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

Michael Wiegand and **Christine Bocionek**
Spoken Language Systems
Saarland University
D-66123 Saarbrücken, Germany
michael.wiegand@lsv.uni-saarland.de
cbocionek@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**$_{verb}$ [Mary $_{OT}$].
(2) [Mary $_{OH}$] was **skeptical**$_{adj}$ [about the plan $_{OT}$].
(3) [Peter $_{OH}$] had firm **intentions**$_{noun}$ [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

Proceedings of NAACL-HLT 2016, pages 800–810,
San Diego, California, June 12-17, 2016.

compounds	*user rating*; *victim support*; *spring upswing*	
immediate constituents	*user*; *victim*; *spring*	*rating*; *support*; *upswing*
grammatical function	modifier	head

Table 1: Internal structure of opinion compounds.

	Dataset I		**Dataset II**	
	2000 compounds 389 (unique) heads		1000 compounds 247 (unique) heads	
category of modifier	role	no role	holder	target
frequency	937	1063	450	580
proportion (in %)	46.85	53.15	45.00	58.00

Table 2: The two different datasets.

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.

2 Related Work

With regard to opinion role extraction, many features for supervised learning have been explored. They typically address the relationship between opinion word and opinion role on the basis of surface patterns (Choi et al., 2005), part-of-speech information (Wiegand and Klakow, 2010), syntactic information (Kessler and Nicolov, 2009; Jakob and Gurevych, 2010) or semantic role labeling (Johansson and Moschitti, 2013; Deng and Wiebe, 2015). The majority of those features cannot be applied to our task since for opinion compounds, there is no context between opinion role and opinion word.

In the area of noun compound analysis, there are two predominant approaches. On the one hand, lexical resources, such as WordNet (Miller et al., 1990), are employed in order to assign semantic categories to head and modifier and infer from those labels the 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).

3 Data & Annotation

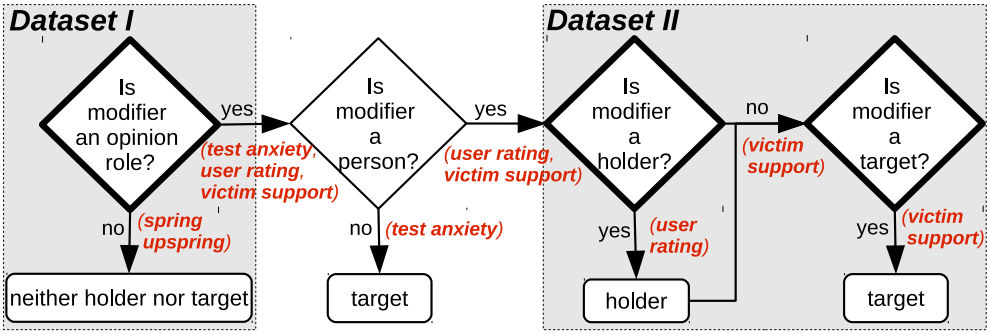
We created a new dataset[1] by retrieving opinion compounds from the ***deWaC***-corpus (Baroni et al., 2009) comprising 1.7 billion words. (Word embeddings (§5.2 & §5.6) and word similarity graphs (§5.7 & §6.4) were also created from this corpus.)

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).[2] 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 *Kundenerwartungen (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 differ-

[1] available at: `www.coli.uni-saarland.de/~miwieg/naacl_2016_op_compounds_data.tgz`

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



Each question (indicated by a rhombus) can be modeled with one binary supervised classifier. We build 3 classifiers, thus excluding the second question because of its simplicity.

Figure 1: Generic pipeline for processing opinion compounds.

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 substeps. 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*.

Global Features	
features used on all three tasks (i.e. Datasets I and II)	PARA (§5.1-5.3), SEM (§5.4), HEAD (§5.5)
Local Features	
feature used only on task *Role* (i.e. Dataset I)	SUBJ (§5.6)
feature used only on task *Holder* and task *Target* (i.e. Dataset II)	VIEW (§5.7)

Table 3: Division of global and local features.

