Semantics-Driven Implicit Aspect Detection in Consumer Reviews

Kim Schouten
schouten@ese.eur.nl

Nienke de Boer
nien_de_boer@hotmail.com

Tjian Lam
tjianlam@gmail.com

Marijtje van Leeuwen
marijtje93@hotmail.com

Ruud van Luijk
ruudvanluijk91@gmail.com

Flavius Frasincar
frasincar@ese.eur.nl

Erasmus University Rotterdam
PO Box 1738, NL-3000 DR
Rotterdam, the Netherlands

ABSTRACT

With consumer reviews becoming a mainstream part of e-commerce, a good method of detecting the product or service aspects that are discussed is desirable. This work focuses on detecting aspects that are not literally mentioned in the text, or implicit aspects. To this end, a co-occurrence matrix of synsets from WordNet and implicit aspects is constructed. The semantic relations that exist between synsets in WordNet are exploited to enrich the co-occurrence matrix with more contextual information. Comparing this method with a similar method which is not semantics-driven clearly shows the benefit of the proposed method. Especially corpora of limited size seem to benefit from the added semantic context.

1. INTRODUCTION

With the advance of e-commerce and the Web 2.0, online consumer reviews are now ubiquitous. Review information is valuable for consumers and producers alike, but the sheer size of it makes its access prohibitive. To be able to leverage the available information, automatic methods for processing and summarizing consumer reviews are needed. In order to provide sufficient insight, consumer review summaries have to present sentiment information on an aspect level, as opposed to just one overall sentiment score [5]. This is usually referred to as aspect-level sentiment analysis, and it roughly consists of two core parts: detecting the aspects and classifying sentiment for each aspect. In this work, only the first part is discussed.

When aspects are mentioned literally in the text, they are called explicit aspects, as opposed to implicit aspects, which are only implied by some fragment of text. With explicit aspects being much more numerous, most research has focused on that type of aspects. The little work [2, 7, 8, 6] that has been done on implicit aspects generally makes use of co-occurrences between words and aspects, either directly or via association rule mining.

> "This product costs too much"

In this example, the aspect 'price' is not literally mentioned, even though it is clearly implied by the word 'costs'. This directly links the word 'costs' to the implicit aspect 'price'. However, this link is not always so obvious.

> "I could not sleep because of the noise"

When reviewing a hotel, this example sentence is obviously a negative comment about the rooms or the location of the hotel, but there is no clear word-to-aspect link that can be used.

We propose a semantics-driven approach, extending the work described in [6], that goes beyond the straightforward word-to-aspect links that have been used in previous work, employing word sense disambiguation and utilizing the semantic relations between words in a text to provide a broader semantic context for finding implicit aspects.

2. METHOD

The core element of the proposed method is to build a co-occurrence matrix that contains the co-occurrence frequencies of annotated implicit aspects and synsets from WordNet in a training corpus. The synsets are acquired by performing word sense disambiguation, for which the Lesk [4] algorithm is used, and they represent the semantics or meaning of a word given the context. This is done at training time using only the training data.

At runtime, when processing the test data, a score is computed for each potential implicit aspect, based on the co-occurrence frequency between that implicit aspect and each synset in a given sentence. This information is enriched by adding a weighted fraction of the co-occurrence frequencies of all semantically related synsets. This helps create a more complete representation of the semantics that are being conveyed by a given sentence. A formula representation of this runtime scoring process is shown in Eq. 1.

$$score\ a_i = \frac{1}{v}(\sum_{j=1}^{v}(\frac{c_{i,j}}{f_j} + \sum_{r \in R}\sum_{k=1}^{K_r(j)} w(r)\frac{c_{i,k}}{f_k})), \qquad (1)$$

where v is the number of synsets corresponding to a sentence, a_i is the ith implicit aspect in the total list of implicit aspects, j is the jth synset in a sentence, $c_{i,j}$ is the co-occurrence frequency of aspect i and synset j in the training set, f_j is the total frequency of synset j in the training set, r is a semantic relation corresponding to synset j, R is the set of semantic relations available in WordNet, $K_r(j)$ is the set of synsets corresponding to relation r and synset j, $w(r)$ is the optimized weight that corresponds to the semantic relation r, $c_{i,k}$ is the co-occurrence frequency of aspect i with the related synset k from $K_r(j)$ in the training set, and f_k is the total frequency of synset k in the training set.

