Towards Domain-Independent Opinion Target Extraction

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

***Abstract*—In this paper, we investigate the problem of domain-independent opinion target extraction. The only lexical resource used is domain-independent (general) sentiment dictionary. We begin from investigating syntactic descriptions (rules) using dependency parsing jointly with sentiment dictionary. We conclude that such a solution is not sufficient for opinion target extraction due to low precision. To overcome this difficulty, we propose a well-known supervised machine learning method as the second step, after applying syntactic rules. We find that supervised model without lexical features outperforms by large margin a comparable one with lexical features. The results appear promising and contribute to domain-independent opinion target extraction. All experiments were carried out on a publicly available Polish dependency treebank with manually verified opinion and sentiment annotations, as well as opinion target information.**

***Keywords*-opinion target extraction, domain-independent, aspect based sentiment analysis, Polish**

I. Introduction

Opinion target extraction, as defined in this paper, is the task of recognition of words towards which an opinion (sentiment) is expressed. Typically, in the domain of product reviews, they are aspect terms related to the reviewed entity or entity itself. However, they could denote any other object. Our understanding is similar as in [1].

Sometimes the problem of opinion target extraction is formulated in a different manner. For instance, organizers of SemEval's ABSA 2015 competition (Aspect Based Sentiment Analysis) [2], narrow Opinion Target Expression (OTE) to the problem of entity extraction. Opinion target extraction, in the sense of [1] and as used in our paper, appears more similar to the task of attribute and entity recognition (in ABSA 2015 called Slot 1).

Opinion target words are domain-dependent. User generated opinions, such as reviews, are expressed about properties of specific types of products. When reviewing perfumes, people opine various aspects of smell, its durability, bottles, and so on. When reviewing phones, they opine aspects such as battery time, screen, application performance. All these constitute opinion targets, and as the provided examples prove, these lists vary between domains. Similarly in social media, targets of opinions depend on current topic under discussion. Generally, diversity of opinion targets and their domain-dependence appears to be significant. This paper is focused on methods of recognizing opinion targets that are domain-independent and therefore based on syntactic information.

The goal is to investigate the limits of rule-based and purely syntactic methods of opinion target extraction. Then, we compare it with machine learning approach to this problem and finally, propose a hybrid solution of rule-based and machine learning.

The paper is organized as follows. Section II provides a short overview of existing work on opinion target extraction. In Section III we describe the data set (treebank with semi-manual opinion and opinion target labeling) : its origins, structure, annotation procedure and basic statistics. Section IV discusses the design, creation and performance of rule-based approach to opinion target extraction, when using dependency rule descriptions with POS and dependency labels. It ends with a discussion of weaknesses of this method. Then, Section V introduces supervised machine learning solution to the problem and the description of related experiments. Finally, Section VI concludes the paper and discusses the ideas for future work.

II. Existing Work

The first pioneering work on aspect-based sentiment analysis dates back to [3], where a corpus of texts (eg. reviews from one domain) is used to extract opinion aspects. In more recent work, [1] propose an iterative algorihm that uses dependency parse information and seed lexicons, to discover sentiment and opinion target vocabulary in iterative fashion. The problem with this type of corpus-based approaches is that their purpose is creation of opinion targets dictionary, and not tagging of texts with opinion target instances. The tasks are somehow different. Corpus-based lexicon creation requires collections of already gathered texts belonging to a single domain (eg. devoted to one product type) and provide domain-specific dictionaries as a result. Tagging requires context-sensitive analysis, as some words are opinion targets only in specific occurrences. The solution proposed in this paper aims the tagging problem, does not require large corpora and is meant to extract opinion targets in domain-independent manner.

Also [4] extract opinion targets (with aspect extraction) using syntactic patterns. Their work is based on patterns

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

978-1-4673-8493-3/15 $31.00 © 2015 IEEE
DOI 10.1109/ICDMW.2015.255

described by [1]. The extraction is not limited to sentence boundaries as it includes a heuristic for anaphora resolution to identify targets across sentences. The usage of anaphora has been also reported in [5], where it significantly improves the opinion target extraction. Dependency patterns are used in the context of aspect-oriented sentiment analysis not only for opinion target extraction, but also for computing sentiment values of sentences and phrases, as for example in [6].

A rule-based approach to aspect extraction that exploits common-sense knowledge and sentence dependency trees to detect both explicit and implicit aspects, was described in [7]. 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, including propagation method and the baseline described in [3]. Our work can be viewed as an extension of this method in multiple ways: by automatically inferring dependency rules rather than handcrafting and applying CRF algorithm as the second step, to increase overall precision of extracted targets.

