Construction of a Chinese Opinion Treebank

Lun-Wei Ku, Ting-Hao Huang, Hsin-Hsi Chen
Department of Computer Science and Information Engineering
National Taiwan University
No.1, Sec. 4, Roosevelt Rd., Taipei, Taiwan 10617
{lwku, thhuang}.nlg.csie.ntu.edu.tw;hhchen@ntu.edu.tw

Abstract

In this paper, we base on the syntactic structural Chinese Treebank corpus, construct the Chinese Opinon Treebank for the research of opinion analysis. We introduce the tagging scheme and develop a tagging tool for constructing this corpus. Annotated samples are described. Information including opinions (yes or no), their polarities (positive, neutral or negative), types (expression, status, or action), is defined and annotated. In addition, five structure trios are introduced according to the linguistic relations between two Chinese words. Four of them that are possibly related to opinions are also annotated in the constructed corpus to provide the linguistic cues. The number of opinion sentences together with the number of their polarities, opinion types, and trio types are calculated. These statistics are compared and discussed. To know the quality of the annotations in this corpus, the kappa values of the annotations are calculated. The substantial agreement between annotations ensures the applicability and reliability of the constructed corpus.

1. Introduction

Opinion analysis is practical for many applications. Sentiment information can be applied to product recommendation, review summarization, public polling, etc. Product reviews, due to their availability on the web, are often adopted to develop prototyped opinion analysis systems (Bai and Padman, 2005; Ghose and Ipeirotis, 2007). However, documents collected from real world are usually raw, and need some text preprocessing before usages. Besides, only documents and their evaluative stars are available in such corpora. The lack of syntactic or semantic information limits the development of opinion analysis technologies.

This problem is more serious for Mandarin Chinese. Few materials are available. NTCIR MOAT[1] is the most well-known task (Seki *et al.*, 2008) which provides experimental corpora at sentence and clause levels for multilingual opinion analysis. Opinion labels, their polarities, holders, and targets were annotated in MOAT corpus. However, this corpus does not provide linguistic features either.

Linguistic features such as word boundaries, parts of speech, and sentence structures might be important in opinion analysis. Previous research revealed that parts of speech are useful features (Wiebe 2000; Riloff *et al.*, 2003). However, there are limited resources for researchers to do thorough researches on the relations of other linguistic features and opinions. Parts of speeches are the most often available features and other more intensive linguistic features are difficult to get. To further explore how composite components function linguistically to express an opinion needs an opinion corpus with labeled syntactic information. For this purpose, Chinese Treebank is selected as the corpus for annotation in this paper, and opinions were not annotated on it before.

2. Opinion Tagging Scheme

Thoughts expressed by persons or organizations are often considered as opinions. In the example: “this expert said that the authority will not forbid workers to leave”, an expert expresses his thoughts, and it is considered as an opinion.

However, opinions are not always in the form of expressions. Let us consider another example. The sentence, “the government has hesitations and may not do it right away”, indicates that the government is not really supportive. Instead of an expression, the action “has hesitations” reveals the attitude of the subject “government”. In this case, both supportive and oppositional actions should be treated as opinions.

Subjective information, including expressions, statuses, and actions, is considered as opinions. During annotation, we tag whether a sentence is an opinion. If it is, its opinion type, i.e., ***expression***, ***status***, or ***action***, is also determined. In opinion sentences of *expression* type, people reflect their subjective judgment. In opinion sentences of *status* type, the subjective information appears as descriptions. These descriptions could reflect the author’s opinions. In opinion sentences of *action* type, the opinion holder’s attitude is revealed by his action.

For those opinion sentences, we also annotate their polarities including ***positive***, ***neutral***, and ***negative***. Positive opinions express a supportive attitude, while negative opinions express an opposite one. Neutral opinions indicate impartial attitudes.

In addition to label fragments as opinions and deter-

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

mine their polarities, our tagging scheme introduces relations between the composite words of opinions. Linguists have defined five structural relations between two words in the Mandarin Chinese (Cheng and Tian, 1992) as follows.

