Detection of Valid Sentiment-Target Pairs in Online Product Reviews and News Media Articles

Svitlana Vakulenko
Department of New Media Technology
MODUL University Vienna
Vienna, Austria
svitlana.vakulenko@modul.ac.at

Albert Weichselbraun
Swiss Institute for Information Research
University of Applied Sciences Chur
Chur, Switzerland
albert.weichselbraun@htwchur.ch

Arno Scharl
Department of New Media Technology
MODUL University Vienna
Vienna, Austria
arno.scharl@modul.ac.at

***Abstract*—This paper investigates the linking of sentiments to their respective targets, a sub-task of fine-grained sentiment analysis. Many different features have been proposed for this task, but often without a formal evaluation. We employ a recursive feature elimination approach to identify features that optimize predictive performance. Our experimental evaluation draws upon two corpora of product reviews and news articles annotated with sentiments and their targets. We introduce competitive baselines, outline the performance of the proposed approach, and report the most useful features for sentiment target linking. The results help to better understand how sentiment-target relations are expressed in the syntactic structure of natural language, and how this information can be used to build systems for fine-grained sentiment analysis.**

***Keywords*-opinion target; fine-grained; sentiment analysis;**

I. Introduction

Sentiment analysis measures positive or negative sentiment expressed towards a specific product, topic, person or organization (the *target*). Document-level sentiment analysis, for example, is traditionally employed for review mining [1]. Customer reviews are a special document genre, providing customer opinions on a single product, which is the target of the review. However, there can be also several product features mentioned in the review, which the customer may assess differently (e.g. "The design is *outstanding*, but the sound quality is *poor*.") Furthermore, in documents of other genres such as news articles or social media posts, there are often several opinion targets mentioned within the same document or even within the same sentence.

Fine-grained sentiment analysis [2] addresses this problem by distinguishing opinions expressed towards different entities (e.g. organization, people, products or product features). It subsumes three sub-tasks (sentiment extraction, target extraction and sentiment-target linking) that can be pursued in several ways:

- Given a set of targets, extract sentiments expressed towards each of the targets (target-driven analysis).
- Extract all sentiments; for each sentiment, extract the corresponding targets (sentiment-driven analysis).
- Extract sentiments and targets independently, generate candidate sentiment-target pairs, and assess the relation between each pair (combinatorial analysis).
- Simultaneously label both sentiments and targets (joint analysis).

The approach presented in this paper follows the third option. We aim to develop a system that uses keyword extraction, named entity recognition and linking to identify relevant targets and draws upon a comprehensive sentiment lexicon to extract sentiments.

Therefore, this work focuses exclusively on the sentiment-target linking task by making use of the publicly available corpora that was already annotated with the correct sentiments and their targets. Thereby, we avoid errors originating from the incorrect sentiment and target extraction and evaluate the sentiment-target linking task independent from other sub-tasks.

We tackle the sentiment-target linking task as a binary classification problem by evaluating the classification function on each candidate sentiment-target pair that we found within the same sentence boundaries. The introduces approach proposes a set of syntactic features parsed from the input text to discriminate between valid and invalid sentiment-target pairs. Evaluating the feature designs in a set of comprehensive experiments derives the features most useful for target-sentiment classification. Finally, we train several classification models using these features and evaluate their performance. The main contributions of this work can be summarized as follows:

- Our experimental evaluation demonstrates that a simple distance-based approach performs very well in presence of gold-standard sentiment and target annotations.
- Feature engineering and feature selection are important for the classifier-based approach to achieve comparable performance.
- Our results reveal useful syntactic features for the sentiment-target linking task, and suggest new features that were not considered previously.

In the following sections we provide a short overview of the state-of-the-art (Section II), introduce the formal definition

978-1-5090-4470-2/16 $31.00 © 2016 IEEE
DOI 10.1109/WI.2016.24

IEEE computer society

of the sentiment-target linking task (Section III), describe our approach (Section IV) and its evaluation (Section V), summarize our results (Section VI) and the limitations of our approach providing directions for future work (Section VII).

II. Related Work

Rule-based sentiment-target linking uses manually designed heuristics to find valid sentiment-target pairs, e.g. *distance-based* approaches such as sentiment-target proximity [3]. *Syntax-based* approaches tha rely on a handful of patterns are most popular [4]–[7], e.g. the dependency path between a sentiment and its target. Our goal is to infer such patterns automatically based on statistic evidence from the available corpora annotated with target-sentiment pairs.

