

Mining Product Reviews Based on Shallow Dependency Parsing

Qi Zhang
School of Computer Science
Fudan University
Shanghai, P.R.China
qi_zhang@fudan.edu.cn

Mitsunori Ogihara
Dept. of Computer Science
University of Miami
Coral Gables, FL 33146
m.ogihara@miami.edu

Yuanbin Wu
School of Computer Science
Fudan University
Shanghai, P.R.China
ybwu@fudan.edu.cn

Joseph Johnson
Department of Marketing
University of Miami
Coral Gables, FL 33124
jjohnson@miami.edu

Tao Li
School of Computer Science
Florida International University
Miami, FL 33199
taoli@cs.fiu.edu

Xuanjing Huang
School of Computer Science
Fudan University
Shanghai, P.R.China
xjhuang@fudan.edu.cn

ABSTRACT

This paper presents a novel method for mining product reviews, where it mines reviews by identifying product features, expressions of opinions and relations between them. By taking advantage of the fact that most of product features are phrases, a concept of shallow dependency parsing is introduced, which extends traditional dependency parsing to phrase level. This concept is then implemented for extracting relation between product features and expressions of opinions. Experimental evaluations show that the mining task can benefit from shallow dependency parsing.

Categories and Subject Descriptors

I.2.7 [Artificial Intelligence]: Natural Language Processing - text analysis

General Terms

Algorithms, Experimentation

Keywords

Product Mining, Dependency Parsing

1. INTRODUCTION

Managers and researchers acknowledge that product quality is highly correlated to several outcome variables such as market share and stock valuation. However, product quality and its dimensions are difficult to assess objectively. One solution is to mine the World-Wide Web with its ever-increasing repository of product reviews generated by consumers, experts, and rating agencies. Yet this vast richness of content has made it difficult to extract useful information such as product quality from the web. Retrieving this information and analyzing this content is an impossible task if it were to be manually done. However, advances in machine learning and natural language processing present us with a unique opportunity to automate the decoding of quality dimensions from online reviews.

Much work has been done on sentimental classification at different levels [4, 3]. In this paper, we study the problem of mining on-

line reviews, which consists of identifying product features, identifying expressions of opinions, and discovering relations among them. Simple inspection of the data reveals that product features usually contain more than one word. This suggests that dependency parsing might not be the best approach here unfortunately, because it provides dependency relations only between words. To solve this issue, inspired by shallow phrase-structure parsing, we introduce the concept of *shallow dependency parsing* and propose an approach to construct it. Shallow dependency parsing segments an input sentence into “phrases” and links the segments with directed arcs. The parsing focuses on the “phrases” and the relations between them, rather than on the single words inside each phrase.

2. THE APPROACH

The approach performs the opinion mining task in three main steps: (1) constructing a shallow dependency tree based on shallow phrase-structure parsing and dependency parsing; (2) extracting candidate product features and candidate opinion expressions; (3) extracting relations between product features and opinion expressions.

Shallow Dependency Parsing

Shallow dependency parsing is dependency parsing with phrase nodes. A dependency relationship, which is an asymmetric binary relationship, holds between two phrases. One phrase is called the *head*, which is the central phrase in the relation, while the other the *dependent*, which modifies the head. A label representing the relation type is assigned to each dependency relationship, such as *subj* (subject) and *obj* (object). Figure 1 shows an example of shallow dependency parsing.

A bottom-up process is used for shallow dependency parsing. We use a shallow parser, Sundance, for segmenting English sentences into syntactic chunks and Stanford Parser for generating dependency parsing trees. Phrases extracted from Sundance shallow parser are treated as nodes in shallow dependency parsing tree. The dependency relationships among words inside a phrase are reserved. The algorithm is shown in Alg. 1.

Candidate Product Features and Opinion Expressions Extraction

The *product features* are defined as parts, properties, properties of parts, product, company name and related objects which the eval-

Algorithm 1 Pseudocode for constructing the shallow dependency parsing trees

INPUT: Sentence $S = w_1 w_2 \dots w_n$

OUTPUT: Shallow dependency relation $\langle H_i, D_i \rangle$

- 1: $\{ \langle h_i, d_i \rangle \mid h_i, d_i \in S \} = \text{dependency-parse}(S)$
- 2: $P = \text{shallow-parse}(S)$
- 3: **for** each phrase $P_t \in P$ **do**
- 4: Find a word $d_t \in P_t$ whose head $h_t \notin P_t$
- 5: Find a phrase $P'_t, h_t \in P'_t$ such that
- 6: $\langle P'_t, P_t \rangle$ is a shallow dependency relation with P'_t as the head and P_t as the dependent
- 7: Add $\langle P'_t, P_t \rangle$ to *OUTPUTS*
- 8: **end for**
- 9: return *OUTPUTS*

