Comparing Shallow and Dependency Syntactic Analysis for Opinion Target Extraction

Aleksander Wawer

Institute of Computer Science, Polish Academy of Science
ul. Jana Kazimierza 5, 01-238 Warszawa, Poland
axw@ipipan.waw.pl

Abstract

This paper is focused on a specific sub-problem of opinion mining, namely opinion target extraction. Its goal is to find an entity (word, phrase or named entity) that is the target of negative or positive sentiment expression (opinion). In our work, we compare two approaches to opinion target extraction in the Polish language, based on two different available syntactic parsing methods: shallow parsing, as implemented in Spejd parser, and dependency parsing. We use a MaltParser's model trained for the Polish language (Wroblewska, 2014) and set of patterns described in (Wawer, 2015). The aim of our work is to use structures and information available in both syntactic methods to predict relations between sentiments (opinions) and their targets. We compare both approaches on two types of texts: sentences from product reviews and on tweets.

1. Introduction

Opinion extraction can be seen as a process of identifying and categorizing opinions expressed in a piece of text (as for example, word, phrase, sentence or document), in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral. It can be seen as consisting of multiple sub-tasks, such as identifying opinions and their polarities, often called sentiment analysis, and finding their targets (topics or objects of opinions). The goal of the latter sub-task, called opinion target extraction, is the recognition of words towards which an opinion (sentiment) is expressed.

In this study we focus only on opinion-bearing sentiment expressions: those that can be linked to their targets, express an attitude towards something. This clarification is required because not all sentiment words and emotive expressions convey opinions. For instance, there exist words with negative connotative (evaluative) value, such as for example *prison*, that are not opinion-related.

Typically, in the domain of product reviews, opinion targets are aspect terms related to the reviewed entity or entity itself. For instance, users express opinions about batteries or screens of their laptops (for example: *the battery of this laptop is rather poor, but screen has great vivid colors.*) or directly about entities (*this Dell is awesome*). In other types of texts, such as blog entries, news articles or tweets, an opinion target could be any other expression type, including beliefs, decisions or deeds of other people (eg. *It's bad you did this.*), events and even other opinions (eg. *I disagree with our opinion*). In fact, they could denote any type of object.

Unfortunately, relations between opinions and targets are often difficult to identify even for humans, due to their indirect character. For example, in: *I like this perfume's bottle*, should the relation be between *like* (sentiment) and *perfume* or rather *bottle* (two possible targets)? We instructed the annotators to prefer semantic heads of phrases in such cases, but unfortunately this rule was not always easy to follow.

The formulation of opinion target extraction we assume in our paper is similar as in (Qiu et al., 2011), where authors also consider multiple types of opinion targets, including entities and their aspects. To extract opinion targets, they apply two simple dependency patterns that are matched against sentiment words.

Rule-based aspect extraction was described in (Poria et al., 2014). Authors use hand-crafted dependency rules on the parse trees to extract aspects. The method is capable to recognize implicit aspects (defined as aspect expressions that are not nouns or noun phrases) and outperforms multiple other approaches.

In the Polish language, a hybrid approach of dependency patterns and machine learning based on CRF has been recently proposed in (Wawer, 2015). The method performs well using only syntactic information (no lexical information included in the CRF) and therefore is believed to be relatively domain independent.

The paper is organized as follows. In Section 1. we formulate the problem and review existing work. In Section 2. we describe the data sets used for evaluation of the performance of both methods, compared in this paper. Sections 3. and 4. describe two syntactic parsing methods, shallow and dependency based, that are used for opinion target extraction. Sections 5. and 6. discuss results obtained using each of those methods. Finally, Section 7. concludes the paper.

2. Data Set

The data set used in our work consists of two parts. The first one is based on Twitter, while the second one is sentences selected from a corpus of product reviews.

2.1. Twitter

Twitter messages are a special type of texts, taken into account in this comparison due to their distinctive, informal character. This is due to two several reasons.

First, this type of data was rarely studied for Polish, with some rare exceptions such as (Piskorski and Ehrmann,

This work was funded by the National Science Centre of Poland grant nr UMO-2012/05/N/ST6/03587

2013). However, it is increasingly important, due to its growing usage and relatively open access, especially in the context of business and government applications, where monitoring of social media streams plays an important role in decision support.

