Construction of an Evaluation Corpus for Opinion Extraction

Lun-Wei Ku, Tung-Ho Wu, Li-Ying Lee and Hsin-Hsi Chen
Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan
{lwku, lylee, dhwu}@nlg.csie.ntu.edu.tw; hhchen@csie.ntu.edu.tw

Abstract

Opinion retrieval aims to tell if a document is positive, neutral or negative on a given topic. Opinion extraction further identifies the document's supportive and the non-supportive evidence. This paper defines the annotation of opinionated material. The algorithm employs opinion holders, a topic's conceptual words, sentiment words, opinion operators, and negation operators to recognize opinions. An opinion extraction system is developed and then reflects the major views of selected information sources. The text-based evidence extracted is ready for opinion summarization and opinionated question answering.

Keywords: *Opinion Extraction, Sentiment Mining*

1 Introduction

Documents discussing public affairs, common themes, interesting products, *etc*. are reported and distributed over the Internet. Positive and negative opinions embedded in the documents are useful references or feedbacks for governments or companies helping them improve their services or products [3].

Opinion extraction touches on opinion specification in both Chinese (sentence) and English (passage). It captures finer semantics than traditional relevance retrieval, and is more challenging. An opinion retrieval system classifies the relevant documents into three types: positive, neutral and negative. A neutral document provides facts and specifications for a topic, or contains a balance between positive and negative opinions.

In a relevant document, there may be passages supporting or opposing a topic. Hereafter, these are called *supportive* or *non-supportive* evidence. Intuitively, the amount of supportive and non-supportive evidence determines a document's type. But this is not always the case, i.e., limited evidence may overwhelm extensive, weak opposing evidence. For example, an expert's opinion may be more significant than the public's point of views.

Recently, several works dealt with opinion retrieval or opinion extraction. Wiebe, Wilson and Bell recognized opinionated documents [12]. Pang, Lee, and Vaithyanathan classified documents by overall sentiment instead of topics [8]. Dave's and Hu's researches both focused on extracting opinions of reviews [3][4]. However, the smallest unit of opinions is surely not a document. Riloff and Wiebe distinguished subjective sentences from objective ones [9]. Kim and Hovy proposed a sentiment classifier for English words and sentences, which utilized thesauri [5].

Many techniques of NLP were applied. Machine learning approaches such as Naive Bayes, maximum entropy classification, and support vector machines have been investigated, however, Pang, Lee and Vaithyanathan showed that they do not perform as well on sentiment classification as on traditional topic-based categorization [8]. Both information retrieval [3] and information extraction [1] technologies have also been explored. A statistical model was used for sentiment words too, but the experiment material was not described in detail [10]. The results for various metrics and heuristics also varied depending on the testing situations.

Chinese is an ideogram. Characters of a word have contributions to its meaning. Based on the characteristic of Chinese, this paper proposes a statistical method to analyze the sentiment degree of Chinese characters. In this approach, the sentiment of a Chinese word is the function of characters.

Building a testing set is always important. This paper defines a set of annotation tags and sets up the experiment material for opinionated tasks. From the basic word level, sentence/passage level, to document level, this paper proposes methods to extract opinion evidence from relevant documents and summarize the results. Then an opinion extraction system, which utilizes these techniques, is demonstrated. Finally, the opinion summarization will serve as a good source to provide information for opinionated question answering.

2 Data Acquisition and Corpus Formation

The collection of data gathered for this work is a subset of NTCIR[1]. NTCIR is one of three major information retrieval evaluation forums in the world. Chen and Chen developed a test collection CIRB010 for Chinese information retrieval in NTCIR 2 [2]. The test collection consists of 50 topics and 132,173 Chinese documents. Each topic in CIRB010 test collection is in TREC[2] style.

2.1 Topic Selection

Of the 50 topics in CIRB010, 6 opinionated topics are chosen for experiments of the opinion extraction in this work. These topics are shown in Table 1.

Topic ID	Total	Topic Title
ZH021	37	Civil ID Card
ZH024	55	The Abolishment of Joint College Entrance Examination
ZH026	30	The Chinese-English Phonetic Transcription System
ZH027	14	Anti-Meinung Dam Construction
ZH028	23	Hewing Down of Chinese Junipers in Chilan
ZH036	33	Surrogate Mother

Table 1. Opinionated Topics in CIRB010

As an example, topic number ZH027 related to environmental protection is presented below. For simplicity, only <title> denoting a request subject, and <concepts> consisting of relevant keywords are shown in Figure 1.

```
<topic>
<number> CIRB010TopicZH027 </number>
<title> 反美濃水庫興建 (Anti-Meinung Dam Construction) </title>
<concepts>
美濃 (Meinung), 水庫 (Dam), 美濃水庫 (Meinung Dam),反水庫(Anti-Dam),抗爭(resist),興建(build),斷層 (fault), 污染 (pollution), 地質 (geology), 工業 (industry), 環境(enviroment),安全(safety),水資源 (water resources), 生態 (ecology), 水質 (water quality), 替代方案 (displace program)
</concepts>
</topic>
```

Figure 1. A NTCIR topic description

Annotators then annotate all documents related to these 6 topics. Annotation tags and rules are introduced in the following section.

