Opinion Holder and Target Extraction on Opinion Compounds – A Linguistic Approach

Michael Wiegand and Christine Bocione
Spoken Language Systems
Saarland University
D-66123 Saarbrücken, Germany
michael.wiegand@lsv.uni-saarland.de
cbocione@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 to the new task of opinion holder and target extraction on opinion compounds. Opinion compounds (e.g. user rating or victim support) are noun compounds whose head is an opinion noun. We do not only examine features known to be effective for noun compound analysis, such as paraphrases and semantic classes of heads and modifiers, but also propose novel features tailored to this new task. Among them, we examine paraphrases that jointly consider holders and targets, a verb detour in which noun heads are replaced by related verbs, a global head constraint allowing inferencing between different compounds, and the categorization of the sentiment view that the head conveys.

1 Introduction

One of the key subtasks in sentiment analysis is opinion role extraction. It can be divided into the extraction of opinion holders (OH), i.e. entities expressing an opinion, and the extraction of opinion targets (OT), i.e. entities or propositions at which sentiment is directed. This task is vital for various applications involving sentiment analysis, e.g. opinion summarization or opinion question answering.

Opinion role extraction is commonly regarded as a task in lexical semantics. An opinion is evoked by some opinion word, e.g. criticized in (1), skeptical in (2) or intentions in (3), and its opinion roles are usually realized as syntactic dependents. Opinion words come in many shapes, the most frequent types being opinion verbs (1), opinion adjectives (2) and opinion nouns (3). These types of opinion words have extensively been studied in various sentiment-related corpora, such as MPQA (Wiebe et al., 2005).

(1) [Peter OH] criticized [Mary OT].
(2) [Mary OH] was skeptical [about the plan OT].
(3) [Peter OH] had firm intentions [to quit his job OT].

In this work, we examine opinion roles that are realized in opinion compounds. We define an opinion compound (Table 1) as a noun compound, i.e. a sequence of two nouns, where the second noun, i.e. the head, is an opinion expression. The first noun, i.e. the modifier, can represent an opinion holder (4)-(5), an opinion target (6)-(7) or neither (8)-(9). Our aim is to automatically classify the modifier into these categories. This task is challenging as, unlike with opinion roles expressed in the syntax (1)-(3), the immediate context of compounds does not contain explicit cues as to the relation between head and modifier. Moreover, due to the high productivity of compounding, this task cannot be solved by compiling a (finite) compound lexicon that encodes for each compound the category of its modifier.

(4) [user OH] rating (i.e. user rates something)
(5) [consumer OH] uncertainty (i.e. consumers are uncertain)
(6) [victim OT] support (i.e. support for victims)
(7) [test OT] anxiety (i.e. having anxiety towards test taking)
(8) spring upswing (i.e. economic upswing in spring)
(9) phone harassment (i.e. harassment inflicted via phone)

Notice that we focus exclusively on opinion role extraction. We do not try to detect the polarity associated with the compound. Neither do we consider implicature-related information about effects (Deng and Wiebe, 2014), but only inherent sentiment.

We study opinion role extraction on opinion compounds in German. German is known for its frequent...
use of noun compounds. In the STEPS-corpus, the benchmark dataset for German opinion role extraction (Ruppenhofer et al., 2014), almost every other sentence contains an opinion compound.

Compounds can also be commonly found in other key languages, such as English. Since the methods we apply to this task and the issues that they address are not language specific, our approach can be replicated on other languages.

Apart from examining traditional features from noun compound analysis, in this paper, we also introduce novel features specially designed for the analysis of opinion compounds.

We also created a new gold standard for this task (see also §3). The STEPS-corpus, as such, is fairly small and only contains about 200 unique compounds. We considered this amount insufficient for producing a gold standard. Also, none of the existing datasets on noun compounds (Lauer, 1995; Barker and Szpakowicz, 1998; Nastase and Szpakowicz, 2003; Girju et al., 2009; Kim and Baldwin, 2005; Tratz and Hovy, 2010; Dima et al., 2014) contain any information regarding opinion roles.