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 lists which feature is used in which task. If a feature is restricted to a specific task (i.e. it is a local feature), then this is motivated below in the relevant subsection introducing the respective feature.

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).[4] *(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) objp$_{unter(among)}$(<opinion noun>, <holder>): Auffassung unter Experten (view among experts)

(11) objp$_{auf(towards)}$(<opinion noun>, <target>): Hass auf Christen (hatred towards Christians)

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 *objp*$_{von(of)}$ 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*.)

(12) *paraphrases for* Leserkommentar (reader comments):
  a) Kommentar$_{noun}$ [von Lesern $_{objp_{von}}$]
  (comment$_{noun}$ [of readers $_{objp_{of}}$]).
  b) Leser$_{subj}$ kommentieren$_{verb}$ ein Ereignis.
  (Readers$_{subj}$ comment$_{verb}$ on an event.)

(13) *paraphrases for* Schülerbeurteilung (student assessment):
  a) Beurteilung$_{noun}$ [von Schülern $_{objp_{von}}$]
  (assessment$_{noun}$ [of students $_{objp_{of}}$])
  b) Lehrer beurteilen$_{verb}$ Schüler$_{obj}$.
  (Teachers assess$_{verb}$ students$_{obj}$.)

Even though the disambiguation of deverbal noun compounds with the help of verb relations has been

[4] We obtain dependency parses by *ParZu* (Sennrich et al., 2009).

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).[5] 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$_{joint}$)

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$_{von/of}$(Widerstand/resistance, Bauern/farmers)* is ambiguous, in the first example of Table 4, it can only represent a holder, since the second relation *objp$_{gegen/against}$(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$_{mit/with}$(Zufriedenheit/satisfaction, Unternehmen/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 $objp_{von}^{\textbf{modifier}}\_objp_{gegen}$.

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 hypernymy 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)*.

[5] We used the `cbow`-model with 200 dimensions. All remaining parameters are set to their respective default values.

Pattern Type	Example Compound	Label	Example Sentence
<head> <holder> <target>	Bauern**widerstand** (farmer **resistance**)	holder	**Widerstand** [von Bauern $_{objp_{von}}$] [gegen das Gesetz $_{objp_{gegen}}$] (**resistance** [of farmers $_{objp_{of}}$] [against the regulation $_{objp_{against}}$])
	Schüler**beurteilung** (student **assessment**)	target	**Beurteilung** [der Lehrer $_{gmod}$] [von Schülern $_{objp_{von}}$] ([teachers' $_{possessive}$] **assessment** [of students $_{objp_{of}}$])
<compound> <target>	Mitarbeiter**zufriedenheit** (staff **satisfaction**)	holder	Mitarbeiter**zufriedenheit** [mit dem Unternehmen $_{objp_{mit}}$] (staff **satisfaction** [with their company $_{objp_{with}}$])
	Sprengstoff**anschlag** (bomb **attack**)	no role	Sprengstoff**anschlag** [auf Touristen $_{objp_{auf}}$] (bomb **attack** [on tourists $_{objp_{on}}$])
<compound> <holder>	Prüfungs**angst** (test **anxiety**)	target	Prüfungs**angst** [unter Schülern $_{objp_{unter}}$] (test **anxiety** [among students $_{objp_{among}}$])
<possessive> <head> <target>	Kinder**freundlichkeit** (child **friendliness**)	target	[seine $_{possessive}$] **Freundlichkeit** [gegenüber Kindern $_{objp_{gegenueber}}$] ([his $_{possessive}$] **friendliness** [towards children $_{objp_{towards}}$])

Table 4: Illustration of patterns for joint paraphrases.

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

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

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.[6] 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).

[6]On average, a head occurs in 5 different compounds on Dataset I, and in 4 different compounds on Dataset II.

Dataset I	88.86	Dataset II	91.36

Table 6: Role-purity of compounds with the same head.

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

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

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.

