicates. Diverse resources are considered for obtaining such words, however, they are only evaluated on the entire MPQA corpus.

The only cross-domain evaluation of opinion holder extraction is reported in (Li et al., 2007) using the MPQA corpus as a training set and the NTCIR collection as a test set. A low cross-domain performance is obtained and the authors conclude that this is due to the very different annotation schemes of those corpora.

8 Conclusion

We examined different generalization methods for opinion holder extraction. We found that for in-domain classification, the more labeled training data are used, the smaller is the impact of generalization. Robust learning methods, such as convolution kernels, benefit less from generalization than weaker classifiers, such as conditional random fields. For cross-domain classification, generalization is always helpful. Distant domains are problematic for learning-based methods, however, rule-based methods provide a reasonable recall and can be effectively combined with the learning-based methods. The types of generalization that help best are manually compiled lexicons followed by an induction method inspired by distant supervision. Finally, we examined the case of opinion holders as patients and also presented a novel automatic extraction method that proved effective. Such dedicated extraction methods are important as common labeled datasets (from the news domain) do not provide sufficient training data for these constructions.

Acknowledgements

This work was funded by the German Federal Ministry of Education and Research (Software-Cluster) under grant no. “01IC10S01”. The authors thank Alessandro Moschitti, Benjamin Roth and Josef Ruppenhofer for their technical support and interesting discussions.

References

Steven Abney. 1991. Parsing By Chunks. In Robert Berwick, Steven Abney, and Carol Tenny, editors, *Principle-Based Parsing*. Kluwer Academic Publishers, Dordrecht.

Alina Andreevskaia and Sabine Bergler. 2008. When Specialists and Generalists Work Together: Overcoming Domain Dependence in Sentiment Tagging. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL/HLT)*, Columbus, OH, USA.

Steven Bethard, Hong Yu, Ashley Thornton, Vasileios Hatzivassiloglou, and Dan Jurafsky. 2004. Extracting Opinion Propositions and Opinion Holders using Syntactic and Lexical Cues. In *Computing Attitude and Affect in Text: Theory and Applications*. Springer-Verlag.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18:467–479.

Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP)*, Vancouver, BC, Canada.

Yejin Choi, Eric Breck, and Claire Cardie. 2006. Joint Extraction of Entities and Relations for Opinion Recognition. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Sydney, Australia.

Charles. J. Fillmore, Christopher R. Johnson, and Miriam R. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16:235 – 250.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Ann Arbor, MI, USA.

Thorsten Joachims. 1999. Making Large-Scale SVM Learning Practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press.

Richard Johansson and Alessandro Moschitti. 2010. Reranking Models in Fine-grained Opinion Analysis. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, Bejing, China.

Richard Johansson and Alessandro Moschitti. 2011. Extracting Opinion Expressions and Their Polarities – Exploration of Pipelines and Joint Models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Portland, OR, USA.

Jason S. Kessler, Miriam Eckert, Lyndsay Clarke, and Nicolas Nicolov. 2010. The ICWSM JDPA 2010 Sentiment Corpus for the Automotive Domain. In *Proceedings of the International AAAI Conference on Weblogs and Social Media Data Challange Workshop (ICWSM-DCW)*, Washington, DC, USA.

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*, Sydney, Australia.

Paul Kingsbury and Martha Palmer. 2002. From TreeBank to PropBank. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, Las Palmas, Spain.

Dan Klein and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Sapporo, Japan.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Dinko Lambov, Gaël Dias, and Veska Noncheva. 2009. Sentiment Classification across Domains. In *Proceedings of the Portuguese Conference on Artificial Intelligence (EPIA)*, Aveiro, Portugal. Springer-Verlag.

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press.

Yangyong Li, Kalina Bontcheva, and Hamish Cunningham. 2007. Experiments of Opinion Analysis on the Corpora MPQA and NTCIR-6. In *Proceedings of the NTCIR-6 Workshop Meeting*, Tokyo, Japan.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant Supervision for Relation Extraction without Labeled Data. In *Proceedings of the Joint Conference of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL/IJCNLP)*, Singapore.

Katharina Morik, Peter Brockhausen, and Thorsten Joachims. 1999. Combining Statistical Learning with a Knowledge-based Approach - A Case Study in Intensive Care Monitoring. In *Proceedings the International Conference on Machine Learning (ICML)*.

Andreas Stolcke. 2002. SRILM - An Extensible Language Modeling Toolkit. In *Proceedings of the In-*

ternational Conference on Spoken Language Processing (ICSLP), Denver, CO, USA.

Veselin Stoyanov and Claire Cardie. 2011. Automatically Creating General-Purpose Opinion Summaries from Text. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, Hissar, Bulgaria.

Veselin Stoyanov, Claire Cardie, Diane Litman, and Janyce Wiebe. 2004. Evaluating an Opinion Annotation Scheme Using a New Multi-Perspective Question and Answer Corpus. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text*, Menlo Park, CA, USA.

Cigdem Toprak, Niklas Jakob, and Iryna Gurevych. 2010. Sentence and Expression Level Annotation of Opinions in User-Generated Discourse. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Uppsala, Sweden.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word Representations: A Simple and General Method for Semi-supervised Learning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Uppsala, Sweden.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning Subjective Language. *Computational Linguistics*, 30(3).

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation*, 39(2/3):164–210.

Michael Wiegand and Dietrich Klakow. 2010. Convolution Kernels for Opinion Holder Extraction. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, Los Angeles, CA, USA.

Michael Wiegand and Dietrich Klakow. 2011a. Prototypical Opinion Holders: What We can Learn from Experts and Analysts. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, Hissar, Bulgaria.

Michael Wiegand and Dietrich Klakow. 2011b. The Role of Predicates in Opinion Holder Extraction. In *Proceedings of the RANLP Workshop on Information Extraction and Knowledge Acquisition (IEKA)*, Hissar, Bulgaria.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP)*, Vancouver, BC, Canada.