An Effective Approach to Automatic Feature Based Opinion Lexicon Expansion

Myat Su Wai[1], Sint Sint Aung[2]

[1]Web Data Mining Lab, University of Computer Studies, Mandalay, Myanmar

[2,] Department of Academic Affairs, University of Computer Studies, Mandalay, Myanmar

***Abstract**:*

In many applications related to opinion mining and sentiment classification, it is necessary to compute the semantic orientation of certain opinion expressions on an object. Many researchers suggest that semantic orientation depends on application domains. Moreover, semantic orientation depends on the specific feature that an opinion is expressed on it. In this paper, we introduce an effective approach to opinion lexicon expansion automatically. We use small set of seed lexicon and dependency relations to extract opinion words and then, we expand it automatically from a larger set of unannotated documents. To do this, we proposed an unsupervised algorithm based on double propagation. Our method was evaluated in three different domains (headphones, hotels and car), using a corpus of product reviews which opinions were annotated at the feature level. We conclude that our method produces feature-level opinion lexicons with better precision and recall that domain-independent opinion lexicons without using annotated documents.

Keywords

Opinion Extraction, Opinion Lexicon Expansion, Dependency Relations.

1. Introduction

Sentiment analysis, also referred to as opinion mining, encompasses a broad area of natural language processing, computational linguistics, and text mining.

There are two fundamental problems in opinion mining; opinion lexicon expansion and opinion target extraction (Liu 2006; Pang and Lee 2008). An opinion lexicon is a list of opinion words such as good, excellent, poor, and bad which are used to indicate semantic orientation such as positive or negative [9]. Although there are several opinion lexicons publicly available, it is hard to maintain a universal opinion lexicon to cover all

domains because opinion expressions are different significantly from one domain to another. For example, a word can be positive in one domain but it has no opinion or even negative opinion in another domain.

Opinion lexicons have proven a valuable resource for opinion mining tasks. With large amounts of data readily available on the Internet, gathering user sentiments and opinions is a relatively effortless and inexpensive undertaking. As an example, it is possible to easily check whether or not a product is positively received. This information is useful to both potential new customers and the manufacturers or the suppliers of said product. While customers may wish to inform themselves whether the product lives up to the desired quality or value, the company has the opportunity to quickly gather the general consensus on the product and is therefore able to react accordingly.

Other examples include politicians or political parties that are able to quickly gather their voters' opinions by using opinion mining. Some of the most well-known lexicons include Senti WordNet (Esuli and Sebastiani, 2006), the Bing Liu Opinion Lexicon (Hu and Liu, 2004) and the General Inquirer (Stone et al., 1966). In recent years there has been increasing interest in building opinion lexicons for other languages as well. For German, Remus et al. (2010) built the Sentiment Wortschatz Lexicon, short SentiWS, using semi-automatic translations of English sentiment resources combined with information about word co-occurrences and word collocations. Banea et al. (2008) use raw data and a bootstrapping method to construct a subjectivity lexicon for languages with scarce resources such as Romanian and Wan (2009) exploits the large amount of annotated English data available to classify Chinese reviews.

The approach is supervised and allows classification without any annotated Chinese data. To the best of our knowledge there are no publicly available lexical resources for sentiment analysis for the Swedish language. Our goal is therefore to lay the groundwork for a Swedish sentiment lexicon.

Therefore, it is necessary to expand a known opinion lexicon in order to use in different domains [8]. There are three main approach to construct the opinion lexicon. Among them, manual approach are very time consuming and thus only usually used as ground through data to validate the automatic approaches. And then, dictionary based approach are unable to find opinion words with domain and context-specific orientations, which is quite common. Finally, corpus based approach are rely on syntactic or co-occurrence patterns and a seed list of opinion words. The proposed system includes in corpus-based approach [1].

There has been many research in opinion extraction. However, machine-learning approaches usually require annotated text and they are known to be domain-dependent. In [3].

In this work, we proposed an unsupervised lexicon based approach to extract opinions and product features simultaneously in three domains without training examples. This paper contributes the following points.

- This system introduced verb opinions and verb product features.
- This system used dependency relations to extract opinion words automatically.

The main advantages of our approach are that there is no need for training data and it has domain independency.

