Automatic Summarization in Chinese Product Reviews

Lizhen Liu, Wandi Du*, Hanshi Wang, Wei Song
Information and Engineering College, Communication Systems, Capital Normal University,
Beijing 100048, China
*Corresponding author, e-mail: dwandi@163.com

Abstract

With the increasing number of online comments, it was hard for buyers to find useful information in a short time so it made sense to do research on automatic summarization which fundamental work was focused on product reviews mining. Previous studies mainly focused on explicit features extraction whereas often ignored implicit features which hadn't been stated clearly but containing necessary information for analyzing comments. So how to quickly and accurately mine features from web reviews had important significance for summarization technology. In this paper, explicit features and "feature-opinion" pairs in the explicit sentences were extracted by Conditional Random Field and implicit product features were recognized by a bipartite graph model based on random walk algorithm. Then incorporating features and corresponding opinions into a structured text and the abstract were generated based on the extraction results. The experiment results demonstrated the proposed methods out preferred baselines.

Keywords: *bipartite graph, implicit features, conditional random field, random walk algorithm, automatic summarization*

1. Introduction

Nowadays, the degree of activities in Chinese online market is still high and it's time-consuming for customers to read a flood of comments. Currently, a few typical Chinese e-commerce websites have done several inductive statistics, for example Tmall.com gives phrases and its quantity to other users for giving reference, Amazon.cn gives star ratings to goods based on user reviews, but all of these are coarse-grained extraction, resulting in interpreting out of context which are limited to objectively understand reviews for users, for example some extracted labels can only represent the experience of a certain people, and some phases express incompletely [1]. When the number of users is large, the problem will be more prominent. Therefore, generating summaries accurately and concisely has great significance to analyze and conduct product reviews, and it will improve the efficiency of online shopping and help others to obtain important information quickly.

The technique of automatic summarization has developed rapidly in recent years. The summary generation can be divided into extraction summarization and generation summarization. By selecting the sentences in the original text to form summary, the extraction summarization usually estimates the sentences in the document according to pre-defined feature sets or machine learning algorithms, then the sentences with high scores are output as summary [2]. The generation summarization includes words and phrases that are not occur in the original text, and typically based on entity information and compression techniques and so on. Due to the generation summarization is still in its infancy and has huge challenge to the natural language processing technology; there has a considerable distance for generating a practical summary [3]. Hence this paper focuses on the former method.

For the extraction summarization, the effect of the comments opinion mining will directly affect the quality of the generated summary. Hu and Liu [8] present two kinds of features in product features mining, namely explicit and implicit features. Many people are aware of the existence of implicit features in [4, 5], whereas the existing methods for mining implicit features are not very mature. Su and Xiang [6] mainly use Point-wise Mutual Information (PMI) to associate semantic analysis with product features and opinion words which match probability in training set. In [7], a co-occurrence association rule mining (CoAR) algorithm is proposed to select implicit product features. But above all sorts of implicit product features extraction

Received September 15, 2016; Revised December 7, 2016; Accepted January 10, 2017

methods can be evaluated only for special words, it is not ideal for general words. Therefore, our study focuses on implicit opinion mining and in order to get high-quality summaries.

Many scholars have put forward a variety of summary methods since Luhn [9] defined automatic summarization in 1958. Reference [10] used hierarchical clustering for documents, and then calculated the relevance of text units by using the LexRank to extract important sentences from each category. The Interdependent Latent Dirichlet Allocation (ILDA) model was used in [11] which took the shallow semantics of the documents into account but ignored the text structure information. Sequence annotation model was used to solve this problem in [12], in which Hidden Markov Model(HMM) with less independence assumptions was used while HMM had limited ability to describe features of the relationships between sentences. Reference [13] combined Hidden Topic Markov Model with LDA topic model, breaking the theme independent hypothesis, but ignoring the semantic synonymy and relevance. Multi-document summary was built based on sentence distribution in [14] which calculated the frequency of occurrence of words forming the sentences. Clustering approach was used in [15] to extract information but ignored the readability of the summary. In our paper, we train models automatically by using the machine learning algorithms and the given feature sets. Conditional Random Field (CRF) and semi-supervised learning method are used to extract features opinions and “feature-opinion” pairs in comments. The methods of this paper are suitable for regular sentences and short comment texts, needing to label parts of the training corpus manually. Based on the existing results of word segmentation, the semi-supervised learning method is used to extract features and opinions in this paper. Besides, the paper combines the features of the merchandising function, capability and components which are gained from comments to construct a bipartite graph, then the highest probability implicit features would be computed by random walk algorithm. Thus, the summary will be generated based on the lowest cost value calculated by the probability distributions of pairs.