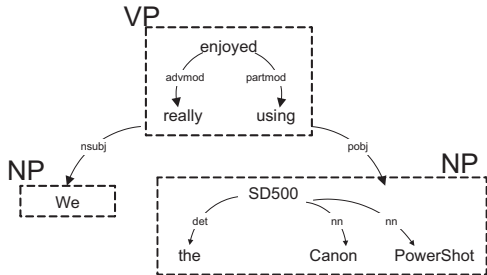


Figure 1: Example of Shallow Dependency Parsing Tree.

uation is made on. We introduce a language model to shrink the candidates based on the assumption that the more likely a phrase to be a product features, the more closely it related to the product reviews. In practice, for a certain domain of product reviews, the language model is built on unlabeled data and each candidate NP or VP chunk in the output of shallow parser is scored by the model and cut off by a threshold.

Opinion expressions are spans of text that express an evaluation or attitude of the opinion holder, which are usually evaluative or subjective phrases. We use a dictionary which contains 8221 opinion expressions to select candidates [6]. Another assumption we use to filter candidate opinion expressions is that opinion expressions tend to appear closely with product features, which is also used to extract product features by Hu and Liu [2]. In our experiments, the tree distance between product feature and opinion expression in a relation should be less than 5 in the shallow dependency parsing tree.

Relation Extraction

The relation extraction is treated as a classification task, and kernel methods is used to address the problem. All combinations between candidate product features and candidate opinion expressions are potential relations. Given a shallow dependency parsing tree, the subtree rooted at the lowest common parent of the opinion expression and the product feature is used to represent the relation. Dependency tree kernels proposed in [1] is used to model this task.

3. EXPERIMENTAL RESULTS

We conducted two sets of experiments to empirically evaluate the performance of the proposed approach. In the first set of experiments, we compare the performances of different relation extraction methods on the corpus used by Hu and Liu [2] which contains reviews of Cell phone and Digital camera domains. Table 1 shows the comparison results on the performances of different methods in

Table 1: Performances of different methods

Methods	Cell Phone			Digital Camera		
	P	R	F	P	R	F
Adjacent	40.3%	60.5%	48.4%	32.7%	59.1%	42.1%
SVM-Linear	50.7%	28.1%	36.2%	48.5%	22.0%	30.3%
SVM-RBF	60.7%	19.7%	29.7%	66.9%	23.3%	34.6%
SVM-Tree	52.7%	52.7%	52.7%	58.9%	51.4%	54.9%

different domains of the corpus. For each domain, we use all the sentences belonged to it, and conduct a 5-fold cross validation on them. The candidate product features and opinion expressions are filtered by the methods discussed in Section 2. All combinations between candidate product features and opinion extractions are potential relations. *SVM-Tree* denotes the results of our tree-kernel based SVM. In comparison with our tree-kernel based approach, we evaluated SVM¹ with two kinds of kernel functions: linear and radial basis function, whose results are denoted by *SVM-Linear* and *SVM-RBF* in the Table 1. The features are the context and part-of-speech tags. We also compared with the *Adjacent* method which extracts relations between the product features and its nearest opinion expression [2]. In both domains, *Adjacent* performed better than *SVM-Linear* and *SVM-RBF* and *SVM-Tree* outperforms the other three methods. The experiment comparison demonstrates the effectiveness of our proposed methods.

In the second set of experiments, we present a case study on applying our proposed approach on fifty reviews written by Walter Mossberg on the Thursday issues of the Wall Street Journal from 1991 to 2001 [5]. Two annotators were asked to independently annotate product features, opinion expressions, and relations between them. We combined the data from cell phone and digital camera domain in the corpus of Hu and Liu [2] and used them for training. The F-score of *SVM-Tree* on fifty reviews drops to 21.6%. The main reason is that the fifty reviews covered different domains including computers, printers, DVD/TV products and MP3 Players. One of future work is to develop robust techniques for identifying opinion expressions across multiple different domains.

Acknowledgments

This work was partially funded by Chinese NSF 60673038, Doctoral Fund of Ministry of Education of China 200802460066, and Shanghai Science and Technology Development Funds 08511500302.

4. REFERENCES

- [1] A. Culotta and J. Sorensen. Dependency tree kernels for relation extraction. In *In Proceedings of ACL 2004*, 2004.
- [2] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the ACM SIGKDD 2004*, 2004.
- [3] N. Kobayashi, K. Inui, and Y. Matsumoto. Extracting aspect-evaluation and aspect-of relations in opinion mining. In *Proceedings of EMNLP-CoNLL 2007*, 2007.
- [4] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In *Proc. of EMNLP 2002*, 2002.
- [5] G. J. Tellis and J. Johnson. The value of quality. *Marketing Science*, 26(6):758–773, 2007.
- [6] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT-EMNLP*, 2005.

¹libsvm 2.88 is used in our experiments