Zhuang [8] and Xu [9] extract dependency patterns between sentiments and their targets using part-of-speech (POS) and dependency relation labels, and determine the frequency of these patterns in the corpus. This approach is very intuitive, but requires a critical mass of patterns and is not able to account for the interplay between several features that may have an influence on the outcome in combination.

Sentiment-target linking is frequently modeled as a classification task - see Section III for the formal definition [2], [10], [11]. The quality of results depends on the features used for the classification, but to the best of our knowledge previous work has not yet systematically evaluated combinations of syntactic features for the task of sentiment-target linking.

As a result researchers employ different feature sets as a part of their pipeline or joint model approaches [2], [11]–[13], which introduce errors at sentiment/target extraction phases. Therefore, we argue for the need of a solid evaluation for the sentiment-target linking task in isolation to identify the set of features that prove helpful in distinguish valid sentiment-target pairs.

The closest work to ours is of Kessler et al. [10], who introduced the J.D. Power and Associates (JDPA) Sentiment corpus, which contains product reviews annotated with sentiment-target pairs [14], and used it to build a classifier for sentiment-target linking. Kessler et al. propose a set of features and run a single evaluation on the full feature set without assessing the performance of the individual features. Hence, it is not clear which features were useful for detecting sentiment-target relations. We expand the feature set proposed by Kessler et al. [10] with features frequently employed for sentiment-target linking [2], [11], [15] and systematically evaluate the performance of different feature subsets using *recursive feature elimination* (RFE).

III. Task Definition

Sentiment-target linking seen as a binary classification problem can be summarized as follows: given a set of sentiment tokens $S_m = \{t_{s_i}\}$ and a set of target tokens $T_m = \{t_{t_j}\}$ extracted from a sentence m, return a set of valid sentiment-target pairs: $\{(t_{s_i}, t_{t_j})\}$, where $y(t_{s_i}, t_{t_j})$ = True. This task can be represented as a bipartite graph that consists of a set of sentiment tokens S_m and a set of target tokens T_m (see Figure 1). The goal is to find the correct matching (the edges between the two sets S_m and T_m), which indicates the valid sentiment-target pairs i.e. $\{(t_{s_i}, t_{t_j})\}$, where $y(t_{s_i}, t_{t_j})$ = True. The classification function y reflects whether sentiment t_{s_i} and target t_{t_j} tokens constitute a valid sentiment-target pair (Valid: True|False).

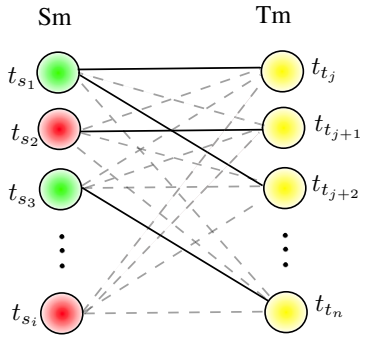

Figure 1. The sentiment-target linking task represented as a bipartite graph that consists of a set of sentiment tokens S_m and a set of target tokens T_m extracted from a sentence m. The dashed edges show the maximum matching, i.e. all candidate sentiment-target pairs, which is the input to the classifier. Bold edges connect valid sentiment-target pairs corresponding to the correct matching, i.e. the desired output from the classifier. Sentiments in Sm may have different polarity: positive (green) or negative (red).

IV. Method

To train the classifier for the task of sentiment-target linking we collect observations from a corpus annotated with words and phrases expressing sentiments $\{t_{s_i}\}$, targets $\{t_{t_j}\}$ and relations between them $\{(t_{s_k}, t_{t_l})\}$. We mark valid sentiment-target pairs based on the corpus annotations A: $y(\{(t_{s_k}, t_{t_l})\})$ = True, where $\{(t_{s_k}, t_{t_l})\} \subseteq A_m$. The rest of the pairs are considered to be invalid: $y(\{(t_{s_o}, t_{t_p})\})$ = False, where $\{(t_{s_o}, t_{t_p})\} \not\subseteq A_m$.