In general, our contributions are the following:

(1) A novel model is used to solve the problem of implicit features extraction, and verify the feasibility of this model under some tests.

(2) We experimentally evaluate our methods against with some existed methods on feature extraction for both precision and recall, and current techniques on automatic summarization for ROUGE.

(3) We focus on product reviews and get summary sentences according to the probability distribution of product “feature-opinion” pairs.

The remaining parts of this paper are organized as follows: Section 2 proposes related knowledge and our approach; the experimental results are presented, evaluated and discussed in Section 3; Section 4 presents our conclusions and future work.

2. Model Design

A number of studies shown that it's essential to use a special text processing technology for web produce comments with brief text, diverse language, sparse data and high in noise, which is different from traditional documents [16-18]. Therefore, we propose the approach in this chapter mostly considering opinion mining. The main content of this chapter are “feature-opinion” pairs identification and collocation, implicit features extraction and automatic summarization. System flowchart is shown in Figure 1.

Product reviews are climbed from e-commerce sites and all reviews can be seen as a document in which each sentence is a comment. In order to obtain the high-quality and reliable experimental data, we firstly proceed review datasets preprocessing, including segmentation, denoising which covers comments emotions, special characters, or off-topic sentences (for example “I am very happy to receive the goods”, “This style is what I want”) and so on.Because of the particularity of Chinese grammar, we need to do word segmentation using ICTCLA segmentation system. Training data is labeled by HowNet [19] and trained by models in order to extract product features and opinions, and comment sentences can be divided into explicit sentences and implicit sentences according to the extraction results. Then we cluster “feature-opinion” pairs in explicit sentences to construct a bipartite graph, using random walk algorithm to calculate the probability of implicit features and achieving the extraction of “feature-opinion” pairs. Finally, we provide the summary for users.

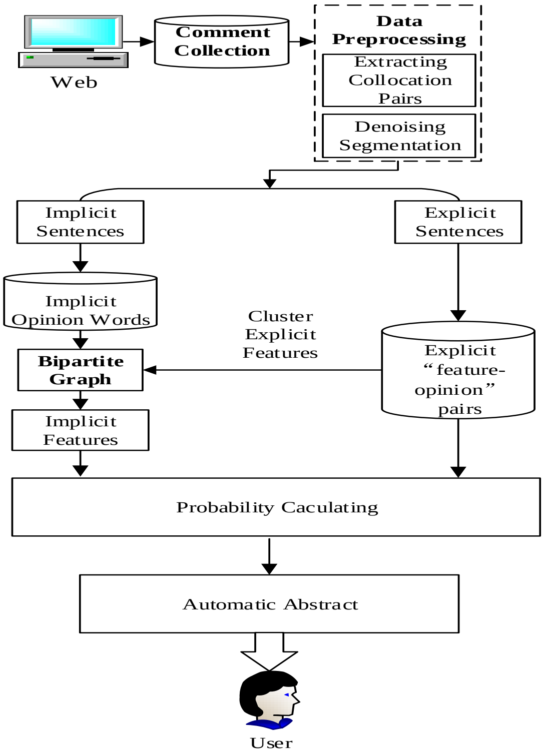

Figure 1. System Flowchart

2.1. Collocation Extraction Based on CRF

The main content of this section is to extract "feature-opinion" pairs based on the Conditional Random Field. As CRF applied in Chinese word segmentation, sentiment analysis and part-of-speech tagging, we transform the problem of collocation extraction into the sequence annotation task.

Collocation extraction is defined as extracting commodity features and opinions which are expressed as <product features, opinion words> in the comment text, like the <pixel, high> in the comment "*苹果手机的像素很高*" ("The pixel of iPhone is high."). The process of identifying features and opinions can be seen as under the condition that input a string of words $w_1, w_2, w_3, \cdots, w_n$, the maximum probability labeled sequence $L_1, L_2, L_3, \cdots, L_n$ is outputted. Here we introduce seven mark symbols $\{B-F, I-F, E-F, B-O, I-O, E-O, OFF\}$, in which "$B-F$" represents the initial word describing property features, "$I-F$" represents the intermediate term describing property features, "$E-F$" represents the end term describing property features, "$B-O$" represents the opinion word which adjacent to the feature, "$I-O$" represents the intermediate term of opinion word, "$E-O$" represents the end term of opinion word, "OFF" represents the unrelated word.