### Table 1: Internal structure of opinion compounds.

<table>
<thead>
<tr>
<th>compounds</th>
<th>user rating; victim support; spring upswing</th>
</tr>
</thead>
<tbody>
<tr>
<td>immediate constituents</td>
<td>user; victim; spring; rating; support; upswing</td>
</tr>
<tr>
<td>grammatical function</td>
<td>modifier; head</td>
</tr>
</tbody>
</table>

3 **Data & Annotation**

We created a new dataset by retrieving opinion compounds from the deWaC-corpus (Baroni et al., 2009) comprising 1.7 billion words. (Word embeddings ($S.2$ & $S.6$) and word similarity graphs ($S.7$ & $S.4$) were also created from this corpus.)

In German, noun compounds are typically realized as single tokens. In order to obtain a set of opinion compounds, we extracted all noun compounds from deWaC whose second morpheme is an opinion noun. Morphological analysis was carried out using morphisto (Zielinski and Simon, 2009). As opinion nouns, we used the nouns from the PolArt sentiment lexicon (Klenner et al., 2009). Unfortunately, this lexicon is lacking in neutral opinion nouns, such as Meinung (opinion) or Erwartung (expectation) which frequently occur in compounds, e.g. Expertenmeinung (expert opinion) or Kundenwartungen (customer expectations). Therefore, we translated the 235 neutral opinion nouns from the (English) Subjectivity Lexicon (Wilson et al., 2005) into German.

From the opinion compounds extracted from deWaC, we created two manually annotated datasets (Table 2). We use more than one dataset as we consider our task as a multi-stage task as shown in Figure 1. We believe that this is necessary as different underlying relation (Rosario and Hearst, 2001; Kim and Baldwin, 2005; Girju et al., 2005; Girju et al., 2009). On the other hand, paraphrases that contain co-occurrences of head and modifier are exploited (Girju et al., 2009; Nakov and Hearst, 2013). In order to increase coverage, paraphrases can be automatically acquired (Butnariu and Veale, 2008; Kim and Nakov, 2011). Cross-lingual information has also been harnessed for this task (Girju, 2007).

### Table 2: The two different datasets.

<table>
<thead>
<tr>
<th></th>
<th>Dataset I</th>
<th>Dataset II</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of compounds</td>
<td>2000</td>
<td>1000</td>
</tr>
<tr>
<td>unique heads</td>
<td>389</td>
<td>247</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>category of modifier</th>
<th>role</th>
<th>no role</th>
<th>holder</th>
<th>target</th>
</tr>
</thead>
<tbody>
<tr>
<td>frequency proportion (in %)</td>
<td>937</td>
<td>1063</td>
<td>450</td>
<td>580</td>
</tr>
<tr>
<td></td>
<td>46.85</td>
<td>53.15</td>
<td>45.00</td>
<td>58.00</td>
</tr>
</tbody>
</table>

available at: www.coli.uni-saarland.de/~miwieg/naacl_2016_op_compounds_data.tgz

The data release provides more details regarding the gold standard, e.g. how compound instances were sampled.

801
ent types of knowledge are required for the different steps. In the first step (Dataset I), the compounds containing some opinion role (4)-(7) are separated from those not containing any role at all (8)-(9). At this stage, holders are not distinguished from targets. This is done in the second step which exclusively focuses on opinion roles. This step is further divided into two sub-steps. First, one checks whether the modifier denotes a person. A modifier representing an opinion role but not denoting a person (e.g. test anxiety) can only be a target. Since this is a simple classification step (provided a lexical resource is available which tells persons apart from non-persons, e.g. WordNet), we have no dataset for it. The greater challenge lies in all those compounds whose modifier is a person and for which we already know that it is either holder or target (e.g. user rating or victim support). Only for those cases do we produce another dataset (Dataset II). Note that in this dataset the two roles are not completely disjoint. In 3% of the compounds, the modifier represents both holder and target. Prominent examples are reciprocal relationships, e.g. Geschwisterneid (sibling jealousy).

