A Political News Corpus in Chinese for Opinion Analysis

Benjamin K. Tsou, Bin Lu
Language Information Sciences Research Centre
City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong
rlbtsou@cityu.edu.hk, lubin2@student.cityu.edu.hk

Abstract

In this paper, we present an annotated corpus of political election news in Chinese for opinion analysis, and discuss some issues in the manual annotation process. The annotation scheme is described with examples, and inter-annotator agreement is explored for different levels of annotation: expression, sentence and document.

Keywords: *Chinese opinion extraction, political election news, corpus annotation, agreement study.*

1 Introduction

Opinions incorporated in factual news reports represent a common phenomenon, and many applications would benefit from being able to automatically identify and analyze opinions in the political and commercial domains. Although many researchers have studied opinion analysis from various perspectives, few annotated corpora are available. Wiebe et al. [4] described the MPQA corpus of 10,000 sentences, in which they annotated texts at the word- and phrase-level in context. For the opinion analysis task at NTCIR-6 [1] and NTCIR-7, news documents in Chinese, Japanese and English were annotated as well.

In this paper, we describe an annotated corpus of political election news in Chinese. A novel annotation scheme was used to annotate opinions at different levels simultaneously: the expression level (including word, phrase or clause), the sentence level and the document level. This annotated corpus is expanding and has enabled previous research like [3]. The rest of this paper is organized as follows. Section 2 presents the annotation scheme. Data collection is described at Section 3. Section 4 gives the results of an inter-annotator agreement study and finally, section 5 concludes this paper.

2 Annotation scheme

The aim of this annotation scheme is to analyze the opinions towards the election candidates in the Chinese political news. Three levels of opinion information were annotated in our corpus: the expression level, the sentence level, and the document level, which are introduced in the following subsections, respectively.

2.1 Expression level annotation

In our annotation scheme, polar expressions are categorized into two classes: SPW (Salient Polar Word) and chunk (Polar Chunk). An SPW is a word which is inherently positive or negative while expressions more than a word are excluded. A *chunk* refers to a polar expression which is more than a word. It can be a compound word, a phrase or a clause, and its opinion and polarity cannot be expressed by only a word under normal circumstances. There are three kinds of *chunks*:

- **Collocations** - two or more words combined together to form a polar expression. For example, 豎起 (erect) & *拇指* (thumb) are two neutral Chinese words when separated. If combined together, the expression *豎起拇指 (thumbs up)* is positive in the sentence below: *陳先生豎起拇指大贊曾蔭權 (Mr. Chen gave thumbs up to and praised Donald Tsang)….*
- **Context-dependent expressions** - an expression whose polarity depends on the context, for example *有經驗 (experienced), 好/壞的經驗 (good/bad experience)*. The Chinese word *經驗 (experience)* is contextual and its polarity depends on the contexts.
- **SPW with contextual valence shifter** - an SPW together with a contextual valence shifter, for example *很成功 (very successful)*. The word 很 (very) is a Chinese contextual valence shifter which would change the intensity to which an expression is positive or negative.

For each opinion expression, we employed a common frame, including *expression itself, opinion holder, opinion target, polarity, intensity of the polarity*. Each SPW or chunk is marked with a polarity (i.e. positive, negative and neutral) and a score of polarity intensity (0-3). The opinion target and the opinion holder are also identified for each *SPW* and *chunk,* if available. The *opinion target* is the person or entity toward whom the opinion is targeted. The *opinion holder* is the one (other than the writer) who expresses the opinion. Opinion holder and opinion target are marked in the form as they appear in the sentence, whether it is a proper noun, a pronoun, a nickname, etc.

2.2 Sentence level annotation

If a sentence as a whole conveys any positive or negative information toward an entity, annotators would mark it at the sentence level. The above frame, including *opinion holder, opinion target, polarity, intensity of the polarity,* was also employed here. Opinion target and opinion holder are also identified for each sentence, if available.

