THE UNIFIED COLLOCATION FRAMEWORK FOR OPINION MINING

YUN-QING XIA[1], RUI-FENG XU[2], KAM-FAI WONG[3], FANG ZHENG[1]

[1]Center for Speech and Language Technologies, Tsinghua University, Beijing 100084, China
[2]Department of Computing, The Hong Kong Polytechnic University, Kowloon, Hong Kong
[3]Department of SE&EM, The Chinese University of Hong Kong, Shatin, Hong Kong
E-MAIL: {yqxia, fzheng}@tsinghua.edu.cn, csrfxu@comp.polyu.edu.hk, kfwong@se.cuhk.edu.hk

Abstract:

Opinion mining is a complicated text understanding technology involving opinion extraction and sentiment analysis. State-of-the-art techniques adopt idea of attribute-driven or sentiment-driven, leading to low opinion mining coverage. This paper proposes the unified collocation framework (UCF) and describes a novel unified collocation-driven (UCD) opinion mining method. The UCF incorporates attribute-sentiment collocations as well as their syntactical features to achieve reasonable generalization ability. Preliminary experiments show that 0.245 on averages improve recall of opinion extraction without obvious loss on opinion extraction precision and sentiment analysis accuracy.

Keywords:

Unified collocation; Opinion extraction; Sentiment analysis; Opinion mining

1. Introduction

E-commerce has created billions of online product reviews and the volume is increasing nowadays. The reviews in turn bring huge advantages to manufacturers and potential consumers. The manufacturers reply on the comments from the experienced costumers to grasp common demands and market trend. For the potential consumers, the comments from the experienced costumers provide important references to their shopping decision. With a large volume of product reviews available on the Internet, people are eager to find reviews on certain products. A popular way to locate relevant product reviews is via searching engine using product name as searching query. Unfortunately, current-searching engines, including Google, cannot provide accurate and friendly service on this task due mainly to two reasons. Firstly, very few returned links are relevant to the input product, which are mostly specifications, news reports, sales advertisements, etc. It is a huge waste of time to click each returned link to check whether a relevant review is included. Secondly, current searching engines suppose that users should take time to go over the links to find useful information. However, some popular products are mentioned in hundreds of reviews. Worst of all, some reviews are very long while only a few opinions on the product are mentioned. As a result, an informed decision is pretty hard to make.

Opinion mining is a text understanding technology in assisting people to locate relevant opinions within a large volume of review collection quickly and friendly. A searching engine based on opinion mining technology shows great potentials to address the aforementioned dilemma. An opinion-mining tool goes over product reviews to extract opinion units and saves them in an opinion database. When user inputs an opinion-searching query, the searching engine extracts product name and attribute(s) from the query, sends a complicated SQL query to the opinion database, and displays the output on the Web interface. The opinion-searching engine arranges information based on opinion rather than document. Therefore information in product reviews can be accessed quickly and friendly. Due to such commercial potentials, opinion-mining research becomes hot nowadays.

Two key tasks are involved, i.e. opinion extraction and sentiment analysis. State-of-the-art opinion mining techniques can be classified into two camps, i.e. attribute-driven methods and sentiment-driven methods. Basic idea of these methods is that they make use of either attributes or sentiment keyword to locate opinion candidates and apply certain opinion patterns (involving attributes and sentiment keywords) to extract sentiment expressions and filter false opinion candidates. One drawback is notable in theses methods. That is, they tend to yield higher *precision* with price of huge *recall* loss as no generalization ability is implied. The problem dues largely to the out-of-vocabulary (OOV) attributes and OOV sentiment keywords being frequently encountered in natural language review text.

Enlightened by collocation theory of natural language, we propose the collocation-driven opinion mining method, in which attribute-sentiment collocations are applied in opinion extraction. However, restrictions within attribute-sentiment collocations are much stronger than that either within attribute-driven methods or within sentiment-driven methods. As a result, precision is improved with price of more *recall* loss. To achieve reasonable generalization, we propose the unified collocation framework (UCF), in which part-of-speech (POS) tags for both attributes and sentiment keywords are incorporated in general collocation framework (GCF). As POS tags can improve generalization ability of attribute-sentiment collocations, *recall* of opinion mining based on UCF can be significantly improved. To avoid over-generalization problem that leads to precision loss, collocations of attribute and sentiment keywords are included in UCF. Preliminary experiments are conducted in this paper to prove our claims.