2.2 Corpus Annotation

Tag			
Level	Attribute	Value	Description
<DOC_ATTITUDE></DOC_ATTITUDE>			
Document	TYPE	POS NEG NEU	Document Attitude: Define the opinion polarity of the whole document
<SEN_ATTITUDE></SEN_ATTITUDE>			
Sentence	TYPE	SUP NSP NEU	Sentence Attitude: Define the opinion polarity of one sentence
<OPINION_SEG></OPINION_SEG>			
Sub-sentence	TYPE	PSV	Opinion Segment: Define the scope of one opinion
<OPINION_SRC></OPINION_SRC>			
Sub-sentence	TYPE	EXP IMP	Opinion Source: Define the opinion holder of a specific opinion
<SENTIMENT_KW></ SENTIMENT_KW >			
Word	TYPE	POS NEG NEU	Sentiment Keyword: Define the opinion polarity of a single word
<OPINION_OPR></OPINION_OPR>			
Word	TYPE	PSV	Opinion Operator: Define the keyword of expressing an opinion

Table 2. Tag description

Value Abbreviation	Meaning
EXP	explicit
IMP	implicit
NEG	negative
NEU	neutral
NSP	non-supportive
POS	positive
PSV	preserved
SUP	supportive

Table 3. Abbreviations of attribute values and their meanings

Given the documents relevant to the 6 topics, human annotators then assign *positive*, *neutral*, and *negative* tags (<DOC_ATTITUDE>) to opinionated document. In addition to document opinion, the annotators also assigns sentence opinion (<SEN_ATTITUDE>), including supportive, neutral, and non-supportive.

Furthermore, the annotators add the tags *positive neutral*, *negative keyword* (<SENTIMENT_KW>), and *opinion operator* (<OPINION_OPR>) to the critical words in the passages. *Positive keyword*s like “成功” (succeed), *etc*., and *negative keyword*s like “質疑” (question), *etc*., are sentiment words that express positive and negative attitudes. In contrast, the *opinion operators* like “表示” (express), *etc*.,

[1] http://research.nii.ac.jp/ntcir/index-en.html

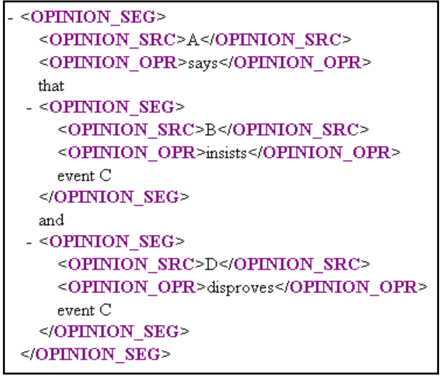
[2] http://trec.nist.gov/

only signal opinions, but do not indicate a clear sentiment tendency. Table 2 lists the annotation tags and their corresponding descriptions. Every element has an opening and closing tag as the XML language.

The tag <OPINION_SEG> is especially useful in dealing with multi-perspective or opinion holder related issues. Consider the following example:

A says that B insists event C and D disproves event C.

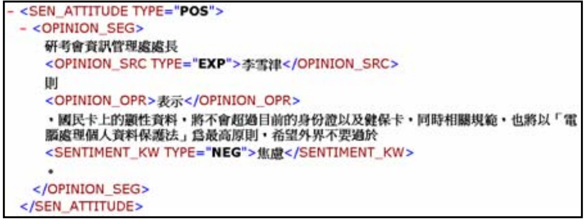
It is tagged as:



```
- <OPINION_SEG>
    <OPINION_SRC>A</OPINION_SRC>
    <OPINION_OPR>says</OPINION_OPR>
    that
  - <OPINION_SEG>
      <OPINION_SRC>B</OPINION_SRC>
      <OPINION_OPR>insists</OPINION_OPR>
      event C
    </OPINION_SEG>
    and
  - <OPINION_SEG>
      <OPINION_SRC>D</OPINION_SRC>
      <OPINION_OPR>disproves</OPINION_OPR>
      event C
    </OPINION_SEG>
  </OPINION_SEG>
```

