Construction of a Chinese Opinion Treebank

Lun-Wei Ku, Ting-Hao Huang, Hsin-Hsi Chen
CSIE, National Taiwan University
Taipei, Taiwan
Overview

- Introduction
- Tagging Scheme
- Sources of Corpus
- Annotation Tools
- Statistics of Chinese Opinion Treebank
- Potential Applications
- Conclusion
Why opinion processing is important?

- Documents discussing public affairs, common themes, interesting products, and other topics are reported and distributed on the Web.
  - review sites, forum, discussion groups, blogs, news, …
- Watching specific information sources and summarizing the newly discovered opinions are important
  - for governments to improve their services,
  - for companies to market their products, and
  - for customers to purchase their objects.
Motivation

- Bag of Word Model is not Enough
  - 提高薪資 (increase salary) is positive for individuals
  - 提高稅率 (increase tax rate) is negative for individuals, but positive for government

- Morphological Structures
  - Predicting Morphological Types of Chinese Bi-Character Words by Machine Learning Approaches (Huang, Ku, and Chen, LREC 2010)

- Syntactic Structures
Opinionated Sentences

- Three formal types of opinion are defined based on Wiebe et al. (2005)
  - Explicit mentions of opinion
    - Psychologists argue that teenagers are not old ...
  - Speech events expressing opinion
    - Ito said the government must concentrate now ...
  - Expressive subjective elements
    - Japan must seem to be a country full of antiquated rules.
      (Opinion holder: author)
MOAT in NTCIR

- Multilingual Opinion Analysis Task held at NTCIR-6, -7, and -8 (2006-current)
- Languages
  - English, Chinese, and Japanese
- Subtasks:
  - Opinion detection, polarity judgment, holder, target, and relevance judgment, etc.
- No opinion type and syntactic information
Tagging scheme

- Opinion: yes or no
- Opinion Polarity: positive, neutral, negative
- Opinion Types:
  - expression: people reflect their subjective judgment.
  - status: the subjective information appears as descriptions.
  - action: the opinion holder’s attitude is revealed by his action.
A Tagging scheme

- Structural relations between words in terms of trio

  - Parallel Type: 美麗 (beautiful) 而 (and) 聰慧 (smart)
  - Substantive-Modifier Type: 淒涼地 (sadly) 笑著 (laugh)
  - Subjective-Predicate Type: 討論 (discussion) 熱烈 (enthusiastic)
  - Verb-Object Type: 恢復 (overcome) 疲勞 (tiredness)
  - Verb-Complement Type: 收拾 (put things) 乾淨 (in order)
Source of Annotation Corpus

- Chinese Treebank 5.1
  - Purchased from Linguistic Data Consortium
  - 507,222 words
  - 824,983 Hanzi
  - 18,782 sentences
  - 890 data files
Annotation Tool (OAT)

With industrialization and the development of modern science, the average life span has extended and, naturally, people have come to want even longer and healthier lives.

Acupuncture, acupressure therapy... These Oriental medical therapies, which have been practiced for thousands of years, haven’t really been given the attention they deserve from the rest of the world.

That’s because Oriental medicine was not well explained in terms of Western science.

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 been published in Korea.

Ji Man-ho, publisher of the monthly magazine and president of Maet Health Magazine Co., said, “In the 21st century, Korean traditional medicine, while improving people’s health, also needs to make a great effort to re-examine its role as an independent medical science.”
(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)
### Chinese Opinion Treebank

- **Statistics of opinion sentences**

| Type   | Opinion | Non-Opinion |
|--------|---------|-------------|
| Polarity | Positive | Neutral | Negative | Total # |
| #      | 6,916   | 1,824     | 1,937     | 10,677 |
| %      | 64.78   | 17.08     | 18.14     | 8,108  |
| Type   | Exp.    | Status    | Act.      | N/A    |
| #      | 4,240   | 4,072     | 722       | 1,643  |
| %      | 39.71   | 38.14     | 6.76      | 15.39  |

Total %: 56.84 | 43.16
Chinese Opinion Treebank

- average kappa value:
  - 0.49 (moderate agreement) between two annotators
  - 0.73 (substantial agreement) between one annotator and the lenient gold standard
黄河“金三角”成为新的投资热点

| Node ID, POS, node content, node depth | Node ID: children | Trio ID, trio head, trio left node, trio right node, trio type |
|--------------------------------------|-------------------|---------------------------------------------------------------|
| 0,,0                                 | 0:1, 1:2, 9, 2:3, 5, 3:4, 4, 5:6, 7, 8, 6:7, 8 | 2,1,2,9,3, 3,22,23,24,2 |
| 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                        |                   |                                                               |

Opinion labels of three annotators:
- annotators

Opinion gold standard:
- chtb_020.raw,230,Y,POS,STATE
## Chinese Opinion Treebank

| 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       |
Discussion

- Statistics of opinions by type
- 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.
**Discussion**

- Statistics of structural trios by polarity
- The distributions of trios appearing in positive, neutral, and negative opinion sentences are similar.

![Bar chart showing the distribution of trios by polarity](chart.png)
Discussion

- Statistics of opinion polarities by structural trio
- There are more Type 4 (Verb-Object) trios in positive opinion sentences.
Discussion

- Statistics of structural trios by opinion type
- The distributions of opinions are similar in all four trio types.
Discussion

- Statistics of opinion types by structural trio.
- In action opinion sentences, Type 4 trios appear more often, while in expression and status opinion sentences, Type 2 trios are the majority.
### Potential Applications in Opinion Mining (EMNLP 2009)

| Setting | Word [w] | Sentence [s] | f-Score (opinion) | f-Score (polarity) |
|---------|----------|--------------|-------------------|-------------------|
| 1       | bag      | bag          | 0.7073            | 0.4988            |
| 2       | struc    | bag          | 0.7162            | 0.5117            |
| 3       | bag      | struc        | 0.8000            | 0.5361            |
| 4       | struc    | struc        | 0.7922            | 0.5297            |
| 5       | struc    | struc        | 0.7993            | 0.5187            |
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.
Thanks & Comments