On a sample of 200 compounds extracted from each of the two datasets we measured inter-annotation agreement. On the first dataset, we obtained Cohen’s $\kappa = 0.60$, while on the second, we obtained $\kappa = 0.60$ for holders and $\kappa = 0.62$ for targets, respectively. These scores can be interpreted as substantial agreement (Landis and Koch, 1977).

4 Classifiers and the Three Different Tasks

We solve the given task as a supervised classification problem. As a classifier, we employ Markov Logic Networks (MLNs). We use this classifier because it allows us to integrate all of our features, including global constraints (see discussion in §5.5).

We consider 3 different tasks (bold rhombuses in Figure 1): the detection of opinion roles (Dataset I), the detection of opinion holders (Dataset II) and the detection of opinion targets (Dataset II). Each task is modeled as a binary classifier. Even though the latter two tasks use the same dataset, we cannot train just one single binary classifier as there are compounds whose modifiers represent both holder and target, e.g. Geschwisterneid (sibling jealousy).3

5 Feature Design

Our core global features, which are used for all three tasks (§4), include the two predominant approaches for compound analysis, i.e. (plain) paraphrases (§5.1) and semantic knowledge (§5.4). We extend the paraphrase approach with two major innovations. First, we examine a verb detour (§5.2) by which we gain important information regarding the syntactic relationship between the modifier and the head of the compound. Secondly, we show that joint paraphrases (§5.3) considering both holder and

---

3 For the holder-detection task, the modifier of such compounds are considered a holder, while for the target-detection task, they are considered a target. For the holder-detection task, we have the two classes holder and no holder, while for the target-detection task, the classes are target and no target.
target are better than paraphrases focusing on only one role. We argue that for our task, (syntactic) ambiguity rather than lack of coverage is the pressing problem. Therefore, we do not focus on paraphrase acquisition but introduce new disambiguation features. Beside the extensions to paraphrases mentioned above, we introduce a global head constraint (§5.5) as an additional global feature. As a local feature for the initial role classification, we perform subjectivity detection on the compound (§5.6). And finally, we use the sentiment view that the head of the compound evokes (§5.7) as a local feature in the holder and target classification tasks.

Table 3: Division of global and local features.

<table>
<thead>
<tr>
<th>Global Features</th>
<th>Local Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>features used on all three tasks</td>
<td>feature used only on task Role (i.e. Dataset I)</td>
</tr>
<tr>
<td>(i.e. Datasets I and II)</td>
<td>SUBJ (§5.6)</td>
</tr>
<tr>
<td>PARA (§5.1-5.3), SEM (§5.4), HEAD (§5.5)</td>
<td>feature used only on task Holder and task Target (i.e. Dataset II)</td>
</tr>
<tr>
<td></td>
<td>VIEW (§5.7)</td>
</tr>
</tbody>
</table>

Table 3: Division of global and local features.

5.1 Plain Paraphrases (PARA$_{plain}$)

An established method for computing the relation expressed by a compound is to consider paraphrases, that is, co-occurrences of the head and modifier as individual constituents accompanied by some predictive context. For example, the compound Expertenauffassung (expert view) can be paraphrased by Auffassung unter Experten (view among experts). The preposition unter (among) is an explicit lexical clue for the (implicit) relation holding between head and modifier in the compound. As paraphrases we manually collected 18 frequent dependency relations that typically hold between an opinion noun and its opinion holder (10) or its opinion target (11). (The data release provides more information including a full list of all paraphrases.) For each compound, we check in deWaC whether head and modifier can be observed in any of those relations.

(10) subj$_{unter}(among)$(<opinion noun>, <holder>): Auffassung unter Experten (view among experts)

We consider each of those selected dependency relations as an individual feature, i.e. we do not explicitly group the chosen relations to holder and target. Assuming that the predictiveness of the different relations varies, this encoding allows a supervised classifier to appropriately weight each relation.

