Journal of Siberian Federal University. Humanities & Social Sciences 1 (2008) 115-123
~ ~ ~

УДК 495.1

Basic Units for Chinese Opinionated Information Retrieval

Taras Zagibalov*
University of Sussex
Falmer Brighton BN1 9QJ, UK [1]

Received 1.09.2007, received in revised form 1.12.2007, accepted 15.01.2008

This paper presents the results of experiments in which the authors tested different types of features for retrieval of Chinese opinionated texts. We assume that the task of retrieval of opinionated texts (OIR) can be regarded as a subtask of general IR, but with some distinct features. The experiments showed that the best results were obtained from combinating character-based processing, dictionary look up (maximum matching) and a negation check.

Keywords: Chinese world, Cross-Language Opinion Extraction system, opinion extraction, opinion summarization, opinion tracking.

1. Introduction

The extraction of opinionated information has recently become an important research topic. Business and governmental institutions often need to have information about how their products or actions are perceived by people. Individuals may be interested in other people's opinions on different items ranging from political events to consumer products.

At the same time globalization has made the whole world smaller, and a notion of the world as a 'global village' does not surprise people nowadays. Indeed, we buy products which are also being sold overseas, we are dependent on political and economic processes which have a global dimension, and we want to know what people feel about certain events, personalities or products worldwide.

In this context we assume information in Chinese to be of particular interest. The Chinese world (the mainland China, Taiwan, Hong Kong, Singapore and numerous Chinese communities all over the world) is getting more and more influential over the world economy and politics. China itself is not just a country that happens to have the world's biggest population, but is also a fast-growing market and, as some observers indicate, a possible candidate for the role of a new world super power.

We therefore believe that a system capable of providing access to opinionated information in other languages (especially in Chinese) might be of great use for individuals as well as for institutions involved in international trade or international relations.

The experiments presented in this paper were done in the context of Opinionated Information

* E-mail address: information@sussex.ac.uk

Retrieval which is planned to be a module in a Cross-Language Opinion Extraction system (CLOE). The main goal of this system is to provide access to opinionated information on any topic ad-hoc in a language different to the language of a query.

To implement the idea the CLOE system which is the context for the experiments described in the paper will consist of four main modules:

1. Query translation
2. Opinionated Information Retrieval
3. Opinionated Information Extraction
4. Results presentation

The OIR module will process complex queries consisting of a word sequence indicating a topic and sentiment information. An example of such a query is: "Asus laptop + OPINIONS", another, more detailed query, might be "Asus laptop + POSITIVEOPINIONS". Thus the proposed module will process ad-hoc queries, which means it is more closely related to IR than to traditional text classification[1]. This paper will discuss only the OIR component of the CLOE system.

2. Related Work

2.1. The problem of a basic unit definition for Chinese NLP

One of the central problems in Chinese NLP is what the basic unit of processing should be. The problem is caused by a distinctive feature of the Chinese language- absence of explicit word boundaries, while it is widely assumed that a word is of extreme importance for any NLP task. This problem is also crucial for the present study as the basic unit definition affects the kinds of features to be used.

Chinese is an ideographic language, and Chinese characters (or hieroglyphs, or hanzi in Chinese) are the main units of written language. The characters are perceived by native speakers as basic units of their language which entitles the character to be a sociological word. However, the character can not be equal to the word as there are units in the language more than one character long, which are not decomposable and cannot be regarded as compounds consisting of independent words (characters).

But most of the word-level units in the Chinese language are compounds consisting of meaningful components and quite often these units are constructed according to the syntactic models of the language and are structurally "transparent" for native speakers. These units semantically and structurally are very close to phrases, which makes it very hard to attribute this kind of compound to either a word or a phrase class. For example: chi fan "eat + food"can be one word 'to eat', or can be a phrase as both parts of it can be separated by attributes (as well as by any other words or phrases): 'chi hao fan' "eat good food".

Another problem is that there are quite a lot of compounds which can be constructed by productive models and thus can not be exhaustively covered by any dictionary[2].

