Opinion Holder and Target Extraction based on the Induction of Verbal Categories

Michael Wiegand
Spoken Language Systems
Saarland University
D-66123 Saarbrücken, Germany
michael.wiegand@lsv.uni-saarland.de

Josef Ruppenhofer
Dept. of Information Science
and Language Technology
Hildesheim University
D-31141 Hildesheim, Germany
ruppenho@uni-hildesheim.de

Abstract

We present an approach for opinion role induction for verbal predicates. Our model rests on the assumption that opinion verbs can be divided into three different types where each type is associated with a characteristic mapping between semantic roles and opinion holders and targets. In several experiments, we demonstrate the relevance of those three categories for the task. We show that verbs can easily be categorized with semi-supervised graph-based clustering and some appropriate similarity metric. The seeds are obtained through linguistic diagnostics. We evaluate our approach against a new manually-compiled opinion role lexicon and perform in-context classification.

1 Introduction

While there has been much research in sentiment analysis on subjectivity detection and polarity classification, there has been less work on the extraction of opinion roles, i.e. entities that express an opinion (*opinion holders*), and entities or propositions at which sentiment is directed (*opinion targets*). Previous research relies on large amounts of labeled training data or leverages general semantic resources which are expensive to construct, e.g. FrameNet (Baker et al., 1998).

In this paper, we present an approach to induce opinion roles of verbal predicates. The input is a set of opinion verbs that can be found in a common sentiment lexicon. Our model rests on the assumption that those verbs can be divided into three different types. Each type has a characteristic mapping between semantic roles and opinion holders and targets. Thus, the problem of opinion role induction is reduced to automatically categorizing opinion verbs.

We frame the task of opinion role extraction as a triple $(pred, const, role)$ where $pred$ is a predicate evoking an opinion (we exclusively focus on opinion verbs), $const$ is some constituent bearing a semantic role assigned by $pred$, and $role$ is the opinion role that is assigned to $const$.

Our work assumes the knowledge of opinion words. We do not cover polarity classification. Many lexicons with that kind of information already exist. Our sole interest is the assignment of opinion holder and target given some opinion verb. There does not exist any publicly available lexical resource specially designed for this task.

For the induction of opinion verb types, we consider semi-supervised graph clustering with some appropriate similarity metric. We also propose an effective method for deriving seeds automatically by applying some linguistic diagnostics.

Our approach is evaluated in a supervised learning scenario on a set of sentences with annotated opinion holders and targets. We employ different kinds of features, including features derived from a semantic parser based on FrameNet. We also compare our proposed model based on the three opinion verb types against a new manually-compiled lexicon in which the semantic roles of opinion holders and targets for each individual verb have been explicitly enumerated.

We also evaluate our approach in the context of cross-domain opinion holder extraction. Thus we demonstrate the importance of our approach in the context of previous datasets and classifiers.

This is the first work that proposes to induce both opinion holders and targets evoked by opinion verbs with data-driven methods. Unlike previous work, we are able to categorize all verbs of a pre-specified set of opinion verbs. Our approach is a low-resource approach that is also applicable to languages other than English. We demonstrate this on German. A by-product of our study are new resources including a verb lexicon specifying

semantic roles for holders and targets.

2 Lexicon-based Opinion Role Extraction

Opinion holder and target extraction is a hard task (Ruppenhofer et al., 2008). Conventional syntactic or semantic levels of representation do not capture sufficient information that allows a reliable prediction of opinion holders and targets. This is illustrated by (1) and (2) which show that, even with common semantic roles, i.e. *agent* and *patient*[1], assigned to the entities, one may not be able to discriminate between the opinion roles.

(1) Peter$_{agent}$ **criticized** Mary$_{patient}$.
$(criticize, Peter, holder)$ & $(criticize, Mary, target)$

(2) Peter$_{agent}$ **disappoints** Mary$_{patient}$.
$(disappoint, Peter, target)$ & $(disappoint, Mary, holder)$

We assume that it is lexical information that decides what semantic role an opinion holder or opinion target takes. As a consequence, we built a gold-standard lexicon for verbs that encodes such information. For example, it states that the target of *criticize* is its patient, while for *disappoint*, the target is its agent. This **fine-grained lexicon** also accounts for the fact that a constituent can have several roles given the same opinion verb. An extreme case is:

(3) [Peter]$_1$ **persuades** [Mary]$_2$ [to accept his invitation]$_3$.

The sentence conveys that:

- Peter wants Mary to do something. *(view$_1$)*
- Mary is influenced by Peter. *(view$_2$)*
- Peter has some attitude towards Mary accepting his invitation. *(view$_3$)*
- Mary has some attitude towards accepting Peter's invitation. *(view$_4$)*

This corresponds to the role assignments:

- view$_1$: $(persuade, [1], holder)$, $(persuade, [2], target)$
- view$_2$: $(persuade, [2], holder)$, $(persuade, [1], target)$
- view$_3$: $(persuade, [1], holder)$, $(persuade, [3], target)$
- view$_4$: $(persuade, [2], holder)$, $(persuade, [3], target)$

(in short: 2 opinion holders and 3 opinion targets).

Our lexicon also includes another dimension neglected in many previous works. Many opinion verbs predominantly express the sentiment of the speaker of the utterance (or some nested source) (4). This concept is also known as *expressive subjectivity* (Wiebe et al., 2005) or *speaker subjectivity* (Maks and Vossen, 2012). In such opinions, the opinion holder is not realized as a dependent of the opinion verb.

(4) At my work, [they]$_1$ are constantly **gossiping**.
$(gossip, \texttt{speaker}, holder)$ & $(gossip, [1], target)$

[1]By *agent* and *patient*, we mean constituents labeled as $A0$ and $A1$ in PropBank (Kingsbury and Palmer, 2002).