5.2 Verb Detour Paraphrases (PARA$_{verb}$)

Some of the paraphrases from §5.1 are ambiguous. This particularly concerns $subj$_{nom}(of) which occurs with approx. 40% of the compounds of our dataset. On the first reading illustrated by (12)a), we observe a modifier being a holder, while, on the second reading shown by (13)a), the modifier is a target.

For heads being deverbal nouns (e.g. comment or assessment), this ambiguity can often be resolved by considering morphologically related verbs. In (12)b) and (13)b), the two modifiers no longer share the same dependency relation to the opinion word. Opinion holders tend to occur in subject position (12)b) while targets occur in object position (13)b). Wiegand and Klakow (2012) identify these dependency relations for the two different opinion roles as the most frequent ones. So for deverbal nouns, which make up 57% of the heads of our compounds, we add a feature that checks in deWaC whether the modifier is more often observed as a subject or an object of a verb related to the head. (Wiegand and Klakow (2012) actually consider semantic roles, i.e. agent and patient, instead of dependency relations. Due to the lack of robust semantic role-labeling for German, we use dependency relations as a proxy. That is, we identify agents with the dependency relation subj and patients with the relation obj.)

Even though the disambiguation of deverbal noun compounds with the help of verb relations has been
examined before (Lapata, 2002), it has not been exploited for an actual application, such as opinion role extraction. Neither has it been compared against plain paraphrases, which use the head noun of the compound directly (§5.1).

Our use of verb semantics for compound analysis is also different from its predominant use in previous work (Kim and Baldwin, 2006; Nakov and Hearst, 2013) where noun compounds are considered whose parts represent arguments of an abstract verbal relation (e.g. malaria mosquito are arguments of relation 'mosquito causes malaria'). Thus, the aim has been to predict verbs for those compounds that match those abstract relations (e.g. to cause). We are looking for different verbs, namely those that are the morphological basis for the head noun.

For this verb detour, we produce a mapping from nouns (i.e. the heads of our opinion compounds) to verbs by combining distributional and string similarity. We extracted the verbs most similar to each of these nouns (we use top 100). For that we induce vector representations of all head nouns of our gold standard and all existing German verbs using the embedding toolkit Word2Vec (Mikolov et al., 2013).<sup>5</sup> For each noun, we select the verb with the highest cosine-similarity that has at least a Levenshtein (string) similarity (Levenshtein, 1966) of 3. This high threshold ensures that nouns which are not deverbal nouns are not mapped to any verb. Against a manual mapping, our automatic method produced an F-score of 76.1 (at a precision of 77.1).

5.3 Joint Paraphrases (PARA<sub>joint</sub>)

Another way of reducing the ambiguity of paraphrases is to employ paraphrases that jointly consider opinion holder and target (Table 4). We assume that the presence of one ambiguous dependency relation is less problematic in the presence of another less ambiguous relation. The ambiguity can be resolved by method of elimination. For instance, even though objp<sub>von/of</sub> (Widerstand/resistance, Bauern/farmers) is ambiguous, in the first example of Table 4, it can only represent a holder, since the second relation objp<sub>gegen/against</sub> (Widerstand/resistance, Gesetz/regulation) implies a target.

We also use paraphrases in which the compound itself occurs (second and third pattern type of Table 4). Since, in the first example of the second pattern type, only the relation objp<sub>mit/with</sub> (Zufriedenheit/satisfaction, Unternehmens/company) is indicative of a target, the modifier is likely to be a holder. (The example of the third pattern type follows an analogous pattern to extract a target.) The second example (of the second pattern type) Sprengstoffanschlag (bomb attack) illustrates that paraphrases can also be used to infer the absence of opinion roles. Sprengstoff (explosive) cannot be a target because of the other target relation that is present. It cannot be a holder either as it is not a person.