Second, tweets are interesting object of study in the context of natural language processing due to their special form. They are limited to 140 characters and therefore are written in concise manner. Many of them contain abbreviated word forms and acronyms for common messages. They contain serious amounts of all types of noise, such as misspellings and grammatical errors, mixing words with other types of non-linguistic data such as hashtags, emoticons or urls. Issues specific to Polish include omitting diacritical marks. Due to all these problems, real-world applications for tweets rarely employ deep syntactic analysis methods, sensitive to input noise, such as dependency parsing. To overcome this issue, dedicated Twitter dependency tools have been developed (Kong et al., 2014). However, they are not available for the Polish language.

The tweets were acquired from the data set gathered during the Trendminer project[1]. They contain messages acquired from twitter channels of Polish politicians, journalists and other public figures. The selection of messages for this comparison was purely random. Initially, the input data set consisted of 1000 tweets. We then filtered them for presence of at least one known sentiment word using the lexicon described by (Wawer and Rogozińska, 2012) and available from `http://zil.ipipan.waw.pl/SlownikWydzwieku`. The lexicon was created automatically using supervised learning techniques. It contains four colums that describe word sentiment. The first three columns reflect sentiment scores computed using an SVM classifier, where input features consists of up to 300 word co-occurrence (concordance) vectors generated from the National Corpus of Polish (www.nkjp.pl). The three columns are:

- Neutral (0) vs positive or negative (1).
- Negative (-1), neutral (0), positive (1).
- Very negative (-2), negative (-1), neutral (0), positive (1), very positive (2).

The last column in the sentiment dictionary is the SO-PMI score calculated using the svd-based paradigm words selection described in (Wawer, 2012).

For the purpose of this comparison, we selected only words where at least two of the automated predictors agreed with regard to word polarity. This selection resulted in 300 tweets, submitted to manual labeling of relations between opinion (sentiment) words and their targets.

The purpose of labeling was to (1) verify sentiment of word indicated by the dictionary and (2) match it to its possible target. Step (1) was necessary, because the actual polarity of many of the words in the sentiment dictionary is context-dependent - verification was needed if in the context of a given tweet, the word is still negative or positive.

[1]`http://www.trendminer-project.eu`

2.2. Reviews

The sentences selected from the corpus of product reviews are the data set used in (Wawer and Gołuchowski, 2012). It consists of reviews, downloaded from one of the biggest Polish opinion aggregation websites, for two types of products: clothes and perfumes.

We selected the sentences with known sentiment words, as identified by manually adjusted version of the domain-independent Polish sentiment lexicon (available from `http://zil.ipipan.waw.pl/SlownikWydzwieku`), and known opinion target words, identified using the lexicon obtained in (Wawer and Gołuchowski, 2012). We parsed sentences using the MaltEval dependency parser and model for the Polish language.

The basic statistics in terms of number of sentences are presented in Table 1.

	perfume	clothes
sentences	946	418

Table 1: Sentences by product type.

For each dependency tree with automatically labeled candidates for opinion words and candidates for their targets, linguists annotated the correctness of both words (whether they are really opinions and opinion targets) and their relationship as a valid opinion-target pair.

The data set with all annotations, as a json file, is available from `http://zil.ipipan.waw.pl/OPTA`.

3. Opinion Target Extraction using Shallow Parsing

For the shallow parsing framework, we selected the Spejd (Buczyński and Przepiórkowski, 2009). In this approach parsing rules are defined using a cascade of regular grammars which match against orthographic forms or morphological interpretations of particular words, including grammatical categories (or part-of-speech). Spejd's pattern-matching language supports regular expression alike operators defined for token sequences that fall into main matched expression, its left and right sides. It also supports a variety of actions to perform on the matching fragments: accepting and rejecting morphological interpretations, agreement of entire tags or particular grammatical categories, grouping.

For this experiment, we used multiservice Spejd configuration, based on a grammar of Polish developed by K. Glowinska within NKJP [2].

When performing opinion target extraction, we hypothesize that syntactic groups are the relevant level of syntactic descriptions. We assume that sentiment phrase matches opinion target if both occur in the same syntactic group.

[2]`http://clip.ipipan.waw.pl/LRT?action=AttachFile&do=view&target=gramatyka_Spejd_NKJP_1.0.zip.`

4. Opinion Target Extraction using Dependency Parsing

For dependency parsing, we used MaltParser (Nivre et al., 2007), a system for data-driven dependency parsing, which can be used to induce a parsing model from treebank data and to parse new data using an induced model. MaltParser is based on a transition-based dependency parsing method: it consists of a transition system for deriving dependency trees, coupled with a classifier for deterministically predicting the next transition given a feature space created from representation of the current parser state.