(1) Parallel Type: Two word sequences play coordinate roles in a sentence. For example, “美麗 (beautiful) 而 (and) 聰慧 (smart)”，the words “美麗” (beautiful) and “聰慧” (smart) are of the parallel relation.

(2) Substantive-Modifier Type: A modified word sequence follows a modifying word sequence. For example, “淒涼地 (sadly) 笑著(laugh)”.

(3) Subjective-Predicate Type: One word sequence is an expresser and the other is described. For example, “討論 (discussion) 熱烈 (enthusiastic)”. *Be* verb is sometimes omitted in this case.

(4) Verb-Object Type: The first word sequence usually plays the role of verb which governs the second one, making these two word sequences similar to a verb followed by its object. For example, “恢復 (overcome) 疲勞 (tiredness)”.

(5) Verb-Complement Type: The first word sequence usually plays the role of verb but sometimes adjective, and the second word sequence explains the first from different aspects. For example, “收拾 (put things) 乾淨 (in order)”.

When constructing a Chinese Opinion Treebank, all relations except Parallel Type are annotated on opinion sentences. Relations of Parallel Type are not annotated because their two components are of equal importance, and the total opinion score can be calculated by simple addition.

Structural relations are represented by structural trios as follows.

(1) A structure trio contains two child nodes which bear a relation.

(2) A structure trio contains one head node which is the nearest common parent of these two child nodes.

Figure 1 shows a partial parsing tree containing words “取得” (obtain), “可喜” (happy), “成果” (results) and two annotated structure trios. The lower one contains two child nodes “可喜” (happy) and “成果” (results), and is labeled as Type 2, i.e., Substantive-Modifier (S-M (2)) in their nearest common parent node, while the upper one contains two child nodes “取得” (obtain) and “可喜成果” (happy results) and is labeled as Verb-Object (V-O (4)).

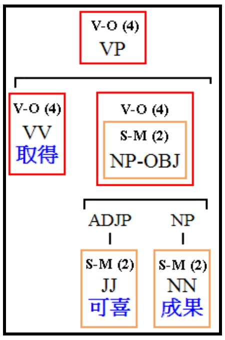


Figure 1: An Example of annotated structural trios

Some tagging schemes were adopted for annotating opinions in previous researches. The MOAT corpus adopted the same schemes for the opinions and the polarities as ours, but opinion sentence types and structural trios are not labeled in it. The well known English opinion corpus MPQA (Wiebe *et al.*, 2002) annotated at the sub-sentence level. The type of attitude (positive, negative, uncertain,) the basis for the opinions (supporting beliefs, experiences, etc.,) and the expressive style of the sentences (sarcastic and vehement, neutral, etc.) were annotated. However, information related to syntactic structures was not annotated either. SentiWordnet[2] annotated the polarity of words and their sentiment weights considering the concept net provided by Wordnet[3]. It was annotated at the word level and the syntactic structure is not available in it.

3. Corpus and Annotation Tools

We adopt Chinese Treebank 5.1 obtained from Linguistic Data Consortium (LDC) as our experimental corpus. It contains 507,222 words, 824,983 Hanzi, 18,782 sentences, and 890 data files. At first, opinion related labels are annotated on all sentences in Chinese Opinion Treebank. Then, structural trios are annotated on the parsing trees of opinion sentences.

Two tools, OAT and PAN, are developed for the annotations of opinion information and structural trios, shown in Figure 2 and Figure 3, respectively. OAT supports multiple languages by using the bilingual (English and the domestic language) command mapping file from users. Figure 2 is shown in English mode to give a better illustration. With OAT, we can browse a document sentence by sentence, and annotate opinion related labels. In addition, cues at the sub-sentence level such as opinion holder, opinion target, opinion sections and their polarities can also be annotated. Considering the annotation cost, we focus on the annotations at the sentence level at the current premier stage.