An observation $x(t_{s_i}, t_{t_j})$ is a set of features to capture syntactic relations between the sentiment token t_{s_i} and the target token t_{t_j} (see Section IV-A). To extract features efficiently, we construct an opinion graph G_m for every sentence $m \in M$ (see Figure 2). The nodes of the opinion graph correspond to the tokens extracted from the sentence m and annotated with their POS tags. If the annotation spans several tokens the respective n-grams form the nodes of the opinion graph. The edges are produced by the dependency parser and labeled with dependency relations between the adjacent nodes. Formally, an opinion graph encodes a sentence $m = t_1/p_1, ..., t_n/p_n$ with the tokens t_i labeled with the POS tags p_i in a directed graph $G_m = (V, E)$ with vertices $V = \{1, ..., n\}$ and labeled edges $E \subseteq V \times V$. Every edge (i, j, l_{ij}) represents a directed dependency between the head token t_i and the dependent token t_j labeled l_{ij}.

Table I
DESCRIPTION OF OUR FEATURE SET BASED ON THE EXAMPLE SENTENCE, "I *like* TO DRIVE THE **CAR**," WHERE 'LIKE' IS A SENTIMENT TOKEN t_s AND 'CAR' IS A TARGET TOKEN t_t (SEE FIGURE 2). THE BEST PERFORMING FEATURES ARE HIGHLIGHTED IN GREY - SEE SECTION VI.

#	Group	Label	Description	Example
1	Lexical Path	L_Dist	Number of tokens on the lexical path	3
2	Lexical Path	L_Ngram	The tokens between t_{s_i} and t_{t_j}	to drive the
3	Lexical Path	L_Stems	Stems of the tokens between t_{s_i} and t_{t_j}	to drive the
4	Lexical Path	L_Penn	POS-tags on the lexical path	['TO','VB','DT']
5	Dependency Path	D_Dir	Dependency relations with directions	[('OPRD', True), ('IM', True), ('OBJ', True)]
6	Dependency Path	D_Sentiment	Number of other sentiments on the dependency path	0
7	Dependency Path	D_Target	Number of other targets on the dependency path	0
8	Target	T_Type	Semantic type of target (for JDPA)	Vehicle
9	Sentiment/Target	ST_Penn	POS-tags of t_{s_i} and t_{t_j}	['VBP','NN']
10	Dependency Path	D_StemDir	Stem of t_{s_i} concatenated to directed dependencies	like [('OPRD', True), ('IM', True), ('OBJ', True)]
11	Sentiment/Target	ST_Pre	t_{s_i} precedes t_{t_j} in the sentence	True
12	Sentiment	S_Penn	POS-tag of t_{s_i}	['VBP']
13	Target	T_Penn	POS-tag of t_{t_j}	['NN']
14	Dependency Path	D_Penn	POS-tags on the dependency path	['TO','VB']
15	Sentiment	S_POS	POS-tag group of t_{s_i}	['verb']
16	Target	T_POS	POS-tag group of t_{t_j}	['noun']
17	Lexical Path	L_POS	POS-tag groups on the lexical path	['to','verb','det']
18	Dependency Path	D_POS	POS-tag groups on the dependency path	['to','verb']
19	Lexical Path	L_Sentiment	Number of other sentiments on the lexical path	False
20	Lexical Path	L_Target	Number of other targets on the lexical path	False
21	Lexical Path	L_Sentiment	Other sentiments on the lexical path	False
22	Lexical Path	L_Target	Other targets on the lexical path	False
23	Dependency Path	D_Sentiment	Other sentiments on the dependency path	False
24	Dependency Path	D_Target	Other targets on the dependency path	False
25	Dependency Path	D_Dist	Length of the dependency path	3
26	Dependency Path	D_Rels	Dependency relations	['OPRD','IM','OBJ']

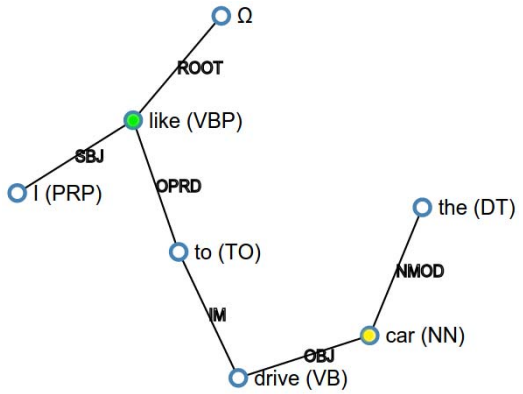

Figure 2. Opinion graph for a sample sentence "I *like* to drive the **car**," where 'like' is a *sentiment token* t_s and 'car' is a **target token** t_t.

