Supplementary Notes to *Opinion Holder and Target Extraction on Opinion Compounds – A Linguistic Approach (NAACL-HLT 2016)*

Michael Wiegand Christine Bocionek Josef Ruppenhofer

March 9, 2016

1 Introduction

This document provides more detailed information regarding certain aspects of our research for which there was not sufficient space in the main paper. We focus on two aspects, the creation of the gold standard (§2) and the paraphrase patterns (§3).

2 Creation of the Gold Standard

2.1 Getting Candidates from a Text Corpus

All compounds included in our gold standard were extracted from the deWaC-corpus [Baroni et al., 2009] which comprises 1.7 billion words. We only consider *opinion compounds*, by which we mean noun compounds whose head is an opinion noun. Other noun compounds are not relevant for our task, as the corresponding modifiers will not represent either opinion holder or opinion target. For the sake of simplicity, we only take into account *simple* opinion compounds. By that we understand compounds containing an atomic modifier and an atomic head. That is, we would include *Entwicklungskonzept (development concept)* or *Bombenangriff (bomb attack)* but we would not include *Gemeindeentwicklungskonzept (community development concept)* or *IRA-Bombenangriff (IRA bomb attack)*. Noun compounds in German are typically realized as closed compound nouns, that is, the compound is represented as one single token (mostly without a hyphen). In order to extract candidates for our gold standard, we checked whether the head of the compound is included in the sentiment lexicon of the PolArt-system [Klenner et al., 2009]. For identifying the head of a closed compound noun, morphological analysis was carried out. (We used *morphisto* [Zielinski and Simon, 2009].) In order to ensure that we ended up with simple opinion compounds, non-atomic opinion nouns were removed from that list of extracted opinion compounds.

2.2 Constraints Imposed upon the Compounds for the Gold Standard

For the opinion compounds included in our gold standard, we imposed several restrictions to ensure an unbiased and usable dataset. For one thing, we limit the compounds having identical heads to 10. Additionally, each compound must have been observed at least 5 times in *deWaC*.

The former restriction has been applied in order to ensure that our set of compounds is well-balanced and not biased towards a few frequently occurring heads. If for one particular head, there were more than 10 compounds, we included the 10 compounds that occurred most frequently in *deWaC*.

The latter restriction (i.e. observing each compound at least 5 times) is necessary in order to be able to carry out experiments taking into account contextual information of the compounds (and not their immediate constituents). We made use of this information in our *distributional baseline* in which we clustered the compounds according to their contextual information. This property of our dataset may actually favour that baseline (making it even harder for our proposed compositional method to beat). In other words, there exist opinion compounds which are rare in *deWaC* and therefore did not end up in our final gold standard. While our proposed compositional approach would have been able to process them, the distributional baseline would not be able to cope with them.

2.3 The Two Datasets of the Gold Standard

Our gold standard comprises two different datasets. The first dataset includes 2000 compounds in which we distinguish between modifiers representing some opinion role, e.g. *Nutzerwertung (user rating)*, from modifiers representing no opinion role, e.g. *Bombenattentat (bombing attack)*. That dataset does not distinguish between opinion holders, e.g *Verbraucherunsicherheit (consumer uncertainty)*, and opinion targets, e.g. *Prüfungsangst (test anxiety)*. For that distinction, the second dataset has been created.

The second dataset includes 1000 compounds. It is smaller than the first dataset since, given our constraints from above, we could not produce the same amount of compounds from *deWaC*. The modifiers in the second dataset exclusively represent opinion holders or opinion targets (i.e. only opinion roles). Additionally, we restricted ourselves to modifiers denoting persons. This is so because every modifier representing an opinion role which does not denote a person can be trivially classified as an opinion target. (The automatic distinction between persons and non-persons is a semantic classification which can be easily accomplished with a lexical resource, such as *WordNet* [Miller et al., 1990] – in our case we could make use of its German counterpart *GermaNet* [Hamp and Feldweg, 1997].) In contrast to that, the categorization of opinion compounds where the modifier denotes a person and represents some opinion role is the hard case. That is why we chose those compounds to be the instances of our second dataset.

