Test Collection Selection and Gold Standard Generation for a Multiply-Annotated Opinion Corpus

Lun-Wei Ku, Yong-Shen Lo and Hsin-Hsi Chen
Department of Computer Science and Information Engineering
National Taiwan University
{lwku, yslo}@nlg.csie.ntu.edu.tw; hhchen@csie.ntu.edu.tw

Abstract

Opinion analysis is an important research topic in recent years. However, there are no common methods to create evaluation corpora. This paper introduces a method for developing opinion corpora involving multiple annotators. The characteristics of the created corpus are discussed, and the methodologies to select more consistent testing collections and their corresponding gold standards are proposed. Under the gold standards, an opinion extraction system is evaluated. The experiment results show some interesting phenomena.

1 Introduction

Opinion information processing has been studied for several years. Researchers extracted opinions from words, sentences, and documents, and both rule-based and statistical models are investigated (Wiebe *et al*., 2002; Pang *et al*., 2002). The evaluation metrics precision, recall and f-measure are usually adopted.

A reliable corpus is very important for the opinion information processing because the annotations of opinions concern human perspectives. Though the corpora created by researchers were analyzed (Wiebe *et al*., 2002), the methods to increase the reliability of them were seldom touched. The strict and lenient metrics for opinions were mentioned, but not discussed in details together with the corpora and their annotations.

This paper discusses the selection of testing collections and the generation of the corresponding gold standards under multiple annotations. These testing collections are further used in an opinion extraction system and the system is evaluated with the corresponding gold standards. The analysis of human annotations makes the improvements of opinion analysis systems feasible.

2 Corpus Annotation

Opinion corpora are constructed for the research of opinion tasks, such as opinion extraction, opinion polarity judgment, opinion holder extraction, opinion summarization, opinion question answering, etc.. The materials of our opinion corpus are news documents from NTCIR CIRB020 and CIRB040 test collections. A total of 32 topics concerning opinions are selected, and each document is annotated by three annotators. Because different people often feel differently about an opinion due to their own perspectives, multiple annotators are necessary to build a reliable corpus. For each sentence, whether it is relevant to a given topic, whether it is an opinion, and if it is, its polarity, are assigned. The holders of opinions are also annotated. The details of this corpus are shown in Table 1.

	Topics	Documents	Sentences
Quantity	32	843	11,907

Table 1. Corpus size

3 Analysis of Annotated Corpus

As mentioned, each sentence in our opinion corpus is annotated by three annotators. Although this is a must for building reliable annotations, the inconsistency is unavoidable. In this section, all the possible combinations of annotations are listed and two methods are introduced to evaluate the quality of the human-tagged opinion corpora.

3.1 Combinations of annotations

Three major properties are annotated for sentences in this corpus, i.e., the relevancy, the opinionated issue, and the holder of the opinion. The combinations of relevancy annotations are simple, and annotators usually have no argument over the opinion holders. However, for the annotation of the opinionated issue, the situation is more com-

Proceedings of the ACL 2007 Demo and Poster Sessions, pages 89–92,
Prague, June 2007.

plex. Annotations may have an argument about whether a sentence contains opinions, and their annotations may not be consistent on the polarities of an opinion. Here we focus on the annotations of the opinionated issue. Sentences may be considered as opinions only when more than two annotators mark them opinionated. Therefore, they are targets for analysis. The possible combinations of opinionated sentences and their polarity are shown in Figure 1.