A. Features

We construct a set of features (see Table I) to be extracted for each observation of a sentiment-target pair $x(t_{s_i}, t_{t_j})$. The redundancy of this feature set helps evaluate different configurations, and to determine the cost-benefit ratio of extracting complex, computationally expensive features (e.g., dependency relations). We aim to reduce the initial broad set of features to an essential subset that maximizes classification performance and minimizes the computation time for each candidate sentiment-target pair.

As a 'warm-start', we initialize our feature set with 10 features proposed by Kessler et.al for JDPA corpus [10] (see Table I: 1-10) and further extend it with other popular syntactic features (see Table I: 11-26). We focused on the syntactic properties of the sentiment-target relation and avoided features containing semantic information, such as n-grams, lemmas, stems or synonym sets. Features are classified into the following groups:

- **Sentiment/Target** (S/T): features that reflect properties of the sentiment t_{s_i} and/or target t_{t_j} tokens, e.g. their POS-tags (p_{s_i} and p_{t_j}) or POS-tag groups[1].
- **Lexical Path** (L): features of the tokens that occur between t_{s_i} and t_{t_j} in the sentence, e.g. number of words, other sentiment or target tokens on the path.
- **Dependency Path** (D): features of the shortest path from t_{s_i} to t_{t_j} in the opinion graph, e.g. labels and directions of the corresponding dependency edges. If a dependency leads from sentiment to target (t_{s_i} to t_{t_j}), its direction is set to *True*, otherwise to *False*.

V. EXPERIMENTAL EVALUATION

The evaluation uses two public corpora as a gold standard to train and evaluate our classification model and the

[1] 'POS-tag groups' features are constructed from the corresponding 'POS-tags' features using a custom mapping from the Penn Treebank POS-tag set to the higher-level groups, such as 'noun': {'NN', 'NNS', 'NNP' , 'NNPS'}, 'verb': {'VB', 'VBD', ... }

Table II
STATISTICS OF THE DATASETS USED IN THE EXPERIMENTS.

Dataset	Documents	Sentences	Targets	Sentiments	Observations (Valid / Invalid)
JDPA	637	21,799	16,218	17,496	36,712 (17,853 / 18,859)
MPQA	685	12,966	4,482	4,988	7,942 (4,198 / 3,744)

baseline approaches. Both corpora contain English-language texts manually annotated with sentiment expressions and their targets (see more details in Table II):

- **JDPA** (J.D. Power and Associates) Sentiment Corpus [2] containing blog posts with customer reviews of digital cameras and car models [14], [16].
- **MPQA** Opinion Corpus Version 2.0 containing news articles and other text documents manually annotated for opinions and sentiments [17], [18].

These corpora represent not only different genres: customer reviews versus news articles, but also different sentiment and target annotation styles. JDPA corpus contains simple granular annotations with the average of a single word per target and sentiment annotation e.g. "car", "optical lens", "great", "new", etc. MPQA annotations are span-based capturing the entire phrases with 6 words per target and 8 words per sentiment annotation, such as "the southern African country", "it is absolutely inadmissible for", etc. Such differences in the sentiment-target annotation approaches are also likely to result in different patterns extracted from these corpora.

MPQA annotations often overlap, e.g the sentence: "We report on the recent events," may contain two overlapping annotations: "events" and "the recent events". In this evaluation we consider only one of the annotations provided (the last one) to ensure the unique token assignment and produce one opinion graph per sentence.

We parsed the sentences using the Stanford POS-tagger [19] and a syntactic parser to extract dependency relations [20] and produced observations for every annotated sentiment-target pair. Each observation contains the full set of 26 features that we proposed in Section IV-A. We reproduced the features proposed in Kessler et al. [10] using our POS- and dependency parsers and extract them also from MPQA dataset (apart from the target type feature: T_Type, which is specific to JDPA annotations).

