Sentiment Classification for Movie Reviews in Chinese Using Parsing-based Methods

Wen-Juan Hou and **Chuang-Ping Chang**
Department of Computer Science and Information Engineering,
National Taiwan Normal University,
No.88, Sec. 4, Tingzhou Road, Wenshan Distinct, Taipei, Taiwan
emilyhou@csie.ntnu.edu.tw; lukechang@std.ntnu.edu.tw

Abstract

Sentiment classification is able to help people automatically analyze customers' opinions from the large corpus. In this paper, we collect some Chinese movie reviews from Bulletin Board System and aim at making sentiment classification so as to extract several frequent opinion words in some movie elements such as plots, actors/actresses, special effects, and so on. Moreover, we result in a general recommendation grade for users. Focusing on the movie reviews in Chinese, we propose a novel procedure which can extract the pairs of opinion words and feature words according to dependency grammar graphs. This parsing-based approach is more suitable for review articles with plenty of words. The grading results will be presented by a 5-grade scoring system. The experimental results show that the accuracy of our system, with the deviation of grades less than 1, is 70.72%, and the Mean Reciprocal Rank (MRR) value is 0.61. When we change the 5-grade scoring system into producing two values: one for recommendation and the other for non-recommendation, we get precision rates 71.23% and 55.88%, respectively. The result shows an exhilarating performance and indicates that our system can reach satisfied expectancy for movie recommendation.

1 Introduction

As the rapid growth of text data, text mining has been applied to discover hidden knowledge from text in many applications and domains. Nowadays, reviews are increasing with a rapid speed and are available over internet in natural languages. Sentiment analysis tries to identify and extract subject information from reviews. The problem of automatic sentiment analysis has received significant attention in recent years, largely due to the explosion of online social-oriented content (e.g., user reviews, blogs, etc). Furthermore, sentiment analysis can be used in various ways and in many applications such as suggestion systems based on the user likes and ratings, recommendation systems, or insisting in election campaigns. As one of the important applications, sentiment classification targets to rate the polarity of a given text accurately towards a label or a score, predicting whether the expressive opinion in the text is positive, negative, or neutral.

Identifying the sentiment polarity is a complex task. To address the problem of sentiment classification, various methodologies have been applied earlier. Generally, there are two types of approaches tackling the sentiment classification task according to the knowledge the systems used. One is corpus-based and the other is lexicon-based. Corpus-based approaches are usually supervised, i.e., requiring training sets, and performing well when the training set is large enough and correctly labeled. The approaches are studied in (Bakliwal *et al*., 2011; Bespalv *et al*., 2011; Kennedy and Inkpen, 2006; Jin *et al*., 2009; Narayanan *et al*., 2009; Pang *et al*., 2002; Schuller and Knaup, 2011). On the contrary, the lexicon-based approaches are mostly unsupervised, requiring a dictionary or a lexicon of pre-tagged words. Each word that is present in a text is compared against the dictionary. If a word is present in the dictionary, then the polarity value in the dictionary is added to the polarity score. The recent related works to lexicon-based approaches include (Baloglu and Aktas, 2010; Hu and Liu, 2004; Montejo-Raez *et al*., 2012; Qiu *et al*., 2009; Taboada *et al*., 2011; Thet *et al*., 2010; Zhao and Li, 2009). Additionally, some researchers use natural language processing techniques to discover statistical and/or linguistic patterns in the text in order to reveal the sentiment polarity. We can find such works in (Bonev

International Joint Conference on Natural Language Processing, pages 561–569,
Nagoya, Japan, 14-18 October 2013.

et al., 2012; Harb *et al*., 2008; Lin and He, 2009; Su *et al*., 2008; Turney, 2002).

From the previous studies, we know the problem of sentiment analysis is of much attention by many researchers. Most of researchers investigate English reviews rather than ones with other languages. Therefore, we motivate to explore the movie reviews on Chinese Bulletin Board System (BBS). The aim of the study is to classify the sentiment of Chinese articles, and it will help readers understand the sentiment orientation. Besides, we are concerned with matching opinion with movie elements, called feature words in this paper. It is a finer-grained classification comparing to the article view.

The rest of this paper is organized as follows. Section 2 presents the overview of our system architecture. We describe the proposed method in details, i.e., the components of the system, in Section 3. The experimental data used by the system and the results achieved by the proposed methods are shown and discussed in Section 4. Finally, we express our main conclusions and the possible future directions.