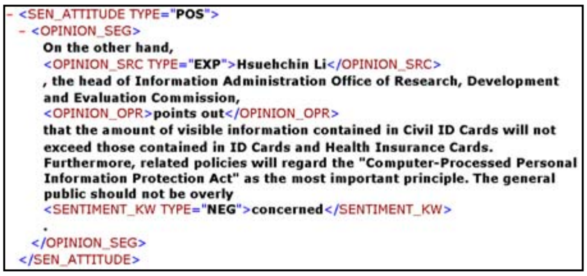
Figure 2. Sample of nested tags

Nested relations of opinion holders are critical to identify the belonging of opinions, that is, multi-perspective issues. XML-like tags can easily represent nested relations and it could co-exist with the original tags for traditional IR purpose. A Chinese and an English tagging examples of views from the XML browser are illustrated in the following figures.



```
- <SEN_ATTITUDE TYPE="POS">
  - <OPINION_SEG>
      研考會資訊管理處處長
      <OPINION_SRC TYPE="EXP">李雪津</OPINION_SRC>
      則
      <OPINION_OPR>表示</OPINION_OPR>
      ，國民卡上的顯性資料，將不會超過目前的身份證以及健保卡，同時相關規範，也將以「電
      腦處理個人資料保護法」爲最高原則，希望外界不要過於
      <SENTIMENT_KW TYPE="NEG">焦慮</SENTIMENT_KW>
      。
    </OPINION_SEG>
  </SEN_ATTITUDE>
```

Figure 3. Civil ID card example in Chinese



```
- <SEN_ATTITUDE TYPE="POS">
  - <OPINION_SEG>
      On the other hand,
      <OPINION_SRC TYPE="EXP">Hsuehchin Li</OPINION_SRC>
      , the head of Information Administration Office of Research, Development
      and Evaluation Commission,
      <OPINION_OPR>points out</OPINION_OPR>
      that the amount of visible information contained in Civil ID Cards will not
      exceed those contained in ID Cards and Health Insurance Cards.
      Furthermore, related policies will regard the "Computer-Processed Personal
      Information Protection Act" as the most important principle. The general
      public should not be overly
      <SENTIMENT_KW TYPE="NEG">concerned</SENTIMENT_KW>
      .
    </OPINION_SEG>
  </SEN_ATTITUDE>
```

Figure 4. Civil ID card example in English

Figure 3 and Figure 4 show a passage opinion for topic ZH021 in Chinese and in English. This topic concerns the personal privacy issue for a government policy. The opinion operator "表示" (point out) shows that this passage may be an opinion, and a negation "不要" (**should not**) that modifies a non-supportive keyword "焦慮" (concerned) transforms a negative passage to a positive passage. The opinion holder is "李雪津" (Hsuehchin Li). Documents of 6 topics are annotated using these tags for experiments.

2.3 Inter-annotator Agreement

To validate tags defined in section 2.2, the agreement of annotations must be tested. The tag of the smallest granularity <SENTIMENT_KW> is tested here. Because it is not cost-effective to examine all the proposed sentiment candidates by all annotators, only the words that are *noun, verb, adjective and adverb* parts-of-speech, and co-occur with one of the seeds (defined in section 4.1) are sampled for agreement test. A total of 838 words were selected. The metrics of the inter-annotator agreement is shown is Formula 1.

$$Agreement(A,B) = \frac{A \cap B}{samples} \qquad (1)$$

Three annotators examined the samples. Tables 4 and 5 show the agreement of the three annotators under strict and lenient metrics. Under lenient metrics, neutral sentiment words and positive sentiment words are in the same category. Strict metrics treats all three categories (positive, neutral, and negative) as distinct. The average agreement between two annotators is 68.62% and 69.69% and the agreement among three annotators is 54.06% and 55.13%, respectively.

Annotators	A vs. B	B vs. C	C vs. A	Ave
Percentage	78.64%	60.74%	66.47%	68.62%
All agree	54.06%			

Table 4. Agreement of annotators under strict metrics

Annotators	A vs. B	B vs. C	C vs. A	Ave
Percentage	79.47%	62.05%	67.54%	69.69%
All agree	55.13%			

Table 5. Agreement of annotators under lenient metrics

The biggest category of samples is "non-sentiment" (400 of 838 words, that is, 47.73%). The agreement of two annotators is far more than this percentage, and the agreement of all annotators is still significant higher. However, we do not define the strength tag as what Wiebe *et al* did in English [11], since the agreement is decreasing when more annotators are involved. Also, the strength of sentiment words is too subjective, which may cause much lower inter-annotator agreement and make the experiment not reliable.

We define the annotations strongly inconsistent if positive polarity and negative polarity are both assigned to one word by different annotators. In a total of 385 inconsistent answers, only 16 words are strongly inconsistent (4.16%). In the contrary, deciding the opinionated degree of a word is hard for human annotators. Annotations are highly inconsistent in weak opinionated words of the form positive/negative vs. neutral (30) and sentiment vs. non-sentiment (339).