[2] http://sentiwordnet.isti.cnr.it/
[3] http://wordnet.princeton.edu/

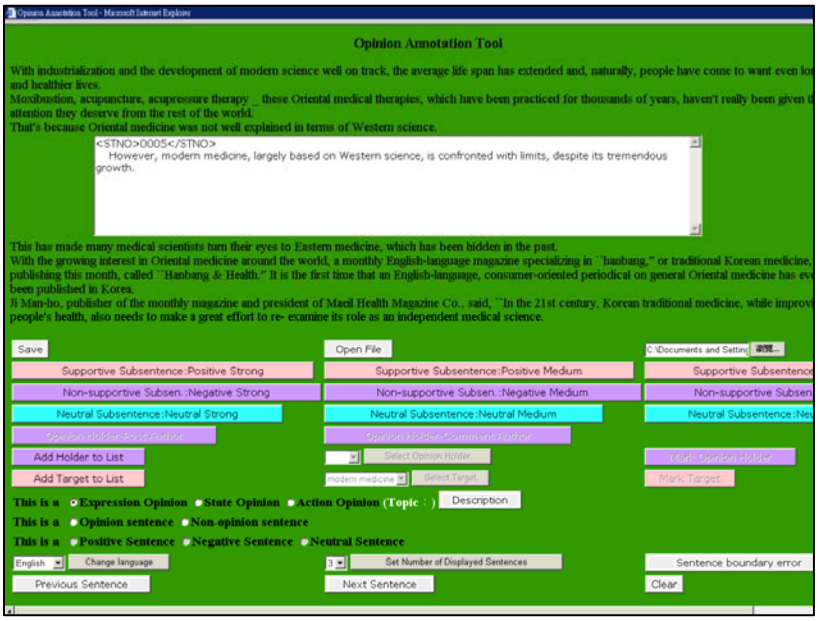
To annotate structural trios, the parsing tree must be displayed at first. PAN analyzes the given parsing tree structure and draws the tree for annotation. OAT and PAN are both designed as browser interface to build a friendly annotation environment.

Each sentence is annotated by three annotators and the gold standard is set up by majority voting. In this way, we can generate ground truth for all sentences. The details of generating the gold standard by the lenient metric are described by Ku *et al.* (2007).

4. A Chinese Opinion Treebank

For opinion annotations, labels including opinion (opinion, or non-opinion), polarity (positive, negative, or, neutral), and type (expression, status, or action) are provided for each sentence. For structural annotations, three kinds of files with file extensions “node”, “tree”, and “trio” are generated. Table 1 shows the annotations of an example sentence. The quality of annotation is satisfactory: the average kappa value, which indicates the agreement between two annotation sets, is 0.49 (moderate agreement) between two annotators and 0.73 (substantial agreement) between one annotator and the lenient gold standard.

Table 2 shows the distribution of the opinion labels, polarity labels, and type labels in the lenient gold standard. Moreover, Figure 4 shows the statistics of opinions by type. It reveals an interesting result: the distribution of the action type is different from the other two types, and the percentage of positive opinions of the action type is overwhelming. In other words, most action opinions encourage people. Opinion expressions which stop someone doing something are rare.



With industrialization and the development of modern science well on track, the average life span has extended and, naturally, people have come to want even lo
and healthier lives.
Moxibustion, acupuncture, acupressure therapy _ these Oriental medical therapies, which have been practiced for thousands of years, haven't really been given t
attention they deserve from the rest of the world.
That's because Oriental medicine was not well explained in terms of Western science.

```
<STNO>0005</STNO>
  However, modern medicine, largely based on Western science, is confronted with limits, despite its tremendous
growth.
```

This has made many medical scientists turn their eyes to Eastern medicine, which has been hidden in the past.
With the growing interest in Oriental medicine around the world, a monthly English-language magazine specializing in ``hanbang,'' or traditional Korean medicine,
publishing this month, called ``Hanbang & Health.'' It is the first time that an English-language, consumer-oriented periodical on general Oriental medicine has ev
been published in Korea.
Ji Man-ho, publisher of the monthly magazine and president of Maeil Health Magazine Co., said, ``In the 21st century, Korean traditional medicine, while improv
people's health, also needs to make a great effort to re- examine its role as an independent medical science.