Our lexicon covers the 1175 verb lemmas contained in the Subjectivity Lexicon (Wilson et al., 2005). We annotated the semantic roles similar to the format of PropBank (Kingsbury and Palmer, 2002). The basis of the annotation were online dictionaries (e.g. *Macmillan Dictionary*) which provide both a verb definition and example sentences. We do not annotate implicature-related information about effects (Deng and Wiebe, 2014) but inherent sentiment *(the data release[2] includes more details regarding the annotation process and our notion of holders and targets)*.

On a sample of 400 verbs, we measured an interannotation agreement of Cohen's $\kappa = 60.8$ for opinion holders, $\kappa = 62.3$ for opinion targets and $\kappa = 59.9$ for speaker views. This agreement is mostly substantial (Landis and Koch, 1977).

3 The Three Verb Categories

Rather than induce the opinion roles for individual verbs, we group verbs that share similar opinion role subcategorization. Thus, the main task for induction is to decide which type an opinion verb belongs to. Once the verb type has been established, the typical semantic roles for opinion holders and targets can be derived from that type. The verb categorization is motivated by the semantic roles of the three common views (Table 1) that an opinion holder can take. In our lexicon, all of the opinion holders were observed with either of these semantic roles. For facilitating induction, we assume that those types are disjoint (see also §3.4).

3.1 Verbs with Agent View (AG)

Verbs with an agent view, such as *criticize*, *love* and *believe*, convey the sentiment of its agent. Therefore, those verbs take the agent as opinion holder and the patient as opinion target. Table 1 also exemplifies semantic role labels as a suitable basis to align opinion holders and targets within a particular verb type. For example, targets of AG-verbs align to the patient, yet the patient can take the form of various phrase types (i.e. NPs, PPs or infinitive/complement phrases[3]).

[2]available at: `www.coli.uni-saarland.de/~miwieg/conll_2015_op_roles_data.tgz`

[3]Note that infinitive and complement clauses may represent a semantic role other than *patient* (e.g. the infinitive clause in (3)). As these types of clauses are fairly unambiguous, we marked them as targets even if they are no patients.

Type	Example	Holder	Target
AG	[They]$_{agent}$ **like** [the idea]$_{patient}$.	agent	pat.
	[The guests]$_{agent}$ **complained** [about noise]$_{patient}$.		
	[They]$_{agent}$ **argue** [that this plan is infeasible]$_{patient}$.		
PT	[The noise]$_{agent}$ **irritated** [the guests]$_{patient}$.	pat.	agent
	[That gift]$_{agent}$ **pleased** [her]$_{patient}$ very much.		
SP	[They]$_{agent}$ **cheated** [in the exam]$_{adjunct}$.	-N/A-	agent,
	[He]$_{agent}$ **besmirched** [the King's name]$_{patient}$.		(pat.)

Table 1: Verb types for opinion role extraction.

Type	Freq	Type	Freq
verbs with **AG**-view	868	verbs with **PT**-view	392
verbs with exclusive **AG**-view	371	verbs with exclusive **PT**-view	117
verbs with **AG**- and SP-view	352	verbs with **PT**- and AG-view	226
verbs with **AG**- and PT-view	226	verbs with **PT**- and SP-view	139
verbs with **SP**-view	537		
verbs with exclusive **SP**-view	134		
verbs with **SP**- and AG-view	352		
verbs with **SP**- and PT-view	139		

Table 2: Verb types in the fine-grained lexicon.

Agent (AG)		Patient (PT)		Speaker (SP)	
Freq	Percent	Freq	Percent	Freq	Percent
450	38.3	188	16.0	537	45.7

Table 3: Verb types in the coarse-grained lexicon.

3.2 Verbs with Patient View (PT)

Verbs with a patient view (*irritate*, *upset* and *disappoint*) are opposite to AG-verbs in that those verbs have the patient as opinion holder and the agent as opinion target.

3.3 Verbs with Speaker View (SP)

The third type we consider comprises all verbs whose perspective is that of the speaker. That is, these are verbs whose sentiment is primarily that of the speaker of the utterance rather than persons involved in the action to which is referred. Typical examples are *gossip*, *improve* or *cheat*.

While the agent is usually the target of the sentiment of the speaker, it depends on the specific verb whether its patient is also a target or not (in Table 1, only the patient of the second SP-verb, i.e. *besmirch*, is considered a target[4]). Since we aim at a precise induction approach, we will always (only) mark the agent of an induced SP-verb as a target.

3.4 Relation to Fine-Grained Lexicon

Table 2 provides statistics as to how clear-cut the three prototypical verb types are in the manually-compiled fine-grained lexicon. These numbers suggest that many verbs evoke several opinion views (e.g. a verb with an AG-view may also evoke a PT-view). While the fine-grained lexicon is fairly exhaustive in listing semantic roles for opinion holders and targets, it may also occasionally overgenerate. One major reason for this is that we do not annotate on the sense-level (word-sense disambiguation (Wiebe and Mihalcea, 2006) is still in its infancy) but on the lemma-level. Accordingly, we attribute all views to all senses, whereas actually certain views pertain only to specific senses. However, we found that usually one view is conveyed by most (if not all) senses of a word. For example, the lexicon lists both an AG-view and a PT-view for *appease*. This is correct for (5) but wrong for (6). The AG-view is derived from a definition *to give your opponents what they want*. (6) does not convey an agent's volitional action. Here, the verb just conveys *make someone feel less angry*. Similarly, the lexicon lists an SP-view and an AG-view for *degrade*, which is right for (7) but wrong for (8). The AG-view is derived from a lexicon definition *to treat someone in a way that makes them stop respecting themselves*. (8) does not convey an agent's volitional action. The verb just conveys *to make something worse*. That is, neither (6) nor (8) evoke an AG-view. We found that these variations regularly occur. We adopt the heuristic that verbs with an SP-view and AG- or PT-view preserve the SP-view across their uses (7)-(8). Verbs with both PT- and AG-view preserve their PT-view (5)-(6). Following these observations, we converted our fine-grained lexicon into a gold standard **coarse-grained lexicon** (only 3% of the verbs needed to be manually corrected after the automatic conversion) in which a verb is classified as AG, PT or SP according to its **dominant view**. The final class distribution of this lexicon is shown in Table 3. In §5.2, we show through an in-context evaluation that our coarse-grained representation preserves most of the information captured by the fine-grained representation.