Later in this paper we will propose a sentiment miner to mine sentiment words. The majority of annotations of a word is the gold standard for evaluation.

Annotators	A	B	C	Average
Recall	94.29%	96.58%	52.28%	81.05%
Precision	80.51%	88.87%	73.17%	80.85%
f-measure	86.86%	92.56%	60.99%	80.14%

Table 6. Annotators' performance considering gold standard

Table 6 shows the annotation results of three annotators with respect to the gold standard. None of the annotators is able to assign 100% same answers with gold standard, that is, the majority. This result reveals an interesting observation. In the opinion extraction task, one annotator cannot tell the opinion of the whole. The statistics tell us on average that one annotator's opinion and the majority is around 80%. The other 20% is inconsistent because of annotators' perspective. This characteristic makes the opinion extraction different from other work in the NLP area.

3 Opinion Extraction Model

To build an opinion detection system, we have to simulate the behaviors of the annotators, that is, identifying the supportive and non-supportive evidence from a document, determining if the document is opinionated, and if so, determining its polarity. Test collection CIRB010 of NTCIR serves as a test bed. The flowchart of our opinion detection system is introduced together with an illustration of the system in this section.

3.1 System Flowchart

Figure 5 shows the flowchart of our system. Relevant documents to the opinion request are retrieved at the first step. Documents are Temporally Processed for Opinion Extraction. With this temporal information, the Summarization component identifies which events are correlated with the designated opinions.

The core of this system is the Opinion Extraction component. It decides the sentiment tendency of every relevant document, and extracts the supportive/non-supportive evidence based on words from a Sentiment Dictionary. In addition, the Sentiment Dictionary is used to compute opinion scores. Opinion scores are numerical assignments of the relative strength of each opinion, and are determined for each word, passage and document by formulas in section 4. Using the sentiments along with temporal information, the Opinion Summarization component generates a statistical report and summarizes the opinions, and, in addition, places the resulting major events along a timeline.

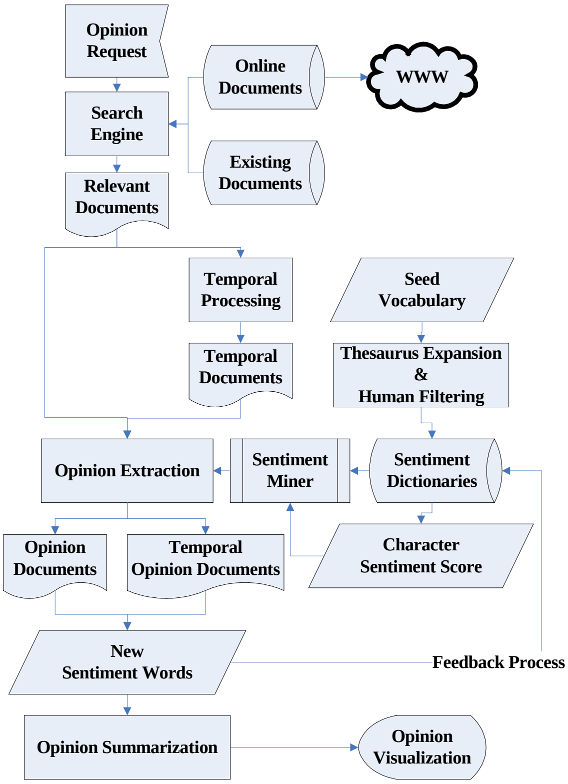


Figure 5. Architecture of an opinion detection and summarization system

3.2 An Illustration of Opinion Detection System

Consider ZH021 (Civil ID Card) as an illustration of the operations of the opinion detection system. A document with the document identifier (chd_pol_19980826_0020) shown in Figure 6 has a "negative opinion score" of -2.48.

Document number: chd_pol_19980826_0020
Document title: 學者質疑國民卡有洩露隱私的權利
(Scholars question the legality of revealing private information on Civil ID Cards.)
Opinion Score: -2.48
Opinion: negative

Figure 6. A total opinion score of an Example Document

Non-supportive/supportive passages in this document are extracted and two of them are shown in Figure 7 in Chinese and in Figure 8 in English, respectively. Passage 5 demonstrates non-supportive evidence (score: -0.95) and Passage 7 demonstrates supportive evidence (score: 1.67). The values enclosed in parentheses denote the sentiment scores of the words in front, and last scores denote the total sentiment scores for the passages. These scores are all determined by the opinion detection system. The algorithm of determining the scores of a sentiment word, a passage, and a document are discussed in Section 4.

