# Using Anaphora Resolution to Improve Opinion Target Identification in Movie Reviews

**Niklas Jakob**
Technische Universität Darmstadt
Hochschulstraße 10, 64289 Darmstadt

**Iryna Gurevych**
Technische Universität Darmstadt
Hochschulstraße 10, 64289 Darmstadt

http://www.ukp.tu-darmstadt.de/people

## Abstract

Current work on automatic opinion mining has ignored opinion targets expressed by anaphorical pronouns, thereby missing a significant number of opinion targets. In this paper we empirically evaluate whether using an off-the-shelf anaphora resolution algorithm can improve the performance of a baseline opinion mining system. We present an analysis based on two different anaphora resolution systems. Our experiments on a movie review corpus demonstrate, that an unsupervised anaphora resolution algorithm significantly improves the opinion target extraction. We furthermore suggest domain and task specific extensions to an off-the-shelf algorithm which in turn yield significant improvements.

## 1 Introduction

Over the last years the task of opinion mining (OM) has been the topic of many publications. It has been approached with different goals in mind: Some research strived to perform subjectivity analysis at the document or sentence level, without focusing on what the individual opinions uttered in the document are about. Other approaches focused on extracting individual opinion words or phrases and what they are about. This aboutness has been referred to as the *opinion target* or opinion topic in the literature from the field. In this work our goal is to extract *opinion target - opinion word* pairs from sentences from movie reviews. A challenge which is frequently encountered in text mining tasks at this level of granularity is, that entities are being referred to by anaphora. In the task of OM, it can therefore also be necessary to analyze more than the content of one individual sentence when extracting opinion targets. Consider this example sentence: *"Simply put, it's unfathomable that this movie cracks the Top 250. It is absolutely awful."*. If one wants to extract what the opinion in the second sentence is about, an algorithm which resolves the anaphoric reference to the opinion target is required.

The extraction of such anaphoric opinion targets has been noted as an open issue multiple times in the OM context (Zhuang et al., 2006; Hu and Liu, 2004; Nasukawa and Yi, 2003). It is not a marginal phenomenon, since Kessler and Nicolov (2009) report that in their data, 14% of the opinion targets are pronouns. However, the task of resolving anaphora to mine opinion targets has not been addressed and evaluated yet to the best of our knowledge.

In this work, we investigate whether anaphora resolution (AR) can be successfully integrated into an OM algorithm and whether we can achieve an improvement regarding the OM in doing so. This paper is structured as follows: Section 2 discusses the related work on opinion target identification and OM on movie reviews. Section 3 outlines the OM algorithm we employed by us, while in Section 4 we discuss two different algorithms for AR which we experiment with. Finally, in Section 5 we present our experimental work including error analysis and discussion, and we conclude in Section 6.

## 2 Related Work

We split the description of the related work in two parts: In Section 2.1 we discuss the related work on OM with a focus on approaches for opinion target identification. In Section 2.2 we elaborate on findings from related OM research which also worked with movie reviews as this is our target domain in the present paper.

### 2.1 Opinion Target Identification

The extraction of opinions and especially opinion targets has been performed with quite diverse

approaches. Initial approaches combined statistical information and basic linguistic features such as part-of-speech tags. The goal was to identify the opinion targets, here in form of products and their attributes, without a pre-built knowledge base which models the domain. For the target candidate identification, simple part-of-speech patterns were employed. The relevance ranking and extraction was then performed with different statistical measures: Pointwise Mutual Information (Popescu and Etzioni, 2005), the Likelihood Ratio Test (Yi et al., 2003) and Association Mining (Hu and Liu, 2004). A more linguistically motivated approach was taken by Kim and Hovy (2006) through identifying opinion holders and targets with semantic role labeling. This approach was promising, since their goal was to extract opinions from professionally edited content i.e. newswire.

Zhuang et al. (2006) present an algorithm for the extraction of *opinion target - opinion word* pairs. The opinion word and target candidates are identified in the annotated corpus and their extraction is then performed by applying possible paths connecting them in a dependency graph. These paths are combined with part-of-speech information and also learned from the annotated corpus.

To the best of our knowledge, there is currently only one system which integrates coreference information in OM. The algorithm by Stoyanov and Cardie (2008) identifies coreferring targets in newspaper articles. A candidate selection or extraction step for the opinion targets is not required, since they rely on manually annotated targets and focus solely on the coreference resolution. However they do not resolve pronominal anaphora in order to achieve that.