(5) [Chamberlain]$_{agent}$ **appeased** [Hitler]$_{patient}$.
(6) [The orange juice]$_{agent}$ **appeased** [him]$_{patient}$ for a while.
(7) [Mary]$_{agent}$ **degrades** [Henrietta]$_{patient}$.
(8) [This technique]$_{agent}$ **degrades** [the local water supply]$_{patient}$.

4 Induction of Verb Categories

The task is to categorize each verb as a predominant AG-, PT-, or SP-verb. Our approach comprises two steps. In the first step, seeds for the different verb types are extracted (§4.1). In the sec-

[4]We consider the patient a target since the speaker has a positive (non-defeasible) sentiment towards that entity.

AG	argue, contend, speculate, fear, doubt, complain, consider, praise, recommend, view, acknowledge, hope
PT	interest, surprise, please, excite, disappoint, delight, impress, shock, trouble, embarrass, annoy, distress
SP	murder, plot, incite, blaspheme, bewitch, bungle, despoil, plagiarize, prevaricate, instigate, molest, conspire

Table 4: The top 12 extracted verb seeds.

ond step, a similarity metric (§4.2) is employed in order to propagate the verb type labels from the seeds to the remaining opinion verbs (§4.3). The *North American News Text Corpus* is used for seed extraction and computation of verb similarities.

Wiegand and Klakow (2012) proposed methods for extracting AG- and PT-verbs. We will re-use these methods for generating seeds. A major contribution of this paper is the introduction of the third dimension, i.e. SP-verbs, in the context of induction. We show that in combination with this third dimension, one can categorize all opinion verbs contained in a sentiment lexicon. Furthermore, given this three-way classification, we also obtain better results on the detection of AG-verbs and PT-verbs than by just detecting those verbs in isolation without graph clustering (this will be shown in Table 7 and discussed in §5.1).

A second major contribution of this work is that we show that these methods are also equally important for *opinion target extraction*. So far, the significance of AG- and PT-verbs has only been demonstrated for opinion holder extraction.

In this work, we exclusively focus on the set of 1175 opinion verbs from the Subjectivity Lexicon. However, this is owed solely to the effort required to generate larger sets of evaluation data. In principle, our induction approach is applicable to any set of opinion verbs of arbitrary (e.g. larger) size.

4.1 Pattern-based Seed Initialization

For AG-verbs, we rely on the findings of Wiegand and Klakow (2012) who suggest that verbs predictive for opinion holders can be induced with the help of *prototypical opinion holders*. These common nouns, e.g. *opponents* (9) or *critics* (10), act like opinion holders and, therefore, can be seen as a proxy. Verbs co-occurring with prototypical opinion holders do not represent the entire range of opinion verbs but coincide with AG-verbs.

(9) **Opponents** *claim* these arguments miss the point.
(10) **Critics** *argued* that the proposed limits were unconstitutional.

For PT-verbs, we make use of the *adjective heuristic* proposed by Wiegand and Klakow (2012). The authors make use of the observation that morphologically related adjectives exist for PT-verbs, unlike for AG- and SP-verbs. Therefore, in order to extract PT-verbs, one needs to check whether a verb in its past participle form, such as *upset* in (11), is identical to some predicate adjective (12).

(11) He had $upset_{verb}$ me.
(12) I am $upset_{adj}$.

We are not aware of any previously published approach effectively inducing SP-verbs. Noticing that many of those verbs contain some form of reproach, we came up with the patterns *accused of* X_{VBG} and *blamed for* X_{VBG} as in (13) and (14).

(13) He was **accused of** *falsifying* the documents.
(14) The UN was **blamed for** *misinterpreting* climate data.

Table 4 lists for each of the verb types the 12 seeds most frequently occurring with the respective patterns. We observed that the SP-verb seeds are exclusively negative polar expressions. That is why we also extracted seeds from an additional pattern *help to* X_{VB} producing prototypical *positive* SP-verbs, such as *stabilize*, *allay* or *heal*.

4.2 Similarity Metrics

4.2.1 Word Embeddings

Recent research in machine learning has focused on inducing vector representations of words. As an example of a competitive word embedding method, we induce vectors for our opinion verbs with *Word2Vec* (Mikolov et al., 2013). Baroni et al. (2014) showed that this method outperforms count vector representations on a variety of tasks. For the similarity between two verbs, we compute the cosine-similarity between their vectors.

4.2.2 WordNet::Similarity

We use *WordNet::Similarity* (Pedersen et al., 2004) as an alternative source for similarity metrics. The metrics are based on WordNet's graph structure (Miller et al., 1990). Various relations within WordNet have been shown to be effective for polarity classification (Esuli and Sebastiani, 2006; Rao and Ravichandran, 2009).

4.2.3 Coordination

Another method to measure similarity is obtained by leveraging coordination. Coordination is known to be a syntactic relation that also preserves great semantic coherence (Ziering et al., 2013), e.g. (15). It has been successfully applied

not only to noun categorization (Riloff and Shepherd, 1997; Roark and Charniak, 1998) but also to different tasks in sentiment analysis, including polarity classification (Hatzivassiloglou and McKeown, 1997), the induction of patient polarity verbs (Goyal et al., 2010) and connotation learning (Kang et al., 2014). We use the dependency relation from Stanford parser (Klein and Manning, 2003) to detect coordination (16).