The evaluation used the following approaches to establish a baseline for the sentiment-target linking task:

- **Sentence-based** baseline corresponds to sentence-level sentiment analysis approaches – all sentiments correspond to all targets within the same sentence. Hence, all observations we record within the same sentence evaluate to True.
- **Distance-based** approaches: *Closest Target* – each sentiment link to the closest target within the same sentence; *Closest Sentiment* – each target link to the closest sentiment within the same sentence. When there is only one target/sentiment in the sentence, the result is identical to the sentence-based approach. If there are several candidates with the same closest distance, the first one is selected. For example, in the sentence "*Roses* are *Red* and *Violets* are Blue", while both *Roses* and **Violets** have the same word distance to *Red*, only *Roses* will be selected as the valid target.
- **Kessler 10** is a classifier-based approach using 10 features proposed by Kessler et al. [10] (see Table I).

We used logistic regression classifier from *scikit-learn* library [21] with the default parameter settings that also provides regression coefficients for each of the input features, which guided our feature selection procedure and helped interpret the resulting model.

The evaluation procedure constructs a separate model for each of the datasets (**JDPA** and **MPQA**) and for the joint dataset (**JDPA+MPQA**). We evaluate the performance of different subsets of features from the set proposed in Section IV-A using recursive feature elimination (RFE) and select the subsets that optimize the classification performance.

We evaluate the sentence-based and distance-based baseline approaches on the whole dataset as one test fold and the classifiers using stratified 10-fold cross-validation with random shuffling of the input observations and report the average performance. For the classifier-based baseline (**Kessler 10**) we evaluate the performance of the feature set with the same logistic regression classification algorithm. Our evaluation report includes the standard classification performance metrics in terms of Precision (**P**), Recall (**R**) and F-score (**F**).

VI. RESULTS

The results of the baseline approaches and the Logistic Regression classifier trained with different feature sets are summarized in Table III. Both distance-based approaches (**Closest Target** and **Closest Sentiment**) with their simple heuristics turned out to be surprisingly strong baselines performing well on both datasets. The classifier baseline (Classifier: **Kessler 10**) trained on the subset of features proposed by Kessler et al. [10] outperforms only the sentence-based baseline and fails to reach the results of distance-based approaches due to low recall.

[2]The JDPA mentions, coreference, meronymy and sentiment corpus has been developed by J.D. Power and Associates (www.jdpower.com) and is the sole and exclusive intellectual property of J.D. Power and Associates.

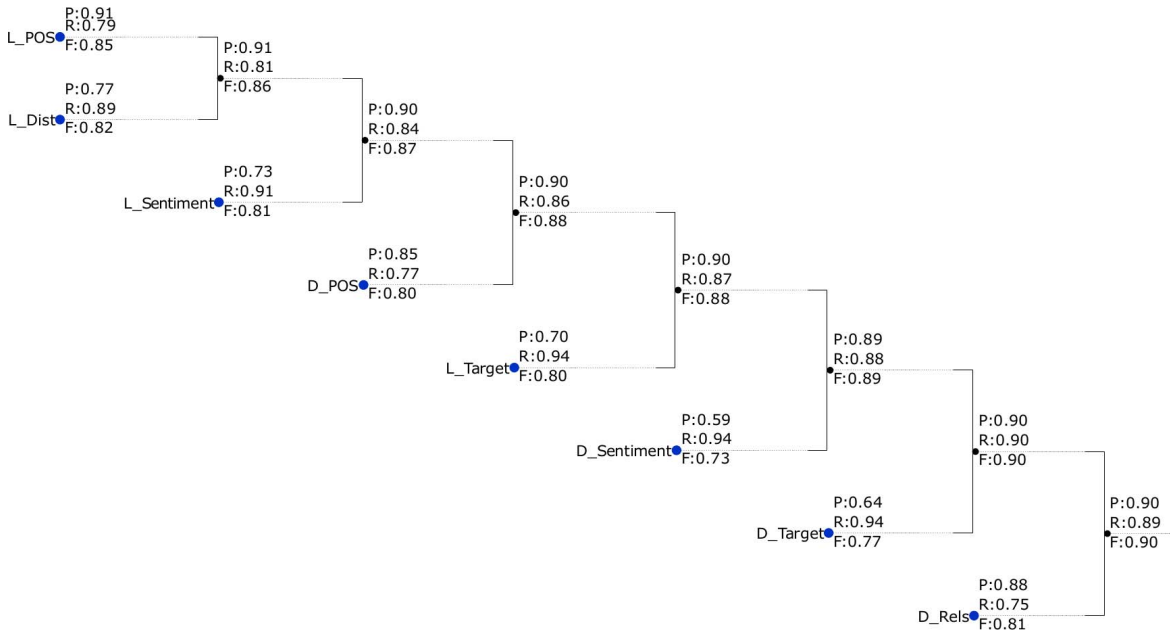

Figure 3. RFE Tree for **Selected 8** features.