We used a model prepared for the Polish language, described by (Wroblewska, 2014), and implemented as a part of Multiservice project available at `http://zil.ipipan.waw.pl/Multiservice`. As the author claims, it achieves 84.7% LAS and 90.5% UAS when tested against the Polish Dependency Bank validation set and 68.5% LAS/72.2% UAS when tested against the set of 50 manually annotated test sentences (see `http://zil.ipipan.waw.pl/PolishDependencyParser`).

The process of opinion-target relation recognition has been described in more detail in (Wawer, 2015). This section contains only an overview. The procedure involves generating dependency patterns as dependency path descriptions by starting off from the opinionated word and traversed dependency tree using the shortest possible path to its associated opinion target. A dependency pattern is produced by remembering POS of intermediate tokens and dependency labels of traversed edges. The method, when applied to the review data set, generated 173 dependency patterns available from `http://zil.ipipan.waw.pl/OPTA`.

5. Results: Shallow Parsing

5.1. Twitter corpus

In twitter data set corpus, out of 122 manually marked, true pairs of opinion and target, shallow parsing methods correctly identified 45 pairs (true positives or TPs), generated 27 false indications (false positives of FPs) and missed 50 pairs that fell out of the scope of any Spejd grammar rule (false negatives or FNs). The precision of this approach can be estimated at 0.62, while recall at 0.47.

Rule-level frequencies of twitter corpus analysis are presented in Table 2. Generally, it appears that nominal groups (various forms of NG) are much better indicators of a valid opinion-target relation than other types of relations. However, some form of additional filtering is still needed, perhaps using machine learning from word-level (lexical) information, as their presence is not a very strong indicator of the presence of valid opinion-target relation.

Some types of groups are not usable for opinion-target relation identification. Their appearance is not related to opinion-target relations. The most apparent example of these are adjective groups (AdjG).

5.2. Review corpus

In this corpus, out of 1315 marked, possible pairs of opinion and target, shallow parsing methods correctly identified 467 pairs (true positives or TPs), generated 325 false indications (false positives of FPs) and missed 523 pairs that fell out of the scope of any Spejd grammar rule (false negatives or FNs). The precision of this approach can be estimated at 0.59, while recall at 0.47.

Rule Name	TP	FP
NGx: pronoun + Adj gen	1	1
NGg: Noun + n-Noun w gen	7	2
NGg: Noun + n-Noun (gen)	1	0
NGs: Noun + n-Noun (nom)	0	1
NGa: Adj + Noun	21	6
NGk: N|NG i N|NG (coordination)	1	0
PrepNG: Prep + NG	9	6
PrepNG with a group in quotes	1	1
NG with adjunct in nom (1)	1	0
PrepAdjG	0	1
AdjG: 2*Adj	0	1
AdjG: Adv + Adj	0	5
AdvG: Adv + Adv	0	3
CG: subordinate clause with że, żeby (2)	1	0
CG: subordinate clause with że, żeby (1)	2	0
Total	45	27

Table 2: Rule level frequencies of twitter opinion target matching using shallow parsing, presented as True Positives (TPs) and False Positives (FP).

Rule-level frequencies of review corpus analysis are presented in Table 3. The "Adj + Noun" rule generated as much as 173 true positives, its "NGa" variant - only 61 false positives. The observation made for tweets, holds also in case of reviews: nominal groups are an indicator of opinion-target relation, but their sole presence isn't usually enough. Very few rules provide only true positives and the notable exception is "Adj + Noun". As in the case of tweets, adjective groups (AdjG) did not provide any useful information for opinion-target relation.

5.3. Discussion

The method yielded slightly higher precision on the twitter data set. The explanation of this fact is possibly related to lower syntactic complexity of twitter language, due to its concise character, so that a larger proportion of the opinion-target pairs may be captured using shallow grammar rules. However, due to small size of Twitter data, comparisons should be made with caution because error margins exceed actual difference between observed values.

Overall, we hypothesize that the shallow parsing approach to opinion target extraction presented here may be possibly improved by removing rules not indicative of the relation, such as adjective groups, and by introducing some form of machine learning to increase precision.