(15) They *criticize* **and** *hate* him.
(16) `conj(criticize,hate)`

As a similarity function, we simply take the absolute frequency of observing two words w_1 and w_2 in a conjunction, i.e. $sim(w_1, w_2) = freq(conj(w_1, w_2))$.

4.2.4 Dependency-based Similarity

The metric proposed by Lin (1998) exploits the rich set of dependency-relation labels in the context of distributional similarity. Moreover, it has been effectively used for the related task of extending frames of unknown predicates in semantic parsing (Das and Smith, 2011).

The metric is based on dependency triples (w, r, w') where w and w' are words and r is a dependency relation (e.g. `(argue-V, nsubj, critics-N)`). The metric is defined as: $sim(w_1, w_2) = \frac{\sum_{(r,w)\in T(w_1)\cap T(w_2)}(I(w_1,r,w)+I(w_2,r,w))}{\sum_{(r,w)\in T(w_1)} I(w_1,r,w)+\sum_{(r,w)\in T(w_2)} I(w_2,r,w)}$ where $I(w, r, w') = log\frac{\|w,r,w'\|\times\|*,r,*\|}{\|w,r,*\|\times\|*,r,w'\|}$ and $T(w)$ is defined as the set of pairs (r, w') such that $log\frac{\|w,r,w'\|\times\|*,r,*\|}{\|w,r,*\|\times\|*,r,w'\|} > 0$.

4.3 Propagation Methods

We use the k **nearest neighbour classifier (kNN)** (Cover and Hart, 1967) as a simple method for propagating labels from seeds to other instances. Alternatively, we consider verb categorization as a **clustering task on a graph** $G = (V, E, W)$ where V is the set of nodes (i.e. our opinion verbs), E is the set of edges connecting them with weights $W : E \rightarrow \mathbb{R}^+$. W can be directly derived from any of the similarity metrics (§4.2.1-§4.2.4). The aim is that all nodes $v \in V$ are assigned a label $l \in \{AG, PT, SP\}$. Initially, only the verb seeds are labeled. We then use the Adsorption label propagation algorithm from *junto* (Talukdar et al., 2008) in order propagate the labels from the seeds to the remaining verbs.

		Acc	Prec	Rec	F1
Baselines	**Majority Class**	45.7	14.2	33.3	20.9
	Only Seeds	8.9	87.0	9.8	17.6
Coordination	**kNN**	45.2	61.5	47.3	53.4
	graph	42.7	68.7	39.7	50.4
WordNet	**kNN**	52.8	51.5	50.7	51.1
	graph	51.1	51.9	51.5	51.5
Embedding	**kNN**	59.3	58.4	61.0	59.7
	graph	64.0	70.5	59.4	64.5
Dependency	**kNN**	65.7	63.8	65.4	64.5
	graph	**70.3**	**72.0**	**68.0**	**70.6**

Table 5: Eval. of similarity metrics and classifiers.

5 Experiments

5.1 Evaluation of the Induced Lexicon

Table 5 compares the performance of the different similarity metrics when incorporated in either kNN or graph clustering. The resulting categorizations are compared against the gold standard coarse-grained lexicon (§3.4). For kNN, we set $k = 3$ for which we obtained best performance in all our experiments.

As seeds, we took the top 40 AG-verbs, 30 PT-verbs and 50 SP-verbs produced by the respective initialization methods (§4.1). The seed proportions should vaguely correspond to the actual class distribution (Table 3). Large increases of the seed sets do not improve the quality (as shown below). 10 of the 50 SP-verbs are extracted from the positive SP-patterns, while the remaining verbs are extracted from the negative SP-patterns (§4.1).

As baselines, we include a classifier only employing the seeds and a majority class classifier always predicting an SP-verb. For word embeddings (§4.2.1) and WordNet::Similarity (§4.2.2), we only report the performance of the best metric/configuration, i.e. for embeddings, the continuous bag-of-words model with 500 dimensions and for WordNet::Similarity, the *Wu & Palmer* measure (Wu and Palmer, 1994).

Table 5 shows that the baselines can be outperformed by large margins. The performance of the different similarity metrics varies. The dependency-based metric performs notably better than the other metrics. Together with word embeddings, it is the only metric for which graph clustering produces a notable improvement over kNN.

Table 6 illustrates the quality of the similarity metrics for the present task. The table shows that the dependency-based similarity metric provides the most suitable output. The poor quality of coordination may come as a surprise. That

Coordin.	appear, believe, refuse, vow, want, offend, shock, help, exhilarate, challenge, support, distort
WordNet	appal, scandalize, anger, rage, sicken, temper, hate, fear, love, alarm, dread, tingle
Embedd.	anger, dismay, disgust, protest, alarm, enrage, shock, regret, concern, horrify, appal, sorrow
Depend.	anger, infuriate, alarm, shock, stun, enrage, incense, dismay, upset, appal, offend, disappoint

Table 6: The 12 most similar verbs to *outrage* (PT-verb) according to the different metrics (verbs other than PT-verbs are underlined).

AG-verbs		**PT-verbs**		**SP-verbs**	
no graph (Wiegand 2012)	graph	no graph (Wiegand 2012)	graph	no graph	graph
55.45	**69.12**	38.59	**67.66**	52.16	**72.03**

Table 7: F-scores of entire output of pattern-based extraction (§4.1) where no propagation is applied (*no graph*) vs. best proposed induction method from Table 5 (*graph*).

method suffers from data-sparsity. In our corpus, the frequency of verbs co-occurring with *outrage* in a conjunction is 5 or lower.[5] The table also shows that WordNet may not be appropriate for our present verb categorization task. However, it may be suitable for other subtasks in sentiment analysis, particularly polarity classification. If we consider the similar entries of *outrage* provided by that metric, we find that polarity is largely preserved (10 out of 12 verbs are negative). This observation is consistent with Esuli and Sebastiani (2006) and Rao and Ravichandran (2009).