The rest of this paper is organized as follows. In Section 2, some related works are presented. In Section 3, the unified collocation framework is described. In Section 4, key issues on unified collocation-driven opinion mining method are addressed. Experiment results are presented in Section 5 as well as comparisons and discussions. Finally the paper concludes in Section 6.

2. Related works

Opinion mining research originates from sentiment analysis on natural language text. A great deal of research works has been devoted to sentiment analysis. For example, Pang et al. apply document classification methods such as Naïve Bayes, Maximum Entropy and Support Vector Machines to find document sentiment polarity [1]. Turney and Littman, on the other hand, proposes an unsupervised learning algorithm for inferring semantic orientation by using two search engine-based strategies, i.e. one based on mutual information and the other based on Latent Semantic Analysis [2]. Research on sentiment analysis is important to opinion mining. But there remains another important work, i.e. opinion target extraction.

Mainstreams of opinion mining research can be grouped into two camps, i.e. attribute-driven methods and sentiment-driven methods. Basic idea of these methods is that they make use of attributes or sentiment keywords to locate opinion candidates and apply certain opinion patterns (involving attributes and sentiment keywords) to filter false candidates. Several opinion-mining systems are developed based on this idea. In Web Fountain system, Yi and Niblack make use of POS heuristics to find attribute using likelihood ratio calculation [3]. In FBS system, Hu and Liu apply lexicons and patterns in attribute extraction [4]. In OPINE system, Popescu and Etzioni make use of conceptual analysis to extract attributes [5]. Lou and Yao apply domain ontology to extract attribute from Chinese automobile reviews [6]. In OPINMINE system, Xu et al. apply sentiment lexicon in sentiment keyword extraction and make use of domain ontology to find attributes [7] associated with the sentiment keyword.

Drawback of these methods is addressed as follows. On the one hand, these methods tend to yield high *precision* as very strong conditions are applied. On the other hand, *recall* achieved in these methods is pretty low because of the exact matching. As a result, out-of-vocabulary (OOV) attributes and sentiment keywords frequently encountered in review text are missed in these methods.

A straightforward idea rises in our mind that we may make use of collocation relation between attribute and sentiment and offer them equal priority in opinion mining. Literatures on Chinese collocation assert that attribute and sentiment comply with the form of general collocation [8, 9]. Then the collocation-driven methods come into being, in which attribute-sentiment collocations are applied in opinions extraction. However, much stronger conditions occur within attribute-sentiment collocations, leading to more *recall* loss. To improve generalization ability, we propose the unified collocation framework (UCF), in which part-of-speech (POS) tags for attribute and sentiment keyword in a collocation are incorporated in collocation framework.

3. Unified collocation framework

3.1. Formal definition

The unified collocation is an extension of general collocation. So definition of the general collocation framework (GCF) is given first. In general, GCF is defined as

(GCF) {[*attribute*], [*sentiment keyword*], <*polarity*>}.

Attribute and sentiment keyword in GCF are optional, i.e. with value *NULL*. This reflects two facts: attribute can be implicit and sentiment can be mentioned in context. They are not allowed to be omitted at the same time.

In GCF, *polarity* can be strong positive (2), positive (1), negative (-1), strong negative (-2) or neutral (0). Examples of general collocation are given as follows.

General collocation examples

{外观, 时尚, 1} ({appearance, fashionable, 1})
{*NULL*, 贵, -1} ({*NULL*, expensive, -1})
{运行, 稳定, 1} ({run, stably, 1})

The general attribute-sentiment collocations can be used in opinion mining. However, restrictions within attribute-sentiment collocations are much stronger than either that in attribute-driven methods or in sentiment-driven methods because opinions are now matched in two dimensions, i.e. attributes and sentiment keywords. As word itself has no generalization ability, opinion mining based on general collocation tends to lose more recall in opinion mining.