6. Results: Dependency Parsing

The other method evaluated in this paper to extract opinion targets, assuming a known set of correctly labeled sentiment (opinion) words, is dependency-based. In this method, we use dependency patterns to extract opinion targets. This section describes the results of applying depen-

Rule Name	TP	FP
[pos:adj] <pd [pos:fin] >subj [pos:subst]	1	4
[pos:adj] <adjunct [pos:subst]	16	8
[pos:adj] <adjunct [pos:subst] >app [pos:subst]	1	8
[pos:adj] <adjunct [pos:subst] >adjunct [pos:subst]	0	2
[pos:adj] <adjunct [pos:subst] <adjunct [pos:subst]	0	1
[pos:adv] <comp [pos:fin] >subj [pos:subst]	1	1
[pos:adv] <adjunct [pos:fin] >subj [pos:subst]	0	1
[pos:ppas] <adjunct [pos:subst]	2	1
[pos:fin] >obj [pos:subst] >adjunct [pos:subst]	0	1
[pos:fin] >subj [pos:subst] >app [pos:subst]	0	2
[pos:fin] <adjunct [pos:subst]	0	1
[pos:subst] <adjunct [pos:subst]	0	1
[pos:subst] <comp [pos:prep] <adjunct [pos:subst]	0	2
[pos:subst] <obj [pos:fin] >subj [pos:subst]	0	3
[pos:adj] <conjunct [pos:conj] >conjunct [pos:fin] >subj [pos:subst]	0	2
[pos:adj] <adjunct [pos:subst] <obj [pos:fin] >subj [pos:subst]	0	1
[pos:adj] <adjunct [pos:subst] <comp [pos:prep] <adjunct [pos:subst]	0	4
[pos:adj] <adjunct [pos:subst] <conjunct [pos:conj] >conjunct [pos:subst]	0	1
Total	21	44

Table 4: Rule level frequencies of twitter dataset opinion target matching using dependency parsing, presented as True Positives (TPs) and False Positives (FP).

dency patterns to tweets corpus. The patterns were originally created for review data set. The procedure of their creation has been illustrated in (Wawer, 2015).

In dependency-based opinion target extraction as outlined in our approach, one starts from sentiment word, and by following a sequence of moves on dependency tree (a tweet, in this case), described according to some formal system, to "arrive" at an opinion target word. Two simple syntactic structures of this kind are described in (Qiu et al., 2011) and used for double propagation of sentiments and opinion targets in a corpus. Patterns used in our work are more complex, as they take into account dependency labels and POS tags.

In the pattern matching system, tokens are expressed as enclosed in [..] and dependency relations as < or >, depending on the direction. We may specify dependency label type, following dependency relation mark. We may also specify that encountered tokens belong to specified POS type (eg. [pos:verb] to specify verbs).

We selected all manually labeled sentiment words as inputs (left side of each pattern) from the twitter data set and used as starting points for pattern matching. Every matched pattern generated either a true positive (a target word, a rightmost result of pattern matching, was really an opinion target) or a false positive (word that was indicated by matched pattern was not an opinion target for the opinion word used as starting point). We treated unmatched targets as false negatives.

In the twitter data set, out of 122 manually marked, true pairs of opinion and target, dependency patterns correctly identified 21 pairs (true positives or TPs), generated 44 false indications (false positives of FPs) and missed 57 pairs that fell out of the scope of any dependency pattern (false negatives or FNs). The precision of this approach can be estimated at 0.32, while recall at 0.27.

Rule-level frequencies of review corpus analysis are presented in Table 4. Even the most productive rule (in terms of true positives), namely "[pos:adj] <adjunct [pos:subst]", generated about half as many false positives. No rule generated exclusively true positives.

6.1. Discussion

The results of dependency-based opinion target extraction on Twitter data set are not satisfactory. The alternative method, based on shallow parsing, performs notably better. The most likely explanation of this fact has to do with relative unsuitability of generic dependency parser model, trained on clean sentences, to use on noisy tweets. We hypothesize that this state could be improved upon when using Twitter dedicated parser that takes into account twitter-specific phenomena, in a manner similar to (Kong et al., 2014). For the Polish language, such parser has yet to be developed.

7. Conclusions

In this paper, we evaluated two methods of opinion target extraction: based on shallow and dependency parsing.

In general, a conclusion that can be formulated is that nominal groups with noun and adjective are the best indication for the relationship between the aspect and opinion.

The evaluation of shallow-based method indicated very similar performance on both data sets, tweets and reviews. Interestingly, the observed precision on tweets turned out to be similar to that observed on reviews. Actual values are a bit higher for tweets, however due to small size of Twitter data, comparisons should be made with caution due to error margins exceeding actual difference between observed values.