Notice that Dataset II is <u>not</u> a subset of Dataset I. Dataset II <u>also</u> includes instances of opinion compounds that are not contained in Dataset I. Of the 2000 compounds in Dataset I, there are 937 compounds whose modifier represents some opinion role. From those 937 instances, however,

	Dataset I	Dataset II
Number of compounds	2000	1000
Number of different heads	389	247
Average number of occurrences of head	5.14	4.05

Table 1: Statistics on Head Distribution.

a large fraction are compounds whose modifier does not represent a person. Such compounds are always opinion targets. As discussed above, these compounds are not included in Dataset II since they can easily be identified as targets. Since we felt that the remaining opinion compounds from Dataset I, i.e. those whose modifier represents a person and convey some opinion role, were too few, we sampled more of such instances in order to have a larger gold standard to constitute Dataset II.

2.4 Heads of the Compounds

In our research, we also want to examine to what extent the head of a compound is predictive of the opinion role that the modifier of the compound represents. As mentioned above, in order to have a well-balanced representation of compounds with diverse heads, we imposed a restriction that the same head must not occur with more than 10 compounds. Table 1 provides some statistics regarding heads. It tells how many different heads the two datasets contain and also states the average number of occurrences of a head. With approximately 5 occurrences of each head on average in Dataset I and 4 on Dataset II, we believe we have created a setting in which each head is observed sufficiently often in order to study the interrelationship between head and the opinion role of the modifier.

3 Paraphrase Patterns

In this section we provide more detailed information regarding the paraphrase patterns.

Please note that there is no one-to-one mapping between German and English prepositions, therefore a dependency relation involving a preposition may be translated differently depending on its context.

3.1 The Paraphrase Relation Patterns

Table 2 lists each of the 18 (plain) paraphrase patterns that we use in our work. The patterns have been chosen ad hoc. Each pattern describes a particular dependency relation between the opinion noun (head) and its modifier. The choice of dependency relations is deliberate. Such representation offers us the best possible generalization on surface realizations. For example, both (1) and (2) would match the dependency relation $objp_{vor/for}$*(Angst/anxiety, Prüfungen/exams)*. (The label $objp_{vor/for}$ denotes the dependency relation between a predicate and the prepositional phrase headed by the preposition *vor (for)*.) While a simple token-based sequential pattern (e.g. *noun_*$prep_{vor}$*_noun*)

would also be sufficient to match (1), such pattern would fail to generalize for (2).

(1) **Angst** vor Prüfungen
(**anxiety** for exams)

(2) **Angst** mancher Leute vor wirklich schwierigen Prüfungen
(**anxiety** of some people for really difficult exams)

Ideally, one would use semantic role labels for these patterns (opinion holders mostly tend to be realized as agents while opinion targets tend to be realized as patients [Wiegand and Klakow, 2012]). However, the automatic detection of semantic roles of noun predicates and its arguments is still in its infancy. (This applies to both English and German.)

3.2 Joint Paraphrase Patterns

We use the four joint paraphrase patterns (3)-(6). Each of those patterns comprises two or three slots where <*holder*> can be instantiated by any plain paraphrase pattern from Table 2 that has been assigned to extract opinion holders. Likewise, the slot <*target*> can be instantiated by any plain paraphrase pattern from Table 2 that has been assigned to extract opinion targets.

(3) <head> <holder> <target>

(4) <compound> <target>

(5) <compound> <holder>

(6) <possessive_pronoun> <head> <target>

There are two plain paraphrase patterns $objp_{von}$ and *gmod* that are ambiguous. We deliberately included these two plain patterns because they both occur frequently. With the help of the joint paraphrase patterns, we can use these ambiguous patterns in order to detect holders and target. If such an ambiguous pattern occurs with another pattern (hopefully a less ambiguous pattern), then the ambiguous pattern can be disambiguated. (Apart from that, we found that ambiguous patterns, if used as a plain paraphrase, i.e. they are not part of a joint paraphrase pattern, can be used effectively to distinguish constituents that represent either an opinion holder or an opinion target, from constituents that represent no opinion role at all.)