The fourth pattern type in Table 4 considers patterns involving possessive pronouns. They typically represent holders, so the remaining dependency relation can only represent a target.

Similar to §5.1, we encode the joint-paraphrase patterns by their individual dependency relations. That is, the first example in Table 4 would be represented as the feature objp<sub>von</sub> <sup>modifier</sup> objp<sub>gegen</sub>.

5.4 Semantic Knowledge (SEM)

We use GermaNet (Hamp and Feldweg, 1997), the German version of WordNet, to look up the hypernyms of each modifier and each head. The hypernym relation is the most frequently used semantic relation employed for noun compound analysis (Girju et al., 2005; Nastase et al., 2006; Girju et al., 2009; Tratz and Hovy, 2010). Hypernyms allow some generalization over the lexical units representing the heads and modifiers of our compounds. By manual inspection, we found that there are several hypernyms that correlate with a category we want to predict. For example, heads having the hypernym politische Handlung (political act) typically indicate holders as in Arbeiterunruhe (worker unrest) or Studentenrebellion (student rebellion). Hypernyms may also serve as negative cues. For example, heads having the hypernym Verbrechen (crime) are typically contained in compounds whose modifiers represent neither a holder nor a target, such as Steuervergehen (tax offense) or Autodiebstahl (car theft).

<sup>5</sup>We used the cbow-model with 200 dimensions. All remaining parameters are set to their respective default values.
Table 4: Illustration of patterns for joint paraphrases.

<table>
<thead>
<tr>
<th>Pattern Type</th>
<th>Example Compound</th>
<th>Label</th>
<th>Example Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;head&gt; &lt;holder&gt; &lt;target&gt;</td>
<td>Bauern widerstand (farmer resistance)</td>
<td>holder</td>
<td>Widerstand [von Bauern obj_{repp}] gegen das Gesetz obj_{regu}</td>
</tr>
<tr>
<td></td>
<td>Schülerbeurteilung (student assessment)</td>
<td>target</td>
<td>Beurteilung [von Lehrer obj_{rat}] von Schülern obj_{repp}</td>
</tr>
<tr>
<td>&lt;compound&gt; &lt;target&gt;</td>
<td>Mitarbeiterzufriedenheit (staff satisfaction)</td>
<td>holder</td>
<td>Mitarbeiterzufriedenheit [mit dem Unternehmen obj_{repp}]</td>
</tr>
<tr>
<td></td>
<td>Sprengstoffanschlag (bomb attack)</td>
<td>no role</td>
<td>Sprengstoffanschlag [auf Touristen obj_{repp}]</td>
</tr>
<tr>
<td>&lt;compound&gt; &lt;holder&gt;</td>
<td>Prüfungsangst (test anxiety)</td>
<td>target</td>
<td>Prüfungsangst [unter Schülern obj_{repp}]</td>
</tr>
<tr>
<td></td>
<td>Kinderfreundlichkeit (child friendliness)</td>
<td>target</td>
<td>seine Freundschaft [gegenüber Kindern obj_{repp}]</td>
</tr>
</tbody>
</table>

Table 5: Illustration of selectional preferences of heads of opinion compounds.

<table>
<thead>
<tr>
<th>Head</th>
<th>Preference</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Haltung (attitude)</td>
<td>holder</td>
<td>Arbeitgeberhaltung (employer attitude), Autorenhaltung (author attitude), Konsumentenhaltung (consumer attitude), Verbraucherhaltung (customer attitude), Zuschauerhaltung (viewer attitude)</td>
</tr>
<tr>
<td>Verehrung (worship)</td>
<td>target</td>
<td>Ahnenverehrung (ancestor worship), Heldenverehrung (hero worship), Ikonenverehrung (icon worship), Kaiserverehrung (emperor worship), Märtyrerverehrung (martyr worship)</td>
</tr>
<tr>
<td>Attentat (attack)</td>
<td>no role</td>
<td>Bombenattentat (bombing attack), Fliegerattentat (aircraft attack), Selbstmordattentat (suicide attack), Sprengstoffattentat (explosive attack), Säureattentat (acid attack)</td>
</tr>
</tbody>
</table>