Choosing a good feature can greatly improve extraction performance, thus it's important to construct the feature template for labeling sequence based on CRF. Features used in our model including word feature, part of speech feature, position feature, interdependent syntactic relation feature, whether is an explicit comment sentence. After building feature templates and training model using training corpus, we get collocation extraction model and mine the collocation pairs after entering new corpus. For example, the results of labeling and training the sentence "*手机很精致，屏幕显示很细腻，音量有点低，WiFi 信号接受能力很差。*" ("This phone is very delicate and the screen display is fine, but with a little low volume and poor WiFi signal reception.") are shown in Table 1. The example has five columns, which represent "word feature", "part of speech feature", "position feature", "interdependent syntactic relation feature", "whether is an explicit comment sentence". All elements of the model in which we can obtain four groups of collocation as <*手机，精致*>、<*屏幕显示，细腻*>、<*音量，低*>，< *WiFi 信号接受*

能力，差> (<cell phone, delicate>, <screen display, fine>, <volume, low>, <WiFi signal reception, poor>).

Table 1. Processing Results of Example Sentence

手机	n	0	1	B-F
很	zg	1	1	OFF
精致	a	2	1	B-O
，	x	3	1	OFF
屏幕显示	n	4	1	B-F
很	zg	5	1	OFF
细腻	a	6	1	B-O
，	x	7	1	OFF
音量	n	8	1	B-F
有点	n	9	1	OFF
低	a	10	1	B-O
，	x	11	1	OFF
WiFi	eng	12	1	B-F
信号	n	13	1	I-F
接受	v	14	1	I-F
能力	n	15	1	E-F
很	d	16	1	OFF
差	a	17	1	B-O
。	x	18	1	OFF

Different words or phrases are often used to describe the same feature by customers, such as words “facade”, “external”, “aspect” are used to describe the appearance of the mobile phone. In order to make similar features have the same description, we cluster n features of “feature-opinion” pairs $\{< f_1, o >, < f_2, o >, \cdots, < f_n, o >\}$ for matching each opinion word o. Our method of clustering is based on the paper [20] which automatically identify some labeled examples by semi-supervised method, then unlabeled features are assigned to a cluster using naive Bayesian based on EM formulation [21]. When EM converges, the classification labels of all the attributive words give us the final grouping. Thus the implicit feature extraction problem is turned to a classification problem.

2.2. Implicit Features Extraction

It's not hard to find explicit features in the reviews, but the number of them is limited. Based on CRF model we can extract explicit features accurately whereas extracting implicit features using rule-based methods with full coverage is difficult. For the implicit features extraction problems, we mine implicit features via calculating the results of random walk algorithm and the probability of candidate features.

In this section, our main task is to extract implicit features. We utilize features and opinions which are collected previously to build a graph. The bipartite graph $G = (F \cup O, E)$ composes of candidate features and opinions, here $F = \{f_1, f_2, \cdots, f_i\}, i = 1,2, \cdots, m$ represents candidate features and seed features, $O = \{o_1, o_2, \cdots, o_j\}, j = 1,2, \cdots, n$ and represents opinions. The edge of E connects the vertex F and O, W_{ij} is the edge weight of connecting the vertex f_i and o_j in the weight matrix $W = \{w_{ij}\}_{m\times n}$, implicit features are represented by b_j and the seed set of F is denoted by F_s where the feature belonging to the extraction feature b_j is signed as a positive example and the others are signed as negative examples. According to graph G and seed feature set F_s, our algorithm calculates the probability of implicit feature b_j assigned to the candidate feature set $\{F - F_s\}$.

Taking some cellphone reviews for example, the more co-occurrence of product features and opinion words, the greater relevance between them. As shown in Figure 2, the opinion word “very big” is associated with the features “screen” and “memory”, whereas the connection with “memory” is closer than “screen” and the edge weight will be higher, the feature described by “very” is more likely to “memory”. A small amount of artificial features can be seen as the seed set based on our graph model, we can obtain implicit features from corresponding candidate features using random walk model with opinions in the implicit sentences.