Unlike most European languages,the process of word production is very active in the Chinese language. The constituents of such"newly-born" compounds can be short forms of two or three-character long words. Example: han zai "drought" can be reduced to han and become a part of chun han "drought in spring" or han qu "drought affected area" (examples are taken from the work by Peng (2002))

All these phenomena of the Chinese language makes it is nearly impossible to exhaustively define what a word is in the Chinese language. It

[1] It well corresponds to the difference between IR and classification as it is stated by Jackson and Moulinier (Jackson and Moulinier, 2002). The main difference between IR and classification is that an IR system is supposed to process almost any query of a user (ad-hoc), while a classification task is usually more rigid, with the objective of obtaining and classifying information for more long-living tasks, such as archiving.

[2] This kind of words is often regarded as grammar words.

results in absence of a widely accepted definition of wordhood in Chinese (Xue, 2003).

2.2. The basic units used in the experiments

Sproat et al. (1996) showed that the rate of agreement between two human judges doing manual word segmentation of Chinese texts is less than 80%. Peng et al. (2002) reported that at around 70% word segmentation accuracy an over-segmentation phenomenon begins to occur which leads to a reduction in information retrieval performance.

These observations inspired us to use a mixed approached,based both on words (tokens consisting of more then one character) and characters as basic units. It is also important to note, that we use notion of words in sense of Vocabulary Word as it was stated by Li (2000). It means that we use only tokens that are listed in a dictionary, and do not look for all words (including grammar words).

2.3. Opinion extraction

Processing of subjective texts and opinions has received a lot of interest recently. This research uses one of three paradigms: classification, information retrieval (IR) or information extraction (IE).Sentiment classification using machine learning was studied by Pang et al. (2002). The authors showed that machine learning methods (NaiveBayes, maximum entropy classification, and support vector machines) do not perform as well on sentiment classification as on traditional topic-based categorization. The authors also showed that bigrams are not effective at capturing context in sentiment extraction, while attempt to model the potentially important contextual effect of negation had some positive influence on performance.

Turney (2002) proposed an unsupervised learning algorithm for classifying a review where the sentiment direction of a phrase is calculated as the mutual information between the given phrase and the word ‘excellent’ minus the mutual information between the given phrase and the word ‘poor’.

Kim and Hovy (2004) used sentiment dictionary based approach. In the work the authors present a system capable of defining not only sentiment polarity,but also the holder of opinion.

Das and Chen (2006) designed an algorithm which comprises different classifier algorithms coupled together by a voting scheme for extracting small investor sentiment from stock message boards. Among the others they use a classifier algorithm based on a word count of positive and negative connotation words.

Some papers report studies of different aspects of opinionated texts classification. For example, Aueand Gamon (2005) and Read (2005) paid special attention to the problem of domain dependency in sentiment classification. Pang and Lee (2004) reported better accuracy comparing to traditional classification when only subjectivity extracts (subjective portions of the document, not the whole text) were processed by the polarity classifier.

Some authors have tried to use more linguistic information (thus more context) to improve classification accuracy. Mullen and Collier (Mullen and Collier, 2004) used several classes of features based upon the proximity of the topic with phrases which have been assigned favour ability values in order to take advantage of situations in which the topic of the text may be explicitly identified. Whitelaw et al. (2005) used appraisal groups, a set of attribute values in several task-independent semantic taxonomies based on Appraisal Theory (for example,very good or not terribly funny). Subasic and Huettner (2001) proposed the fuzzy-affect lexicon,from which a fuzzy thesaurus and affect category groups are generated for analysing the affect content in free text.

Several sentiment information retrieval models were proposed in the framework of probabilistic language models by Eguchi and Lavrenko (2006). The setting for the study was a situation when a user's query specifies not only terms expressing a certaintopic and also specifies a sentiment polarity of interest in some manner. Dave et al. (Dave et al.,2003)described a tool for sifting through and synthesizing product reviews, automating the sort of work done by aggregation sites or clipping services. A number of studies (Riloff et al., 2005; Wiebe and Riloff, 2005; Choi et al., 2005; Riloff et al.,2006) use an information extraction paradigm for sentiment extraction and automatic feature selection for this task.