5.5 Head Constraint (HEAD)

We observed that many heads have a strong selectional preference as to what type they select as a modifier. This is illustrated in Table 5. There are heads that prefer opinion holders as modifiers (e.g. Haltung (attitude)), heads that prefer targets (e.g. Verehrung (worship)) or heads that prefer no role (e.g. Attentat (attack)). This is further substantiated by Table 6 showing the high average role-purity of compound groups sharing the same head. Purity is measured by the proportion of the most frequent role occurring within each group of compounds sharing the same head. Given this selectional preference, we formulate a global head constraint (Table 7) that if two compounds have the same head, their modifiers should convey the same opinion role.

In order to implement this constraint in a supervised classifier we employ Markov Logic Networks (MLNs), which combine first-order logic with probabilities. As a tool, we use thebeast (Riedel, 2008). MLNs have been effectively used in various related NLP tasks, such as discourse-based sentiment analysis (Zirn et al., 2011), semantic-role labeling (Meza-Ruiz and Riedel, 2009), anaphora resolution (Hou et al., 2013) or question answering (Khot et al., 2015).

MLNs are a set of pairs $(F_i, w_i)$ where $F_i$ is a first-order logic formula and $w_i$ an associated real-valued weight. They build a template for constructing a Markov network given a set of constants $C$. The probability distribution that is estimated is a log-linear model

$$P(X = x) = \frac{1}{Z} \exp \left( \sum_{i=1}^{k} w_i n_i(x) \right)$$

where $n_i(x)$ is the number of groundings in $F_i$ in $x$ and $Z$ is some normalization constant.

5.6 Subjectivity Disambiguation (SUBJ)

Many opinion words are known to be ambiguous. Some of their senses convey subjectivity while others do not (Akkaya et al., 2009). 13% of the compounds in Dataset I (Figure 1) are not subjective due to an ambiguous head. The modifier of such compounds neither represents a holder or a target. Examples are Luftdruck (air pressure) or Strömungswiderstand (flow resistance). Dataset II exclusively contains compounds whose modifiers are holders or targets. By definition, all those compounds are subjective. So a subjectivity feature may only be useful for the role-detection task, which uses Dataset I.

6On average, a head occurs in 5 different compounds on Dataset I, and in 4 different compounds on Dataset II.
For a feature indicating the subjectivity of a compound, we cannot look up the compounds in a sentiment lexicon since they are rarely included. Instead, we compute the 100 most similar German nouns for every compound and use as a feature the proportion of opinion nouns (according to the PolArt sentiment lexicon) on that list. Opinion nouns on that similarity list are less likely to be compounds and therefore more likely to be found in a sentiment lexicon. As in §5.2, similarity is measured by the cosine between two Word2Vec-vector embeddings. As a result, we find, for example, for Luftdruck (air pressure), other non-subjective terms, such as Temperatur (temperature) or Luftfeuchtigkeit (humidity), while for the subjective compound Hexenglaube (witch belief), we find the subjective expressions Abergläube (superstition) or Häresie (heresy).

5.7 Sentiment Views (VIEW)

Our final feature considers the sentiment view (Wiegand and Ruppenhofer, 2015) that an opinion noun, in our case the head of the compound, conveys. We distinguish between speaker views, expressions conveying sentiment of the speaker of the utterance (e.g. mistake, finesse, noise), and actor views, expressions conveying sentiment of the entities participating in the event denoted by the opinion noun (e.g. support, criticism, rating). Nouns conveying speaker views have an implicit opinion holder (i.e. the speaker). Therefore, if such a noun is the head of an opinion compound, the modifier cannot be a holder but only a target, e.g. A rztfehler (doctor’s mistake), Kinderlüm (children’s noise) or Neonazipropaganda (neonazi propaganda). Only heads conveying an actor view can take modifiers (Nutzerwertung/user rating) or a target (Opferunterstützung/victim support). Sentiment views may be helpful on Dataset II (Figure 1), where we have to decide between holders and targets. 40.3% of those heads convey a speaker view.