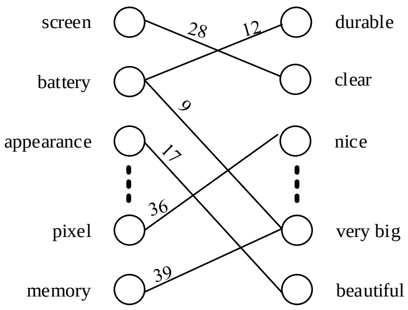

Figure 2. The Bipartite Graph of Some Cellphone Reviews

The size of state matrix $X(t)$ is $m \times 2$ and the number of matrix iterative is defined by t. When $t = 0$, the initial state of the candidate feature set is denoted by $X(0)$, and when $t = c$ iteration will stop, $X(c)$ represents the final state of all candidate features via random walk algorithm. The probability of feature f_i belongs to cluster category b_j is expressed by each entry $X_{i,j}$ in the matrix X which is non-negative and can be calculated as shown in (1).

$$X(t) = \lambda HX(t-1) + (1-\lambda)X(0), \lambda \epsilon (0,1) \quad (1)$$

Here, $H = D^{-\frac{1}{2}} R D^{-\frac{1}{2}}$ is a normal matrix in which diagonal matrix D is to normalize relational matrix R and each diagonal entry d_{ii} in matrix D is the sum of each element in matrix R, whereas others in matrix D are zero. The function of λ is to adjust the degree of depending on the initial state or bipartite graph when distribute candidate features. We define $j = 0,1$ when $j = 0$ is the first column of X corresponding the category b_j (positive example) and $j = 1$ is the second column of X corresponding the non-implicit features (negative examples). When reaching the final state, the probability of each feature f_i belongs to the category b_j is calculated by $P(b_j|f_i)$ as shown in formula (2):

$$P(b_j|f_i) = \frac{X_{i,j}}{X_{i,0}+X_{i,1}} \quad (2)$$

For the above definition, this paper uses random walk algorithm to extract implicit features is shown in Table 2.

Table 2. Random Walk Algorithm Based on Bipartite Graph

Algorithm
Input: weight matrix W, category b_j, seed word set F_s, candidate set $\{F - F_s\}$ **Output**: $P(b_j\|f_i)$ 1. $R = WW^T$, $H = D^{-\frac{1}{2}}RD^{-\frac{1}{2}}$. 2. Initialize X by $X(0)$. 3. Repeat. 4. $X(t) = \lambda HX(t-1) + (1-\lambda)X(0), \lambda\epsilon(0,1)$. 5. Until $X(t)$ converges to $X(c)$.

Our algorithm firstly build the relational matrix $R = WW^T$ between the candidate features according to weight matrix W to obtain diagonal matrix D and structure normal matrix

$H = D^{-1/2} R D^{-1/2}$. For the state matrix initialization, there are three cases: we set $X_{i,1} = 1$ and $X_{i,0} = 0$ if f_i is the positive example of F_s; $X_{i,1} = 0$ and $X_{i,0} = 1$ if f_i is the negative example of F_s; $X_{i,1} = 0$, $X_{i,0} = 0$ if f_i belongs to $\{F - F_s\}$. Until $X(t)$ convergence to the state $X(c)$ after iterative calculation, and finally the probability of each f_i belonging to the category b_j is calculated by $\frac{X_{i,j}}{X_{i,0}+X_{i,1}}$, we believe that the word with the highest probability is the implicit feature related to the opinion word according to the probability is arranged from high to low.

2.3. Automatic Summarization

After opinion mining, we need to extract some candidate sentences which are related to products with most keywords, and then calculate the importance of each sentence. The process of generating summaries can be divided into three steps:

(1) Training the CRF model, extracting keywords and the collocation of them;

(2) Calculating the probability distributions of “feature-opinion” pairs;

(3) Comparing the probability distributions of the comment sentences with the pairs, and extracting the candidate sentences.