Recently Ku et al. (Ku et al., 2006a; Ku et al.,2005a; Ku et al., 2006c; Ku et al., 2006b; Ku et al.,2005b) published several works on sentiment extraction from Chinese texts (opinion extraction, opinion summarization and opinion tracking).

3. Experiments

In this paper we present the results of experiments in which we tested different kinds of features (based on our definition of the basic unit, see 2.2) for retrieval of Chinese opinionated information.

As stated earlier (see 1), we assume that the task of retrieval of opinionated texts (OIR) can be regarded as a subtask of general IR with a query consisting of two parts: (1) words indicating topic and (2) a semantic class indicating sentiment (OPINIONS).The latter part of the query cannot be specified in terms that can be instantly used in the process of retrieval.

The sentiment part of the query can be further detailed into subcategories such as POSITIVE OPINIONS,NEGATIVE OPINIONS, NEUTRAL OPINIONS each of which can be split according to sentiment intensity (HIGHLY POSITIVE OPINIONS,SLIGHTLY NEGATIVE OPINIONS etc.). But whatever level of categorisation we use, the query is still too abstract and cannot be used in practice. It therefore needs to be put into words and most probably expanded.

To test the proposed approach we designed two experiments.

The purpose of the first experiment was to find the most effective kind of features for sentiment polarity discrimination (detection) which can be used for OIR[1].

Nieet al. (2000) found that for Chinese IR the most effective kinds of features were a combination of dictionary look up (longest-match algorithm) together with unigrams (single characters). The approach was tested in the context of OIR in the first experiment.

The second experiment was designed to test the found set of features with OIR query of the first level (retrieves opinionated information) and in OIR query of the second level (retrieves opinionated information with sentiment direction detection). Interims of IR the experimental system for the second test can be formulated as the system capable of retrieving texts with the following two kinds of queries: 1. OPINIONS and 2. POSITIVE OPINIONS and NEGATIVE OPINIONS.

For the "wording" and expansion of the sentiment part of the query in the second experiment we use the NTU sentiment dictionary (NTUSD) (by Ku et al. (2006b))[2] as well as a list of sentiment scores of Chinese characters obtained from processing of the same dictionary. Dictionary look up used the longest-match algorithm. The dictionary has 2809 items in the "positive" part and 8273 items in the"negative" one. The same dictionary was also used as a corpus for calculating

[1] For simplicity we used only binary polarity in both experiments. Thus terms "sentiment polarity"and "sentiment direction" are used interchangeably in this work.

[2] Ku et al. (2006b) automatically generated the dictionary by enlarging an initial manually created seed vocabulary by consulting two thesauri, including tong2yi4ci2ci2lin2 and the Academia Sinica Bilingual Ontological Wordnet.

the sentiment scores of Chinese characters. The use of the dictionary as a training corpus for obtaining the sentiment scores of characters is justified by two reasons: 1) it is domain-independent and 2) it contains only relevant(sentiment-related) information. The above mentioned parts of the dictionary used as the corpus comprised 24308 characters in ”negative” part and7898 characters in ”positive”. The dictionary does not provide any linguistic information on its entries, which results in possible ambiguity.

3.1. Experiment 1

A corpus of E-Bay[1] customers' reviews of products and services was used as a test corpus. The total number of reviews is 128, of which 37 are negative (average length 64 characters) and 91 are positive.

We computed two scores for each item (a review)- one for positive sentiment value, another for negative sentiment value. The decision about an item's sentiment polarity was made every time by finding the biggest score of the two.

For every phrase (a chunk of characters between punctuation marks) a score was calculated as:

$Scphrase = SUM(Scdictionary) + SUM(Sccharacter)$

where Scdictionary is a dictionary based score calculated using following formula:

$Scdictionary = Ld/Ls * 100$

where Ld - length of a dictionary item, Ls - length of a phrase. The constant value 100 is used to weight the score, obtained by a series of preliminary tests as a value that most significantly improved the accuracy.