The rest of this paper is organized as follow; section 2 describes some related works with our approach. Section 3 describes detail of the proposed system. Section 4 shows experimental results and analysis of the proposed system and section 5 concludes the proposed system.

2. Related Works

Many research has been done about opinion word extraction. In general, the existing work can be categorized as corpora-based and dictionary-based approaches. Our work falls into the corpora-based approaches.

Hatzivassiloglou and McKeown (1997) proposed the first method for determining adjective polarities. The method predicts orientations of adjectives by detecting pairs of such words conjoined by conjunctions like and, and or in a large document set. The underlying intuition is that the orientations of conjoined adjectives are subject to some linguistic constraints. The weakness of this method is that as it relies on the conjunction relations it is unable to extract adjectives that are not conjoined.

Wiebe (2000), Wiebe et al. (2004) proposed an approach to finding subjective adjectives using the results of word clustering according to their distributional similarity. However, they did not tackle the prediction of sentiment polarities of the found subjective adjectives.

Turney and Littman (2003) compute the point wise mutual information (PMI) of the target term with each seed positive and negative term as a measure of their semantic association. However, their work requires additional access to the Web (or any other corpus similar to the Web to ensure sufficient coverage) which is time consuming.

Another recent corpora-based approach is proposed by Kanayama and Nasukawa (2006). Their work first uses clause level context coherency to find candidates, then uses a statistical estimation method to determine whether the candidates are appropriate opinion words. However, their method for finding candidates would have low recall if the occurrences of seed words in the data are infrequent or an unknown opinion word has no known

opinion words in its context. Besides, the statistical estimation can be unreliable if the corpus is small, which is a common problem for statistical approaches.

In dictionary-based approaches, Kamps et al. (2004) take advantage of WordNet to construct a synonymy network by connecting pairs of synonymous words. The semantic orientation of a word is decided by its shortest paths to two seed words "good" and "bad" which are chosen as representatives of positive and negative orientations.

Esuli and Sebastiani (2005) use text classification techniques to classify orientations. Their method is based on the glosses (textual definitions) in an online "glossary" or dictionary.

The work of Takamura, Inui, and Okumura (2005) also exploits the gloss information from dictionaries. The method constructs a lexical network by linking two words if one appears in the gloss of the other. The weights of links reflect if these two connected words are of the same orientation. The works of (Hu and Liu 2004; Kim and Hovy 2004) are simpler as they simply used synonyms and antonyms. However, all dictionary-based methods are unable to find domain dependent sentiment words because entries in dictionaries are domain independent.

Guang et al. (2011) used a bootstrapping based method to expand opinion words and to extract targets. To perform the tasks, they considered syntactic relations between opinion words and targets. However, the authors only considered adjective opinions. The authors did not consider verb opinion. The extraction rules used in their system are only direct relations between product features and opinions. So, some dependency relations are still missing [6].

Ebrahim et al. (2012) presented a method for sentiment classification of online product reviews using product features. They used association rule mining to extract product features and also used support vector machine to classify sentiment orientation. However, since their approach is supervised, this method required a set of annotated review sentences as training examples. Some polarities are incorrect for another domain, i.e., their method is domain dependent [5].

Qian Liu (2013) proposed a logic programming approach for aspect extraction. In their system, they implemented double propagation in Answer Set Programming using 8 ASP rules. The recall is low because correct aspects were pruned as incorrect features and they considered only direct relations. Moreover, their approach may miss some infrequent features because this method extracted frequent noun or noun phrases as product features [14].

Yahui Xi (2013) developed an approach for extracting Chinese product features from Chinese product reviews. The authors also emphasize only on product features not on opinions [16] [17].

Zhao et al (2015) presented a new method called joint propagation and refinement for mining opinion words and targets. The authors used frequency based threshold to prune incorrect targets. So, targets that are not occurred frequently, i.e. infrequent features are removed in their system. Threshold need to be raised to improve the precision which will affect the recall [18].

Our approach extracts not only domain independent opinion words but also context dependent opinion words.

3. Seed Lexicon

The proposed system uses 1000 words as seed opinion lexicon from original words of 6789 positive and negative opinion lexicons provided by Hu and Bing Liu, in KDD-2004.