In Table 5 we only used the top 40/30/50 verbs from the initialization methods as seeds. We can also compare the output of these methods (combined with propagation, i.e. graph clustering) with the *entire* verb lists produced by these pattern-based initialization methods where no propagation is applied. As far as AG- and PT-verbs are concerned, the entire lists of these initialization methods correspond to the original approach of Wiegand and Klakow (2012). Table 7 shows the result. The new graph-induction always outperforms the original induction method by a large margin.

In Table 8, we compare our automatically generated seeds using the patterns from §4.1 (*Pattern*) with seeds extracted from our gold standard (*Gold*). We rank those verbs by frequency. Size and verb type distribution are preserved. We also examine what impact doubling the size of seeds (*Gold*$|$*Pattern*$_{double}$) and halving them (*Gold*$|$*Pattern*$_{half}$) has on classification. Dependency-based similarity and graph clustering is used for all configurations. Only if we double the amount of seeds are the gold seeds notably better than the automatically generated seeds.

[5] We found that for frequently occurring opinion verbs, this similarity metric produces more reasonable output.

Pattern$_{half}$	**Pattern**	**Pattern**$_{double}$	**Gold**$_{half}$	**Gold**	**Gold**$_{double}$
68.71	70.59	66.50	66.21	70.31	73.77

Table 8: Comparison of automatic and gold seeds *(evaluation measure: macro-average F-score)*.

		Coordin.		**WordNet**		**Embedd.**		**Depend.**	
	Major.	**kNN**	**graph**	**kNN**	**graph**	**kNN**	**graph**	**kNN**	**graph**
English	20.9	53.4	50.4	51.1	51.5	59.7	64.5	64.5	**70.6**
German	22.9	43.8	48.9	53.2	59.9	54.3	60.9	58.3	**63.1**

Table 9: Comparison of English and German data *(evaluation measure: macro-average F-score)*.

Since our induction approach just requires a sentiment lexicon and aims at low-resource languages, we replicated the experiments for German, as shown in Table 9. We use the PolArt-sentiment lexicon (Klenner et al., 2009) (1416 entries). (As a gold standard, we manually annotated that lexicon according to our three verb types.) As an unlabeled corpus, we chose the *Huge German Corpus*[6]. As a parser, we used ParZu (Sennrich et al., 2009). Instead of WordNet, we used *GermaNet* (Hamp and Feldweg, 1997). The automatically generated seeds were manually translated from English to German. Table 9 shows that as on English data, dependency-based similarity combined with graph clustering performs best. The fact that we can successfully replicate our approach in another language supports the general applicability of our proposed categorization of verbs into three types for opinion role extraction.

5.2 In-Context Evaluation

We now evaluate our induced knowledge in the task of extracting opinion holders and targets from actual text. For this in-context evaluation, we sampled sentences from the North American News Corpus in which our opinion verbs occurred. We annotated all holders and targets of those verbs. (A constituent may have several roles for the same verb (§2).) The dataset contains about 1100 sentences. We need to rely on this dataset since it is the only corpus in which our opinion verbs are

[6] www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/hgc.html

Features	Description
cand_lemma	head lemma of candidate (phrase)
cand_pos	part-of-speech tag of head of candidate phrase
cand_phrase	phrase label of candidate
cand_person	is candidate a person
verb_lemma	verb lemmatized
verb_pos	part-of-speech tag of verb
word	bag of words: all words within the sentence
pos	part-of-speech sequence between cand. and verb
distance	token distance between candidate and verb
const	path from constituency parse tree from cand. to verb
subcat	subcategorization frame of verb
srl$_{propbank}$/dep	semantic role/dependency path between cand. and verb (semantic roles based on PropBank)
brown	Brown-clusters of cand_word/verb_word/word
srl$_{framenet}$	frame element name assigned to candidate and the frame name (to which frame element belongs)
fine-grain_lex	is candidate holder/target/target$_{speaker}$ according to the fine-grained lexicon
coarse-grain_lex	is candidate holder/target/target$_{speaker}$ according to the coarse-grained lexicon
induc$_{graph}$	is candidate holder/target/target$_{speaker}$ according to the coarse-grained lexicon automatically induced with graph clustering (and induced seeds (§4.1))

Table 10: Feature set for in-context classification.

widely represented and both holders and targets are annotated.

We solve this task with supervised learning. As a classifier, we employ Support Vector Machines as implemented in SVM^{light} (Joachims, 1999). The task is to extract three different entities: opinion holders, opinion targets and opinion targets evoked by speaker views. All those entities are always put into the relation to a specific opinion verb in the sentence. The instance space thus consists of tuples $(verb, const)$, where $verb$ is the mention of an opinion verb and $const$ is any possible (syntactic) constituent in the respective sentence. The dataset contains 753 holders, 745 targets and 499 targets of a speaker view. Since a constituent may have several roles at the same time, we train three binary classifiers for either of the entity types. On a sample of 200 sentences, we measured an interannotation agreement of Cohen's $\kappa = 0.69$ for holders, $\kappa = 0.63$ for targets and also $\kappa = 0.63$ for targets of a speaker view.

Table 10 shows the features used in our supervised classifier. They have been previously found effective (Choi et al., 2005; Jakob and Gurevych, 2010; Wiegand and Klakow, 2012; Yang and Cardie, 2013). The ***standard features*** are the features from *cand_word* to *brown*. For semantic role labeling of PropBank-structures, we used *mate-tools* (Björkelund et al., 2009). For person detection, we employ named-entity tagging (Finkel et al., 2005) and WordNet (Miller et al., 1990).