Table III
BASELINE EVALUATION AND FEATURE SELECTION RESULTS.

Methods \ Datasets	JDPA			MPQA			JDPA+MPQA		
	P	R	**F**	P	R	**F**	P	R	**F**
Sentence-based	0.53	1	**0.69**	0.49	1	**0.65**	0.49	1	**0.66**
Closest Target	0.81	0.96	**0.88**	0.90	0.88	**0.89**	0.88	0.89	**0.89**
Closest Sentiment	0.89	0.95	**0.91**	0.94	0.85	**0.89**	0.92	0.87	**0.90**
Classifier: **Kessler 10**	0.91	0.85	**0.88**	0.91	0.88	**0.89**	0.92	0.84	**0.88**
Classifier: **All 26**	0.91	0.89	**0.90**	0.90	0.91	**0.90**	0.91	0.89	**0.90**
Classifier: **Selected 8**	0.91	0.89	**0.90**	0.90	0.91	**0.90**	0.90	0.89	**0.90**
Classifier: **Selected 4**	0.90	0.84	**0.87**	0.90	0.88	**0.89**	0.90	0.84	**0.87**

The redundant set of features enabled us to discover the most useful feature configurations for linking sentiments to their targets; e.g., directions of dependency relations (D_Dir in Table I) do not improve performance of the classifier; the exact sentiment/target counts along the path (see Table I: 6-7 & 19-20) do not provide an improvement over simple Boolean indicators (Table I). On the contrary, grouping POS tags into semantically equivalent groups (see 'POS tag groups' in Table I) provides a slight improvement in recall compared to the original 'POS tags' features. We also observe that dependency path features help predicting sentiment-target pairs - but POS tags along the dependency path suffice to uncover the relations, and dependency labels appear to be redundant.

We found a subset of eight features (highlighted in grey in Table I), which achieves nearly the same performance as the classifier trained on the whole feature set (see Classifier: **Selected 8** versus Classifier: **All 26** in Table III). Only a single feature (L_Dist: lexical distance) out of 10 proposed by Kessler et al. appears in this subset.

For each feature from the subset of **Selected 8** we report individual and combined performance results on the joint JDPA+MPQA dataset by providing a snapshot of the recursive feature elimination (RFE) procedure in Figure 3. The leaf-nodes of the RFE tree correspond to the features and the labels next to them show the performance of the classifier trained on this single feature as an input. The internal (non-leaf) nodes show how the features can be combined to gradually increase the classification performance.

Classifier: **Selected 4** in Table III corresponds to the subset of features from Classifier: **Selected 8**, which do not rely on the annotations of other targets in sentiments: L_POS, L_Dist, D_POS and D_Rels. This results are important to account for errors in sentiment/target extraction.

L_POS feature, which corresponds to the sequence of POS-tag groups for the tokens located on the lexical path between a sentiment-target pair showed the best performance on the joint JDPA+MPQA dataset. On its own, it achieves an F-score of 0.85 and 0.91 in Precision (see Figure 3). Table IV lists the top-10 POS patterns ordered by their predictive

Table IV
TOP-10 MOST USEFUL POS PATTERNS ON THE LEXICAL PATH.
BOLD FONT INDICATES THE CORRESPONDING **TARGET TOKENS**, AND ITALICS – THE *sentiment tokens*.