2 Architecture Overview

Figure 1 shows the overall architecture of our methods for the sentiment analysis on movie reviews in the Chinese language. At first, we segment the words where the Chinese Knowledge Information Processing (CKIP) word segmentation system is utilized.[1] We then divide the document collection into two parts: one for training and one for testing. For the training corpus, we manually annotate opinion words to be positive, negative or neutral. In the following, we manually annotate feature words which are related to some movie elements such as plots, actors/actresses, special effects, and so on, and hence we get a list of feature words and their corresponding categories. After that, a Chinese parser is applied and we propose an algorithm to relate feature words to opinion words based on the parsing information. Subsequently, we apply an approach to determine the classification of opinion words and thus build an opinion word database. For the testing corpus, besides opinion and feature word lists, we use three more databases to help extract opinion and feature words. The rest processes are like ones of the training corpus, including parsing the documents and making an association between feature words and opinion words. Finally, we calculate the scores of reviews and produce the movie recommendation results.

3 Methods

As shown in Figure 1, the methods of classifying sentiment are separated into several parts. The details of each part are explained in the following.

3.1 Word segmentation

We use CKIP word segmentation system in this phase. When we input a Chinese sentence, CKIP word segmentation system will segment the word and show the part of speech for each word. For example, if we input a Chinese sentence 我覺得很好看 wo-jue-de-hen-hao-kan 'I thought it was very good to see' to CKIP, the result will be "我(Nh) 覺得 (Vt) 很(ADV) 好看(Vi)." It segments the sentence into four words 我 wo 'I', 覺得 jue-de 'thought', 很 hen 'very' and 好看 hao-kan 'good to see'. The parts of speech are pronoun, transitive verb, adverb and intransitive verb for states, respectively. From the above example, we see a word in the Chinese language can be regarded as a "lexical item," which is a sequence of one or more Chinese characters. In this paper, we use "word" to represent the "lexical item," not limiting to the Chinese character numbers.

3.2 Opinion word manual annotation

In the Chinese language, the parts of speech of opinion words are subcategories of verbs, for example, Vi (intransitive verb for states), Vt (transitive verb for states or actions), and so on. There is no "adjective" tagged in Chinese, but the corresponding part of speech is "verb." Therefore, we extract vocabularies that are verbs and treat them as opinion words. Meanwhile, we give the sentiment polarity as positive, negative and neutral for each opinion word. We also observe that adverbs can imply the strength of sentiment polarity. For example, 很 hen 'very' and 非常 fei-chang 'very much' can put emphasis on the opinion words. The negation words, 不 bu 'no' and 沒有 mei-you 'not', are able to put oppositeness on the opinion words. The parts of speech of above words are adverbs (ADV). Hence we extract the adverbs from the documents and annotate them as emphasis, oppositeness and irrelevance. Some examples of positive opinion words, negative opinion words, adverbs for emphasis and adverbs for oppositeness are listed in Table 1.