In this paper, we calculate the probability distribution of “feature-opinion” pairs based on CRF model and bipartite graph, supposing that summaries have the similar probability distribution with high frequency pairs. Calculating the probability distribution of the comment sentence is based on the collocation of product features and opinions. The probability distribution of the comment sentence S which has “feature-opinion” pair Y is calculated as:

$$P(Y|S) = \frac{a_{\square int} \times \sum_{k=1}^{n} Sim(S_{ij}, S_{ik})}{len(S)}. \tag{3}$$

Where n is the number of sentences in pair Y and $\sum_{k=1}^{n} Sim(S_{ij}, S_{ik})$ is the similarity sum of the comment sentence and other sentences in pair Y, which reflects the representative between the comment sentence with pair Y. The higher value of it means the more information and more representative the comment sentence has. The value of $a_{\square int}$ is determined by some prompt words in the sentences like “*我认为*”(“I think”), “*虽然……但是……*” (“not only……but also……”) and so on. The more words like these the comment sentences has, the higher score it will has. Since long sentences can be easily recognised, we use the $len(S)$ which is the sentence length using the word as the unit to eliminate the perference of long sentences.

The similarity of the comment sentence and the corresponding pair is showed by KL divergence which is calculated as:

$$D_{KL} = \sum_i P(Y|S) \times log \frac{P(Y|S)}{P(Y|X)} \tag{4}$$

Here, P and Q are probability distributions. When the KL divergence is lower, the difference between the comment sentence and the corresponding pair will lower, and the degree of similarity will higher, that is the lowest $D_{KL}(P(Y|S) \| P(Y|X))$. Since we select the sentences as summary sentences with the minimum KL divergence value and the maximum $P(Y|X)$ value, the cost of generating summary sentences is calculated as shown in (5) in order not to be bound by the KL divergence value:

$$cost = |P(Y \vdash |S \dashv\ ) - S(D\_KL)| \tag{5}$$

Where $S(D\_KL\ ) = 1/(1 + e\hat{}(-D\_KL\ )\ )$ is the sigmoid function of D_KL. Finally, the text summary is generated by extracting the sentences with the lowest cost value.

3. Results and Analysis

We conduct the experiments based on the approach we proposed. The experiment results and analyses are as follows.

3.1. Experimental Data

In this paper, the 121790 pieces of comments which are crawled from three Chinese e-commerce sites are adopted as experimental data in two areas including 79855 pieces of comments from mobile phones and 41935 pieces of comments from computers. Via observing the corpus of information, it can be concluded that most of the syntactic structure in the experimental data are short texts, then the comments are segmented by the Chinese punctuation, which leads to 368963 pieces of comment clauses. And after eliminating some irrelevant comment sentences, we deal with the remaining 311870 clauses. In this paper, 200 pieces of comments from mobile phones and computers respectively are manually selected as the set, which contains 100 pieces of explicit comments and 100 pieces of implicit comments.

3.2. Experimental Results and Analysis

This paper uses the accuracy and recall as the evaluation criteria, we extract explicit features and opinions as well as their collocation, comparing the results with Hu and Liu's research in [8] it's shown in the Table 3. Hu's approach is defined as Method One.

Table 3. Explicit Extraction Results Comparison

		Phone			Computer	
Methods	Precision	Recall	F-measure	Precision	Recall	F-measure
Method One	81.5%	67.5%	73.8%	63.4%	70.3%	66.7%
CRFs	90.3%	75.4%	82.2%	83.1%	71.2%	76.7%

Hu and Liu represent that the more important commodity features are, the higher frequency they have. Thus, the association rules are used to extract the high-frequency terms and noun phrases to mine commercial features according to setting the text window and extracting non-frequent features depending on the adjective collocations around the frequent features. This method is easy and efficient, but the effect partly lies on the selective correlation of frequent item sets. Both of the extraction results of F-value from the two areas mobile phones and computers are lower than ours. Because features and opinions are associated in comments where the speech tags are completed based on CRF model. The higher F-value can be gotten when we deal with the sentences which are short sentences and strong regularity comment corpus.

Convergence probability has been calculated after several iterations based on random walk algorithm, in our experiment the iterative time$t = 5$. This experiment investigates the accuracy of “mobile phone” and “computer” these two kinds of goods when λ within the scope of different values from 0.1 to 1.0. As shown in Figure 3, it describes the accuracy of the first 100 results from two types of commodity evaluation sets. With the in cease of λ, the accuracy of two types of comments changes gradually from high to low, reaching a peak at a certain point. From the Figure 3 we can see that when$\lambda = 0.63$, relatively high accuracy of extracting implicit features on both two types of comments has obtained.