In SemEval 2014 competition on Aspect-Based Sentiment Analysis (ABSA 2014 [8]), opinion target extraction in the sense of our paper overlaps partially with subtask SB1 (aspect term extraction), since as the organizers put it, it is to identify all aspect terms present in each sentence. However, the notable difference is the fact that also aspect terms for which no sentiment is expressed (neutral polarity) are to be extracted. Results reveal that the best performing systems are all based on Conditional Random Fields (CRF) algorithm, which became a default solution for this type of tagging.

III. Dataset

This section describes the annotation procedure and data set used in the experiments. The starting point for our work was the corpus and opinion target lexicon used in [9]. 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 [9]. We parsed sentences using the MaltEval dependency parser and model for the Polish language (briefly described in [10] and [11]).

The basic statistics in terms of number of texts, sentences with opinion target words (T)[1] and sentiment words (S) are presented in Table I.

	perfume	clothes
T words	311	222
sentences	946	418

Table I

.

For each dependency tree with automatically labeled candidates for opinion words (S) and candidates for their targets (T), taken from sentiment lexicon and T-word lexicon respectively, linguists annotated:

1) Correct or erroneous dependency structure between S and T
2) Whether T-word is an opinion target in the context of specific sentence
3) Whether S-word is an opinion (has sentiment) in the context of specific sentence
4) If conditions 1-3) are positively met:
   - S is related to T (in other words, S describes or modifies T)

We verified annotation quality by double annotation of small, randomly selected subset of the treebank. Results presented in Table II demonstrate high levels of agreement, with relatively the highest value for structure correctness and relatively the lowest for relation between sentiments (S) and targets (T). The analysis of reasons of behind difficulties in annotating relations between S and T reveals a number of borderline cases, where the relation is weak or indirect. For example, in: "I like(S) this perfume(T)'s bottle", the relation between perfume (target) and like (sentiment) is indirect, and it is disputable whether exists or not. Obviously, this could be alleviated by introducing phrase-level annotation for opinion targets with marked phrase heads, as well as by multiple relation types, stronger and weaker.

	Total	Agreed	% agreement
structure correctness	82	75	91%
correctness of T	75	64	85%
correctness of S	75	70	93%
S related to T	54	42	77%

Table II
Inter-annotator agreement for each annotation sub-task.

IV. Results

This section describes the procedure of creating dependency patterns (paths) to link opinion targets with sentiments and the results of associated experiments. The overall idea here is that a sentiment dictionary (could be replaced by any phrase-level sentiment recognition method) combined with information how to traverse dependency tree produced by a parser (a set of patterns), is sufficient to recognize opinion targets. In this scenario, one could start from sentiment word S, and by following a sequence of moves on dependency

[1] The presented number represents only the size of dictionary used as a starting point. It has been extended by pattern application as described further.

path	precision	matched	not matched	total
[pos:adj] <adjunct [pos:subst]	0.886	396	51	447
[pos:fin] >comp [pos:prep] >comp [pos:subst]	0.814	48	11	59
[pos:adj] >adjunct [pos:prep] >comp [pos:subst]	0.906	48	5	53
[pos:adj] <adjunct [pos:subst] >adjunct [pos:prep] >comp [pos:subst]	0.333	16	32	48
[pos:adj] <pd [pos:fin] >subj [pos:subst]	0.909	40	4	44
[pos:adj] <adjunct [pos:subst] <conjunct [pos:conj] >conjunct [pos:subst]	0.333	11	22	33
[pos:adj] <conjunct [pos:interp] <adjunct [pos:subst]	0.939	31	2	33
[pos:fin] >adjunct [pos:prep] >comp [pos:subst]	0.433	13	17	30
[pos:adj] <adjunct [pos:subst] >adjunct [pos:subst]	0.64	16	9	25
[pos:adj] <conjunct [pos:conj] >conjunct [pos:subst]	0.625	15	9	24
[pos:fin] <conjunct [pos:conj] >conjunct [pos:fin] >subj [pos:subst]	0.304	7	16	23

Table III
MOST FREQUENT DEPENDENCY PATTERNS: PRECISION, CORRECT, INCORRECT AND TOTAL MATCHES.

tree, described according to some formal system, "arrive" at an opinion target word T. Two syntactic structures of this kind are described in [1] and used for double propagation of sentiments and opinion targets in a corpus. Their structure is rather simple and consists only of upward and downward traversal. No dependency labels or POS tags are taken into account, and neither recall nor precision are evaluated.