4. Datasets and Annotation

Amazon product reviews datasets are used for the experiment. The first three datasets are annotated by Qian Liu and Bing Liu, University of Illinois at Chicago, (IJCAI, 2015). Opinions are manually collected from Dataset according to Bing Liu's lexicon, Vader lexicon and Sent WordNet. To get consistency, we check whether the collected words are contained in these three lexicons. If collected word contains in one of these lexicons, it is regarded as opinion word. Otherwise, it is ignored. Figure 1 shows the xml for of input datasets.

```
<sentence id="1">
  <text> My overall experience with this monitor was very poor. </text>
    <aspectTerms>
   <aspectTerm from="32" to="38" polarity="negative" term="monitor" pos="nn"/>
   </aspectTerms>
</sentence>
```

Figure 1. Input data format.

5. Rules for Opinion Extraction

In this section, we describe how to extract opinion and product features using extraction rules. They are the most important tasks for text sentiment analysis, which has attracted much attention from many researchers. Based on the relations between features and opinions, there are four main rules in the double propagation;

1. Extracting features using opinion words
2. Extracting features using the extracted features
3. Extracting opinion words using the extracted features

4. Extracting opinion words using both the given and the extracted opinion words

Table 1. Extraction rules using dependency relations

Rule	Observation	Constraint	Output
R11	O →O-Dep →F F→F-Dep→O	O∈{O} O-Dep ∈{DR} F-Dep ∈ {DR} POS(F)∈ {NN, VB}	F=Feature
R12	O→O-Dep→H←O-Dep←F	O∈{O} O-Dep∈{DR} F-Dep∈{DR} POS(F)∈ {NN, VB}	F=Feature
R13	O→O-Dep→H→F-Dep→F O←O-Dep←H←F-Dep←F	O∈{O} O-Dep∈{DR} F-Dep∈{DR} POS(F)∈ {NN, VB}	F=Feature
R21	Fi →Fi -Fep→Fj	Fj ∈{F} Fi-Dep=Fj-Dep POS(Fi) ∈ {NN}	Fi=Feature
R22	Fi →Fi -Dep→H←Fj Dep←Fj	Fj ∈{F} Fi-Dep=Fj-Dep POS (Fi) ∈{NN}	Fi=Feature
R31	O → O-Dep → F F→F-Dep→O	F∈{F} O-Dep∈{DR} F-Dep∈{DR} POS(O) ∈{JJ}	O=Opinion
R32	O→O-Dep→H←O-Dep←F	F∈{F} O-Dep∈{DR} F-Dep∈{DR} POS(O) ∈ {JJ, VB}	O=Opinion
R33	O→O-Dep→H→F-Dep→F O←O-Dep←H←F-Dep←F	F∈{F} O-Dep∈ {DR} F-Dep∈ {DR} POS(O) ∈ {JJ}	O=Opinion
R41	Oi →Oi-Dep →Oj	Oj ∈{O} Oi-Dep∈{CONJ} POS (Oj) ∈{JJ}	Oi=Opinion
R42	Oi→Oi-Dep→H←Oj-Dep←Oj	Oi ∈{O} Oi-Dep=Oj-ep, POS (Oj) ∈{JJ}	Oi=Opinion

In the extraction rules shown in table 1, O is opinion word, H is the third word, {O} is a set of seed lexicon, F is product feature, and O-Dep is part-of-speech information and dependency relations. {JJ}, {VB} and {NN} are sets of POS tags of potential opinion words and features, respectively. And {DR} contains dependency relations between features and opinions such as mod, pnmod, subj, s, obj, obj2, and conj.

5.1. Additional Patterns

In this section, we will describe some proposed patterns of this work because rules are not covered to extract opinions in some cases.

5.1.1. "Verb + Adjective" Pattern. This pattern means that subject is followed by verb and then followed by adjective. For example, "The output image gets worse if we block more of the frequencies." In this sentence, "worse" is an opinion word. So, we proposed a pattern to extract this kind of sentences.

6. Automatic Opinion Lexicon Expansion

The extraction process uses a rule-based approach using the relations defined in above. The system assumed opinion words to be adjectives, adverbs and verbs in some cases.

The primary idea is that opinion words are usually associated with product features in some ways. Thus, opinion words can be recognized by identified features, and features can be identified by known opinion words. So, the extracted opinion words and product features are used to identify new opinion words and new product features. The extraction process ends when no more opinion words or product features can be found.