For example, if we instantiate the joint paraphrase pattern <*head*> <*holder*> <*target*> (3) with the ambiguous $objp_{von/of}$ for <*holder*> and the less ambiguous $objp_{gegen/against}$ for <*target*>, we could match (7). Even though we include the ambiguous pattern $objp_{von/of}$ in that particular joint paraphrase pattern, it is more likely to convey some holder since it precedes another pattern most likely to convey a target.

(7) **Widerstand** [von Bauern $_{objp_{von}}$] [gegen die Bestimmung $_{objp_{gegen}}$]
(**resistance** [of farmers $_{objp_{of}}$] [against the regulation $_{objp_{against}}$])

Two major assumptions underlie our paraphrase patterns:

Pattern	Role	Context	Compound
objp_{unter} (objp_{among})	holder	Ansicht unter Experten (view among experts)	Expertenansicht (expert view)
$\text{objp}_{zwischen}$ ($\text{objp}_{between}$)	holder	Bündnis zwischen Staaten (alliance between states)	Staatenbündnis (state alliance)
objp_{durch} (objp_{by})	holder	Kontrolle durch das Militär (control by the military)	Militärkontrolle (military control)
subj_{haben} (subj_{have})	holder	Konsumenten haben Vertrauen in etwas. (Consumers have trust in something.)	Konsumentenvertrauen (consumer trust)
subj_{machen} (subj_{make})	holder	Der Vorstand macht einen Vorschlag. (The management makes a proposal.)	Vorstandsvorschlag (management proposal)
subj_{geben} (subj_{give})	holder	Ein Jurist gibt einen Ratschlag. (A lawyer gives some advice.)	Juristenratschlag (lawyer's advice)
objp_{an} (objp_{in})	target	Glaube an Hexen (belief in witches)	Hexenglaube (witch belief)
objp_{auf} (objp_{on})	target	Aussicht auf Erfolg (perspective on success)	Erfolgsaussicht (success perspective)
objp_{fuer} (objp_{for})	target	Unterstützung für Opfer (support for victims)	Opferunterstützung (victim support)
objp_{gegen} ($\text{objp}_{against}$)	target	Boykott gegen die Wahl (boycott against the election)	Wahlboykott (election boycott)
$\text{objp}_{gegenueber}$ ($\text{objp}_{towards}$)	target	Freundlichkeit gegenüber Kunden (friendliness towards customers)	Kundenfreundlichkeit (customer friendliness)
objp_{mit} (objp_{with})	target	Solidarität mit Flüchtlingen (solidarity with refugees)	Flüchtlingssolidarität (refugee solidarity)
objp_{nach} (objp_{to})	target	Sucht nach Alkohol (addiction to alcohol)	Alkoholsucht (alcohol addiction)
objp_{vor} (objp_{for})	target	Angst vor Prüfungen (anxiety for exams)	Prüfungsangst (exam anxiety)
objp_{um} (objp_{about})	target	Sorge um Geld (worries about money)	Geldsorgen (money worries)
objp_{zu} ($\text{objp}_{towards}$)	target	Verhältnis zum Vater (relationship towards one's father)	Vaterverhältnis (father relationship)
objp^{*}_{von} (objp_{of})	holder	Zufriedenheit von Kunden (satisfaction of customers)	Kundenzufriedenheit (customer satisfaction)
objp^{*}_{von} (objp_{of})	target	Unterstützung von Opfern (support of victims)	Opferunterstützung (victim support)
gmod^{*}	holder	Zufriedenheit der Kunden (customers' satisfaction)	Kundenzufriedenheit (customer satisfaction)
gmod^{*}	target	Unterstützung der Opfer (victims' support)	Opferunterstützung (victim support)

*: marks a dependency relation which is ambiguous, i.e. it may indicate an opinion holder or an opinion target *(these patterns are mainly included because they can be usefully exploited in joint paraphrases, see also §3.2; these patterns are also used in order to distinguish constituents representing some role from those which do not represent any role at all)*

Table 2: The full set of (plain) paraphrase patterns.