Passage 5：行政院研考會即將推出結合高科技的IC「國民卡」引發各界對隱私外洩的質疑(**opinion operator/-0.30**)，中研院資訊科學研究所研究員何建明昨（二十五）日表示(**opinion operator**)，「電腦處理個人資料保護法」當中明文(**0.50**)規範(**0.06**)民眾的行為，但是卻沒有(**-0.14**)對政府的行為做規範(**0.06**)，因此行政權將有被濫用(**-0.29**)的可能，民眾的隱私權將嚴重(**-0.28**)被侵害(**-0.55**).
(Score: -0.95; not supportive)

Passage 7：但與會的研考會資訊管理處處長李雪津則表示(**opinion operator**)，國民卡上的顯性資料，將不會(**negation**)超過(**-0.21**)目前的身份證以及健保卡，同時相關(**0.23**)規範(**0.06**)，也將以「電腦處理個人資料保護法」為最高(**0.61**)原則，希望(**0.38**)外界不要**(negation)**過於焦慮(**-0.18**).
(Score: 1.67; supportive)

Figure 7. Two sample passages of a document (in Chinese)

Passage 5：The Information Administration Office of Research, Development and Evaluation Commission will soon release the Civil ID Cards, which incorporate state-of-the-art IC technology. This plan makes the general public worry about (**opinion operator/-0.30**) their privacy. Chienming He, a researcher at the Institute of Information Science Academia Sinica, stated (**opinion operator**) yesterday (the 25th) that the "Computer-Processed Personal Information Protection Act" provides rules (**0.50**) regulating (**0.06**) the general public's behaviors, but does not (**-0.14**) restrict (**0.06**) the government's actions. So, it is possible for the government to abuse (**-0.29**) its authority and seriously (**-0.28**) violate (**-0.55**) the public's right of privacy.
(Score: -0.95; not supportive)

Passage 7：On the other hand, Hsuehchin Li, the head of Information Administration Office of Research, Development and Evaluation Commission, points out (**opinion operator**) that the amount of visible information contained in Civil ID Cards will not (**negation**) exceed (**-0.21**) those contained in ID Cards and Health Insurance Cards. Furthermore, related (**0.23**) policies **(0.06)** will regard the "Computer-Processed Personal Information Protection Act" as the most important (**0.61**) principle. Li hopes (**0.38**) that the general public would not **(negation)** be overly concerned (**-0.18**)
(Score: 1.67; supportive)

Figure 8. Two sample passages of a document (in English)

All the passage scores for the document are then summed resulting in the first document score of –2.48 shown in Figure 9. The above process is then repeated for all the documents in the test collection, 6 topics of CIRB010. Figures 9 and 10 then show the summary reports of documents relevant to the Civil ID Card and the supportive and non-supportive evidence detected. The evidence comes from different documents, and the values before the evidence indicate the degrees of support/non-support evidence in each document. This then results in an overall total score for the Civil ID Card opinions.

The extraction of supportive and non-supportive evidence is very important in further opinionated work, for example, opinion summarization and opinionated question answering. Considering relevance and opinion degree together produces a text based opinion summarization [6]. The answer of why-type opinionated question, such as why people want the Civil Card, can also be answered by supportive and non-supportive evidence extracted.

-2.48 學者質疑國民卡有洩露隱私的權利
(Scholars question the legality of revealing private information on Civil ID Cards.)
-11.51 國民卡6成4民眾怕洩底
(64% of people are concerned about Civil ID Cards' security.)
-11.39 國民卡計畫應懸崖勒馬徹底檢討
(Plans for Civil ID Cards should be stopped and reevaluated.)
-10.10 適法性存疑國民卡政策恐有漏洞
(Civil ID Card legitimation is questionable. Civil ID Card policies may have loopholes.)
…etc.

Figure 9. Partial total of non-supportive evidence in Civil ID Card documents

10.77 王令台：國民卡安全性絕對沒問題
(Lingt'ai Wang: Civil ID Cards will not have any security problem.)
10.12 業者：營運不會踰越政府委託範圍
(Business owners: Operations will not exceed the extent of government's authorization.)
9.88 國民卡健保資料層層保護
(Health insurance information on Civil ID Cards is tightly protected.)
…etc.

Figure 10. Partial total of supportive evidence in Civil ID Card documents

4 An Opinion Extraction Algorithm

The opinion passage/sentence is the basic unit from which opinions are extracted. Four factors are considered when extracting opinion passages and determining their tendency, the topic concepts, the sentiment words, the opinion operators, and the contextual information. In this paper, the words in the concept field are used to represent the content of a topic. An opinion passage must contain at least one concept word as well as one sentiment word.

We postulate that the opinion of the whole is a function of the opinions of the parts. That is, a summary report is a function of all relevant documents, the opinion of a document is a function of all the supportive/non-supportive evidence, and the degree of supportive/non-supportive evidence is a function of an opinion holder together with sentiment words. The sentiment miner determines the opinion scores of words.