The mean absolute error (MAE) is used to measure the accuracy of implicit feature extraction in our experiment, equals to the difference of implicit features extracted by machine identification and human annotation, which is calculated as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|M_i - A_i|. \tag{6}$$

Where M and A respectively represent the implicit features extracted by machine recognition and by human annotation, the number of implicit features is denoted by n. The higher value of MAE represents the lower extraction quality or vice versa. Comparing MAE results with PMI [6] algorithm and CoAR [7] algorithm are shown in Figure 4 and the values of Precision, Recall, F-measure in three methods are shown in Table 4.

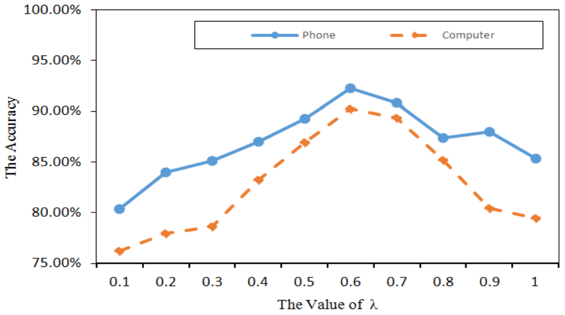

Figure 3. The Accuracy Comparison of Extracting Implicit Features Using Random Walk Algorithm Based on A Bipartite Graph At Different Values of λ

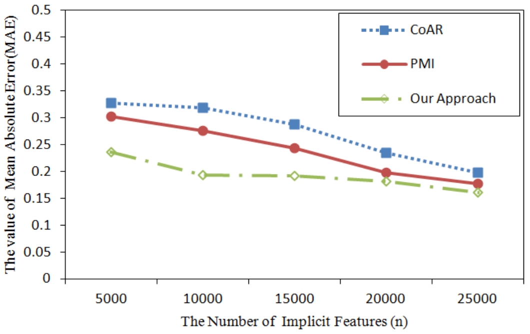

Figure 4. The MAE Value of Three Methods on Implicit Features Extraction

Table 4. The Comparison of Implicit Feature Extraction

	Phone			Computer		
Methods	Precision	Recall	F-measure	Precision	Recall	F-measure
CoAR	76.29%	72.71%	74.46%	70.59%	69.11%	69.84%
PMI	81.34%	79.51%	80.41%	79.16%	77.31%	78.22%
Our Approach	90.33%	85.62%	87.91%	86.49%	80.42%	83.34%

In the experiment, we find that product features modified by PMI algorithm and CoAR algorithm with fixed category, like through “便宜”(cheap), “实惠”(benefit) can get the appropriate product feature “price”, but for some strong generality words, such as “不错”(nice), “一般”(just so so) etc. are treated unsatisfactory in effect, because these general opinions can be used to modify almost all features. The proposed method in dealing with these opinion words has achieved good results, the MAE values of our method are lower than other two methods. Moreover, we also find that the precision and recall blended in implicit features are higher than extraction results that only considering explicit features.

The ROUGE [22] automatic evaluation tools are used to analyze and evaluate the experiment results of automatic summarization. In this paper, the methods in [10-13] are used as the baselines, and the experiment results show that the generated summary based on a bipartite graph and the CRF model are better than baselines not only in the key information coverage index (ROUGH-1) but also in the summary readability evaluation index (ROUGH-2, ROUGH-SU) as shown in Table 5.

The quality of summarization depends on the extraction performance. Therefore, the quality of summarization based on the extraction with higher precision in our study outperforms existing methods. The hidden semantic information in the comments is obtained and the lack of shallow semantic analysis is filled，our summaries can express the feelings of users adequately and present closer to the expert summaries.

Table 5. The Comparison of Automatic Summarization

Methods	ROUGE-1	ROUGE-2	ROUGE-SU4
LexRank	0.3784	0.0857	0.1312
ILDA	0.3891	0.0732	0.1263
HTMM	0.3608	0.0651	0.1183
LDA+HTMM	0.3713	0.0769	0.1267
Our Approach	0.3972	0.0920	0.1421

4. Conclusion

In this work, we present extraction models respectively for explicit and implicit features according to their characteristics. Using CRF model to mine explicit features and “feature-opinion” pairs in the explicit sentences, then we propose a bipartite graph based on random walk algorithm to extract implicit features, combining features and corresponding opinions into binary collocation that is turning the unstructured or semi-structured text into structured text. At last, we select comment sentences as summary by calculating the cost value. Experimental results show that our method is reasonable and effective, the two models and automatic summarization proposed achieve good results. Opinion mining based on Chinese product reviews is a difficult subject which reflects the flexibility and uncertainty of natural language processing. It can also provide useful information for sentiment analysis with great research value.