[1] http://ckipsvr.iis.sinica.edu.tw/

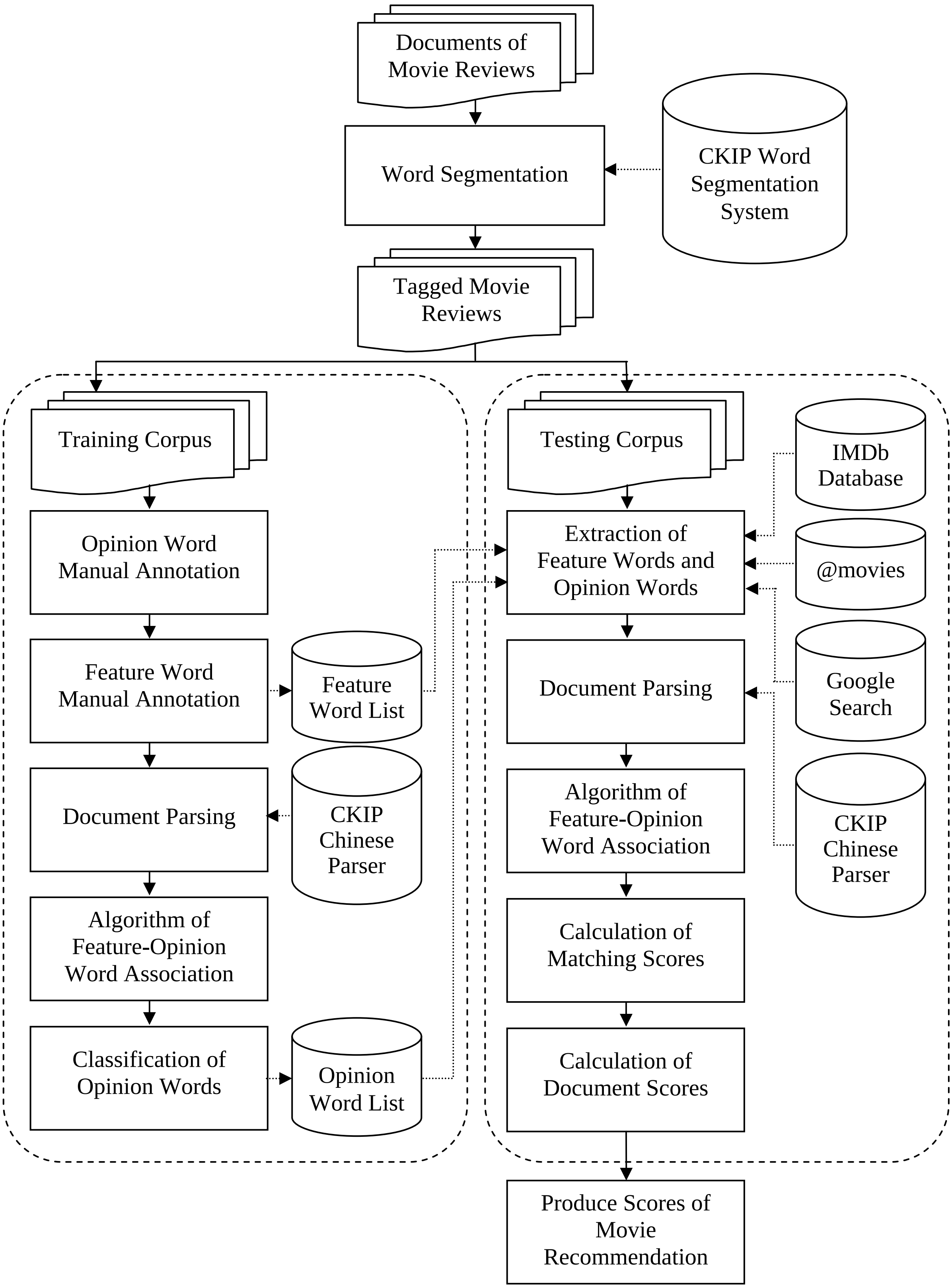

Figure 1. System architecture for the sentiment analysis on Chinese movie reviews.

Category	Examples
Positive opinion words	一氣呵 成 yi-qi-he-cheng ‘accomplish something at one go’, 鮮明 xian-ming ‘bright’, 討喜 tao-xi ‘satisfactory’
Negative opinion words	不清不楚 bu-qing-bu-chu ‘unclear’, 落伍 luo-wu ‘superannuated’, 莫名其妙 mo-ming-qi-miao ‘odd’
Adverbs for emphasis	十分 shi-fen ‘perfectly’, 愈來愈 yu-lai-yu ‘even more’, 格外 ge-wai ‘especially’
Adverbs for oppositeness	不可能 bu-ke-neng ‘impossible’, 尚未 shang-wei ‘not yet’, 無法 wu-fa ‘unable’

Table 1. Examples of opinion word annotations.

3.3 Feature word manual annotation

At this phase, we first manually annotate vocabularies related to movies from the training corpus and then build our own general feature word list. From the training corpus, we do not include some special proper noun related to movies such as names of actors/actresses (e.g., 湯姆克魯斯 tang-mu-ke-lu-si ‘Tom Cruise’) and character names (e.g., 哈利波特 ha-li-po-te ‘Harry Potter’) because it is not complete enough to cover all of the latest movies.

To include more complete movie-related people names, we reference to IMDb[2] and @movies[3] from which we get Chinese and English names of directors, playwrights and stars. Because authors often use hypocorisms or different translated Chinese names in the review articles, we search for hypocorisms and translated Chinese names from Google.[4] Finally, we combine searching data from Google with information from IMDb and @movies, and build a specific feature word list.

For classifying sentiment words at a finer-grained level, we reference to the study of Zhuang *et al*. (2006) and give the classification to each feature word. In this paper, we design four categories, including (1) entirety of movie, (2) story, (3) people (directors, playwrights, actors/actresses and characters), and (4) special effect and others. Table 2 shows some annotated feature word examples for each category.