### 2.2 Opinion Mining on Movie Reviews

There is a huge body of work on OM in movie reviews which was sparked by the dataset from Pang and Lee (2005). This dataset consists of sentences which are annotated as expressing positive or negative opinions. An interesting insight was gained from the document level sentiment analysis on movie reviews in comparison to documents from other domains: Turney (2002) observes that the movie reviews are hardest to classify since the review authors tend to give information about the storyline of the movie which often contain characterizations, such as *"bad guy"* or *"violent scene"*. These statements however do not reflect any opinions of the reviewers regarding the movie. Zhuang et al. (2006) also observe that movie reviews are different from e.g. customer reviews on Amazon.com. This is reflected in their experiments, in which their system outperforms the system by Hu and Liu (2004) which attributes an opinion target to the opinion word which is closest regarding word distance in a sentence. The sentences in the movie reviews tend to be more complex, which can also be explained by their origin. The reviews were taken from the Internet Movie Database[1], on which the users are given a set of guidelines on how to write a review. Due to these insights, we are confident that the overall textual quality of the movie reviews is high enough for linguistically more advanced technologies such as parsing or AR to be successfully applied.

## 3 Opinion Target Identification

### 3.1 Dataset

Currently the only freely available dataset annotated with opinions including annotated anaphoric opinion targets is a corpus of movie reviews by Zhuang et al. (2006). Kessler and Nicolov (2009) describe a collection of product reviews in which anaphoric opinion targets are also annotated, but it is not available to the public (yet). Zhuang et al. (2006) used a subset of the dataset they published (1829 documents), namely 1100 documents, however they do not state which documents comprise this subset used in their evaluation. In our experiments, we therefore use the complete dataset available, detailed in Table 1. As shown, roughly $9.5\%$ of the opinion targets are referred to by pronouns. Table 2 outlines detailed statistics on which pronouns occur as opinion targets.

Table 1: Dataset Statistics

| | |
|---|---|
| # Documents | 1829 |
| # Sentences | 24918 |
| # Tokens | 273715 |
| # Target + Opinion Pairs | 5298 |
| # Targets which are Pronouns | 504 |
| # Pronouns | $>$ 11000 |

### 3.2 Baseline Opinion Mining

We reimplemented the algorithm presented by Zhuang et al. (2006) as the baseline for our

[1] http://www.imdb.com (IMDB)

Table 2: Pronouns as Opinion Targets

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| it | 274 | he | 58 | she | 22 | they | 22 |
| this | 77 | his | 26 | her | 10 | | |
| | | him | 15 | | | | |

experiments. Their approach is a supervised one. The annotated dataset is split in five folds, of which four are used as the training data. In the first step, opinion target and opinion word candidates are extracted from the training data. Frequency counts of the annotated opinion targets and opinion words are extracted from four training folds. The most frequently occurring opinion targets and opinion words are selected as candidates. Then the annotated sentences are parsed and a graph containing the words of the sentence is created, which are connected by the dependency relations between them. For each *opinion target - opinion word* pair, the shortest path connecting them is extracted from the dependency graph. A path consists of the part-of-speech tags of the nodes and the dependency types of the edges.
In order to be able to identify rarely occurring opinion targets which are not in the candidate list, they expand it by crawling the cast and crew names of the movies from the IMDB. How this crawling and extraction is done is not explained.

## 4 Algorithms for Anaphora Resolution

As pointed out by Charniak and Elsner (2009) there are hardly any freely available systems for AR. Although Charniak and Elsner (2009) present a machine-learning based algorithm for AR, they evaluate its performance in comparison to three non machine-learning based algorithms, since those are the only ones available. They observe that the best performing baseline algorithm (OpenNLP) is hardly documented. The algorithm with the next-to-highest results in (Charniak and Elsner, 2009) is MARS (Mitkov, 1998) from the GuiTAR (Poesio and Kabadjov, 2004) toolkit. This algorithm is based on statistical analysis of the antecedent candidates. Another promising algorithm for AR employs a rule based approach for antecedent identification. The CogNIAC algorithm (Baldwin, 1997) was designed for high-precision AR. This approach seems like an adequate strategy for our OM task, since in the dataset used in our experiments only a small fraction of the total number of pronouns are actual opinion targets (see Table 1). We extended the CogNIAC implementation to also resolve *"it"* and *"this"* as anaphora candidates, since off-the-shelf it only resolves personal pronouns. We will refer to this extension with **[id]**. Both algorithms follow the common approach that noun phrases are antecedent candidates for the anaphora. In our experiments we employed both the MARS and the CogNIAC algorithm, for which we created three extensions which are detailed in the following.