$\forall c_1[\forall c_2[\forall h[\forall r_1[\forall r_2[[isCompound(c_1) \wedge isCompound(c_2) \wedge isHeadOf(h, c_1) \wedge isHeadOf(h, c_2) \wedge isRoleOfModifierOf(r_1, c_1) \wedge isRoleOfModifierOf(r_2, c_2)] \rightarrow (r_1 == r_2)]]]]]$

Table 7: Head constraint as logic formula.

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 *Aberglaube (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. *Arztfehler (doctor's mistake)*, *Kinderlärm (children's noise)* or *Neonazipropaganda (neonazi propaganda)*. Only heads conveying an actor view can take modifiers to represent a holder *(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.[7] 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 (protoOHs)*. 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).

(14) *Pattern*: obji(*beschuldigt/blamed for*, <**speaker-view verb**>)

(15) Die UNO wurde *beschuldigt*, [die Klimadaten **fehlgedeutet**$_{verb}$ zu haben $_{obji}$]. (The UN was *blamed for* **misinterpreting**$_{verb}$ climate data.)

(16) *Pattern*: objg(*beschuldigt/blamed for*, <**speaker-view noun**>)

(17) Die UNO wurde [der **Fehldeutung**$_{noun\ \ objg}$] von Klimadaten *beschuldigt*. (The UN was *blamed for* the **misinterpretation**$_{noun}$ of climate data.)

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.

[7] 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.

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.

	Tasks		
Features	**Role**	**Holder**	**Target**
SEM	54.75	58.82	58.10
SEM+HEAD	56.33°	60.88°	60.33°
PARA	62.62	57.01	57.46
PARA+HEAD	63.82*†	59.07*	60.64*
PARA+SEM	63.92†	60.28	62.20‡
PARA+SEM+HEAD	**65.26***†‡	**61.58***†	**63.27**$^{\circ\ddagger}$

statistical significance testing (paired t-test): $^{\circ}$: better than w/o **+HEAD** ($p < 0.1$); *: better than w/o **+HEAD** ($p < 0.05$); †: better than **SEM+HEAD** ($p < 0.05$); ‡: better than **PARA+HEAD** ($p < 0.05$)

Table 8: F-scores of features applicable to all tasks.

	Tasks		
Features	**Role**	**Holder**	**Target**
PARA$_{plain}$	58.34	52.55	51.64
PARA$_{plain+joint}$	62.34*	54.87*	54.96*
PARA$_{plain+verb}$	58.85	**57.51***†	**58.43***†
PARA$_{plain+joint+verb}$	**62.62***	57.01*†	57.46*†

statistical significance testing (paired t-test, significance level $p < 0.05$) *: better than **PARA**$_{plain}$; †: better than **PARA**$_{plain+joint}$

Table 9: F-scores of paraphrase features.

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).

Features	**SEM**		**PARA**		**PARA+SEM**	
		+HEAD		**+HEAD**		**+HEAD**
	54.75	56.33†	62.62	63.82‡	63.92	65.26‡
+SUBJ	56.37°	58.57$^{\circ\dagger}$	63.07	64.76*‡	64.57	66.42$^{\circ\ddagger}$

statistical significance testing (paired t-test) $^{\circ}$: better than w/o **+SUBJ** ($p < 0.1$); *: better than w/o **+SUBJ** ($p < 0.05$); †: better than w/o **+HEAD** ($p < 0.1$); ‡: better than w/o **+HEAD** ($p < 0.05$)

Table 10: Comparison of SUBJ and HEAD evaluated on task *Role* (Dataset I); *evaluation measure: F-score*.

		PARA$_{plain}$			PARA+SEM+HEAD		
			+VIEW			+VIEW	
Task	**VIEW$_{gold}$**		**boot**	**gold**		**boot**	**gold**
Holder	42.4	52.6	59.5*	64.8*	61.6	64.7*	**71.2***†
Target	43.6	51.6	61.7*	65.1*	63.3	66.5*	**73.4***†

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

Table 11: F-scores of sentiment view features.

all words in the sentences (bag of words)
brown clusters of all words in the sentences (bag of clusters)
part-of-speech sequences between head and modifier mentions
part-of-speech tags before/after modifier mentions
part-of-speech tags before/after head mentions
dependency paths between head and modifier mentions
proportion of opinion words in the sentences

each training/test instance represents the set of all sentences in which head and modifier of a specific compound co-occur

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^{light} (Joachims, 1999).