4.1 Sentiment Miner

Sentiment words are employed to compute the tendency of a passage, and then a document. Intuitively, a Chinese sentiment dictionary is indispensable. However, a small dictionary may suffer from the problem of coverage. We develop a method to learn sentiment words and their strengths, here represented by weights, from multiple resources.

First we collect two sets of sentiment words, including General Inquirer[3] (abbreviated as GI) and Chinese Network Sentiment Dictionary[4] (abbreviated as CNSD). The former is in English. We translate those words into Chinese. The latter, whose sentiment words are collected from the Internet, is in Chinese. Table 7 shows the statistics of the revised dictionaries. Words from these two resources become the "seed vocabulary" in our dictionary.

Dictionary	Positive	Negative	Total
GI	2,333	5,830	8,163
CNSD	431	1,948	2,379
Total	2,764	7,778	10,542

Table 7. Qualified seeds

Then, we enlarge the seed vocabulary by consulting two thesauri, including tong2yi4ci2ci2lin2 (abbreviated as Cilin) [7] and the Academia Sinica Bilingual Wordnet[5] (abbreviated as BOW). Cilin is composed of 12 large categories, 94 middle categories, 1,428 small categories, and 3,925 word clusters. BOW is a Chinese thesaurus with a similar structure as WordNet[6]. However, words in the same clusters may not always have the same opinion tendency. For example, 「寬恕」 (forgive: positive) and 「姑息」 (appease: negative) are in the same synonym set (synset), but they do not have the same opinion tendency. How to distinguish this polarity within the same cluster/synset is the major issue of using thesauri to expand the seed vocabulary and is addressed below.

We postulate that the meaning of a Chinese sentiment word is a function of the composite Chinese characters. This is exactly how people read ideogram when they come to a new word. A sentiment score is then defined for a Chinese word by the following formula. This formula, not only tells us the possible opinion tendency of an unknown word, but also indicates their strength. Moreover, using these equations, synonyms of different polarities are distinguishable while doing thesaurus expansion. We start the discussion from the definition of the formulas of Chinese characters.

$$P_{c_i} = \frac{fp_{c_i}}{fp_{c_i} + fn_{c_i}} \qquad (2)$$

$$N_{c_i} = \frac{fn_{c_i}}{fp_{c_i} + fn_{c_i}} \qquad (3)$$

Where fp_{ci} and fn_{ci} denote the frequencies of a character c_i in the positive and negative words, respectively; n and m denote total number of unique characters in positive and negative words, respectively.

Formulas (2) and (3) utilize the probability of a character in positive/negative words to show its sentiment tendency. However, there are more negative words than positive ones in the human tagged dictionary. Hence, the frequency of a character in a positive word may tend to be smaller than that in a negative word. That causes bias for learning, so formulas (2) and (3) are normalized into formulas (4) and (5).

$$P_{c_i} = \frac{fp_{c_i} / \sum_{j=1}^{n} fp_{c_j}}{fp_{c_i} / \sum_{j=1}^{n} fp_{c_j} + fn_{c_i} / \sum_{j=1}^{m} fn_{c_j}} \qquad (4)$$

$$N_{c_i} = \frac{fn_{c_i} / \sum_{j=1}^{m} fn_{c_j}}{fp_{c_i} / \sum_{j=1}^{n} fp_{c_j} + fn_{c_i} / \sum_{j=1}^{m} fn_{c_j}} \qquad (5)$$

Where P_{ci} and N_{ci} denote the weights of c_i as positive and negative characters, respectively. The difference of P_{ci} and N_{ci}, i.e., P_{ci} - N_{ci} in Formula (6), determines the sentiment tendency of character c_i. If it is a positive value, then this character occurs more often in positive Chinese words than negative ones. and vice versa. A value close to 0 means that it is not a sentiment character or it is a neutral sentiment character.

$$S_{c_i} = (P_{c_i} - N_{c_i}) \qquad (6)$$

Formula (7) defines that a sentiment tendency of a Chinese word w is the average of the sentiment scores of the composing characters c_1, c_2, …, c_p.

$$S_w = \frac{1}{p} \times \sum_{j=1}^{p} S_{c_j} \qquad (7)$$

4.2 Opinion Extraction from Documents

With the sentiment words extracted, we are able to tell the opinion tendencies of passages and documents. Algorithms to determine the opinions at the passage and document levels are shown below.