The sentiment scores for characters were obtained by the formula:

$Sci = Fi/F(i+j)$

where Sci is the sentiment score for a character for a given class (i), Fi - the character's relative frequency in a class (i), F(i+j) - the character's relative frequency in both classes (i) and (j) taken as one unit.

The relative frequency of character 'c' is calculated as

$Fc = Pnc / PN(1...n)$

where PNc is a number of the character's occurrences in the corpus, and PN(1...n) is the number of all characters in the same corpus.

Preliminary tests showed that inverting all the characters for which Sci < 1 improves accuracy. The inverting is calculated by formula:

$Scinverted = Sci - 1$

The sentiment score (rather than the probability) was chosen as a more compatible measure with the score obtained by dictionary look up.

In addition to the features specified (characters and dictionary items) we also used a simple negation check, very similar to the technique described by Das and Chen (Das and Chen, 2001) and Pang etal (Pang et al., 2002). The system checked two most widely used negations in Chinese: bu and mei. Every phrase was compared with the following pattern:negation+ 0-2 characters+ phrase. The scores of all the unigrams in the phrase that matched the pattern were multiplied by -1.

Finally, the score was calculated for an item as the sum of the phrases' scores modified by the negation check:

$Scitem = SUM(Scphrase * NegCheck)$

For sentiment polarity detection the item scores for each of the two polarities were compared to each other: the polarity with bigger score was assigned to the item.

$SentimentPolarity = argmax(Sci|Scj)$

where Sci is an item score for one polarity and Scj is an item score for another one.

The main evaluation measure was accuracy of sentiment identification expressed in percent.

[5] http://www.ebay.com.cn/

3.1.1 Results of Experiment 1

To find out which kinds of features perform best for sentiment polarity detection the system was run several times with different settings.

Running without character scores (with dictionary longest-match only) gave following results: almost 65% of negative and near 64% for positive reviews were detected correctly, which is 64% accuracy for the whole corpus. We shall consider this result as a baseline.

Characters with sentiment scores alone performed much better on positive reviews (84% accuracy)rather than on negative (65%), but overall performance was still better – 70%. Both methods combined gave a significant increaseon negative reviews (73%) and no improvement on positive (84%), 77% overall.

The last run was with the dictionary look up, the characters and the negation check. The results were:77% for negative and 89% for positive, 80% corpus wide,with t-Test score against the baseline 3.36 (see Table 1).

Judging from the results it is possible to suggest that both the word-based dictionary look up method and character-based method contributed to the final result. It also corresponds to the results obtained by Nie et al. (2000) for Chinese information retrieval,where the same combination of features (characters and words) also performed best.

The negation check increased the performance by 3% overall, up to 80%. Although the performance gain is not very high, the computational cost of this feature is very low.

3.2. Experiment 2

The second experiment included two parts: processing of the OPINION part of the query to retrieve texts that contain opinionated information; and processing a more detailed form of this query-POSITIVE/NEGATIVE OPINION to retrieve texts with specified sentiment direction. We used the features that showed the best performances described in section 3.1 to implement and expand the queries. The expansion of the sentiment part of the query was done by means of the dictionary items and the characters with the sentiment scores.

The test corpus for this experiment consisted of 282 items, where every item is a paragraph. We used paragraphs as basic items in this experiment because of two reasons: 1. opinionated texts (reviews) are usually quite short (in our corpus all of them are one paragraph), while texts of other genres are usually much longer and 2. for IR tasks it is more usual to retrieve units longer then a sentence. The test corpus has following structure: 128 items are opinionated, of which 91 are positive and 37 are negative (all the items are the reviews used in the first experiment, see 3.1). 154 items are not opinionated, of which 97 are paragraphs taken from a book on Chinese linguistics and 57 items are from articles taken form a Chinese on-line encyclopaedia Baidu Baike .