Features	Holder	Target	Target$_{Speaker}$
standard	63.59	54.18	40.06
+srl$_{framenet}$	65.44*	55.70*	42.14
+induc$_{graph}$	68.06$^{*\circ}$	59.61$^{*\circ}$	46.66$^{*\circ}$
+srl$_{framenet}$+induc$_{graph}$	69.70$^{*\circ}$	60.47$^{*\circ}$	47.33$^{*\circ}$
+coarse-grain_lex	68.56$^{*\circ}$	59.89$^{*\circ}$	54.31$^{*\circ\dagger}$
+srl$_{framenet}$+coarse-grain_lex	69.70$^{*\circ}$	60.68$^{*\circ}$	54.06$^{*\circ\dagger}$
+fine-grain_lex	69.83$^{*\circ\dagger}$	62.89$^{*\circ\dagger}$	**56.71**$^{*\circ\dagger}$
+srl$_{framenet}$+fine-grain_lex	**70.80**$^{*\circ\dagger}$	**63.72**$^{*\circ\dagger}$	56.64$^{*\circ\dagger}$

statistical significance testing (paired t-test, significance level $p < 0.05$) *: better than *standard*; $^{\circ}$: better than +*srl$_{framenet}$*; †: better than +*induc$_{graph}$*

Table 11: In-context evaluation (*eval.: F-score*).

For semantic role labeling of FrameNet-structures (*srl$_{framenet}$*), we used *Semafor* (Das et al., 2010) with the argument identification based on dual decomposition (Das et al., 2012). We run the configuration that also assigns frame structures to unknown predicates (Das and Smith, 2011). This is necessary as 45% of our opinion verbs are not contained in FrameNet (v1.5). FrameNet has been shown to enable a correct role assignment for AG- and PT-verbs (Bethard et al., 2004; Kim and Hovy, 2006). For instance, in (17) and (18), the opinion holder is assigned to the same frame element EXPERIENCER. However, the PropBank representation does not produce a correct alignment: In (17), the opinon holder is the agent of the opinion verb, while in (18), the opinion holder is the patient of the opinion verb.

(17) Peter$^{\text{EXPERIENCER}}_{agent}$ **dislikes** Mary$_{patient}$.
$(dislike, Peter, holder)$

(18) Peter$_{agent}$ **disappoints** Mary$^{\text{EXPERIENCER}}_{patient}$.
$(disappoint, Mary, holder)$

With the feature ***fine-grain_lex***, we want to validate that our manually-compiled opinion role lexicon for verbs (§2), i.e. the lexicon that also allows multiple opinion roles for the same semantic roles, is effective for in-context evaluation. ***Coarse-grain_lex*** is derived from the fine-grained lexicon (§3.4). With this feature, we measure how much we lose by dropping the fine-grained representation. ***Induc***$_{graph}$ induces the verb types of the coarse representation automatically by employing the best induction method obtained in Table 5.

Table 11 compares the different features on 10-fold crossvalidation. The table shows that the features encoding opinion role information, including our induction approach, are more effective than *srl$_{framenet}$*. Even though the fine-grained lexicon produces the best results, we almost reach that performance with the coarse-grained lexicon. This is

		manual lexicons	
Features	**induc**$_{graph}$	**coarse-grain**	**fine-grain**
lexicon feature only ($\approx$ *unsuperv.*)	52.38	55.81	60.68
lexicon with all other features	59.90*	61.71*	63.92°

statistical significance testing (paired t-test, significance level $p < 0.05$) *: better than *lexicon feature only*; $^{\circ}$: only significant on 2 out of 3 roles

Table 12: Lexical resources and the impact of other (not lexicon-based) features (*evaluation measure: macro-average F-score*).

	Distribution of Verb Types			
Corpus	**AG**	**PT**	**SP**	**# sentences**
MPQA *(training+test)*	77.5	7.7	14.8	15,753
FICTION *(test)*	67.5	15.1	17.3	614
VERB *(test)*	34.8	14.0	52.7	1,073

Table 13: Statistics on the different corpora used.

further evidence that our proposed three-way verb categorization, which is also the basis of our induction approach, is adequate.

Table 12 compares the performance of the different lexicons in isolation (this is comparable with an *unsupervised* classifier, as each lexicon feature has three values each predicting either of the opinion roles) and in combination with the standard (+srl$_{framenet}$) features. The table shows that all lexicon features are strong features on their own. The score of induction is lowest but this feature has been created without manual supervision. Moreover, the improvement by adding the other features is much larger for induction than for the manually-built fine-grained lexicon. This means that we can compensate some lexical knowledge missing in induction by standard features.

Since we could substantially outperform the features relying on FrameNet with our new lexical resources, we looked closer at the predicted frame structures. Beside obvious errors in automatic frame assignment, we also found that there are problems inherent in the frame design. Particularly, the notion of SP-verbs (§3.3) is not properly reflected. Many frames, such as SCRUTINY, typically devised for AG-verbs, such as *investigate* or *analyse*, also contain SP-verbs like *pry*. This observation is in line with Ruppenhofer and Rehbein (2012) who claim that extensions to FrameNet are necessary to properly represent opinions evoked by verbal predicates.

5.3 Comparison to Previous Cross-Domain Opinion Holder Extraction

We now compare our proposed induction approach with previous work on opinion holder extraction. We replicate several classifiers and compare them to our new approach. (Because of the limited space of this paper, we cannot also address cross-domain opinion target extraction.) We consider three different corpora as shown in Table 13. **MPQA** (Wiebe et al., 2005) is the standard corpus for fine-grained sentiment analysis. **FICTION**, introduced in Wiegand and Klakow (2012), is a collection of summaries of classic literary works. **VERB** is the new corpus used in the previous evaluation (§5.2). VERB and MPQA both originate from the news domain but VERB is sampled in such a way that mentions of *all* opinion verbs of the Subjectivity Lexicon are represented. The other corpora consist of contiguous sentences. They will have a bias towards only those opinion verbs frequently occurring in that particular domain. This also results in different distributions of verb types as shown in Table 13. For example, SP-verbs are rare in MPQA. However, there exist plenty of them (Table 3). Other domains may have much more frequent SP-verbs (just as FICTION has more PT-verbs than MPQA). A robust domain-independent classifier should therefore be able to cope equally well with all three verb types.