[2] http://www.imdb.com
[3] http://www.atmovies.com.tw
[4] http://www.google.com.tw

Category	Examples
Entirety of movie	電影 dian-ying ‘movie’, 影片 ying-pian ‘film’, 場面 chang-mian ‘scene’
Story	伏筆 fu-bi ‘foreshadow’, 劇本 ju-ben ‘scenario’, 情節 qing-jie ‘action’
People	主角 zhu-jiao ‘leading role’, 人物 ren-wu ‘figure’, 配角 pei-jiao ‘minor actor’
Special effect and others	主題曲 zho-ti-qu ‘theme song’, 動畫 dong-hua ‘animation’, 布景 bu-jing ‘stage settings’

Table 2. Examples of feature word annotations.

3.4 Document parsing

The structure of a sentence is important for understanding and analysis of sentence semantics. In this phase, we utilize CKIP Chinese parser for the parsing task.[5] The parser is based on probabilistic context-free grammar and is refined with probabilities of word-to-word association in disambiguation. After parsing, we get the syntactic structure trees of the documents in the corpus.

3.5 Algorithm of feature-opinion word association

In a sentence of movie reviews, some feature word is associated with some opinion word. Generally, most of algorithms (e.g., Hu and Liu, 2004) relate a feature word to an opinion word if they collocate closely in a sentence. For example, a sentence 這部電影很吸引人 zhe-bu-dian-ying-hen-xi-yin-ren ‘The movie is very attractive’ includes an association between the feature word (電影 dian-ying ‘movie’) and the opinion word (吸引人 xi-yin-ren ‘attractive’). But the algorithm is too simple to find all correct associations. Let us look at the following sentence.

再好看的電影也會變得無聊 zai-hao-kan-de-dian-ying-ye-hui-bian-de-wu-liao ‘The even more interesting film will become boring too.’ (1)

If we use a simple rule of only considering collocation, the opinion word 好看 hao-kan ‘interesting’ will associate with the feature word 電影 dian-ying ‘film’. But the semantics of the sentence means there should be an association be-

[5] http://godel.iis.sinica.edu.tw/CKIP/parser.htm

tween the feature word 電影 dian-ying 'film' and the opinion word 無聊 wu-liao 'boring'. It is due to no syntactic information is adapted. Consequently, we introduce a Chinese parser and try to solve this problem.

From analyzing the parsing tree, we find the feature word 電影 dian-ying 'film' and the opinion word 無聊 wu-liao 'boring' are at the same level where the feature word 電影 dian-ying 'film' is related to the opinion word 無聊 wu-liao 'boring'. It is an important clue in the study.

The other problem occurs if there are two feature words appearing in a sentence because we have to decide which feature word (or both words) should be related to the opinion word. Let us see the following sentences.

周杰倫的電影實在不吸引人 zhou-jie-lun-de-dian-ying-shi-zai-bu-xi-yin-ren 'The film of Jay Chou is not attractive at all.' (2)

周杰倫和電影都不吸引人 zhou-jie-lun-han-dian-ying-dou-bu-xi-yin-ren 'Jay Chou and his film are not attractive at all.' (3)

In Sentence (2), the opinion word 不吸引人 bu-xi-yin-ren 'not attractive at all' is to modify the feature word 電影 dian-ying 'film'. But in Sentence (3), the opinion word 不吸引人 bu-xi-yin-ren 'not attractive at all' is to modify the feature words 周杰倫 zhou-jie-lun 'Jay Chou' and 電影 dian-ying 'film'. We investigate this problem using the information of the syntactic structure tree.

By the above analysis, we propose the following algorithm to make an association between feature words and opinion words.

1. Traverse the syntactic structure tree by bread-first search, and get the levels for all nodes.
2. Starting from the root, find whether there exists any feature word or opinion word.
   2.1 If feature words and opinion words are found, associate each feature word to all opinion words. For example, if there are three feature words and two opinion words, there will be six pairs of association. Stop searching in the tree.
   2.2 If only feature words exist, search for a subtree rooted with VP (verb phrase). If there is, search all nodes in the VP subtree for opinion words. If there is no VP subtree or no opinion word exists, stop searching in the tree.
   2.3 If only opinion words exist, search for a subtree rooted with NP. If there is, search all nodes in the NP subtree for feature words. If there is no NP subtree or no feature words, stop searching in the tree.
   2.4 If there are no opinion words and feature words, but we can find a subtree rooted with NP and a subtree rooted with VP, then search for feature words in the NP subtree and opinion words in the VP subtree. Make associations with all feature words and opinion words.
   2.5 If no association is found in Steps 2.1 – 2.4, recursively, search the subtree at the different level and repeat Step 2.
3. At the levels of existing feature words and opinion words, find if there is any adverb built in our list. If there is, add the adverb to the feature-opinion word pair.
4. When the system stops searching for feature words and opinion words, only opinion words are extracted from the tree. Associate the opinion words with the special feature word NULL.
5. According to the category of feature words, the pair of feature-opinion words is put into the proper category. The special feature word NULL is classified to a new category.