To improve generalization ability, we propose the unified collocation framework (UCF), in which part-of-speech (POS) tags for attributes and sentiment keywords are incorporated. UCF is defined as

(UCF) {[a*ttribute*], <*pos_attribute*>, [*sentiment_keyword*], <*pos_sentiment_keyword*>, <*polarity*>}

in which *pos_attribute* represents POS tag of the attribute and *pos_sentiment_keyword* POS tag of the sentiment keyword. In our case, POS tag is obtained using ICTCLAS [10]. Examples of unified collocation are given below.

Unified collocation examples

{外观, *noun*, 时尚, *adjective*, 1}
({appearance, *noun*, fashionable, *adjective*, 1})
{*NULL*, *noun*, 贵, *adjective*, -1}
({*NULL*, *noun*, expensive, *adjective* -1})
{运行, *verb*, 稳定, *adjective*, 1}
({run, *verb*, stably, *adjective*, 1})

(Remark: "稳定(stably)" is considered as an adjective in Chinese.)

POS tags within the unified collocations release matching restrictions thus improve generalization ability of attribute-sentiment collocations. As OOV problem can be addressed appropriately, recall of opinion mining based on UCF is significantly improved. In order to avoid over-generalization problem that leads to precision loss, collocation of attribute and sentiment keyword are included in the unified collocation framework.

3.2. Parameters

In UCF, three types of parameters should be estimated, i.e. attribute-sentiment collocations, POS tags and some statistics. These parameters are obtained on OPINMINE corpus [11].

Attribute-Sentiment Collocations

Annotation in OPINMINE corpus involves opinion features such as *attribute*, *sentiment keyword*, *modifier*, *negation*, *polarity*, *degree*, etc. It can be safely assumed that all opinions within OPINMINE corpus are annotated correctly. To extract attribute-sentiment collocations, we merely use *attribute*, *sentiment keyword* and *polarity* in each annotation.

Note that negations might be used in opinions. So polarity of the sentiment keyword should be opposite to that of the opinion if negation exists. As attributes and sentiment keywords are already annotated in OPINMINE corpus, no collocation extraction tool is demanded to identify the attribute and sentiment keyword collocation.

POS tags

We apply ICTCLAS on the sentence that each attribute-sentiment collocation resides and then extract POS tags for the attribute and sentiment keyword, respectively. This guarantees that the POS tags are grammatically correct.

Some attributes are some multiple-word expressions. So ICTCLAS segments them into a few words and assigns each word a POS tag. For these cases, we define simple rules to merge the POS tags so that each attribute hold only one POS tag. As attributes are normally *nouns* and *verbs* in product reviews, the following two rules are defined to merge POS tags in our method.

(1) [*variable*] + *noun* → *noun*
(2) [*variable*] + *verb* → *verb*

The first rule merges a *noun* and any preceding POS tag(s) into one *noun* tag. Similarly, the second rule merges a verb and any preceding POS tag(s) into one *verb* tag.

Statistics

With OPINMINE corpus, three types of statistics are produced for each unified collocation.

(1) Frequencies of <attribute string, sentiment keyword string> collocations
(2) Frequencies of <attribute string, sentiment keyword POS tag> (unified) collocations
(3) Frequencies of <attribute POS tag, sentiment keyword string> (unified) collocations

Statistics of type (2) and type (3) can group the attributes and sentiment keywords, respectively, in terms of POS tags. As one POS tag represents a set of words, appropriate generalization ability is thus obtained in the unified collocations. Note that statistics on <attribute POS tag, sentiment keyword POS tag> are not considered in our method. So the over-generalization problem can be controlled.

4. Opinion mining with UCF

4.1. Architecture

The UCF-based opinion mining method comprises of two modules, i.e. learning component and working component. The learning component goes through the OPINMINE corpus to find all general collocations and produces statistics of type (1). Then a POS tagger is applied to assign a POS tag to each attribute and each sentiment keyword in a sentence. At last, the unified collocations are obtained together with statistics of type (2) and type (3).

In the working component, ICTCLAS is first applied to split the review text into words and to assign each word a unique POS tag. Then the general attribute-sentiment collocations are applied to match known opinions and the unified collocations to find OOV opinions. The working component also involves extraction of modifier and negation, which is important to determine sentiment polarity of each OOV opinion. A modifier list and a negation list are handcrafted to do this work.