Passages that are relevant to the topic and also express opinions are extracted. There are two clues to extraction, i.e., concept keywords and sentiment

[3] http://www.wjh.harvard.edu/~inquirer/
[4] http://134.208.10.186/WBB/EMOTION_KEYWORD/Atx_emtwordP.htm
[5] http://bow.sinica.edu.tw/
[6] http://wordnet.princeton.edu/

words. The former determines the relevance of a passage to the topic, and the latter identifies the degree of passage opinion. In our experiments, concept keywords come from the "concepts" field in NTCIR corpus. Sentiment words are extracted by the sentiment miner. Algorithms for opinion extraction at passage and document level are shown below.

```
[Passage Level]
For every passage
  If it contains concept words
    For every sentiment word in the passage
      If a negation operator appears before, then
      reverse the sentiment tendency.
  Decide the tendency of the passage opinion by the
  function of sentiment words and opinion
  holder as follows.
```

$$S_p = S_{opinion-holder} \times \sum_{j=1}^{n} S_{w_j} \qquad (8)$$

Where S_p, $S_{opinion\text{-}holder}$, and S_{wj} are sentiment score of passage *p*, weight of *opinion holder*, and sentiment score of word w_j, and *n* is total number of sentiment words in *p*.

```
[Document level]
For every document
  Decide the tendency of the document opinion by
  the function of the tendencies of the passage
  opinions as follows.
```

$$S_d = \sum_{j=1}^{m} S_p \qquad (9)$$

Where S_d and S_p are sentiment scores of document *d* and passage *p*, and *m* is the amount of evidence. If the topic is *anti* type, reverse sentiment type.

5 Experiments and Discussion

5.1 Evaluation of Sentiment Miner

Given the test corpus, the seed sentiment words, and two thesauri, our sentiment miner determines if a word is a sentiment word and, if so, its weight. Table 8 shows the performance of our sentiment miner using the formulas without and with normalization. Nouns, adjectives and adverbs are in the Noun category, while verbs are in the Verb category. Results using non-normalized formulas and normalized formulas are compared, and the gold standard in section 2.3 is used here. The f-measure of results using normalized formulas is 73.18% for verbs and 63.75% for nouns, better than that of results using non-normalized formulas.

Recall that the average result of annotators is 80.14%. In other words, with this testing set, our system achieves 91.32% (73.18/80.14) in Verb and 79.55% (63.75/80.14) in Noun with respect to annotators.

	Non-normalized		Normalized	
	Verb	Noun	Verb	Noun
Precision	69.25%	50.50%	70.07%	52.04%
Recall	75.48%	81.45%	76.57%	82.26%
f-measure	72.23%	62.35%	73.18%	63.75%

Table 8. Performance of sentiment word mining

In spite of polarity information, the sentiment miner provides strength information. For example, the Chinese word "富貴" (fù guì) means wealth. Its sentiment score 0.61 is computed from the sum of " 富" (fù, rich, 0.75) and "貴" (guì, expensive, 0.48). To determine the context polarity, "富貴" (fù guì, wealth, 0.61) is stronger than"有錢" (yǒu qián, have money, 0.33), which is another Chinese word describing rich in a subtler degree.

A major issue in a statistic model is the size of training words. To show the influence of the amount of seed vocabulary, we randomly drop seeds from our dictionaries and redo the experiments. The results are shown in Figure 11.

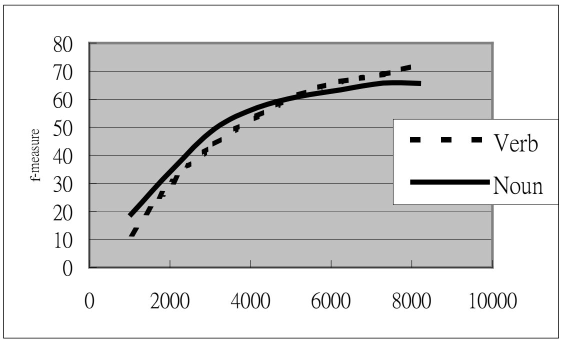


Figure 11. F-Measure using different quantity of seed vocabulary

In Figure 11, the x-axis is f-measure and the y-axis is the quantity of seeds. F-measure increases when training with more seeds. However, we find the improvements decrease when the quantity of seed vocabulary increases. The improvements saturate at around size 8000. As shown in Table 7, the sentiment miner has 10,542 seeds. According to Figure 11, this quantity of seeds is able to provide a reliable performance.

In summary, our sentiment miner effectively identifies both opinions words and their weights in documents. As it mines, it incorporates new information into its sentiment dictionary, as illustrated by Feedback Process in Figure 5.

5.2 Evaluation of Opinion Extraction

Table 9 shows the performance of opinion extraction at the passage level using the sentiment dictionary mined in the last section. Formula (8) computes the passage score. If an opinion holder and the opinion operator must appear in the qualified opinion passage, the average f-measure is 56.11%.

Because opinions may be expressed by the author, the opinion holder may be implicit. The average f-measure without considering opinion holders and opinion operators raises to 62.16%. Both precision and recall are improved in this case and in particular, recall rate increases 12.15%.