For processing of the first query we used the following technique: every item was assigned a score (a sum of the characters' scores and dictionary scores described in 3.1). The score was

Table 1. Results of Experiment 1 (accuracy in percent).

Method	Positive	Negative	All
Dictionary (baseline)	63.7	64.8	64.0
Characters	64.8	83.7	70.3
Characters+Dictionary	73.6	83.7	76.5
Char's+Dictionary+negation	76.9	89.1	80.4

divided by the number of characters in the item to obtain the average score:

averScitem = Scitem/Litem

where Scitem is the item score, and Litem is the length of an item (number of characters in it). A positive and a negative average score is computed for each item.

3.2.1. Results of Experiment 2

To determine whether an item is opinionated (OPINION query), the maximum of the two scores was compared to a threshold value. The best performance was achieved with the threshold value of 1.6- more than 85% of accuracy with the baseline 55% (see Figure 1).

Next query (NEGATIVE/POSITIVE OPINIONS) was processed by comparing the negative and positive scores for each retrieved item (see Table 2).

It is worth noting that we observed significant increase in accuracy of sentiment direction detection in the opinionated texts retrieved by the first query: positive 89.9% against 76.9% (obtained in Experiment1); negative 95.6% against 89.1% (see 3.1.1). The same relation between subjectivity detection and polarity classification accuracy was described by Pang and Lee (2004) and Eriksson (2006).

4. Conclusion and Future Work

These preliminary experiments showed that using single characters and dictionary items modified by the negation check can produce reasonable results: about 78% F-measure for sentiment detection(see 3.1.1) and almost 70% F-measure for sentiment polarity identification (see 3.2.1) in a domain independent opinionated information retrieval task.

However, since the test corpus is very small the results obtained need further validation on bigger corpora. The use of the dictionary as a training corpus helped to avoid domain-dependency, however, using a dictionary as a training corpus makes it impossible to obtain grammar information by means of analysis of punctuation marks and function word frequencies.

More intensive use of context information is regarded as a promising tool for improving the accuracy. The dictionary-based processing may benefit from the use of word relations information: most probably some words have sentiment information being used together only. For example, a noun dongxi (‘a thing’) does not seem to have any sentiment information on its own, although it is tagged to be ‘negative’ in the dictionary.

Also we think that some manual filtering of the dictionary and adding more linguistic information to its entries may also improve the output. It might be promising to test the influence on performance of the different classes of words in the dictionary,for example, to use only adjectives or adjectives and nouns together (excluding adverbials).

Another technique to be tested is computing the positive and negative scores for the characters used only in one class, but absent in another. In the present system the characters are assigned only one score (for the class they present). It might improve the accuracy if such single class bound character shave appropriate negative score for the class they are absent.

Table 2. Results of Experiment 2 (in percent)

Query	Recall	Precision	F-measure
OPINION	71.8	85.1	77.9
POS/NEG OPINION	64.0	75.9	69.4