MPQA is also the largest corpus. Following Wiegand and Klakow (2012), this corpus is chosen as a training set.[7] Despite its size, however, almost every second opinion verb from our set of opinion verbs is not contained in that corpus.

In the evaluation, we only consider the opinion holders of our opinion verbs. (Other opinion holders, both in the gold standard and the predictions of the classifiers are ignored.) Recall that we take the knowledge of what is an opinion verb as given. Our graph-based induction can be arbitrarily extended by increasing the set of opinion verbs.

	in domain	*out of domain*	
Config	**MPQA**	**FICTION**	**VERB**
MultiRel	**72.54**$^{*\circ}$	53.02	44.80
CK	62.98	52.91	43.88
CK + induc$_{Wiegand\ 2012}$	65.15	57.33*	50.83*
CK + induc$_{graph}$	66.06*	**65.03**$^{*\circ}$	60.91$^{*\circ}$
CK + coarse-grain_lex	66.82*	64.13$^{*\circ}$	63.72$^{*\circ\dagger}$
CK + fine-grain_lex	66.16*	64.98$^{*\circ}$	**70.85**$^{*\circ\dagger\ddagger}$

statistical significance testing (permutation test, significance level $p < 0.05$) *: better than *CK*; $^{\circ}$: better than *CK + induc*$_{Wiegand\ 2012}$; †: better than *CK + induc*$_{graph}$; ‡: better than *CK + coarse-grain_lex*

Table 14: Evaluation on opinion holder extraction on various corpora (*evaluation measure: F-score*).

[7] The split-up of training and test set on the MPQA corpus follows the specification of Johansson and Moschitti (2013).

For classifiers, we consider convolution kernels ***CK*** from Wiegand and Klakow (2012) and the sequence labeler from Johansson and Moschitti (2013) ***MultiRel*** that incorporates relational features taking into account interactions between multiple opinion cues. It is currently the most sophisticated opinion holder extractor. *CK* can be combined with additional knowledge. We compare $induc_{graph}$ with $\textbf{\textit{induc}}_{Wiegand\ 2012}$, which employs the word lists induced for AG- and PT-verbs in the fashion of Wiegand and Klakow (2012), i.e. without graph clustering. As an upper bound for the induction methods, *coarse_grain_lex* and *fine_grain_lex* are used.[8] The combination of *CK* with this additional knowledge follows the best settings from Wiegand and Klakow (2012).[9]

Table 14 shows the results. *MultiRel* produces the best performance on MPQA but suffers similarly from a domain-mismatch as *CK* on FICTION and VERB. *MultiRel* and *CK* cannot handle many PT- and SP-verbs in those corpora, simply because many of them do not occur in MPQA. On MPQA, only the new induction approach and the lexicons significantly improve *CK*. The knowledge of opinion roles has a lower impact on MPQA. In that corpus, most opinion verbs take their opinion holder as an agent. Given the large size of MPQA, this information can be easily learned from the training data. The situation is different for FICTION and VERB where the knowledge from induction largely improves classification. In these corpora, opinion holders as agents are much less frequent than on MPQA. The new induction proposed in this paper also notably outperforms the induction from Wiegand and Klakow (2012).

Although the fine-grained lexicon is among the top performing systems, we only note large improvements on VERB. VERB has the highest proportion of PT- and SP-verbs (Table 13). Knowledge about role-assignment is most critical here.

6 Related Work

Most approaches for opinion role extraction employ supervised learning. The feature design is mainly inspired by semantic role labeling (Bethard et al., 2004; Li et al., 2012). Some work also employs information from existing semantic role labelers based on FrameNet (Kim and Hovy, 2006) or PropBank (Johansson and Moschitti, 2013; Wiegand and Klakow, 2012). Although those resources give extra information for opinion role extraction in comparison to syntactic or other surface features, we showed in this work that further task-specific knowledge, i.e. either opinion verb types or a manually-built opinion role lexicon, provide even more accurate information.

There has been a substantial amount of research on opinion target extraction. It focuses, however, on the extraction of topic-specific opinion terms (Jijkoun et al., 2010; Qiu et al., 2011) rather than the variability of semantic roles for opinion holders and targets. Mitchell et al. (2013) present a low-resource approach for target extraction but their aim is to process Twitter messages without using general *syntax* tools. In this work, we use such tools. Our notion of *low resources* is different in that we mean the absence of *semantic* resources helpful for our task (e.g. FrameNet).

7 Conclusion

We presented an approach for opinion role induction for verbal predicates. We assume that those predicates can be divided into three different verb types where each type is associated with a characteristic mapping between semantic roles and opinion holders and targets. In several experiments, we demonstrated the relevance of those three types. We showed that verbs can effectively be categorized with graph clustering given a suitable similarity metric. The seeds are automatically selected. Our proposed induction approach outperforms both a previous induction approach and features derived from semantic role labelers. We also pointed out the importance of the knowledge gained by induction in supervised cross-domain classification.

Acknowledgements

The authors would like to thank Stephanie Köser for annotating parts of the resources presented in this paper. For proofreading the paper, the authors would also like to thank Ines Rehbein, Asad Sayeed and Marc Schulder. We also thank Ashutosh Modi for advising us on word embeddings. The authors were partially supported by the German Research Foundation (DFG) under grants RU 1873/2-1 and WI 4204/2-1.

[8] Wiegand and Klakow (2012) use a lexicon *Lex* which just comprises the notion of AG and PT verbs, so our manual lexicons are more accurate and harder to beat.

[9] For in-domain evaluation (i.e. MPQA) the trees (tree structures are the input to *CK*) are augmented with verb category information. For out-of-domain evaluation (i.e. FICTION and VERB), we add to the predictions of *CK* the prediction of a rule-based classifier using the opinion role assignment according to the respective lexicon or induction method.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of COLING/ACL*.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of ACL*.