### 4.1 Extensions of CogNIAC

We identified a few typical sources of errors in a preliminary error analysis. We therefore suggest three extensions to the algorithm which are on the one hand possible in the OM setting and on the other hand represent special features of the target discourse type: **[1.]** We observed that the Stanford Named Entity Recognizer (Finkel et al., 2005) is superior to the *Person* detection of the (MUC6 trained) CogNIAC implementation. We therefore filter out *Person* antecedent candidates which the Stanford NER detects for the impersonal and demonstrative pronouns and *Location & Organization* candidates for the personal pronouns. This way the input to the AR is optimized. **[2.]** The second extension exploits the fact that reviews from the IMDB exhibit certain contextual properties. They are gathered and to be presented in the context of one particular entity (=movie). The context or topic under which it occurs is therefore typically clear to the reader and is therefore not explicitly introduced in the discourse. This is equivalent to the situational context we often refer to in dialogue. In the reviews, the authors often refer to the movie or film as a whole by a pronoun. We exploit this by an additional rule which resolves an impersonal or demonstrative pronoun to *"movie"* or *"film"* if there is no other (matching) antecedent candidate in the previous two sentences. **[3.]** The rules by which CogNIAC resolves anaphora were designed so that anaphora which have ambiguous antecedents are left unresolved. This strategy should lead to a high precision AR, but at the same time it can have a negative impact on the recall. In the OM context, it happens quite frequently that the authors comment on the entity they want to criticize in a series of arguments. In such argument chains, we try to solve cases of antecedent ambiguity by analyzing the opinions: If there are ambiguous antecedent candidates for a

pronoun, we check whether there is an opinion uttered in the previous sentence. If this is the case and if the opinion target matches the pronoun regarding gender and number, we resolve the pronoun to the antecedent which was the previous opinion target.
In the results of our experiments in Section 5, we will refer to the configurations using these extensions with the numbers attributed to them above.

## 5 Experimental Work

To integrate AR in the OM algorithm, we add the antecedents of the pronouns annotated as opinion targets to the target candidate list. Then we extract the dependency paths connecting pronouns and opinion words and add them to the list of valid paths. When we run the algorithm, we extract anaphora which were resolved, if they occur with a valid dependency path to an opinion word. In such a case, the anaphor is substituted for its antecedent and thus extracted as part of an *opinion target - opinion word* pair.
To reproduce the system by Zhuang et al. (2006), we substitute the cast and crew list employed by them (see Section 3.2), with a NER component (Finkel et al., 2005). One aspect regarding the extraction of *opinion target - opinion word* pairs remains open in Zhuang et al. (2006): The dependency paths only identify connections between pairs of single words. However, almost $50\%$ of the opinion target candidates are multiword expressions. Zhuang et al. (2006) do not explain how they extract multiword opinion targets with the dependency paths. In our experiments, we require a dependency path to be found to each word of a multiword target candidate for it to be extracted. Furthermore, Zhuang et al. (2006) do not state whether in their evaluation annotated multiword targets are treated as a single unit which needs to be extracted, or whether a partial matching is employed in such cases. We require all individual words of a multiword expression to be extracted by the algorithm. As mentioned above, the dependency path based approach will only identify connections between pairs of single words. We therefore employ a merging step, in which we combine adjacent opinion targets to a multiword expression. We have compiled two result sets: Table 3 shows the results of the overall OM in a five-fold cross-validation. Table 4 gives a detailed overview of the AR for opinion target identification summed up over all folds. In Table 4, a true positive refers to an extracted pronoun which was annotated as an opinion target and is resolved to the correct antecedent. A false positive subsumes two error classes: A pronoun which was not annotated as an opinion target but extracted as such, or a pronoun which is resolved to an incorrect antecedent.
As shown in Table 3, the recall of our reimplementation is slightly higher than the recall reported in Zhuang et al. (2006). However, our precision and thus f-measure are lower. This can be attributed to the different document sets used in our experiments (see Section 3.1), or our substitution of the list of peoples' names with the NER component, or differences regarding the evaluation strategy as mentioned above.
We observe that the MARS algorithm yields an improvement regarding recall compared to the baseline system. However, it also extracts a high number of false positives for both the personal and impersonal / demonstrative pronouns. This is due to the fact that the MARS algorithm is designed for robustness and always resolves a pronoun to an antecedent.
CogNIAC in its off-the-shelf configuration already yields significant improvements over the baseline regarding f-measure[2]. Our CogNIAC extension [id] improves recall slightly in comparison to the off-the-shelf system. As shown in Table 4, the algorithm extracts impersonal and demonstrative pronouns with lower precision than personal pronouns. Our error analysis shows that this is mostly due to the *Person* / *Location* / *Organization* classification of the CogNIAC implementation. The names of actors and movies are thus often misclassified. Extension [1] mitigates this problem, since it increases precision (Table 3 row 6), while not affecting recall. The overall improvement of our extensions [id] + [1] is however not statistically significant in comparison to off-the-shelf CogNIAC.
Our extensions [2] and [3] in combination with [id] each increase recall at the expense of precision. The improvement in f-measure of CogNIAC [id] + [3] over the off-the-shelf system is statistically significant. The best overall results regarding f-measure are reached if we combine all our extensions of the CogNIAC algorithm. The results of this configuration show that the positive effects of extensions [2] and [3] are complemen-