Table 10 shows the performance of opinion extraction at the document level. On the average, the opinion extraction system achieves the precision rate 76.56% on document level. Recall that document score is the sum of passage scores according to Formula (9), which is in term the sum of word scores according to Formula (8). Because we could not tell the rank of different opinion holders, the weights of opinion holders in Formula (8) were all set to 1. Compared to the opinion extraction at the passage level, precision actually then goes up at the document level because of the increased amount of data. (P: Precision, R: Recall, f-m: f-measure, Ave: Average)

	With holder (%)			Without holder (%)		
ID	P	R	f-m	P	R	f-m
ZH021	63.92	72.66	68.01	63.19	74.10	68.21
ZH024	51.92	80.60	63.16	52.38	82.09	63.95
ZH026	71.57	41.24	52.33	68.48	63.84	66.08
ZH027	71.11	60.38	65.31	66.67	75.47	70.80
ZH028	47.12	37.98	42.06	50.36	54.26	52.24
ZH036	43.36	57.64	49.50	45.16	65.88	53.59
Ave	57.19	55.08	56.11	57.80	67.23	62.16

Table 9. Opinion extraction (passage)

Topic ID	Total Documents	Precision
ZH021	37	86.49%
ZH024	55	94.55%
ZH026	30	66.67%
ZH027	14	78.57%
ZH028	23	47.83%
ZH036	33	63.64%
Average	32	76.56%

Table 10. Opinion extraction (document)

6 Conclusion and Future Work

A set of tags to describe the basic building blocks of opinionated documents is defined in this work. Experiment material is developed and then the algorithm proposed mines positive and negative sentiment words and their weights on the basis of Chinese word structures. The f-measure is 73.18% for verbs and 63.75% for nouns.

With sentiment words and concept words, this approach identifies supportive and non-supportive evidence. The amount of evidence and the degree of support further determine the polarity of a document. The system achieves f-measure 62.16% at the passage level and 76.56% at the document level. It then reflects the major views of information sources.

Though this approach works well, some issues have to be considered further. First, a negation character may inverse the polarity of a word. For example, a positive word “抗病” (kàng bìng, disease-resistant) consists of a negation character “抗” (kàng, resist, oppose) and a negative character ”病” (bìng, disease). To deal with this, we will collect a special character set, and integrate it to our sentiment miner.

Second, the effects of the collocation of sentiment characters may not always be neglected. For instance, “加” (jiā, increase) in “加薪” (jiā xīn, increase income) is generally a positive sentiment word. However, when it is integrated with the word “税” (shuì, tax), the polarity is reversed. Negations and collocation of semantics at word level are also issues at the passage and document level.

Our experiments on opinion summarization and opinionated question answering are ongoing. The future goal is to monitor the opinions of the masses by enlarging the scale of experiments.

References

[1] C. Cardie, J. Wiebe, T. Wilson and D. Litman. Combining low-level and summary representations of opinions for multi-perspective question answering. *Proceedings of AAAI Spring Symposium Workshop*, pages 20-27. 2004.

[2] K.-H. Chen and H.-H. Chen. The Chinese text retrieval tasks of NTCIR II workshop. *Proceedings of 2nd NTCIR Workshop*, pages 51-72. 2001.

[3] K. Dave, S. Lawrence and D.M. Pennock. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. *Proceedings of 12th International Conference on World Wide Web*, pages 519-528. 2003.

[4] Minqing Hu and Bing Liu. Mining and Summarizing Customer Reviews. *SIGKDD 2004*, pages 168-177. 2004.

[5] Soo-Min Kim and Eduard Hovy. Determining the Sentiment of Opinions. *Coling*, pages 1367-1373. 2004.

[6] L.-W. Ku, L.-Y. Li, T.-H. Wu and H.-H. Chen. Major topic detection and its application to opinion summarization. *SIGIR 2005*, pages. 627-628. 2005.

[7] J. Mei, Y. Zhu, Y. Gao and H. Yin. *tong2yi4ci2ci2lin2*. Shanghai Dictionary Press. 1982.

[8] B. Pang, L. Lee and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. *Proceedings of the 2002 Conference on EMNLP*, pages 79-86. 2002.

[9] E. Riloff and J. Wiebe. Learning extraction patterns for subjective expressions. *Proceedings of the 2003 Conference on EMNLP*, pages 105-112. 2003.

[10] H. Takamura, T. Inui and M. Okumura. Extracting Semantic Orientations of Words Using Spin Model. *ACL 2005*, pages 133-140. 2005.

[11] J. Wiebe, et. al. NRRC summer workshop on multi-perspective question answering, final report. *ARDA NRRC Summer 2002 Workshop*. 2002.

[12] J. Wiebe, T. Wilson and M. Bell. Identify collocations for recognizing opinions. *Proceedings of ACL/EACL2001 Workshop on Collocation*. 2001.