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.

Anders Björkelund, Love Hafdell, and Pierre Nugues. 2009. Multilingual Semantic Role Labeling. In *Proceedings of the CoNLL – Shared Task*.

Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns. In *Proceedings of HLT/EMNLP*.

Thomas Cover and Peter Hart. 1967. Nearest neighbor pattern classification. *Information Theory*, 13(1):21–27.

Dipanjan Das and Noah A. Smith. 2011. Semi-Supervised Frame-Semantic Parsing for Unknown Predicates. In *Proceedings of ACL*.

Dipanjan Das, Nathan Schneider, Desai Chen, and Noah A. Smith. 2010. Probabilistic Frame-Semantic Parsing. In *Proceedings of HLT/NAACL*.

Dipanjan Das, André F.T. Martins, and Noah A. Smith. 2012. An Exact Dual Decomposition Algorithm for Shallow Semantic Parsing with Constraints. In *Proceedings of *SEM*.

Lingjia Deng and Janyce Wiebe. 2014. Sentiment Propagation via Implicature Constraints. In *Proceedings of EACL*.

Andrea Esuli and Fabrizio Sebastiani. 2006. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of LREC*.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of ACL*.

Amit Goyal, Ellen Riloff, and Hal Daume III. 2010. Automatically Producing Plot Unit Representations for Narrative Text. In *Proceedings of EMNLP*.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a Lexical-Semantic Net for German. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the Semantic Orientation of Adjectives. In *Proceedings of EACL*.

Niklas Jakob and Iryna Gurevych. 2010. Extracting Opinion Targets in a Single- and Cross-Domain Setting with Conditional Random Fields. In *Proceedings of EMNLP*.

Valentin Jijkoun, Maarten de Rijke, and Wouter Weerkamp. 2010. Generating Focused Topic-Specific Sentiment Lexicons. In *Proceedings of ACL*.

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*, pages 169–184. MIT Press.

Richard Johansson and Alessandro Moschitti. 2013. Relational Features in Fine-Grained Opinion Analysis. *Computational Linguistics*, 39(3):473–509.

Jun Seok Kang, Song Feng, Leman Akoglu, and Yejin Choi. 2014. ConnotationWordNet: Learning Connotation over the Word+Sense Network. In *Proceedings of ACL*.

Soo-Min Kim and Eduard Hovy. 2006. Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text. In *Proceedings of ACL Workshop on Sentiment and Subjectivity in Text*.

Paul Kingsbury and Martha Palmer. 2002. From TreeBank to PropBank. In *Proceedings of LREC*.

Dan Klein and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. In *Proceedings of ACL*.

Manfred Klenner, Angela Fahrni, and Stefanos Petrakis. 2009. PolArt: A Robust Tool for Sentiment Analysis. In *Proceedings of NoDaLiDa*.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

Shoushan Li, Rongyang Wang, and Guodong Zhou. 2012. Opinion Target Extraction via Shallow Semantic Parsing. In *Proceedings of AAAI*.

Dekang Lin. 1998. Automatic Retrieval and Clustering of Similar Words. In *Proceedings of ACL/COLING*.

Isa Maks and Piek Vossen. 2012. A lexicon model for deep sentiment analysis and opinion mining applications. *Decision Support Systems*, 53:680–688.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of Workshop at ICLR*.

George Miller, Richard Beckwith, Christine Fellbaum, Derek Gross, and Katherine Miller. 1990. Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3:235–244.

Margaret Mitchell, Jacqueline Aguilar, Theresa Wilson, and Benjamin Van Durme. 2013. Open Domain Targeted Sentiment. In *Proceedings of EMNLP*.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. WordNet::Similarity – Measuring the Relatedness of Concepts. In *Proceedings of HLT/NAACL*.

Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2011. Opinion Word Expansion and Target Extraction through Double Propagation. *Computational Linguistics*, 37(1):9–27.

Delip Rao and Deepak Ravichandran. 2009. Semi-Supervised Polarity Lexicon Induction. In *Proceedings of EACL*.

Ellen Riloff and Jessica Shepherd. 1997. A Corpus-Based Approach for Building Semantic Lexicons. In *Proceedings of EMNLP*.

Brian Roark and Eugene Charniak. 1998. Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction. In *Proceedings of COLING*.

Josef Ruppenhofer and Ines Rehbein. 2012. Semantic frames as an anchor representation for sentiment analysis. In *Proceedings of WASSA*.

Josef Ruppenhofer, Swapna Somasundaran, and Janyce Wiebe. 2008. Finding the Source and Targets of Subjective Expressions. In *Proceedings of LREC*.

Rico Sennrich, Gerold Schneider, Martin Volk, and Martin Warin. 2009. A New Hybrid Dependency Parser for German. In *Proceedings of GSCL*.

Partha Pratim Talukdar, Joseph Reisinger, Marius Pasca, Deepak Ravichandran, Rahul Bhagat, and Fernando Pereira. 2008. Weakly-Supervised Acquisition of Labeled Class Instances using Graph Random Walks. In *Proceedings of EMNLP*.

Janyce Wiebe and Rada Mihalcea. 2006. Word Sense and Subjectivity. In *Proceedings of COLING/ACL*.

Janyce Wiebe, Therea 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. 2012. Generalization Methods for In-Domain and Cross-Domain Opinion Holder Extraction. In *Proceedings of EACL*.

Thesesa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. In *Proceedings of HLT/EMNLP*.

Zhibiao Wu and Martha Palmer. 1994. Verb semantics and lexcial selection. In *Proceedings of ACL*.

Bishan Yang and Claire Cardie. 2013. Joint Inference for Fine-grained Opinion Extraction. In *Proceedings of ACL*.

Patrick Ziering, Lonneke van der Plas, and Hinrich Schuetze. 2013. Bootstrapping Semantic Lexicons for Technical Domains. In *Proceedings of IJCNLP*.