References

1. A. Aue and M. Gamon , "Customizing Sentiment Classifiers to New Domains: A Case Study", *Proceedings of RANLP* (2005).
2. Y. Choi, C. Cardie, Riloff E. and Patwardhan S. Identifying sources of opinions with conditional random fields and extraction patterns. In Proceedings of HLT/EMNLP (Vancouver, 2005, October), p. 355–362.
3. Das S.R. and Chen M.Y. Yahoo! for Amazon: Extracting market sentiment from stock message boards. In Proceedings of the Asia Pacific Finance Association Annual Conference (2001).
4. Das S.R. and Chen M.Y. Yahoo! For Amazon: Sentiment extraction from small talk on the Web (2006).
5. Dave K., Lawrence S. and Pennock D.M. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of the International World Wide Web Conference (Budapest, Hungary. ACMPress, 2003), p. 519-528.
6. Eguchi K. and Lavrenko V. Sentiment retrieval using generative models. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (Sydney, 2006, July), p. 345–354.
7. Eriksson B. Sentiment classification of movie reviews using linguistic parsing (2006). http://www.cs.wisc.edu/#apirak/cs/cs838/eriksson final.pdf.
8. Jackson P. and Moulinier I. Natural Language Processing for On-line Applications. Text retrieval,Extraction and Categorization. John Benjamins Publishing Company (2002).
9. Kim Soo-Min and Hovy E.H. Determining the sentiment of opinions. In Proceedings ofCOLING-04 (2004).
10. Ku Lun-Wei, Lee Li-Ying, Wu Tung-Ho and Chen Hsin-Hsi. Major topic detection and its application to opinion summarization. In SIGIR05 (Salvador, Brazil, 2005a, August).
11. Ku Lun-Wei, Wu Tung-Ho, Lee Li-Ying and Chen Hsin-Hsi. Construction of an evaluation corpus for opinion extraction. NTCIR, 1 (2005b), 627–628.
12. Ku Lun-Wei, Ho Hsiu-Wei and Chen Hsin-Hsi. Novel relationship discovery using opinions mined from the web (2006a).
13. Ku Lun-Wei, Liang Yu-Ting and Chen Hsin-Hsi. Opinion extraction, summarization and tracking in news and blog corpora. In Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs, vol. AAAI Technical Report (2006b, March).
14. Ku Lun-Wei, Liang Yu-Ting and Chen Hsin-Hsi. Tagging heterogeneous evaluation corpora for opinionated tasks (2006c).
15. Li Wei. On Chinese parsing without using a separate word segmenter. Communication of COLIPS,10 (2000), 17–67.
16. Mullen T. and Collier N. Incorporating topic information into sentiment analysis models. InACL Poster Session (Barcelona, 2004).
17. Nie Jian-Yun, Gao J., Zhang J. and MingZhou. On the use of words and n-grams for Chinese information retrieval. In Proceedings of the 5th International Workshop Information Retrieval with Asian Languages, ACM Press (2000, November), p. 141–148.

18. Pang B. and Lee L. A sentimental education:Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the 42ndAnnual Meeting of the Association for Computational Linguistics (Barcelona, Spain, 2004).
19. Pang B., Lee L. and Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the2002 Conference on Empirical Methods in Natural Language Processing, University of Pennsylvania (2002).
20. Peng F., Huang X., SchuurmansD. and Cercone N. Investigating the relationship between word segmentation performance and retrieval performance in Chinese IR. In Proceedings of the19th international conference on Computational linguistics, vol. 1 (2002).
21. Read J. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In Proceedings of the Student Research Workshop of ACL-05 (2005).
22. Riloff E., Wiebe J. and Phillips W. Exploiting subjectivity classification to improve information extraction. In Proceedings of the 20th National Conference on Artificial Intelligence (2005).
23. Riloff E., Patwardhan S. and Wiebe J. Feature subsumption for opinion analysis. In In Proceedings of Conference on Empirical Methods in Natural Language Processing (2006).
24. Sproat R., Shih C., Gale W. and Chang N. A stochastic finite-state word segmentation algorithm for Chinese. Computational Linguistics, 22(3) (1996, September), 377–404.
25. Subasic P. and Huettner A. Affect analysis of text using fuzzy semantic typing. IEEE TRANSACTIONSON FUZZY SYSTEMS, 9(4) (2001, August), 483–496.
26. Turney P.D. Thumbs up or thumbs down?Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (Philadelphia, Pennsylvania, 2002), p. 417–424.
27. Whitelaw C., Garg N. and Argamon Sh. Using appraisal groups for sentiment analysis. In Proceedings of CIKM-05 (2005), p. 625-631.
28. Wiebe J. and Riloff E. Creating subjective and objective sentence classifiers from unannotated texts. In Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing (2005).
29. Xue N. Chinese word segmentation as character tagging. Computational Linguistics and Chinese Language Processing, 8(1) (2003, February), 29–48. The Association for Computational Linguistics and Chinese Language Processing.