C:\Documents and Settin

modern medicine

This is a ● Expression Opinion ○ State Opinion ○ Action Opinion (Topic :)

This is a ○ Opinion sentence ○ Non-opinion sentence

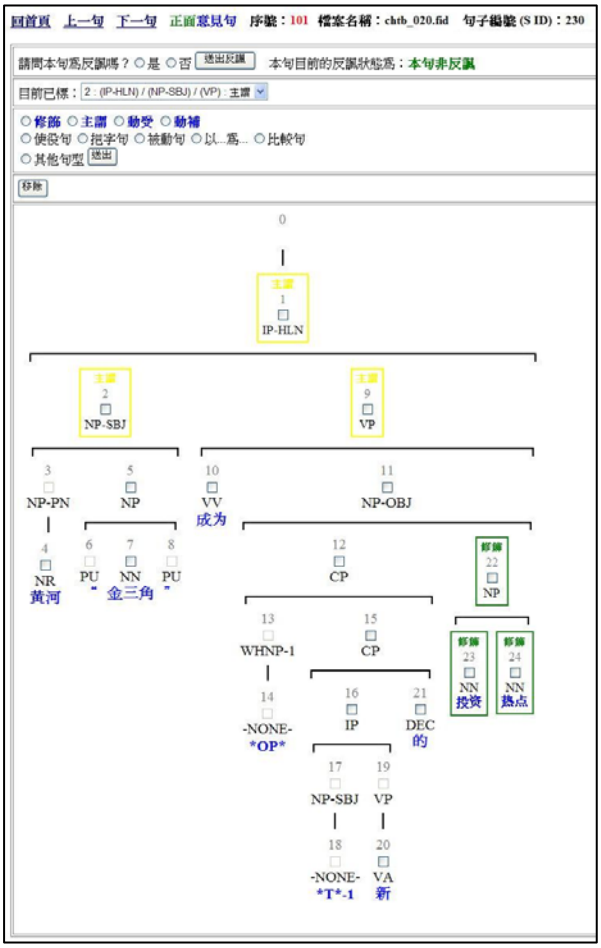
This is a ○ Positive Sentence ○ Negative Sentence ○ Neutral Sentence

English

3

Save
Open File
瀏覽...
Supportive Subsentence:Positive Strong
Supportive Subsentence:Positive Medium
Supportive Subsentence
Non-supportive Subsen.:Negative Strong
Non-supportive Subsen.:Negative Medium
Non-supportive Subsen
Neutral Subsentence:Neutral Strong
Neutral Subsentence:Neutral Medium
Neutral Subsentence:Ne
Opinion Holder-Post Author
Opinion Holder-Comment Author
Add Holder to List
Select Opinion Holder
Mark Opinion Holder
Add Target to List
Select Target
Mark Target
Description
Change language
Set Number of Displayed Sentences
Sentence boundary error
Previous Sentence
Next Sentence
Clear

Figure 2: An OAT interface in English mode



回首頁 上一句 下一句 正面意見句 序號：101 檔案名稱：chtb_020.fid 句子編號 (S ID)：230

請問本句爲反諷嗎？○是 ○否 本句目前的反諷狀態爲：本句非反諷

目前已標：2 : (IP-HLN) / (NP-SBJ) / (VP) : 主謂

○修飾 ○主謂 ○動受 ○動補
○使役句 ○把字句 ○被動句 ○以...爲... ○比較句
○其他句型

Figure 3: PAN interface
(S ID=230: 黄河 “金三角” 成为新的投资热点)
(Golden Triangle of Yellow River becomes a new invest hotspot)

Table 3 shows the statistics of structural trios. Type 2 (Substantive-Modifier) and Type 4 (Verb-Object) trios are the majority in opinion sentences, while Type 5 (Verb-Complement) trios are comparably few. Figure 5 analyzes trios by polarity. The distributions of trios appearing in positive, neutral, and negative opinion sentences are similar. If we further check the ratios of four types in each polarity shown in Figure 6, we can find that there are more Type 4 (Verb-Object) trios in positive opinion sentences, compared to Type 4 trios in neutral and negative sentences.