[2]Significance of improvements was tested using a paired two-tailed t-test and $p \leq 0.05$ ($^*$) and $p \leq 0.01$ ($^{**}$)

Table 3: Op. Target - Op. Word Pair Extraction

| Configuration | Reca. | Prec. | F-Meas. |
|---|---|---|---|
| Results in Zhuang et al. | 0.548 | 0.654 | 0.596 |
| Our Reimplementation | 0.554 | 0.523 | 0.538 |
| MARS off-the-shelf | 0.595 | 0.467 | 0.523 |
| CogNIAC off-the-shelf | 0.586 | **0.534** | 0.559** |
| CogNIAC+[id] | 0.594 | 0.516 | 0.552 |
| CogNIAC+[id]+[1] | 0.594 | 0.533 | 0.561 |
| CogNIAC+[id]+[2] | 0.603 | 0.501 | 0.547 |
| CogNIAC+[id]+[3] | 0.613 | 0.521 | 0.563* |
| CogNIAC+[id]+[1]+[2]+[3] | **0.614** | 0.531 | **0.569*** |

Table 4: Results of AR for Opinion Targets

| Algorithm | Pers.[1] | | Imp. & Dem.[1] | |
|---|---|---|---|---|
| | TP[2] | FP[2] | TP | FP |
| MARS off-the-shelf | 102 | 164 | 115 | 623 |
| CogNIAC off-the-shelf | 117 | 95 | 0 | 0 |
| CogNIAC+[id] | 117 | 95 | 105 | 180 |
| CogNIAC+[id]+[1] | 117 | 41 | 105 | 51 |
| CogNIAC+[id]+[2] | 117 | 95 | 153 | 410 |
| CogNIAC+[id]+[3] | 131 | 103 | 182 | 206 |
| CogNIAC+[id]+[1]+[2]+[3] | 124 | 64 | 194 | 132 |

[1] **pers**onal, **imp**ersonal & **dem**onstrative pronouns
[2] **t**rue **p**ositives, **f**alse **p**ositives

tary regarding the extraction of impersonal and demonstrative pronouns. This configuration yields statistically significant improvements regarding f-measure over the off-the-shelf CogNIAC configuration, while also having the overall highest recall.

### 5.1 Error Analysis

When extracting opinions from movie reviews, we observe the same challenge as Turney (2002): The users often characterize events in the storyline or roles the characters play. These characterizations contain the same words which are also used to express opinions. Hence these combinations are frequently but falsely extracted as *opinion target - opinion word* pairs, negatively affecting the precision. The algorithm cannot distinguish them from opinions expressing the stance of the author. Overall, the recall of the baseline is rather low. This is due to the fact that the algorithm only learns a subset of the opinion words and opinion targets annotated in the training data. Currently, it cannot discover any new opinion words and targets. This could be addressed by integrating a component which identifies new opinion targets by calculating the relevance of a word in the corpus based on statistical measures.