References

[1] Tuarob Suppawong, Conrad S Tucker. Automated discovery of lead users and latent product features by mining large scale social media networks. *Journal of Mechanical Design.* 2015; 137(7): 071-402.

[2] Mani Inderjeet, Mark T Maybury. Advances in automatic text summarization. Cambridge, MA: MIT press. 1999.

[3] Mani I, Bloedorn E. *Machine learning of generic and user focused summarization*. Proceedings of the Fifteenth National Conference on Artificial Inteligence. 1998: 821-826.

[4] González-Ibáñez R, Muresan S, Wacholder N. *Identifying Sarcasm in Twitter: A Closer Look*. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2. Association for Computational Linguistics. 2011: 581-586.

[5] Reyes A, Rosso P, Veale T. A multidimensional approach for detecting irony in Twitter. *Language Resources & Evaluation*. 2013; 47(1): 239-268.

[6] Su Q, Xiang K, Wang H. Using Pointwise Mutual Information to Identify Implicit Features in Customer Reviews. Computer Processing of Oriental Languages. Beyond the Orient: The Research Challenges Ahead. Springer Berlin Heidelberg. 2006: 22-30.

[7] Hai Z, Chang K, Kim JJ. Implicit Feature Identification via Co-occurrence Association Rule Mining. *Lecture Notes in Computer Science*. 2011; 6608: 393-404.

[8] Hu M, Liu B. *Mining and summarizing customer reviews.* Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM. 2004: 168-177.

[9] Luhn Hans Peter. The automatic creation of literature abstracts. *IBM Journal of research and development*. 1958; 2(2): 159-165.

[10] Kitagawa Ryota, Katsuhide Fujita. *Automatic Summarization Considering Time Series and Thread Structure in Electronic Bulletin Board System for Discussion*. Advanced Applied Informatics (IIAI-AAI), 2016 5th IIAI International Congress on IEEE. 2016.

[11] Moghaddam Samaneh, Martin Ester. ILDA: I*nterdependent LDA model for learning latent aspects and their ratings from online product reviews*. Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. ACM. 2011.

[12] Darling William M, Fei Song. *Probabilistic document modeling for syntax removal in text summarization*. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2. Association for Computational Linguistics. 2011.

[13] Liu Jiangming, Jin'an Xu, Yujie Zhang. Summarization Based on Hidden Topic Markov Model with Multi-features. *Acta Scientiarum Naturalium Universitatis Pekinensis*. 2014; 1: 027.

[14] Wahib Aminul, Agus Zainal Arifin, Diana Purwitasari. Improving Multi-Document Summary Method Based on Sentence Distribution. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*. 2016; 14(1): 286-293.

[15] Bhole Pankaj Kailas, AJ Agrawal. Extractive Based Single Document Text Summarization Using Clustering Approach. *IAES International Journal of Artificial Intelligence (IJ-AI).* 2014; 3(2).

[16] Zhao WX, Jiang J, Weng J. Comparing Twitter and Traditional Media Using Topic Models. *Lecture Notes in Computer Science*. 2011; 6611(2011): 338-349.
[17] Ya Juan D, WEIF uRu CZ, Heung ZM, Shum Y. *Twitter topic summarization by ranking tweets using social influence and content quality*. In Proceedings of the 24th International Conference on Computational Linguistics. 2012: 763-780.
[18] Wang Y, Wu H, Fang H. An Exploration of Tie-Breaking for Microblog Retrieval. Advances in Information Retrieval. Springer International Publishing. 2014: 713-719.
[19] Dong Zhendong, Qiang Dong, Changling Hao. *Hownet and its computation of meaning*. Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations. Association for Computational Linguistics. 2010.
[20] Zhai Z, Liu B, Xu H. *Clustering Product Features for Opinion Mining.* Fourth ACM International Conference on Web Search & Data Mining. 2011: 347-354.
[21] Nigam K, Mccallum AK, Thrun S. Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*. 2000; 39(2-3): 103-134.
[22] Dang Hoa Trang, Karolina Owczarzak. *Overview of the TAC 2008 update summarization task.* Proceedings of text analysis conference. 2008.