2.3 Document level annotation

Since the political news articles in our corpus are almost on elections, we also marked them in terms of the following two aspects: *focus person* and *focus event*. *Focus person* refers to the candidate(s) or highly related person(s) in the given elections. For news articles on the 2008 US presidential election, the candidates including Barack Obama, John McCain, Joe Biden and Sarah Palin, and highly related celebrities such as George W. Bush or Hillary Clinton, are treated as *focus persons*. *Focus event* refers to the major event(s) discussed in the articles. The *focus person(s)* and *focus event(s)* involved are marked with their overall polarity. If two or more *focus persons or focus events* occur in an article, we would mark all of them.

3 Data collection

Political news documents of three elections were extracted from the LIVAC synchronous corpus [2] which was the 12-year news coverage of Chinese communities, including Hong Kong, Beijing and Taiwan. More than 10 annotators were trained to annotate this corpus. They were asked to annotate sample documents according to the annotation scheme. The annotators met regularly to discuss problems they encountered in order to maintain consistency and agreement, and to revise the annotation scheme as appropriate.

Up to now, more than 1,700 documents containing more than 33,000 sentences have been marked, and the statistics is presented in Table 1. For each document, we made sure that at least three annotators marked it.

Table 1. Statistics of annotated data

Election title	#doc	#sentence
2004 US presidential election	566	11,800
2007 HK chief executive election	1,028	17,880
2008 US presidential election	190	3,379

4 Agreement Study

Three annotators who participated in the agreement study were all trained as described above. They were given 56 documents with a total of 956 sentences to annotate. For the evaluation of expression and document level, we employ the *agr* metric [4] rather than Cohen's Kappa to measure agreement in identifying SPW, chunk, focus person and focus event because the annotators would identify different text spans. For the sentence level, Cohen's Kappa was employed to measure agreement in identifying the opinionated sentences.

For the expression level, the average *agr*s for SPW and chunk are 0.70 and 0.42, respectively. These *arg*s are comparable with 0.72 [4], the *agr* of expressive subjective elements in the MPQA corpus, in which the partial match was used while we only consider the exact match of SPW and chunk as an agreement. For the document level, the average *agr*s for *focus person* and *focus event* are 0.82 and 0.64, respectively.

For the sentence level, if a sentence was marked at the sentence level, or an SPW or chunk was marked in it, we would consider it an opinionated sentence, otherwise non-opinionated. The average pair-wise Cohen's Kappa and *agr* value for opinionated judgment over all sentences are 0.62 and 0.87 respectively, which are slightly lower than 0.77 and 0.90 [4] for sentence level objective or subjective judgment in the MPQA corpus, but much higher than the average Kappa 0.23 [1] for the opinionated tagging tasks in the Chinese corpus of the opinion analysis task at NTCIR-6.

5 Conclusion and future work

In this paper, we describe a novel annotation scheme and an annotated corpus of political election news, which we believe is valuable for the study of opinion analysis in Chinese. The inter-annotator agreement was explored and the agreement results show that the consistency between different annotators is high on several levels.

The multi-level and fine-grained annotation of this corpus would be valuable for many NLP applications. We have planned to make part of it public to the research community in the near future. The future work includes the investigation of how the corpus could be used in the evaluation of Chinese opinion analysis.

Acknowledgement

Research presented here is supported in part by a Earmarked Research Grant (CERG) of the HKSAR Research Grants Council under grant No. CityU149607.

References

[1] Seki Y., Evans D.K., Ku L.W., Chen H.H., Kando N., and Lin C.-Y. 2007. Overview of opinion analysis pilot task at NTCIR-6. *Proc. of the Sixth NTCIR Workshop*. May 2007, Japan.

[2] Tsou B.K.Y., Tsoi W.F., Lai T.B.Y., Hu J., and Chan S.W.K. 2000. LIVAC, A Chinese Synchronous Corpus, and Some Applications. *Proceedings of the ICCLC International Conference on Chinese Language Computing*, Chicago. pp. 233–238.

[3] Tsou B.K.Y., Yuen W.M.R., Kwong O.Y., Lai T.B.Y., Wong W.L. 2005. Polarity classification of celebrity coverage in the Chinese press. In *Proceeding of the 2005 International Conference on Intelligence Analysis*. Virginia, USA.

[4] Wiebe J., Wilson T., Cardie C. 2005. Annotating Expressions of Opinions and Emotions in Language, *Language Resources and Evaluation*, volume 39, issue 2-3, pp. 165-210.