The potential implicit aspect that has the highest score using the above formula is associated to this sentence, as long as it exceeds a certain trained threshold. This enables the algorithm to not choose any implicit aspect for a sentence if scores are too low.

3. EVALUATION

Two standard data sets are used to evaluate the work presented here. The first data set contains restaurant reviews [1]; the second contains electronic product reviews [3]. Both data sets roughly have the same number of sentences. However, whereas the restaurant data has 5 different implicit aspects that all have a relatively high frequency, the product data set has many different implicit aspects and most implicit aspects have a frequency below 10.

For all algorithms, we use a 10-fold cross-evaluation. Optimization of the weights associated with the semantic relations and the threshold are all performed on the training data only. In [6], the effect of Part-of-Speech (POS) filters was shown to have a significant effect on the results. With POS filters, certain categories of words can be filtered out of the co-occurrence matrix. For this evaluation, the presented results are acquired by using the best performing POS filter for each algorithm.

In Table 1, the results of the different methods are given. The original word-based co-occurrence method from [6] is reported, as well as the 'Synsets' method which only replaces the words by the synsets using word sense disambiguation, followed by the results of the 'Relations' method, which is the 'Synsets' method enriched with information from the semantic relations.

Method	F_1**-score**
Restaurant review data set	
Schouten and Frasincar [6]	63.6 (NN+JJ)
Synsets	64.2 (NN)
Relations	79.4 (NN+VB+JJ)
Product review data set	
Schouten and Frasincar [6]	12.5 (NN+VB+JJ+RB)
Synsets	12.9 (NN+JJ)
Relations	70.4 (all)

Table 1: Comparison of the F_1-scores of the 'Relations', 'Synsets', and Schouten and Frasincar [6] methods. In parentheses, the used, optimal, POS filter can be found.

From the above table, one can notice that simply exchanging words for synsets does not significantly improve results. This can be due to the fact that word sense disambiguation is a very hard problem and the used algorithm is likely to introduce errors that will offset any improvements gained by using synsets. However, adding the information from the semantic relations, which is only possible after transitioning from words to synsets, yields a significant increase in performance. This is true for both data sets, but more so on the harder product data set, where only few positive training examples are available. The lack of data is, at least partially, compensated by having a broader semantic context available.

4. CONCLUSION

In this work, a new method to find implicit aspects is proposed that better represents the semantics of a sentence by utilizing synsets and semantic relations in WordNet. Enriching an existing word-based co-occurrence method with this semantic information improved the results, especially for the smaller data set considered. An obvious improvement for future work is to have the algorithm be able to assign more than one implicit aspect to a given sentence. Furthermore, the synsets and semantic relations from WordNet could be complemented with concepts and relations from domain ontologies to yield an even better sentence representation.

Acknowledgment

The authors are partially supported by the Dutch national program COMMIT.

5. REFERENCES

[1] I. Androutsopoulos, D. Galanis, S. Manandhar, H. Papageorgiou, J. Pavlopoulos, and M. Pontiki. SemEval-2014 Task 4, 2014.

[2] Z. Hai, K. Chang, and J. j. Kim. Implicit Feature Identification via Co-occurrence Association Rule Mining. In *12th Int. Conf. on Computational Linguistics and Intelligent Text processing*, volume 6608, pages 393–404. Springer, 2011.

[3] M. Hu and B. Liu. Mining and Summarizing Customer Reviews. In *10th Int. Conf. on Knowledge Discovery and Data Mining*, pages 168–177. ACM, 2004.

[4] M. Lesk. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In *5th Annual Int. Conf. on Systems Documentation*, pages 24–26. ACM, 1986.

[5] B. Liu. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool, 2012.

[6] K. Schouten and F. Frasincar. Finding Implicit Features in Consumer Reviews for Sentiment Analysis. In *14th Int. Conf. on Web Engineering*, volume 8541, pages 130–144. Springer, 2014.

[7] Q. Su, X. Xu, H. Guo, Z. Guo, X. Wu, X. Zhang, B. Swen, and Z. Su. Hidden Sentiment Association in Chinese Web Opinion Mining. In *17th Int. Conf. on World Wide Web*, pages 959–968. ACM, 2008.

[8] Y. Zhang and W. Zhu. Extracting Implicit Features in Online Customer Reviews for Opinion Mining. In *22nd Int. Conf. on World Wide Web Companion*, pages 103–104. IW3C2, 2013.