Figure 7 shows the analysis of structural trios by opinion type. We can find that the distributions of opinions are similar in all four trio types. Figure 8 illustrates an interesting comparison. In action opinion sentences, Type 4 trios appear more often, while in expression and status opinion sentences, Type 2 trios are the majority.

<table>
<tr><td colspan="3">S ID=230: 黄河“金三角”成为新的投资热点
(Figure 3)</td></tr>
<tr><td>.node file</td><td>.tree file</td><td>.trio file</td></tr>
<tr><td colspan="3">Fields</td></tr>
<tr><td>Node ID, POS, node content, node depth</td><td>Node ID: children</td><td>Trio ID, trio head, trio left node, trio right node, trio type</td></tr>
<tr><td colspan="3">Content</td></tr>
<tr><td>0,,,0
1,IP-HLN,,1
2,NP-SBJ,,2
3,NP-PN,,3
4,NR,黄河,4
5,NP,,3
6,PU,“,4
7,NN,金三角,4
8,PU,”,4
9,VP,,2
10,VV,成为,3
11,NP-OBJ,,3
12,CP,,4
13,WHNP-1,,5
14,-NONE-,*OP*,6
15,CP,,5
16,IP,,6
17,NP-SBJ,,7
18,-NONE-,*T*-1,8
19,VP,,7
20,VA,新,8
21,DEC,的,6
22,NP,,4
23,NN,投资,5
24,NN,热点,5</td><td>0:1,
1:2,9,
2:3,5,
3:4,
4:
5:6,7,8,
6:
7:
8:
9:10,11,
10:
11:12,22,
12:13,15,
13:14,
14:
15:16,21,
16:17,19,
17:18,
18:
19:20,
20:
21:
22:23,24,
23:
24:</td><td>2,1,2,9,3
3,22,23,24,2</td></tr>
<tr><td colspan="3">Opinion labels of three annotators
(filename, SID, opinion, polarity, opinion type)</td></tr>
<tr><td colspan="3">chtb_020.raw,230,N,,
chtb_020.raw,230,Y,POS,STATE
chtb_020.raw,230,Y,POS,STATE</td></tr>
<tr><td colspan="3">Opinion gold standard</td></tr>
<tr><td colspan="3">chtb_020.raw,230,Y,POS,STATE</td></tr>
</table>

Table 1: An example annotation in Chinese Opinion Treebank

<table>
<tr><td></td><td colspan="4">Opinion</td><td>Non-Opinion</td></tr>
<tr><td>Polarity</td><td>Positive</td><td colspan="2">Neutral</td><td>Negative</td><td rowspan="6"></td></tr>
<tr><td>#</td><td>6,916</td><td colspan="2">1,824</td><td>1,937</td></tr>
<tr><td>%</td><td>64.78</td><td colspan="2">17.08</td><td>18.14</td></tr>
<tr><td>Type</td><td>Exp</td><td>Status</td><td>Act</td><td>N/A</td></tr>
<tr><td>#</td><td>4,240</td><td>4,072</td><td>722</td><td>1,643</td></tr>
<tr><td>%</td><td>39.71</td><td>38.14</td><td>6.76</td><td>15.39</td></tr>
<tr><td>Total #</td><td colspan="4">10,677</td><td>8,108</td></tr>
<tr><td>Total %</td><td colspan="4">56.84</td><td>43.16</td></tr>
</table>

Table 2: Statistics of opinions

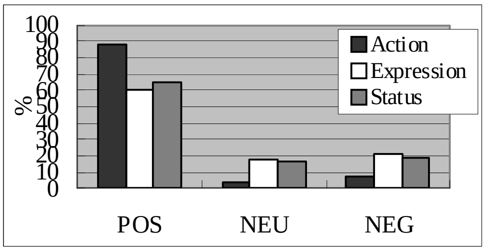


Figure 4: Statistics of opinions by type

Trio Type	Number	Percentage %
2	20,061	36.92
3	15,544	28.61
4	17,580	32.36
5	1,147	2.11
Total	54,332	100.00