Applying the algorithm, the matching pairs of Sentences (1) – (3) are presented in Table 3.

Sentence	Matching Pair
(1)	電影 dian-ying 'film' – 無聊 wu-liao 'boring'
(2)	電影 dian-ying 'film' – 不 bu 'not' – 吸引人 xi-yin-ren 'attractive'
(3)	周杰倫 zhou-jie-lun 'Jay Chou' – 不 bu 'not' – 吸引人 xi-yin-ren 'attractive' 電影 dian-ying 'film' – 不 bu 'not' – 吸引人 xi-yin-ren 'attractive'

Table 3. Matching pairs for Sentences (1), (2) and (3).

3.6 Classification of opinion words

Based on the matching pairs extracting from Section 3.5, we can count the number of opinion words in the different categories. We introduce the concept of Term Frequency – Inverse Document Frequency (TF – IDF) to compute the importance of opinion words in the specific categories. The equations are listed as follows.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (1)$$

$$idf_{i,j}^{\ *} = \log \frac{\sum_p n_{i,p}}{1+\sum_{p,p\neq j} n_{i,p}} \quad (2)$$

$$tf\_idf_{i,j}^{\ *} = tf_{i,j} \times idf_{i,j}^{\ *} \quad (3)$$

where $n_{i,j}$ is the number of opinion word t_i appearing in the category g_j. $tf_{i,j}$ is the normalized frequency of term t_i appearing in the category g_j. $idf_{i,j}^{*}$ is a variation of traditional $idf_{i,j}$. In the traditional $idf_{i,j}$, it is obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient. In this study, there are only four categories, the numerator will be 4 and it causes no discrimination with other terms. Hence, we design a new $idf_{i,j}^{*}$ as Equation (2). The numerator is the total number of opinion word t_i in all categories. The denominator is one plus the total count of opinion word t_i in other three categories that t_i does not belong to (a summand one is for avoiding divided by zero).

According to our previous approach, the category of opinion words is the same as the associated feature word. If there is no feature word in the sentence but with an opinion word in a sentence, we must devise some strategy to decide the category. To solve this problem, we propose two thresholds M_1 and M_2, and Equation (4). In this way, all opinion words t_i can have their mapping categories g_j.

$$t_i \in g_i \text{ if } tf\_idf_{i,j}^{\ *} > M_1 \text{ and } \frac{tf_{i,j}\_idf_{i,j}^{\ *}}{\max\{tf\_idf_{i,k} \mid \forall k, k \neq j\}} > M_2 \quad (4)$$

We assign M_1=2.0 and M_2=0.02. Table 4 lists examples of opinion words using Equation (4).

Category	Opinion Word Examples
Entirety of movie	推薦 tui-jian 'recommend' 叫座 jiao-zuo 'box-office'
Story	動人 dong-ren 'touching' 冗長 rong-chang 'long'
People	迷人 mi-ren 'charming' 成熟 cheng-shou 'mature'
Special effect and others	逼真 bi-zhen 'lifelike' 震撼 zhen-han 'vibrate'

Table 4. Examples of opinion words.

3.7 Calculation of document scores

Now, we have a set of matching pairs between opinion words and feature words. We design an equation to compute the scores for the pairs. The equation is listed in Equation (5).

$$S_{i,j} = opS(Opinion_{i,j}) \times \prod_k advS(Adv_{i,j,k}) \quad (5)$$

where $S_{i,j}$ presents the score of the *i*th matching pair in category g_j. $Opinion_{i,j}$ standards for the opinion word of the *i*th matching pair in category g_j. The function *opS* will output 1 if $Opinion_{i,j}$ is a positive sentiment vocabulary, and output －1 if $Opinion_{i,j}$ is a negative sentiment vocabulary. $Adv_{i,j,k}$ is the adverb of the *i*th matching pair in category g_j、. Since there is perhaps more than one adverb in the matching pair, we give a third index *k* in the equation. The function *advS* has values of 1.2 and －1 as its range. If it produces the value 1.2, it means the adverb $adv_{i,j,k}$ is used for emphasis on the opinion word $Opinion_{i,j}$. If it generates the value －1, it means the adverb $adv_{i,j,k}$ puts oppositeness on the opinion word $Opinion_{i,j}$. The final score of the *i*th matching pair in category g_j is the product of *opS* and *advS*.