One could devise mutiple formal systems to describe such dependency patterns, or rules. We decided to use (implement) our own, closely resembling German TIGERSearch [12] formalism, developed for searching the TIGER treebank, and SemGrex, a modification of Tregex pattern language [13] aimed at dependency structures. Both allow addressing attributes of tokens, even multiple attributes at once (for example, POS and lemma) and expressing the direction of dependencies. In the pattern matching system used in our paper tokens are expressed as enclosed in [..] and dependency relations as < or >, depending on the direction. For example, we may specify that encountered tokens belong to specified POS type (eg. [pos:verb] to specify verbs). We may also specify dependency label type.

For inducing dependency rules, we filtered out sentences with incorrect sentiment and structure errors. For every known S-T pair, we generated dependency path descriptions by starting off from the opinionated word S and traversed dependency tree using the shortest possible path to the opinion target T. We used two types of information: POS and dependency labels. This step generated 173 dependency patterns.

The top frequent 11 patterns (each over 20 occurrences) are reported in Table III, along with pattern precision, numbers of correct, incorrect and total T-S pairs extracted by the pattern.

The patterns should be read from left to right, with the leftmost token indicating sentiment word and the rightmost token opinion target word. The most frequent pattern can be interpreted as an adjective ([pos:adj]) of some sentiment, governed by (using adjunct relation) by a noun ([pos:subst]), an opinion target.

In the second step, we applied the 173 extracted patterns (dependency path descriptions) to the same set of sentences, using all sentiment words as starting points. This was done in order to ensure that all possible opinion target (T) words, even those not present in the dictionary, are captured in our treebank as S-T pairs. This step resulted in additional 668 S-T pairs (and sentences) that were subsequently annotated. Finally, the data set consists of 1737 annotated S-T pairs (sentence-level descriptions).

Analysis of deducted dependency patterns reveals a long tail of descriptions that matched only on one sentence. We performed manual analysis of a sample of involved cases and discovered the following causes:

- Difficult sentences, semantically and syntactically. It is not straightforward even for a human annotator, how should the correct dependency tree be like between A and S.
- Not marked previously parser errors, due to spelling and grammatical mistakes, and possibly due to overall complexity.

The performance of the rule set can be estimated as 0.73 precision assuming the most frequent class (rule matched, rather than not matched) as a baseline.

Consequently, we believe that dependency patterns are useful for the task of opinion target extraction to much larger extent than reported in [1], where only two patterns are applied, but in any case they may not be considered as sufficient for this task. Some additional step is required for two reasons. First, in order to discover opinion targets that are not pointed to by any dependency rule, purely by means of statistical inference based on contextual features of potential opinion targets. Second, for those potential opinion targets that are pointed to an extraction rule, increase precision. The solution we propose to solve these issues is adding another step after rule-based extraction, namely a conditional random fields tagger. It has been described in Section V.

The annotated data set (opinion-target treebank) and all induced dependency patterns with their computed precision, as well as the Python application to extract patterns from sentences in CONLL format, may be downloaded from http:

//zil.ipipan.waw.pl/OPTA.

V. Machine Learning Opinion Target Extraction

For machine learning, we selected a well-known and proven sequence-labeling algorithm, Conditional Random Fields (CRF) [14]. This type of algorithms are often used for structured prediction. Whereas an typical classifier predicts a label for a single word (token) without regard to neighbouring words, a CRF takes context into account. The linear chain CRF popular in natural language processing predicts sequences of labels for sequences of input words.

In our experiment, we used CRFsuite tool with lbfgs algorithm [15]. The objective of CRF was to extract all (labels of) targets of opinions from the dataset, using several groups of features including syntactic, lexical, and sentiment lexicon features, grouped into templates.

Specifically, we test multiple feature feature templates, from T1 to T6, each consisting of several feature sets. For brevity, we provide descriptions of feature sets only once below, and subsequently use their corresponding IDs in square brackets. In parentheses, we denote positional information, zero referring to current token. For example, (-1,0,1) corresponds to window of one tokens left from current token, current token, and one token after it. The feature sets used as building blocks for templates are follows:

- [lemma]: lemma unigrams at (-1,0,1), lemma bigrams at (-1,0), (0,1);
- [POS]: unigrams, bigrams and trigrams of POS from (-2,-1,0,1,2);
- [dep]: unigrams and bigrams of dependency relation labels at (-2,-1,0,1,2);
- [ruleAny]: binary information whether current word (0) is being pointed to by any rule from the rule set described in Section IV;
- [ruleID]: ID of the specific rule that points to current word (0), if any;
- [S]: binary information whether current word (0) has been found in sentiment dictionary.