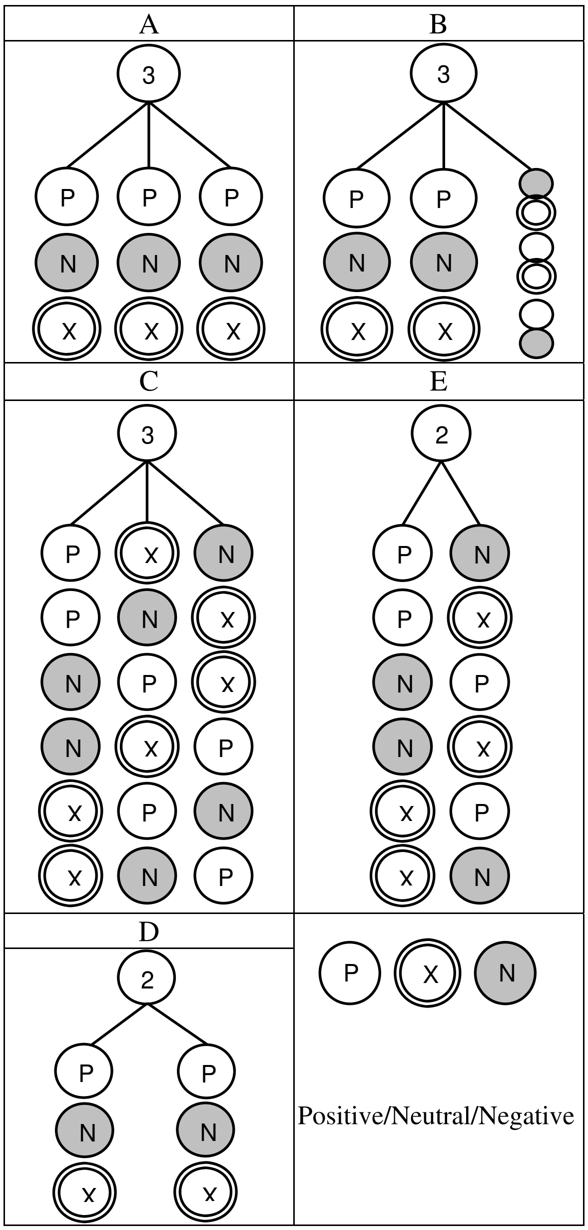

Figure 1. Possible combinations of annotations

In Figure 1, Cases A, B, C are those sentences which are annotated as opinionated by all three annotators, while cases D, E are those sentences which are annotated as opinionated only by two annotators. In case A and case D, the polarities annotated by annotators are identical. In case B, the polarities annotated by two of three annotators are agreed. However, in cases C and E, the polarities annotated disagree with each other. The statistics of these five cases are shown in Table 2.

Case	A	B	C	D	E	All
Number	1,660	1,076	124	2,413	1,826	7,099

Table 2. Statistics of cases A-E

3.2 Inconsistency

Multiple annotators bring the inconsistency. There are several kinds of inconsistency in annotations, for example, relevant/non-relevant, opinionated/non-opinionated, and the inconsistency of polarities. The relevant/non-relevant inconsistency is more like an information retrieval issue. For opinions, because their strength varies, sometimes it is hard for annotators to tell if a sentence is opinionated. However, for the opinion polarities, the inconsistency between positive and negative annotations is obviously stronger than that between positive and neutral, or neutral and negative ones. Here we define a sentence "strongly inconsistent" if both positive and negative polarities are assigned to a sentence by different annotators. The strong inconsistency may occur in case B (171), C (124), and E (270). In the corpus, only about 8% sentences are strongly inconsistent, which shows the annotations are reliable.

3.3 Kappa value for agreement

We further assess the usability of the annotated corpus by Kappa values. Kappa value gives a quantitative measure of the magnitude of inter-annotator agreement. Table 3 shows a commonly used scale of the Kappa values.

Kappa value	Meaning
<0	less than change agreement
0.01-0.20	slight agreement
0.21-0.40	fair agreement
0.41-0.60	moderate agreement
0.61-0.80	substantial agreement
0.81-0.99	almost perfect agreement

Table 3. Interpretation of Kappa value

The inconsistency of annotations brings difficulties in generating the gold standard. Sentences should first be selected as the testing collection, and then the corresponding gold standard can be generated. Our aim is to generate testing collections and their gold standards which agree mostly to annotators. Therefore, we analyze the kappa value not between annotators, but between the annotator and the gold standard. The methodologies are introduced in the next section.

4 Testing Collections and Gold Standards

The gold standard of relevance, the opinionated issue, and the opinion holder must be generated according to all the annotations. Answers are chosen based on the agreement of annotations. Considering the agreement among annotations themselves, the strict and the lenient testing collections and their corresponding gold standard are generated. Considering the Kappa values of each annotator and the gold standard, topics with high agreement are selected as the testing collection. Moreover, considering the consistency of polarities, the substantial consistent testing collection is generated. In summary, two metrics for generating gold standards and four testing collections are adopted.

4.1 Strict and lenient

Namely, the strict metric is different from the lenient metric in the agreement of annotations. For the strict metric, sentences with annotations agreed by all three annotators are selected as the testing collection and the annotations are treated as the strict gold standard; for the lenient metric, sentences with annotations agreed by at least two annotators are selected as the testing collection and the majority of annotations are treated as the lenient gold standard. For example, for the experiments of extracting opinion sentences, sentences in cases A, B, and C in Figure 1 are selected in both strict and lenient testing collections, while sentences in cases D and E are selected only in the lenient testing collection because three annotations are not totally agreed with one another. For the experiments of opinion polarity judgment, sentences in case A in Figure 1 are selected in both strict and lenient testing collections, while sentences in cases B, C, D and E are selected only in the lenient testing collection. Because every opinion sentence should be given a polarity, the polarities of sentences in cases B and D are the majority of annotations, while the polarity of sentences in cases C are given the polarity neutral in the lenient gold standard. The polarities of sentences in case E are decided by rules P+X=P, N+X=N, and P+N=X. As for opinion holders, holders are found in opinion sentences of each testing collection. The strict and lenient metrics are also applied in annotations of relevance.