When we get the scores of all opinion words in the categories, we can compute the score of the opinion word in the review article that is the summation of individual scores. The equation is listed as below.

$$Score = \sum_{j=1}^{4} \sum_k S_{k,j} \quad (6)$$

where *Score* is the sentiment score of the document. We then map *Score* into five levels, i.e., 1, 2, 3, 4 and 5, and call it as "level score". Levels 1 and 2 identify that the document is with negative sentiment, and level 1 is more negative than level 2. Level 3 means that the document is neutral. Levels 4 and 5 identify the positive sentiment polarity in the document, and level 5 is stronger.

3.8 Make movie recommendation

In the last phase, we average the level scores of reviews for each movie, and then the recommendation for each movie is presented with the averaged level scores.

4 Experiments and Results

4.1 Experimental data

The experimental data were obtained from the movie discussion board of website ptt BBS.[6] Users can post their opinions on BBS that acts as a platform for users to share their opinions. From the latest movies, we first select seven movies

[6] telnet://ptt.cc

with different styles. The movies are Mission: Impossible - Ghost Protocol, Treasure Hunter, The A-Team, The Avengers, Toy Story 3, You are the Apple of My Eye, and The Hunger Games. Then we automatically retrieve 50 articles for each movie. To get richer information, the length of retrieved articles is restricted to more than 100 Chinese words. In the following, we filter out the articles that are irrelevant to reviews. For example, some articles are filled with the same words such as 好看 hao-kan 'good to see' and no other words are appeared, so the articles will be filtered out. Finally, we get 321 articles with 379,360 Chinese words.

4.2 Experimental results and discussion

In our experiments, we retrieve 11,837 pairs of feature-opinion word matching where 6,906 pairs are useful and 4,931 pairs are useless. The 6,906 pairs are used as evaluation.

Evaluation of reliability of agreement: First, we invite three humans (A, B and C) who often go to the movies to read the review articles and give the level scores (1, 2, 3, 4, and 5). The average scores will be the gold standard for evaluation. For assessing the reliability of agreement between three scores the humans give, we use the weighted Kappa coefficient (Sim and Wright, 2005) with the quadratic weighting scheme. The formula is given as below.

$$k_w = \frac{\sum(w \cdot f_o) - w \cdot f_c}{n - \sum(w \cdot f_c)},$$
$$quadratic\ weight = 1 - \left(\frac{i-j}{k-1}\right)^2 \qquad (7)$$

where w is the weight, f_0 is the value of the observed disagreement, and f_c is the value of the chance disagreement. k is the number of the ratings and $(i - j)$ is the value of disagreement. Kappa's possible valuels are constrained to the interval [0, 1]; k_w=0 means that agreement is not different from by chance, and k_w=1 means perfect agreement.

The agreement evaluation results are shown in Table 5. It shows that human answers agree with each other almost perfect since the values of the weighted kappa are larger than 0.8.

Human Pairs	Weighted Kappa k_w
A, B	0.87
A, C	0.89
B, C	0.89

Table 5. Human agreement evaluation results.

Evaluation of results with human scores: We use the Mean Reciprocal Rank (MRR) to evaluate the difference between the scores produced by the system and the scores given by the humans. The formula of MRR is as follows.

$$MRR = \frac{1}{|A|}\sum_{i=1}^{|A|}\frac{1}{rank_i},$$
$$rank_i = |human(A_i) - system(A_i)| + 1 \qquad (8)$$

where A is the set of all review documents. $|A|$ is the number of review documents. A_i is the ith review document. $human(A_i)$ is the average score given by humans. $system(A_i)$ is the score given by our proposed system. $rank_i$ means the difference between humans and the system, and a summand 1 is used for avoiding divided by zero. The MRR value to our system is 0.61. Table 6 shows the rank distribution.

$rank_i$	Article Numbers	Ratio
1	110	34.27%
2	117	36.45%
3	67	20.87%
4	18	5.61%
5	9	2.80%

Table 6. Rank distribution of the evaluation.

From Table 6, 34.27% articles get the same scores between the system and humans. If the score difference of the system and humans is less than 1, there are 70.72% articles. Only 8.41% articles have the deviation greater than 2. The result demonstrates that the scores produced by our system are reliable.

If we classify the scores of 4 and 5 as recommendable and others as non-recommendable, then the evaluation results are shown in Table 7.