To start the extraction process, opinion word lexicon O and review data R are provided as the input. Moreover, adjectives that are not opinion words are filtered out during the extraction process in order to increase precision and recall.

```
Input: Seed Word Lexicon {O}, General Word {G}, Review Data R
Output: Extracted Features {F}, Expanded Opinion Lexicon
{O-Expanded}
Function:
1. {O-Expanded} = {O}
2. {F} = ϕ, {Fi} = ϕ, {Oi} = ϕ
3. for each parsed sentence in R
4. {Fi} = Extracted features using R1i and patterns
5. for each f in {Fi}
6. if f not in {F} and {G}
7. Add f into {F}
8. end if
9. end for
10.{Fi} = Extracted features using R21, R22 and patterns
11. for each f in {Fi}
12. if f not in {F} and {G}
13. Add f into {F}
14. end if
15. end for
16. {Oi} = Extracted opinions using R31, R32, R33
17. for each o in {Oi}
18. if o not in {O-Expand}
19. Add o in {O-Expand}
20. end if
21. end for
22. {Oi} = Extracted opinions using R41, R42
23. for each o in {Oi}
24. if o not in {O-Expand}
25. Add o in {O-Expand}
26. end if
27. end for
28. end for
29. Repeat 2 until no se
```

Figure 2. Proposed algorithm

6.1. Context Dependent Opinion Words

Context dependent opinion means that a word may indicate different opinions in the same domain. This system can extract context dependent opinion words by using R31, R32 and R33 with dependency relations of amod and cop.

Table 2. Some context dependent opinion words extracted from the proposed system.

Longer battery life	positive	Longer run time	negative
Low price	positive	Low audio volume	negative
Small cost	positive	My house is small	negative
Big crisp screen	positive	Big problem	negative
Much more	positive	Much lower	negative

7. Experimental Results and Analysis

For the comparison of our approach for experiment, we use core i7 processor, 4GB RAM and 64 bit Ubuntu OS, and, we implement the proposed system with python programming language. In this paper, product review datasets are collected from https://www.cs.uic.edu/~liub/FBS/sentimentanalysis as resources for experiment. We choose the reviews of computer, router, and speaker datasets.

Table 3. Dataset for experiment

Dataset	**No. of sentences**	**Opinion words**	
		Redundant	**Non-redundant**
Computer	241	327	211
Router	245	349	156
Speaker	299	614	224

Table 3 shows the domains according to their names, the number of sentences and the number of opinions. This performance of opinion words expansion is evaluated in term of precision (P), recall (R) and f1-measure (F1). The system used two kinds of evaluation; duplicated words and non-duplicated words to analyze the performance of opinion lexicon expansion.

First, the system is evaluated with 100 duplicated words from as seeds. This is because we intended to consider number of words counts contained in the input datasets.

For example, suppose the word “great” contains 10 times in the annotated words of dataset. If we can extract it as the number of words as shown in annotation, the accuracy is 100%. So, we used duplicated words in the experiment. Table 4 shows performance evaluation with duplicated seed and duplicated words extraction.

Table 4. Experimental results of opinion lexicon expansion on all words

Dataset	P	R	F1
Computer	0.80	0.87	0.83
Router	0.80	0.90	0.85
Speaker	0.78	0.95	0.86

The second one is that newly extracted are only considered in evaluation and it is meaningful to evaluate the performance of the opinion words expansion. Different seed numbers are used to evaluate the performance of opinion words expansion such as 50, 70, 100, 200 and 300. Each of seed are randomly chosen and run, then the results are recorded. The results are shown by averaging these results. Tables 5 and 6 show average precision and recall of opinion lexicon expansion on newly extracted words with different seed numbers.

Table 5. Average precision of opinion expansion on newly extracted words.

Dataset	Seed numbers				
	50	70	100	200	300
Computer	0.61	0.61	0.61	0.59	0.54
Router	0.61	0.61	0.59	0.57	0.56
Speaker	0.67	0.67	0.67	0.63	0.61

Table 6. Average recall of opinion expansion on newly extracted words.

Generally, the results of all datasets are slightly difference. In two datasets (computer and speaker), using the number of seeds 100 achieve the highest F1-score. In router dataset, using the number of seeds 70 gets highest f1-measure. The larger the size of seed numbers, the more decrease in expansion rate. Table 7 also describes the comparative results of newly extracted words between different seed numbers.