So far, the detection of sentiment views on a lexical level has only been examined for opinion verbs. Wiegand and Ruppenhofer (2015) propose a bootstrapping approach in which seed verbs for the different sentiment views are automatically extracted. Then, a label propagation algorithm (Talukdar et al., 2008) is run on a word-similarity graph generated from the opinion verbs. Thus labels from the seeds can be expanded to the remaining opinion verbs. The nodes in the graph correspond to the opinion verbs. The best performing graph is based on the similarity metric introduced in Lin (1998).

A critical step is the seed generation. Wiegand and Ruppenhofer (2015) extract seeds representing actor views by looking for opinion words frequently co-occurring with prototypical opinion holders (proOHs). These are common nouns, such as opponents or critics, that typically act as opinion holders (Wiegand and Klakow, 2011). By definition, such explicit opinion holders indicate an actor view. Seeds for speaker-view verbs are obtained by extracting verbs co-occurring with reproach-patterns, such as obji(beschuldigt/blamed for, <verb>) (14) that matches in (15).

Table 7: Head constraint as logic formula.

<table>
<thead>
<tr>
<th>Head constraint as logic formula.</th>
</tr>
</thead>
<tbody>
<tr>
<td>∀c1[∀c2[∀h[∀r1[∀r2] isCompound(c1) ∧ isCompound(c2) ∧ isHeadOf(h, c1) ∧ isHeadOf(h, c2) ∧ isRoleOfModifierOf(r1, c1) ∧ isRoleOfModifierOf(r2, c2) → (r1 == r2)]]]</td>
</tr>
</tbody>
</table>

This bootstrapping approach can be immediately applied to our setting. In the word-similarity graph, the opinion verbs are replaced by opinion nouns. With protoOHs, not only actor-view verbs but also actor-view nouns can be extracted. Similarly, the reproach-patterns work for both verbs (15) and nouns (17). (Only the dependency relation changes from obji (14) to objg (16).) ProtoOHs and reproach patterns are simply translated from English to German.

Wiegand and Ruppenhofer (2015) consider two types of actor views, agent view and patient view. The former take their opinion holder as an agent (typical verbs are criticize or support), while the latter align holders to patients (typical verbs are disappoint or please). Since this distinction of actor views does not exist among nouns, we combine them into a single category in this paper.

806
6 Experiments

We consider one binary MLN classifier for each of our three tasks (§4). Most of our features are frequently occurring features (e.g. paraphrases (§5.1), subjectivity feature (§5.6), sentiment views (§5.7)). Supervised classifiers only require few training data in order to assign appropriate weights to such features. Therefore, we sample 20% of the instances for each task of the respective dataset as training data. We test on the remaining 80% of the dataset. This procedure is repeated 5 times. The 5 training samples within each task are disjoint. We report macro-average F-score averaged over the 5 test samples.

We will first evaluate global features and then proceed to the local features. A division of our feature set into these groups was presented in Table 3.

6.1 Evaluation of Global Features

Table 8 compares the features that can be applied on all three tasks. On average, PARA (§5.1-§5.3) is slightly better than SEM (§5.4). Since their combination always results in a significant improvement, we conclude that these features contain complementary information. In the majority of cases, HEAD (§5.5) also yields significant improvement.

Table 9 compares the different subtypes of paraphrases (§5.1-§5.3). For all tasks, notable improvements are obtained by adding the other types of paraphrases to the plain paraphrases. While the joint paraphrases improve the plain paraphrases on all tasks, for the verb detour, improvements can be observed only for the extraction of holders and targets. However, this improvement is significantly better than that of the joint paraphrases. In summary, in order to obtain best possible results on all three types of classifications, we need all types of paraphrases.