Rule Name	TP	FP
Adj + Noun	173	0
NGg: Noun + n-Noun in gen	53	24
NGa: 2*Adj + Noun	17	5
NGa: 2*Adj + Noun + Adj	1	0
NGa: Adj + Noun	0	61
NGa: Adj + Noun + Adj	10	3
NGa: Noun + Adj	45	18
NGs: Noun + n-Noun (nom)	2	3
NGs: Noun + n-Noun (acc)	0	1
NGk: N|NG i N|NG (coordination)	22	8
NGk: Noun i Noun (coordination)	0	2
NGx: pronoun + Adj gen	0	5
AdjGe: one of ...	9	7
PrepNG: Prep + NG	101	35
PrepAdjG	2	21
AdvG: Adv + Adv	0	15
AdjGk: Adj i Adj (coordination)	0	19
AdjG: Adv + Adj	0	81
AdjG: 2*Adj	0	4
NumGr: Num + Noun	1	0
CG: subordinate clause with że, żeby (1)	26	12
CG: subordinate clause with że, żeby (2)	5	1
Total	467	325

Table 3: Rule level frequencies of review dataset opinion target matching using shallow parsing, presented as True Positives (TPs) and False Positives (FP).

We compared both methods on the Twitter data set. In this case, the method based on dependency parsing turned out to perform worse, almost by a half in terms of precision and recall. We suspect that this is due to low quality of dependency parsing caused by high amount of noise present on Twitter (many misspellings, hashtags and abbreviations).

Both of the methods suffer from extraction rules that generate a lot of false positives. In the case of shallow parsing, they may simply be removed from the set of rules used to extract opinion targets, increasing the overall precision.

A serious issue that has to be address in further studies is the fact that even the best rules still need additional steps to increase precision (or remove false positives). We may hypothesize that this could be achieved using machine learning techniques, but this in turn can lead to domain dependency, especially if using lexical information to train classifiers.

A promising direction for the further research would be to apply a twitter-specific dependency parser, once an appropriate tool becomes available for the Polish language.

8. References

Buczyński, Aleksander and Adam Przepiórkowski, 2009. Spejd: A shallow processing and morphological disambiguation tool. In Zygmunt Vetulani and Hans Uszkoreit (eds.), *Human Language Technology: Challenges of the Information Society*, volume 5603 of *Lecture Notes in Artificial Intelligence*. Berlin: Springer-Verlag, pages 131–141.

Kong, Lingpeng, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith, 2014. A dependency parser for tweets. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics.

Nivre, Joakim, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi, 2007. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.

Piskorski, Jakub and Maud Ehrmann, 2013. On Named Entity Recognition in Targeted Twitter Streams in Polish. In *Proceedings of the 4^{th} Biennial Workshop on Balto-Slavic Natural Language Processing (BSNLP), collocated with ACL 2013*.

Poria, Soujanya, Erik Cambria, Lun-Wei Ku, Chen Gui, and Alexander Gelbukh, 2014. A rule-based approach to aspect extraction from product reviews. In *Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP)*.

Qiu, Guang, Bing Liu, Jiajun Bu, and Chun Chen, 2011. Opinion word expansion and target extraction through double propagation. *Computational Linguistics*, 37(1):9–27.

Wawer, Aleksander, 2012. Mining co-occurrence matrices for so-pmi paradigm word candidates. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Avignon, France: Association for Computational Linguistics.

Wawer, Aleksander, 2015. Towards domain-independent opinion target extraction. In *2015 IEEE 15th International Conference on Data Mining Workshops (SENTIRE 2015). To appear.*. IEEE Computer Society.

Wawer, Aleksander and Konrad Gołuchowski, 2012. Expanding opinion attribute lexicons. In Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala (eds.), *Text, Speech and Dialogue: 15th International Conference, TSD 2012, Brno, Czech Republic*, volume 7499 of *Lecture Notes in Artificial Intelligence*. Heidelberg: Springer-Verlag, pages 72–80.

Wawer, Aleksander and Dominika Rogozińska, 2012. How Much Supervision? Corpus-based Lexeme Sentiment Estimation. In *Data Mining Workshops, 2012 IEEE 12th International Conference on. SENTIRE 2012.*, ICDMW. Los Alamitos, CA, USA: IEEE Computer Society.

Wroblewska, Alina, 2014. *Polish Dependency Parser Trained on an Automatically Induced Dependency Bank*. Ph.D. dissertation, Institute of Computer Science, Polish Academy of Sciences, Warsaw.