Key issues in OOV opinion extraction are two: (1) OOV attribute & sentiment keyword extraction; and (2) OOV sentiment keyword analysis. We address the two issues in the following sections.

4.2. OOV attribute & sentiment keyword extraction

Observation on digital camera reviews discloses that attributes are expressed flexibly in natural language. For example, *LCD* of digital camera is expressed by “屏幕”, “显示屏”, “LCD”, etc. in Chinese. Manufacturers usually provide standard terms in user manuals. But due to language habit and education background, people use numerous alternatives to represent one term. So it is normal that OOV attributes occur constantly in product reviews written by different users. Situation for OOV sentiment keywords is worse because sentiment is expressed much more flexibly. It is not surprising that opinion-mining methods based on general collocations yield pretty low recall.

The unified collocation framework provides feasible mechanism in handling the OOV attributes and OOV sentiment keywords. Matching algorithm for unified collocation framework is an extension of that for general collocation. Three types of collocation matching are considered.

(1) Attributes and sentiment keywords match respectively in terms of string.
(2) Attributes match in terms of POS tag and sentiment keywords match in terms of string.
(3) Attributes match in terms of string and sentiment keywords match in terms of POS tag.

Collocation matching of the first type is exactly the same one in general collocation. But the latter two matching types are new. We find that matching of the second type is helpful to find OOV attributes using known sentiment keywords and matching of the third type provides a way to find OOV sentiment keywords using known attributes. We apply statistics on unified collocations to improve performance of opinion extraction.

Note that OOV attributes and OOV sentiment keywords extracted in this stage are just candidates. To exclude the false candidates, an SVM classifier is trained on OPINMINE corpus using opinion sentences as positive samples and non-opinion sentences as negative samples. We use word bi-grams as classification features.

4.3. OOV sentiment keyword analysis

Opinion mining system needs to figure out sentiment polarity of each OOV sentiment keywords. In our method, this task is performed based on the context coherency assumption. Kanayama and Nasukawa conclude that the polarities of opinions in intra-sentential context and inter-sentential context are the same unless an adversative conjunction such as “but” and “however” is found to connect the two opinions [12]. Sentiment polarity of each OOV sentiment words is determined as follows.

(1) Locate the OOV opinion O_{OOV} containing the OOV sentiment word W_{OOV}.
(2) Locate O_{OOV}'s neighboring opinion O_l and read its sentiment polarity P_l; If O_l is not found, return UNKNOWN.
(3) Locate adversative conjunction c that connects the two opinions.
(4) If c is found, sentiment polarity of O_{OOV}: P_{OOV_O} =-P_l; otherwise P_{OOV_O} =P_l.
(5) Match negation *neg* within O_{OOV}.
(6) If *neg* is found, sentiment polarity of W_{OOV}: P_{OOV_W} =-P_{OOV_O}; otherwise P_{OOV_W} = P_{OOV_O}.

5. Evaluation

5.1. Experiment setup

Data Description

We use OPINMINE corpus to evaluate our opinion mining method. OPINMINE corpus contains 1,000 annotated reviews on digital camera covering 9,207 opinions on 59 attributes in 6,331 sentences. We use 80%

reviews randomly selected from OPINMINE corpus as training samples and the rest 20% reviews as test samples.

Evaluation Criteria

We evaluate two tasks separately, i.e. opinion extraction and sentiment analysis. For opinion extraction evaluation, we adopt *precision* (p), *recall* (r) and *f-1* score (f), in which *precision* is defined as percentage of correct opinions within all extracted opinions and *recall* percentage of correct opinions within all annotated opinions. For sentiment analysis evaluation, we adopt *accuracy* (a), which represents percentage of opinions assigned correct sentiment polarity within all extracted opinions.

5.2. Experiment I: opinion mining methods

Four opinion-mining methods are implemented in this experiment, i.e. attribute-driven (AD) method, sentiment-driven (SD) method, general collocation-driven (GCD) method and our unified collocation-driven (UCD) method. Details for the implementations are given as follows.