Target	L: POS-tags	Example
+	['noun', 'conj']	*Great* power and **acceleration**.
-	['conj']	Focus drives *well* and **SVT** did a great job.
+	['adverb']	**It's** pretty *neat*.
+	['verb', 'conj']	AF mode *actively* tracks and **focuses** on moving subjects.
+	['prep', 'noun', 'conj']	*High levels* of stiffness and **strength**.
-	['punct', 'conj']	He was *sleeping comfortably*, and **the food** was not too bad either.
-	['conj', 'det']	The convertible received a good **frontal-offset-crash test** and an *acceptable* in the side-crash test.
+	[]	Automatic model with paddle shifters **performed** *well* on the track.
-	['punct', 'conj', 'det']	Combination with an extremely economical combustion **engine**, and the *outstanding* aerodynamic qualities.
+	['punct', 'adj', 'adj']	The new XJ is the epitome of *fluid*, contemporary automotive **style**.

power (absolute value of the Logistic Regression coefficients), distinguishing both positive and negative patterns: column 'Target' indicates whether the pattern correlates with valid (+) or invalid (-) sentiment-target pairs.

VII. DISCUSSION

A. Classifier Performance

Our evaluation results demonstrate that it is not a trivial task to train a machine-learning classifier able to outperform naive distance-based heuristics. In particular, the initial set of features from Kessler et al. showed a drop in recall, which may indicate the issue of overfitting.

Statistics from the two manually annotated corpora indicate that the valid sentiment is most likely to appear closer to its target than other sentiments within the same sentence. Nevertheless, there are examples in which this is not the case (consider "*Roses* are very *Red* and *Violets* are Blue").

B. Parsing Errors

The performance of our approach depends on the performance of the POS-tag and dependency-relation parsers it builds upon. The extracted patterns may contain errors, if the parser fails to produce the correct parse tree of the input sentence. For example, participles were sometimes misclassified as verbs instead of adjectives (e.g. 'perfectly exposed'), gerund (verbal nouns) – as verbs (e.g. 'expensive looking and feeling'), or nouns that function as adjectives (e.g 'storage space').

In this case such patterns are not generalizable and may differ from the results returned by other parsers. Nevertheless, the parser-specific errors may be neglected, if the same parsing algorithm used for training the classification model is employed in production assuming that the errors in training will be replicated on the new data as well.

C. Sentiment and Target Extraction

Our initial assumption was that the sets of sentiment and target tokens already exists. Therefore, the performance on the sentiment-target linking task depends on the correct annotation of sentiments and their targets. For example, our classifier evaluates to False for the pair: "like - car" from the sample sentence: "I *like* to drive the **car**", and to True for the pairs: "like - to drive the car" and "like - to drive".

The high performance demonstrated by the simple proximity-based heuristics (**Closest Target** and **Closest Sentiment**) reveals that lexical distance serves as a major predictor for the sentiment-target relation. However, these approaches to a large extent depend on the correct annotation of targets and sentiments, which has been provided by the ground truth in our experiments. In practice, the correct sentiment target extraction is a hard task on its own.

The results of the feature selection and classification procedures shed the light on the common patterns that correlate with the valid and invalid sentiment-target assignment. The extracted patterns, such as the ones listed in Table IV, can help to extract the valid targets given their sentiments and vice verse.

D. Limitations

The approach proposed in this paper is supervised and requires an annotated corpus for training the classification model. The quality and coverage of the observations contained in the training corpora influence the performance of the model. In our experiments we used two corpora (JDPA and MPQA), which are publicly available and free to use for academic and research purposes. However, extending the suggested approach to other languages requires additional language resources, i.e. new corpora annotated with sentiment-target pairs.

The feature set that we evaluated is by far not exhaustive. We also did not address the interplay between different features that may harm performance of the classifier.

Our results suggest new features, which may improve the classification performance, such as integration of the closest sentiment/target baselines into the classifier-based approach as Boolean features. Compound features combining several of the **Selected 8** features are also good candidates that will harness the interplay between the most productive features. Furthermore, the extracted top-10 POS patterns hint at the importance of conjunctions and punctuation marks on the lexical path between the sentiment-target pair (see Table IV).

VIII. Conclusion

This paper presents a machine-learning approach to the sentiment-target linking task that builds upon an extended set of features extracted using syntactic analysis of an input sentence. Our experimental evaluation demonstrates that a simplistic distance-based approach performs very well in presence of gold-standard sentiment and target annotations. The classifier-based approach struggles to achieve comparable performance using a wide range of features, which highlights the importance of the feature engineering and feature selection phase.