Table 3: Statistics of structural trios

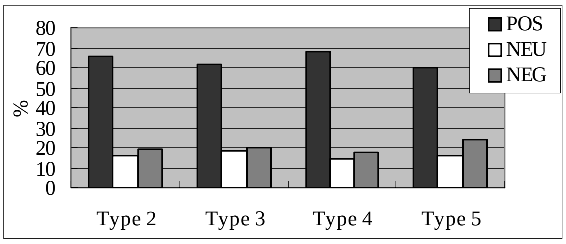


Figure 5: Statistics of structural trios by polarity

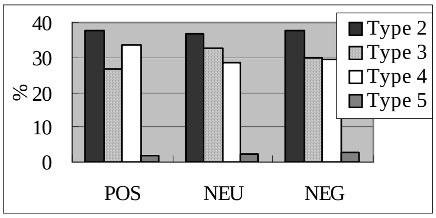


Figure 6: Statistics of opinion polarities by structural trio

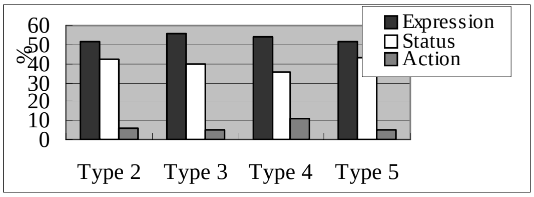


Figure 7: Statistics of structural trios by opinion type

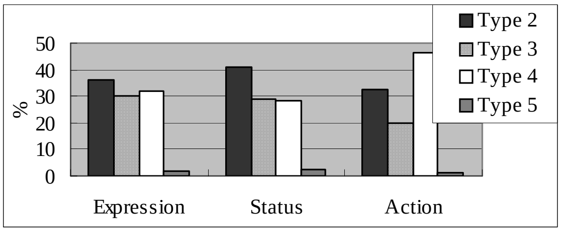


Figure 8: Statistics of opinion types by structural trio

5. Conclusion

We have constructed a Chinese Opinion Treebank, which includes 18,785 sentences. Information including opinions, their polarities, types, and structural trios is annotated. The substantial agreement between annotations ensures the applicability and reliability of the constructed corpus.

We have applied this corpus and obtain a preliminary result (Ku *et al.*, 2009). The influence of the performance of text pre-processing on opinion analysis, and the usages of the linguistic cues for the opinion analysis will be investigated.

6. References

Bai, X., Padman, R. and Airoldi, E. (2005). On learning parsimonious models for extracting consumer opinions. Proceedings of the 38th HICSS, Track 3, Volume 03, page 75.2.

Cheng, X.-H. and Tian, X.-L. (1992). Modern Chinese. Bookman Books Ltd.

Ghose, A. and Ipeirotis, P. (2007). Designing novel review ranking systems: Predicting usefulness and impact of reviews. Proceedings of ICEC, Invited paper.

Ku, L.-W., Lo, Y.-S. and Chen, H.-H. (2007). Test Collection Selection and Gold Standard Generation for a Multiply-Annotated Opinion Corpus. Proceedings of 45th ACL, pp. 89-92.

Ku, L.-W., Huang, T.-H. and Chen, H.-H. (2009). Using Morphological and Syntactic Structures for Chinese Opinion Analysis. Proceedings of EMNLP 2009, pp. 1260-1269.

Seki, Y., Evans, D. K., Ku, L.-W., Sun, L., Chen, H.-H. and Kando, N. (2008). Overview of Multilingual Opinion Analysis Task at NTCIR-7. Proceedings of the 7th NTCIR.

Riloff, E., Wiebe, J. and Wilson, T. (2003). Learning subjective nouns using extraction pattern bootstrapping. Proceedings of the 7th CoNLL, pp. 25-32.

Wiebe, J. (2000). Learning Subjective Adjectives from Corpora. Proceeding of 17th National Conference on Artificial Intelligence, pp. 735-740.

Wiebe, J., Breck, E., Buckly, C., Cardie, C., Davis, P., Fraser, B., Litman, D., Pierce, D., Riloff, E., and Wilson, T. (2002). NRRC summer workshop on multi-perspective question answering, final report. ARDA NRRC Summer 2002 Workshop.