The results for each feature template are reported as mean values from 10-fold stratified cross-validation. In stratified k-fold cross-validation, the folds are selected so that the mean response value is approximately equal in all the folds. In our case (of a dichotomous classification) this means that each fold contains roughly the same percentage of the two types of class labels: correct (valid) and incorrect (invalid) pairs of opinion target and sentiment words. Table IV presents the results obtained with the CRF, reported using following measures:

- opinion target extraction precision (tPrec), recall (tRec) and F1 (tF1),
- macro-average precision (mPrec), recall (mRec) and F1 (mF1).

We also report average number of features (across the folds). Macro-average is the harmonic mean of scores for each class of tokens, opinion target tokens (words) as one class and all other tokens as another.

The results indicate surprisingly weak performance of lexical feature space T1. CRF models based on lexical information perform poorly in terms of precision, but also recall, even despite over 17 thousand features. Extending this feature space with syntactic features, as in T2, brings notable improvement in recall. A purely syntactic feature space of T3 is comparable to both previous ones, which may also surprise. The most substantial influence on results is caused by introducing rule features, as in T4, T5 and T6. It raises not only recall (which is expected), but also precision. Interestingly, a significant increase in precision is obtained by rule ID feature, as in T5 and T6, which indicates that the CRF model captures rule specific, discriminative information. It appears that sentiment dictionary feature S is not really an important one, as it trades off small increase in precision for decreased recall. The best performing feature templates (T4, T5 and T6) are also the ones with moderate number of features, not exceeding 8 thousands.

VI. Conclusions and Future Work

In this paper, we focused on the problem of opinion target extraction with as little domain-dependent information as possible. For that purpose we investigated using non-lexical methods that base on syntactic information. The only lexical resource used was domain-independent (general) sentiment dictionary.

We started from the method of opinion target extraction based on applying syntactic descriptions (rules) jointly with sentiment dictionary. We concluded that this solution could be further developed into more patterns than just two described in [1]. Our paper proposes a method of extraction of these patterns from corpus and their application using a formalism similar to those used in treebank pattern matching and search engine solutions, such as SemGrex or Tiger Search. However, the number of practically usable patterns is limited to no more than few dozens and there is a number of cases (tuples made of opinions and their targets) that fall beyond the reach of any pattern due to high complexity of a sentence (syntactic and semantic) and parse errors. In addition, the rule-based extraction suffers from rather low precision. Due to all these problems, we proposed a supervised machine learning method based on CRF algorithm as the second step after pattern-based extraction, using syntactic rules as input features for the CRF sequence classifier. While the application of CRF in this task has been already proven as state-of-the-art, as for example in SemEval competition, the contribution of this paper is the demonstration that without lexical features, and therefore in a more domain-independent fashion, using the patterns as an input feature, the CRF method turns out to perform very well. What is especially promising and interesting, it outperforms models with lexical features.

template	description	tPrec	tRec	tF1	mPrec	mRec	mF1	features
T1	[lemma]	0.586	0.33	0.421	0.768	0.656	0.693	17435
T2	[lemma]+[POS]+[dep]	0.553	0.466	0.505	0.756	0.719	0.735	25234
T3	[POS]+[dep]	0.548	0.426	0.478	0.752	0.699	0.721	7805
T4	[POS]+[dep]+[ruleAny]	0.783	0.891	0.833	0.887	0.936	0.91	7808
T5	[POS]+[dep]+[ruleAny]+[ruleID]	0.823	0.901	0.859	0.908	0.943	0.924	8048
T6	[POS]+[dep]+[ruleAny]+[ruleID]+[S]	0.829	0.889	0.857	0.91	0.937	0.923	8067

Table IV

CRF OPINION TARGET EXTRACTION RESULTS: AVERAGE VALUES IN 10-FOLD CROSS VALIDATION.

Generally, the reported CRF results, especially the models that use rule-based features, leave little room for further improvements due to their high overall performance. However, one can speculate that further increases in precision could be obtained by more careful corpus annotation: possibly more coherent handling of borderline cases thanks to extended annotator guidelines, perhaps annotation by multiple independent linguists. An improvement could be obtained also if more even efforts were put into feature engineering. These could include for instance experiments with word embeddings (for example as in [16]), used as features for supervised learning.