The results also demonstrate the performance of many syntactic features that were previously employed in the related work and were assumed to be efficient. The paper provides a comprehensive evaluation of the individual features as well as their combinations and suggest several optimal configurations able to maximize the performance of the classification model. In addition, we reveal which features have proven useful for the sentiment-target linking tasks, as well as suggest good candidates for new features, which may improve the classification performance. Although we have chosen corpora with very different annotation styles (JDPA and MPQA) for the experimental setting, more extensive evaluations will be required to confirm that the presented results are applicable across domains. The frequent patterns discovered by the classifier may also assist in extracting opinion targets given the sentiment tokens and vice verse.

Acknowledgment

The presented work was supported by the Pheme Project (www.pheme.eu), funded by the European Unions 7th Framework Program for Research, Technology Development and Demonstration under the Grant Agreement No. 611233, the uComp Project (www.ucomp.eu) with funding support of EPSRC EP/K017896/1, FWF 1097-N23 and ANR-12-CHRI-0003-03, in the framework of the CHIST-ERA ERA-NET Program Line, and the IMAGINE project (www.htwchur.ch/imagine) funded by the Swiss Commission for Technology and Innovation (CTI).

References

[1] A. Weichselbraun, S. Gindl, and A. Scharl, "Extracting and grounding context-aware sentiment lexicons," *IEEE Intelligent Systems*, vol. 28, no. 2, pp. 39–46, 2013.

[2] B. Yang and C. Cardie, "Joint inference for fine-grained opinion extraction," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 13)*, 2013, pp. 1640–1649.

[3] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the Tenth International Conference on Knowledge Discovery and Data Mining (KDD 04)*, 2004, pp. 168–177.

[4] K. Bloom, N. Garg, and S. Argamon, "Extracting appraisal expressions," in *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 07)*, 2007, pp. 308–315.

[5] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent twitter sentiment classification," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL 11)*, 2011, pp. 151–160.

[6] S. Gindl, A. Weichselbraun, and A. Scharl, "Rule-based opinion target and aspect extraction to acquire affective knowledge," in *WWW Workshop on Multidisciplinary Approaches to Big Social Data Analysis (MABSDA 13)*, 2013, pp. 557–564.

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

[8] L. Zhuang, F. Jing, and X.-Y. Zhu, "Movie review mining and summarization," in *Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM 06)*, 2006, pp. 43–50.

[9] L. Xu, K. Liu, S. Lai, Y. Chen, and J. Zhao, "Mining opinion words and opinion targets in a two-stage framework." in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 13)*, 2013, pp. 1764–1773.

[10] J. S. Kessler and N. Nicolov, "Targeting sentiment expressions through supervised ranking of linguistic configurations." in *Proceedings of the Third International AAAI Conference on Weblogs and Social Media (ICWSM 09)*, 2009, pp. 90–97.

[11] L. Deng and J. Wiebe, "Joint prediction for entity/event-level sentiment analysis using probabilistic soft logic models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 15)*, 2015.

[12] N. Jakob and I. Gurevych, "Extracting opinion targets in a single-and cross-domain setting with conditional random fields," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 10)*, 2010, pp. 1035–1045.

[13] 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, 2011.

[14] J. S. Kessler, M. Eckert, L. Clark, and N. Nicolov, "The ICWSM 2010 JDPA sentiment corpus for the automotive domain," in *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media Data Workshop Challenge (ICWSM-DWC 2010)*, 2010.

[15] A. L. Ginsca, "Fine-Grained Opinion Mining as a Relation Classification Problem," in *Proceedings of the Imperial College Computing Student Workshop*, vol. 28, 2012, pp. 56–61.

[16] J. S. Kessler and N. Nicolov, "The JDPA Sentiment Corpus for the Automotive Domain," in *The Handbook of Linguistic Annotation*, 2015.

[17] T. Wilson, "Fine-grained subjectivity and sentiment analysis: recognizing the intensity, polarity, and attitudes of private states," Ph.D. Dissertation, University of Pittsburgh, 2008.

[18] J. Wiebe, T. Wilson, and C. Cardie, "Annotating expressions of opinions and emotions in language," *Language Resources and Evaluation*, vol. 39, no. 2-3, pp. 165–210, 2005.

[19] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 03)*, 2003, pp. 173–180.

[20] A. Weichselbraun and N. Süsstrunk, "Optimizing dependency parsing throughput," in *Proceedings of the 7th International Conference on Knowledge Discovery and Information Retrieval (KDIR 15)*, 2015, pp. 511–516.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.