		Tasks		
Features		**Role**	**Holder**	**Target**
BASELINES	Majority	34.70	35.49	36.71
	Distant Superv.	54.85	47.71	45.72
	Distributional	58.15	52.91	52.72
our approach *(best feature sets)*		**66.42***	**64.71***	**66.50***

*: better than all baselines according to statistical significance testing (paired t-test, significance level at $p < 0.05$)

Table 13: Comparison of our approach against baselines; *evaluation measure: F-score*.

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

Cem Akkaya, Janyce Wiebe, and Rada Mihalcea. 2009. Subjectivity Word Sense Disambiguation. In *Proceedings of EMNLP*, pages 190–199.

Ken Barker and Stan Szpakowicz. 1998. Semi-Automatic Recognition of Noun ModifierRelationships. In *Proceedings of COLING/ACL*, pages 96–102.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetti. 2009. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43(3):209–226.

Cristina Butnariu and Tony Veale. 2008. A Concept-Centered Approach to Noun-Compound Interpretation. In *Proceedings of COLING*, pages 81–88.

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*, pages 355–362.

Lingjia Deng and Janyce Wiebe. 2014. Sentiment Propagation via Implicature Constraints. In *Proceedings of EACL*, pages 377–385.

Lingjia Deng and Janyce Wiebe. 2015. Joint Prediction for Entity/Event-Level Sentiment Analysis using Probabilistic Soft Logic Models. In *Proceedings of EMNLP*, pages 179–189.

Corina Dima, Verena Henrich, Erhard Hinrichs, and Christina Hoppermann. 2014. How to Tell a Schneemann from a Milchmann: An Annotation Scheme for Compound-Internal Relations. In *Proceedings of LREC*, pages 1194–1201.

Roxana Girju, Dan Moldovan, Marta Tatu, and Daniel Antohe. 2005. On the semantics of noun compounds. *Computer Speech and Language*, 19:479–496.

Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2009. Classification of semantic relations between nominals. *Language Resources and Evaluation*, 43(2):105–121.

Roxana Girju. 2007. Improving the Interpretation of Noun Phrases with Cross-linguistic Information. In *Proceedings of ACL*, pages 568–575.

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*, pages 9–15.

Yufang Hou, Katja Markert, and Michael Strube. 2013. Global Inference for Bridging Anaphora Resolution. In *Proceedings of HLT/NAACL*, pages 907–917.

Niklas Jakob and Iryna Gurevych. 2010. Extracting Opinion Targets in a Single- and Cross-Domain Setting with Conditional Random Fields. In *Proceedings of EMNLP*, pages 1035–1045.

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.

Jason S. Kessler and Nicolas Nicolov. 2009. Targeting Sentiment Expressions through Supervised Ranking of Linguistic Configurations. In *Proceedings of ICWSM*.

Tushar Khot, Niranjan Balasubramanian, Eric Gribkoff, Ashish Sabharwal, Peter Clark, and Oren Etzioni. 2015. Exploring Markov Logic Networks for Question Answering. In *Proceedings of EMNLP*, pages 685–694.

Su Nam Kim and Timothy Baldwin. 2005. Automatic Interpretation of Noun Compounds Using Wordnet Similarity. In *Proceedings of IJCNLP*, pages 945–956. Springer.

Su Nam Kim and Timothy Baldwin. 2006. Interpreting Semantic Relations in Noun Compounds via Verb. In *Proceedings of COLING/ACL*, pages 491–498.

Su Nam Kim and Preslav Nakov. 2011. Large-Scale Noun Compound Interpretation Using Bootstrapping and the Web as a Corpus. In *Proceedings of EMNLP*, pages 648–658.