- An opinion noun usually comes with (exactly) one opinion holder and target each.
- The opinion holder is more likely to precede the opinion target.

Even though there are cases in which these conditions may not be fulfilled (for example, Wiegand and Ruppenhofer [2015] address opinion predicates having more than exactly one holder and target), we assume that they are correct for the majority of instances.

These assumptions may even help us in (8) where we face not only one but two ambiguous relations, i.e. $objp_{von}$ and *gmod*. By the order of those relations, we can conclude that *Firma (company)* is the opinion holder and *Kunden (customers)* is the opinion target of *Abhängigkeit (dependency)*.

(8) **Abhängigkeit** [der Firma $_{objp_{gmod}}$] [von Kunden$_{objp_{von}}$]
(**dependency** [of the company $_{objp_{of}}$] [on customers $_{objp_{on}}$])

The second joint paraphrase pattern (4) and third joint paraphrase pattern (5) are instantiated with a mention of the *compound* itself. At first sight, this may look contradictory to the idea to paraphrases for noun compound analysis, since, so far, we always split the compound into its modifier and its head, and looked for occurrences where both expressions were realized as individual constituents. However, for this particular joint paraphrase pattern, we treat the modifier within the compound as an ambiguous constituent, similar to $objp_{von/of}$ in (7).

A typical instantiation of pattern (4) is (9).

(9) Mitarbeiter**zufriedenheit** [mit dem Unternehmen $_{objp_{mit}}$]
(staff **satisfaction** [with their company $_{objp_{with}}$])

We observe the compound with the plain target pattern $objp_{mit/with}$. Since we have already found a target for the opinion noun *Zufriedenheit (satisfaction)*, the modifier *Mitarbeiter (staff)* can only be its opinion holder. The same procedure can be applied to identify targets, as in (10) instantiating pattern (5), where the prepositional phrase headed by *among (unter)* indicates an opinion holder which means that the modifier can only be a target.

(10) Prüfungs**angst** [unter Schülern $_{objp_{unter}}$]
(test **anxiety** [among students $_{objp_{among}}$])

In case we were dealing with a modifier that represents no person (or group of persons), such as *Sprengstoff (bomb)* in (11), and we have also observed some prepositional phrase representing a target, we may have found a case in which the modifier represents neither opinion holder nor target. This is an important property of our joint paraphrase patterns since our set of opinion compounds also includes compounds in which the modifier represents neither an opinion holder nor an opinion target.

(11) Sprengstoff**anschlag** [auf Touristen $_{objp_{auf}}$]
(bomb **attack** [on tourists $_{objp_{on}}$])

Finally, we also include a joint paraphrase pattern (6) containing possessive pronouns, e.g. *sein (his)*, *ihr (her)*, *unser (our)*, as a slot. We assume that possessive pronouns are good candidates for opinion holders. So if we find a paraphrase matching pattern (6), such as (12), we conclude that the constituent following the head represents some opinion target.

(12) seine **Freundlichkeit** [gegenüber Kindern $_{objp_{gegenueber}}$]
(his **friendliness** [towards children $_{objp_{towards}}$])

References

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

Birgit Hamp and Helmut Feldweg. 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, Madrid, Spain, 1997.

Manfred Klenner, Angela Fahrni, and Stefanos Petrakis. PolArt: A Robust Tool for Sentiment Analysis. In *Proceedings of the Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 235–238, Odense, Denmark, 2009.

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

Michael Wiegand and Dietrich Klakow. Generalization Methods for In-Domain and Cross-Domain Opinion Holder Extraction. In *Proceedings of the Conference on European Chapter of the Association for Computational Linguistics (EACL)*, pages 325–335, Avignon, France, 2012.

Michael Wiegand and Josef Ruppenhofer. Opinion Holder and Target Extraction based on the Induction of Verbal Categories. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*, pages 215–225, Beijing, China, 2015.

Andrea Zielinski and Christian Simon. 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, 2009.