Class	Totals	System	Correct	Recall	Precision
Re	201	219	156	77.61%	71.23%
Non-re	120	102	57	47.50%	55.88%

Table 7. Rank distribution of the evaluation.

In Table 7, "Re" means recommendable and "Non-re" means non-recommendable. "Totals" is the article amount that humans give and "System" is the article amount that the system produces. "Correct" presents number of consistency between humans and the system. "Recall" is the value of "Correct" dividing "Totals". "Precision" is the value of "Correct" dividing "System". The result shows that the performance of "recommendable" is better than "non-recommendable".

To explore the difference between humans and the system more detailed, we observe that most of the scores generated by the system and humans are similar. Especially, there is no difference for the movie “Mission: Impossible”. The largest difference exists in the movie “Treasure Hunter”. It is because the reviews have some sentences damn the movie with faint praise, i.e., with negative sentiment but seems positive to the system. Let us see the following sentence.

很意外的發現這是一部中等以上的優良惡搞片 hen-yi-wai-de-fa-xian-zhe-shi-yi-bu-zhong-deng-yi-shang-de-you-liang-e-gao-pian ‘Unexpectedly we find that it is a good spoof movie with an average level or better.’ (4)

For Sentence (4), the system produces a feature-opinion pair of “movie – good” and marks it positive. But in effect, the opinion should be negative. It is the main reason why the performance of non-recommendation is worse.

In addition, the system can retrieve opinion words that reviewers often use for different categories. Table 8 lists some opinion words that are frequently used for the movie “Mission: Impossible - Ghost Protocol”. These extracted opinion words are reasonable for commenting about a Hollywood action movie.

Category	Opinion Words
Story	緊湊 jin-cou ‘compact’ 刺激 ci-ji ‘exciting’ 幽默 you-mo ‘humorous’ 驚人 jing-ren ‘amazing’
People	好 hao ‘good’ 強 qiang ‘strong’ 帥氣 shuai-qi ‘smart’ 鮮明 xian-ming ‘bright’
Special effect and others	經典 jing-dian ‘classic’ 精彩 jing-cai ‘splendid’ 罕見 han-jian ‘rarely seen’ 豐富 feng-fu ‘rich’

Table 8. Some frequently used opinion words.

Evaluation of results with IMDb scores: Except for the scores produced by humans, we also compare the results with the scores in the IMDb website. IMDb is a popular movie database and its registered members can give scores for movies. This aspect of evaluation will tell us the consistence between Taiwanese and members in the IMDb website. Because the scores proposed by IMDb are ranged from 1 to 10, we divide them by 2 to mapping into our 5-graded scores. The result is shown in Table 9.

Movies	System	IMDb
Mission: Impossible - Ghost Protocol	4.1	3.7
Treasure Hunter	3.5	2.0
The A-Team	4.0	3.5
The Avengers	4.1	4.3
Toy Story 3	4.2	4.3
You are the Apple of My Eye	3.9	3.8
The Hunger Games	2.9	3.8

Table 9. Comparison to the system and IMDb.

Although the reviewers from IMDb are different from ones from our movie discussion board, the comparison also demonstrates that the recommendable trend is consistent.

5 Conclusion

In this paper, we propose a sentiment classification system based on parsing models. We present a matching algorithm for feature-opinion words, and make effective analysis for reviews with plenty of words.

The experimental results show 70.72% precision rate under the difference less than 1. If the scores are mapped to two levels (recommendable, non-recommendable), the precision rates are 71.23% and 55.88%, respectively. We also compare the result with a popular movie website IMDb, and we discover most of the score trend is similar. It shows the results are exhilarating and indicates that our system can reach satisfied expectancy for movie recommendation.

In the future, we plan to adapt learning methods for matching feature words and opinion words. Besides, we want to explore word polarity according to opinion holders. It will help users understand sentiment orientation for each review more thoroughly.

Acknowledgments

Research of this paper was partially supported by National Science Council, Taiwan, under the contract NSC 102-2221-E-003-027.

References

Akshat Bakliwal, Piyush Arora, Ankit Patil and Vasudeva Varma. 2011. Towards Enhanced Opinion Classification Using NLP Techniques. *Proceedings of the Workshop on Sentiment Analysis where AI Meets Psychology, IJCNLP 2011*, 101 – 107, Chiang Mai, Thailand, November 13, 2011.

Arzu Baloglu and Mehmet S. Aktas. 2010. BlogMiner: Web Blog Mining Application for Classification of Movie Reviews. *Proceedings of 2010 5th International Conference on Internet and Web Applications and Services*, 77 – 84, Barcelona, Spain, May 5 – 19, 2010.