Potential future work also includes the possibility of using another structural prediction method instead of the CRF algorithm. These methods could include structural Support Vector Machines algorithm, described in [17] and [18]. Unlike regular SVMs, which consider only univariate predictions like in typical classification, structural SVM can predict complex objects like sequences.

A more new alternative is deep learning, a more and more popular type of algorithms typicall based on neural networks. It has already been reported by some recent research as a promising alternative for CRF in tasks related to opinion mining [19].

REFERENCES

[1] G. Qiu, B. Liu, J. Bu, and C. Chen, "Opinion word expansion and target extraction through double propagation," *Computational Linguistics*, vol. 37, no. 1, pp. 9–27, Mar. 2011.

[2] M. Pontiki, D. Galanis, H. Papageogiou, S. Manandhar, and I. Androutsopoulos, "Semeval-2015 task 12: Aspect based sentiment analysis," in *Proceedings of the 9th International Workshop on Semantic Evaluation*, ser. SemEval 2015, 2015.

[3] M. Hu and B. Liu, "Mining opinion features in customer reviews," in *Proceedings of the 19th national conference on Artifical intelligence*, ser. AAAI'04. AAAI Press, 2004, pp. 755–760.

[4] S. Gindl, A. Weichselbraun, and A. Scharl, "Rule-based opinion target and aspect extraction to acquire affective knowledge," in *Proceedings of WWW'13 workshop on Multidisciplinary Approaches to Big Social Data Analysis*, 2013.

[5] N. Jakob and I. Gurevych, "Extracting opinion targets in a single- and cross-domain setting with conditional random fields," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 2010.

[6] S. Poria, E. Cambria, G. Winterstein, and G.-B. Huang, "Sentic patterns: Dependency-based rules for concept-level sentiment analysis," *Knowledge-Based Systems*, vol. 69, pp. 45–63, 2014.

[7] S. Poria, E. Cambria, L.-W. Ku, C. Gui, and A. Gelbukh, "A rule-based approach to aspect extraction from product reviews," in *Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP)*, August 2014.

[8] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, "Semeval-2014 task 4: Aspect based sentiment analysis," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Dublin, Ireland: Association for Computational Linguistics and Dublin City University, August 2014, pp. 27–35. [Online]. Available: http://www.aclweb.org/anthology/S14-2004

[9] A. Wawer and K. Goluchowski, "Expanding opinion attribute lexicons," in *Text, Speech and Dialogue: 15th International Conference, TSD 2012, Brno, Czech Republic*, ser. Lecture Notes in Artificial Intelligence, P. Sojka, A. Horak, I. Kopecek, and K. Pala, Eds. Heidelberg: Springer-Verlag, 2012, vol. 7499, pp. 72–80.

[10] A. Wróblewska, "Polish dependency bank," *Linguistic Issues in Language Technology*, vol. 7, no. 1, 2012. [Online]. Available: http://elanguage.net/journals/index.php/lilt/article/view/2684

[11] A. Wróblewska and M. Woliński, "Preliminary experiments in Polish dependency parsing," ser. Lecture Notes in Computer Science, P. Bouvry, M. A. Kłopotek, F. Leprevost, M. Marciniak, A. Mykowiecka, and H. Rybiński, Eds., vol. 7053. Springer-Verlag, 2011, pp. 279–292.

[12] W. Lezius, "Ein suchwerkzeug für syntaktisch annotierte textkorpora," Ph.D. dissertation, IMS, University of Stuttgart, December 2002, Arbeitspapiere des Instituts fĂĽr Maschinelle Sprachverarbeitung (AIMS), volume 8, number 4.

[13] N. Chambers, D. Cer, T. Grenager, D. Hall, C. Kiddon, B. MacCartney, M.-C. de Marneffe, D. R. E. Yeh, and C. D. Manning., "Learning alignments and leveraging natural logic," in *Proceedings of the Workshop on Textual Entailment and Paraphrasing*, 2007, pp. 165–170.

[14] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 282–289.

[15] N. Okazaki, "Crfsuite: a fast implementation of conditional random fields (crfs)," 2007. [Online]. Available: http://www.chokkan.org/software/crfsuite/

[16] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *Proceedings of Workshop at ICLR*, 2013.

[17] T. Joachims, "Learning to align sequences: A maximum-margin approach," August 2003, online manuscript.

[18] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun, "Support vector machine learning for interdependent and structured output spaces," in *International Conference on Machine Learning (ICML)*, 2004, pp. 104–112.

[19] O. Irsoy and C. Cardie, "Opinion mining with deep recurrent neural networks," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, October 2014, pp. 720–728.