Attribute-Driven (AD) method

The AD method first learns an attribute list and some attribute-driven opinion rules (AOR) from OPINMINE corpus in. The AOR rules comply with the following form.

(AOR) <*sentiment word attribute*, *conditions*> → *opinion*

The AOR rules assert that, provided that an attribute exists, then an opinion is found if a sentiment word also exists and the *conditions* are satisfied.

The AD method uses the attribute list to locate opinion candidates and applies AOR rules to filter false opinions. Some post-processing steps are conducted in AD method, i.e. detection of modifiers and negations, determination of opinion boundary and induction of opinion sentiment polarity. A modifier list and a negation list are applied to find modifiers and negations, respectively. Sentiment polarity of an opinion is inferred as follows.

(1) Given sentiment polarity P_k of sentiment keyword w;
(2) If negation *neg* is found, sentiment polarity of the opinion $P_o = -P_k$; otherwise, $P_o=P_k$;
(3) If modifier *mod* is found, sentiment polarity of the opinion $P_o=P_o*2$.

The opinion boundary is determined as the minimum window that covers all opinion elements including attribute, sentiment keyword, modifier and negation.

Sentiment-Driven (SD) method

Similar to AD method, the SD method first learns a sentiment keyword list and some sentiment-driven opinion rules (SOR) from OPINMINE corpus in. Form of SOR rules is given below.

(SOR) <[*attribugte*]|*sentiment word*, *conditions*>→ *opinion*

Different from AOR rules, SOR rules are sorted according to *sentiment keyword*. Besides, attribute in SOR rules are optional. That is, review text containing no known sentiment keyword will be ignored while attribute is allowed to be omitted. Post-processing steps are similar to that in AD method.

GC-Driven (GCD) method

The GCD method holds similar matching nature as in AD and SD methods. The difference lies in that matching process becomes more restricted. That is, attribute list and sentiment keyword list are learnt from OPINMINE corpus and applied at the same time to locate opinion units. Obviously, more true opinions are missed in this method. Below is form of the collocation-driven opinion rules (COR).

(COR) <*sentiment word*, *attribute*, *conditions*> → *opinion*

The COR rules require that *attribute* and *sentiment keyword* appear at the same time and satisfy *conditions*. Post-processing steps in GCD method are also similar to that in AD method.

UC-Driven (UCD) method

The UCD method is an extension of the GCD method. As described in Section 4, GCD method first extracts general collocations using the first type of statistics (see Section 3.2). Another two types of statistics on unified collocations are applied to extract the OOV opinion candidates. The false OOV opinion candidates is excluded by SVM^{light} [13]. In what follows, the context coherency assumption is made to help determining sentiment polarity of each OOV sentiment keyword. Finally, sentiment polarity of opinion can be inferred by considering sentiment polarity of sentiment keyword, modifier and negation in this opinion.

We train the four opinion mining methods on the training samples and test them on the test samples. Experimental results are presented in Table 1.

Table 1. Results in opinion extraction and sentiment analysis by four opinion mining methods.

Method	Opinion Extraction			Sentiment Analysis
	p	r	f	a
UCD	0.879	0.802	0.839	0.842
GCD	0.903	0.476	0.623	0.844
AD	0.872	0.573	0.692	0.843
SD	0.857	0.623	0.722	0.846

Table 1 shows that UCD outperforms the other three methods by 0.245 on average on recall in opinion extraction. This is because many OOV opinions are extracted using the unified collocations. People may worry about precision loss when the unified collocations are applied. Table 1 shows that precision achieved in UCD is close (+0.0167) to that achieved in GCU, AD and SD. A conclusion can thus be made that UCD method improves recall of opinion extraction significantly without precision loss.

Similar worry is on accuracy in sentiment analysis. As many OOV opinions are extracted, will accuracy in sentiment analysis drop? Encouragingly, Table 2 shows that UCD achieves close accuracy (-0.023). This owes largely to the context coherency assumption in determining sentiment polarity of OOV opinions.

5.3. Experiment II: SVM filtering

This experiment seeks to investigate contribution of the SVM classifier in opinion filtering. We remove the SVM classification module from the opinion mining method and run the UCD method again. Experimental results are presented in Table 2.