6.2 Evaluation of the Local Feature for Role Detection

Table 10 examines the impact of the subjectivity feature (§5.6). We closely compare this feature with the head constraint since we found both features only working in combination with other features. In terms of statistical significance, the head constraint is more effective than the subjectivity feature.

6.3 Evaluation of the Local Feature for the Detection of Holders and Targets

Table 11 examines the impact of the sentiment-view feature (§5.7). We evaluate two variants of this feature. VIEW gold is a manual view annotation of all opinion head nouns. It should be considered an upper bound. The second variant, VIEW boot, employs the views as produced automatically by the bootstrapping approach outlined in §5.7.8

Table 11 shows that this feature has a notable impact on both PARA plain (i.e. the simplest feature set) and SEM+PARA+HEAD (i.e. the most complex feature set). This underlines that sentiment views are an important aspect for opinion role extraction.

8 Note that unlike Wiegand and Ruppenhofer (2015) we manually removed incorrect seeds from the set of automatically generated seeds (this affects less than 9% of the seeds).
Table 11: F-scores of sentiment view features.

<table>
<thead>
<tr>
<th>Task</th>
<th>PARA_plain</th>
<th>PARA+SEM+HEAD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>+VIEW</td>
<td>+VIEW</td>
</tr>
<tr>
<td>Holder</td>
<td>42.4</td>
<td>64.8*</td>
</tr>
<tr>
<td>Target</td>
<td>43.6</td>
<td>65.1*</td>
</tr>
</tbody>
</table>

* indicates statistical significance testing (paired t-test, significance level $p < 0.05$): better than w/o +VIEW, †: better than +VIEW

Table 12: Features for distant supervision (baseline) classifier.

6.4 Comparison against Baselines

Table 13 compares the best result from our previous experiments against 3 baselines. The first is a majority classifier predicting the majority class.

The second baseline is a classifier inspired by distant supervision (Mintz et al., 2009). As in our paraphrase features, this classifier considers the context in which modifier and head of a compound occur as separate constituents. The difference is, however, that we consider every such co-occurrence (within the same sentence) as a context that conveys the same relation as the one that is (implicitly) conveyed by the compound. Even though such an assumption is naive, it has been shown to produce quite reasonable performance in relation extraction (Mintz et al., 2009). The advantage of such an approach is that a generic relation extraction/opinion role extraction classifier can be trained on the resulting data. Unlike our proposed method, it does not require features tailored to the specific task (e.g. manually written paraphrases). Since the resulting feature set (see also Table 12) is fairly high-dimensional, we employ a support vector machine. As an implementation, we use SVM$^\text{light}$ (Joachims, 1999).

The third baseline is a distributional approach in which label propagation is performed on a word-similarity graph for compounds. The fundamental difference between that baseline and our proposed approach is that no relationship between head and modifier is modeled but just the contexts of the compounds themselves. We use the same (distributional) similarity metric to form the word-similarity graph and the same label propagation algorithm for this task as we did for bootstrapping sentiment views in §5.7. The only difference is that the nodes in the graph are opinion compounds instead of opinion nouns. The training data for the second and third baseline are the same compounds as in our previous experiments.

Table 13 shows that our proposed method substantially outperforms the baselines.

7 Conclusion

We presented an approach to the new task of opinion role extraction on opinion compounds. We produced a gold standard and proposed a method for classification. We did not only consider established features for noun compound analysis, i.e. paraphrases and semantic classes of heads and modifiers, but also proposed useful new features tailored to our task. We examined paraphrases that jointly consider holders and targets, a verb detour in which noun heads are replaced by related verbs, a global head constraint, and an auxiliary classification categorizing the sentiment view of the head of the compound. None of these features is language-specific.

Acknowledgements

We would like to thank Ines Rehbein for interesting discussions and helpful feedback on earlier drafts of the paper. The authors were partially supported by the German Research Foundation (DFG) under grants RU 1873/2-1 and WI 4204/2-1.
References