Dmitriy Bespalov, Bing Bai, Yanjun Qi and Ali Shokoufandeh. 2011. Sentiment Classification Based on Supervised Latent n-gram Analysis. *Proceedings of the 20th ACM Conference on Information and Knowledge Management*, 375 – 382, Glasgow, UK, October 24 – 28, 2011.

Boyan Bonev, Gema Ramırez-Sanchez and Sergio Ortiz Rojas. 2012. Opinum: Statistical Sentiment Analysis for Opinion Classification. *Proceedings of the 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, 29 – 37, Jeju, Republic of Korea, July 12, 2012.

CKIP word segmentation system. Available at http://ckipsvr.iis.sinica.edu.tw/.

Ali Harb, Michel Plantié, Gerard Dray, Mathieu Roche, François Trousset and Pascal Poncelet. 2008. Web Opinion Mining: how to Extract Opinions from Blogs? *Proceedings of the 5th International Conference on Soft Computing as Transdisciplinary Science and Technology*, Cergy-Pontoise, France, October 27 – 31, 2008.

Minqing Hu and Bing Liu. 2004. Mining and Summarizing Customer Reviews. *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 168 – 177, Seattle, WA, USA, August 22–25, 2004.

Wei Jin, Hung Hay Ho and Rohini K. Srihari. 2009. OpinionMiner: A Novel Machine Learning System for Web Opinion Mining and Extraction. *Proceedings of the 15th ACM SIGKDD Conference on Knowledge Discovery and Dada Mining*, 1195 – 1204, Paris, France, Jun 28 – July 1, 2009.

Alistair Kennedy and Diana Inkpen. 2006. Sentiment Classification of Movie Reviews Using Contextual Valence Shifters. *Computational Intelligence*, 22 (2): 110 – 125.

Chenghua Lin and Yulan He. 2009. Joint Sentiment/Topic Model for Sentiment Analysis. *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, 375 – 384, Hong Kong, China, November 2 – 6, 2009.

A. Montejo-Raez, E. Martınez-Camara, M. T. Martın-Valdivia, L. A. Urena-Lopez. 2012. Random Walk Weighting over SentiWordNet for Sentiment Polarity Detection on Twitter. *Proceedings of the 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, 3 – 10, Jeju, Republic of Korea, August 12, 2012.

Ramanathan Narayanan, Bing Liu and Alok Choudhary. 2009. Sentiment Analysis of Conditional Sentences. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 180 – 189, Singapore, August 6 – 7, 2009.

Bo Pang, Lillian Lee and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proceedings of 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 79 – 86, University of Pennsylvania, Philadelphia, PA, USA, July 6 – 7, 2002.

Likun Qiu, Weishi Zhang, Changjian Hu and Kai Zhao. 2009. SELC: A Self-Supervised Model for Sentiment Classification. *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, 929 – 936, Hong Kong, China, November 2 – 6, 2009.

Bjorn Schuller and Tobias Knaup. 2011. Learning and Knowledge-based Sentiment Analysis in Movie Review Key Excerpts. *Proceedings of the 3rd COST 2102 International Training School*, 448 – 472, Caserta, Italy, March 15–19, 2011.

Julius Sim and Chris C. Wright. 2005. The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Physical Therapy*, 85:257–268.

Qi Su, Xinying Xu, Honglei Guo, Zhili Guo, XianWu, Xiaoxun Zhang, Bin Swen and Zhong Su. 2008. Hidden Sentiment Association in Chinese Web Opinion Mining. *Proceedings of the 7th International World Wide Web Conference*, 959 – 968, Beijing, China, April 21 – 25, 2008.

Maite Taboata, Julian Brooke, Milan Tofiloski, Kimberly Voll and Manfred Stede. 2011. Lexicon-based Methods for Sentiment Analysis. *Computational Linguistics*, 37(2):267 – 307.

Tun Thura Thet, Jin-Cheon Na and Christopher S.G. Khoo. 2010. Aspect-based Sentiment Analysis of Movie Reviews on Discussion Boards. *Journal of Information Science*, 36 (6): 823 – 848.

Peter D. Turney. 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 417 – 424, July 6 – 12, 2002, Philadelphia, PA, USA.

Lili Zhao and Chunping Li. 2009. Ontology Based Opinion Mining for Movie Reviews. *Proceedings of the 3rd International Conference on Knowledge Science, Engineering and Management*, 204 – 214, University of Vienna, Austria, November 25 – 27, 2009.