Table 2. Results by UCD method with SVM classifier.

UCD Method	Opinion Extraction			Sentiment Analysis
	p	r	f	a
SVM	0.879	0.802	0.839	0.851
No SVM	0.703	0.785	0.742	0.679

Table 2 shows that precision of UCD method drops significantly (by 0.176) when SVM classifier is used. This is because many false opinions are extracted using the unified collocations. Interestingly, recall drops much less than precision, i.e. by 0.017. This is because the SVM classifier assigns correct labels to most (95% in our case) of the true opinions so that they remain in the answer sheet after SVM classifier is applied. Situation for accuracy of sentiment analysis is similar to precision in opinion extraction. It drops by 0.172 when SVM classifier is not used. The reason for such a drop is that two types of errors remain in the final answer sheet, i.e. opinions whose sentiments are incorrectly analyzed and the false opinions. It is again justified that SVM classifier is important for our opinion mining method.

6. Conclusion and future works

This paper investigates a novel unified collocation-driven (UCD) opinion mining method. Compared to the attribute-driven method, sentiment-driven method and general collocation-driven method, the UCD method exhibits reasonable generalization ability. As showed by the experimental results, 0.245 on average improves recall in opinion extraction without obvious loss on opinion extraction precision and sentiment analysis accuracy.

Nevertheless, research work reported in this paper is just preliminary output in our opinion-mining project. There are a few future works though achievements in current stage are encouraging. Firstly, method reported in this paper can be used in opinion lexicon expansion, as it is helpful to find OOV attributes and OOV sentiment keywords. So the first future work is to design algorithm for opinion lexicon expansion. Secondly, the UCD method is deemed applicable in other languages. So the second future work is to develop an opinion mining system for English product reviews.

Acknowledgements

This paper is partially supported by Tsinghua University in project COMVERSE (No. JC2007049).

References

[1] B. Pang and L. Lee and S. Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. Proc. of EMNLP-02, pp.79-86.

[2] P. D. Turney and M. L. Littman. 2002. Unsupervised learning of semantic orientation from a hundred-billion-word corpus. Technical Report EGB-1094, National Research Council Canada.

[3] J. Yi and W. Niblack. 2005. Sentiment Mining in WebFountain. In Proc. of ICDE-2005, pp.1073-1083.

[4] M. Hu and B. Liu. 2004. Mining and summarizing customer reviews. Proc. of KDD-04. pp. 168-177.

[5] A.M. Popescu and O. Etzioni. 2005. Extracting Product Features and Opinions from Reviews. Proc. of EMNLP-05, pp. 339-346.

[6] D. Lou and T. Yao. 2006. Semantic polarity analysis and opinion mining on Chinese review sentences (in Chinese). Journal of Computer Applications, 2006 Vol.26 No.11 P.2622-2625.

[7] R. Xu, K.F. Wong and Y. Xia. 2007. Opinmine – Opinion Analysis System by CUHK for NTCIR-6 Pilot Task. Proc. of NTCIR-6.

[8] Q. Lu, Y. Li, and R.F. Xu. 2003. Improving Xtract for Chinese Collocation Extraction. Proc. of NLPKE-03.

[9] S. Wang, J. Yang and W. Zhang. 2006. Automatic Acquisition of Chinese Collocation (in Chinese).

Journal of Chinese Information Processing, 2006 Vol.20 No.6 P.31-37.

[10] Z. Zhang, H. Yu, D. Xiong and Q. Liu. 2003. HMM-based Chinese Lexical Analyzer ICTCLAS. 2nd SIGHAN workshop in ACL-03, pp. 184-187.

[11] R. Xu, Y. Xia, K.-F. Wong and W. Li. Annotation Opinions in Chinese Product Reviews. Will submit to International Journal of Computational Linguistics and Chinese Langauge Processing.

[12] H. Kanayama and T. Nasukaw. 2006. Fully Automatic Lexicon Expansion for Domain-oriented Sentiment Analysis. Proc. of EMNLP-06, pp.355–363.

[13] T. Joachims. 2001. A Statistical Learning Model of Text Classification with Support Vector Machines. Proc. of SIGIR-01, New Orleans. ACM Press, New York (2001).