The AR introduces new sources of errors regarding the extraction of opinion targets: Errors in gender and number identification can lead to an incorrect selection of antecedent candidates. Even if the gender and number identification is correct, the algorithm might select an incorrect antecedent if there is more than one possible candidate. A non-robust algorithm as CogNIAC might leave a pronoun which is an actual opinion target unresolved, due to the ambiguity of its antecedent candidates.

The upper bound for the OM with perfect AR on top of the baseline would be recall: 0.649, precision: 0.562, f-measure: 0.602. Our best configuration reaches $\sim 50\%$ of the improvements which are theoretically possible with perfect AR.

## 6 Conclusions

We have shown that by extending an OM algorithm with AR for opinion target extraction significant improvements can be achieved. The rule based AR algorithm CogNIAC performs well regarding the extraction of opinion targets which are personal pronouns. The algorithm does not yield high precision when resolving impersonal and demonstrative pronouns. We present a set of extensions which address this challenge and in combination yield significant improvements over the off-the-shelf configuration. A robust AR algorithm does not yield any improvements regarding f-measure in the OM task. This type of algorithm creates many false positives, which are not filtered out by the dependency paths employed in the algorithm by Zhuang et al. (2006).

AR could also be employed in other OM algorithms which aim at identifying opinion targets by means of a statistical analysis. Vicedo and Ferrández (2000) successfully modified the relevance ranking of terms in their documents by replacing anaphora with their antecedents. The approach can be taken for OM algorithms which select the opinion target candidates with a relevance ranking (Hu and Liu, 2004; Yi et al., 2003).

## Acknowledgments

The project was funded by means of the German Federal Ministry of Economy and Technology under the promotional reference "01MQ07012". The authors take the responsibility for the contents. This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806.

## References


Breck Baldwin. 1997. Cogniac: High precision coreference with limited knowledge and linguistic resources. In *Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 38–45, Madrid, Spain, July.

Eugene Charniak and Micha Elsner. 2009. EM works for pronoun anaphora resolution. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 148–156, Athens, Greece, March.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 363–370, Michigan, USA, June.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177, Seattle, WA, USA, August.

Jason Kessler and Nicolas Nicolov. 2009. Targeting sentiment expressions through supervised ranking of linguistic configurations. In *Proceedings of the Third International AAAI Conference on Weblogs and Social Media*, San Jose, CA, USA, May.

Soo-Min Kim and Eduard Hovy. 2006. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the ACL Workshop on Sentiment and Subjectivity in Text*, pages 1–8, Sydney, Australia, July.

Ruslan Mitkov. 1998. Robust pronoun resolution with limited knowledge. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 869–875, Montreal, Canada, August.

Tetsuya Nasukawa and Jeonghee Yi. 2003. Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd International Conference on Knowledge Capture*, pages 70–77, Sanibel Island, FL, USA, October.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 115–124, Michigan, USA, June.

Massimo Poesio and Mijail A. Kabadjov. 2004. A general-purpose, off-the-shelf anaphora resolution module: Implementation and preliminary evaluation. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 663–666, Lisboa, Portugal, May.

Ana-Maria Popescu and Oren Etzioni. 2005. Extracting product features and opinions from reviews. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 339–346, Vancouver, Canada, October.

Veselin Stoyanov and Claire Cardie. 2008. Topic identification for fine-grained opinion analysis. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 817–824, Manchester, UK, August.

Peter Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424, Philadelphia, Pennsylvania, USA, July.

José L. Vicedo and Antonio Ferrández. 2000. Applying anaphora resolution to question answering and information retrieval systems. In *Proceedings of the First International Conference on Web-Age Information Management*, volume 1846 of *Lecture Notes In Computer Science*, pages 344–355. Springer, Shanghai, China.

Jeonghee Yi, Tetsuya Nasukawa, Razvan Bunescu, and Wayne Niblack. 2003. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Proceedings of the 3rd IEEE International Conference on Data Mining*, pages 427–434, Melbourne, FL, USA, December.

Li Zhuang, Feng Jing, and Xiao-Yan Zhu. 2006. Movie review mining and summarization. In *Proceedings of the ACM 15th Conference on Information and Knowledge Management*, pages 43–50, Arlington, VA, USA, November.