Manfred Klenner, Angela Fahrni, and Stefanos Petrakis. 2009. PolArt: A Robust Tool for Sentiment Analysis. In *Proceedings of NoDaLiDa*, pages 235–238.

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

Maria Lapata. 2002. The Disambiguation of Nominalizations. *Computational Linguistics*, 28(3):357–388.

Mark Lauer. 1995. *Designing Statistical Language Learners: Experiments on Noun Compounds*. Ph.D. thesis, Department of Computing, Macquarie University, Australia.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Dekang Lin. 1998. Automatic Retrieval and Clustering of Similar Words. In *Proceedings of ACL/COLING*, pages 768–774.

Ivan Meza-Ruiz and Sebastian Riedel. 2009. Jointly Identifying Predicates, Arguments and Senses using Markov Logic. In *Proceedings of HLT/NAACL*, pages 155–163.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of Workshop at the International Conference on Learning Representations (ICLR)*.

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

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant Supervision for Relation Extraction without Labeled Data. In *Proceedings of ACL/IJCNLP*, pages 1003–1011.

Preslav I. Nakov and Marti A. Hearst. 2013. Semantic Interpretation of Nouns Compounds Using Verbal and Other Paraphrases. *ACM Transactions on Speech and Language Processing*, 10(3).

Vivi Nastase and Stan Szpakowicz. 2003. Exploring Noun-Modifier Semantic Relations. In *Proceedings of IWCS*, pages 285–301.

Vivi Nastase, Jelber Sayyad-Shirabad, Marina Sokolova, and Stan Szpakowicz. 2006. Learning Noun-Modifier Semantic Relations with Corpus-based and WordNet-based Features. In *Proceedings of AAAI*, pages 781–786.

Sebastian Riedel. 2008. Improving the Accuracy and Efficiency of MAP Inference for Markov Logic. In *Proceedings of UAI*, pages 468–475.

Barbara Rosario and Marti Hearst. 2001. Classifying the Semantic Relations in Noun Compounds via a Domain-Specific Lexical Hierarchy. In *Proceedings of EMNLP*.

Josef Ruppenhofer, Julia Maria Struß, Jonathan Sonntag, and Stefan Gindl. 2014. IGGSA-STEPS: Shared Task on Source and Target Extraction from Political Speeches. *Journal for Language Technology and Computational Linguistics*, 29(1):33–46.

Rico Sennrich, Gerold Schneider, Martin Volk, and Martin Warin. 2009. A New Hybrid Dependency Parser for German. In *Proceedings of GSCL*, pages 115–124.

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*, pages 582–590.

Stephen Tratz and Eduard Hovy. 2010. A Taxonomy, Dataset, and Classifier for Automatic Noun Compound Interpretation. In *Proceedings of ACL*, pages 678–687.

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 HLT/NAACL*, pages 795–803.

Michael Wiegand and Dietrich Klakow. 2011. Prototypical Opinion Holders: What We can Learn from Experts and Analysts. In *Proceedings of RANLP*, pages 282–288.

Michael Wiegand and Dietrich Klakow. 2012. Generalization Methods for In-Domain and Cross-Domain Opinion Holder Extraction. In *Proceedings of EACL*, pages 325–335.

Michael Wiegand and Josef Ruppenhofer. 2015. Opinion Holder and Target Extraction based on the Induction of Verbal Categories. In *Proceedings of CoNLL*, pages 215–225.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. In *Proceedings of HLT/EMNLP*, pages 347–354.

Andrea Zielinski and Christian Simon. 2009. Morphisto – An Open Source Morphological Analyzer for German. In *Proceedings of the 2009 Conference on Finite-State Methods and Natural Language Processing: Post-proceedings of the 7th International Workshop FSMNLP 2008*, pages 224–231. IOS Press Amsterdam, The Netherlands.

Cäcilia Zirn, Mathias Niepert, Heiner Stuckenschmidt, and Michael Strube. 2011. Fine-Grained Sentiment Analysis with Structural Features. In *Proceedings of IJCNLP*, pages 336–344.