4.2 High agreement

To see how the generated gold standards agree with the annotations of all annotators, we analyze the kappa value from the agreements of each annotator and the gold standard for all 32 topics. Each topic has two groups of documents from NTCIR: very relevant and relevant to topic. However, one topic has only the relevant type document, it results in a total of 63 (2*31+1) groups of documents. Note that the lenient metric is applied for generating the gold standard of this testing collection because the strict metric needs perfect agreement with each annotator's annotations. The distribution of kappa values of 63 groups is shown in Table 4 and Table 5. The cumulative frequency bar graphs of Table 4 and Table 5 are shown in Figure 2 and Figure 3.

Kappa	<=0	0-0.2	0.21-0.4	0.41-0.6	0.61-0.8	0.81-0.99
Number	1	2	12	14	33	1

Table 4. Kappa values for opinion extraction

Kappa	<=0	0-0.2	0.21-0.4	0.41-0.6	0.61-0.8	0.81-0.99
Number	9	0	7	21	17	9

Table 5. Kappa values for polarity judgment

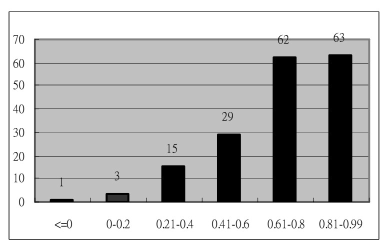

Figure 2. Cumulative frequency of Table 4

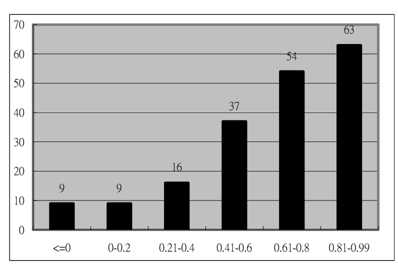

Figure 3. Cumulative frequency of Table 5

According to Figure 2 and Figure 3, document groups with kappa values above 0.4 are selected as the high agreement testing collection, that is, document groups with moderate agreement in Table 3. A total of 48 document groups are collected for opinion extraction and 47 document groups are collected for opinion polarity judgment.

4.3 Substantial Consistency

In Section 3.2, sentences which are “strongly inconsistent” are defined. The substantial consistency test collection expels strongly inconsistent sentences to achieve a higher consistency. Notice that this test collection is still less consistent than the strict test collection, which is perfectly consistent with annotators. The lenient metric is applied for generating the gold standard for this collection.

5 An Opinion System -- CopeOpi

A **C**hinese **op**inion **e**xtraction system for **op**inionated **i**nformation, CopeOpi, is introduced here. (Ku *et al.*, 2007) When judging the opinion polarity of a sentence in this system, three factors are considered: sentiment words, negation operators and opinion holders. Every sentiment word has its own sentiment score. If a sentence consists of more positive sentiments than negative sentiments, it must reveal something good, and vice versa. However, a negation operator, such as ”not” and ”never”, may totally change the sentiment polarity of a sentiment word. Therefore, when a negation operator appears together with a sentiment word, the opinion score of the sentiment word S will be changed to -S to keep the strength but reverse the polarity. Opinion holders are also considered for opinion sentences, but how they influence opinions has not been investigated yet. As a result, they are weighted equally at first. A word is considered an opinion holder of an opinion sentence if either one of the following two criteria is met:

1. The part of speech is a person name, organization name or personal.
2. The word is in class A (human), type Ae (job) of the Cilin Dictionary (Mei *et al.*, 1982).

6 Evaluation Results and Discussions

Experiment results of CopeOpi using four designed testing collections are shown in Table 6. Under the lenient metric with the lenient test collection, f-measure scores 0.761 and 0.383 are achieved by CopeOpi. The strict metric is the most severe, and the performance drops a lot under it. Moreover, when using high agreement (H-A) and substantial consistency (S-C) test collections, the performance of the system does not increase in portion to the increase of agreement. According to the agreement of annotators, people should perform best in the strict collection, and both high agreement and substantial consistency testing collections are easier than the lenient one. This phenomenon shows that though this system’s performance is satisfactory, its behavior is not like human beings. For a computer system, the lenient testing collection is fuzzier and contains more information for judgment. However, this also shows that the system may only take advantage of the surface information. If we want our systems really judge like human beings, we should enhance the performance on strict, high agreement, and substantial consistency testing collections. This analysis gives us, or other researchers who use this corpus for experiments, a direction to improve their own systems.

	Opinion Extraction			Opinion + Polarity		
Measure	P	R	F	P	R	F
Lenient	0.664	0.890	0.761	0.335	0.448	0.383
Strict	0.258	0.921	0.404	0.104	0.662	0.180
H-A	0.677	0.885	0.767	0.339	0.455	0.388
S-C				0.308	0.452	0.367

Table 6. Evaluation results

Acknowledgments

Research of this paper was partially supported by Excellent Research Projects of National Taiwan University, under the contract 95R0062-AE00-02.

References

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

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

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*.

Ku, L.-W., Wu, T.-H., Li, L.-Y. and Chen., H.-H. (2007). Using Polarity Scores of Words for Sentence-level Opinion Extraction. *Proceedings of the Sixth NTCIR Workshop*.