Dataset	Seed numbers

	50	70	100	200	300
Computer	0.63	0.64	0.69	0.61	0.5
Router	0.64	0.65	0.66	0.57	0.52
Speaker	0.72	0.72	0.77	0.66	0.58

Table 7. Comparison of average f1-measure of opinion expansion between different seed numbers.

Dataset	Seed numbers				
	50	70	100	200	300
Computer	0.62	0.62	0.65	0.59	0.53
Router	0.62	0.63	0.62	0.57	0.53
Speaker	0.72	0.7	0.72	0.65	0.59

8. Conclusion

One of the important tasks in the performance of sentiment classification is having a comprehensive sentiment lexicon. However, since sentiment words have different polarities not only in different domains, but also in different contexts within the same domain, constructing such context-specific sentiment lexicons is not an easy task. In this work, we proposed an unsupervised corpora-based approach to automatic opinion lexicon expansion. The biggest advantage of the method is that it requires no additional resources except an initial seed opinion lexicon. According to experimental results, the proposed system works well in opinion extraction and polarity classification.

9. References

[1] Bing Liu, (2011) “Web Data Mining, Springer”, Second Edition, Department of Computer Science, University of Illinois, Chicago, USA

[2] C.J. Hutto and Eric Gilbert, 2014, “VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Texts”, Association for the Advancement of Artificial Intelligent AAAI, www.aaai.org, 2014,.

[3] Esuli, Andrea and Fabrizio Sebastiani. 2005. Determining the semantic orientation of terms through gloss classification. In Proceedings of CIKM’05, pages 617–624.

[4] Guang Qiu, Bing Liu, Jiajun Bu and Chun Chen. "Opinion Word Expansion and Target Extraction through Double Propagation." Computational Linguistics, March 2011, Vol. 37, No. 1: 9.27.

[5] Hatzivassiloglou, Vasileios and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. In Proceedings of ACL’97, pages 174–181.

[6] Hu, Mingqing and Bing Liu. 2004. Mining and summarizing customer reviews. In Proceedings of SIGKDD'04, pages 168–177.

[7] Kamps, Jaap, Maarten Marx, Robert J. Mokken, and Maarten de Rijke. 2004. Using wordnet to measure semantic orientation of adjectives. In Proceedings of LREC'04, pages 1115–1118.

[8] Kanayama, Hiroshi and Tetsuya Nasukawa. 2006. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In Proceedings of EMNLP'06, pages 355–363.

[9] Kim, Soo-Min and Eduard Hovy. 2004. Determining the sentiment of opinions. In Proceedings of COLING'04, pages 1367–1373.

[10] Kobayashi, Nozomi, Kentaro Inui, and Yuji Matsumoto. 2007. Extracting aspect-evaluation and aspect-of relations in opinion mining. In Proceedings of EMNLP'07.

[11] Maria Pontiki et al. (2016) "Aspect Based Sentiment Analysis", Association for Computational Linguistics (ACL), San Diego, California, June 16-17, 2016.

[12] Qian Liu, Zhiqiang Gao, Bing Liu and Yuanlin Zhang. A Logic Programming Approach to Aspect Extraction in Opinion Mining. Proceedings of IEEE/WIC/ACM International Confernece on Web Intelligence (WI-2013), 2013.

[13] Qian Liu, Zhiqiang Gao, Bing Liu and Yuanlin Zhang. "Automated Rule Selection for Aspect Extraction in Opinion Mining." Proceedings of International Joint Conference on Artificial Intelligence (IJCAI-2015), July 25-31, 2015.

[14] Takamura, Hiroya, Takashi Inui, and Manabu Okumura. 2005. Extracting semantic orientations of words using spin model. In Proceedings of ACL'05, pages 133–140.

[15] Turney, Peter D. and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. 21(4):315–346.

[16] Wiebe, Janyce. 2000. Learning subjective adjective from corpora. In Proceedings of AAAI'00, pages 735–740.

[17] Wiebe, Janyce, TheresaWilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. 30(3):277–308.

[18] Zhao et al. (2014) "Joint Propagation and Refinement for Mining Opinion Words and Targets", IEEE Data Mining Workshop, 2014, pp.417-424.