Tagging Heterogeneous Evaluation Corpora for Opinionated Tasks

Lun-Wei Ku, Yu-Ting Liang and Hsin-Hsi Chen

Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan
{lwku, eagan}@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 supportive and the non-supportive evidence of a document. To evaluate the performance of technologies for opinionated tasks, a suitable corpus is necessary. This paper defines the annotations for opinionated materials. Heterogeneous experimental materials are annotated, and the agreements among annotators are analyzed. How human can monitor opinions of the whole is also examined. The corpus can be employed to opinion extraction, opinion summarization, opinion tracking and opinionated question answering.

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 to improve their services or products (Dave *et. al*, 2003).

Opinion extraction, opinion summarization and opinion tracking are three important techniques for understanding opinions (Ku, Liang and Chen, 2006). Opinion extraction mines opinions at word, sentence and document levels from articles. Opinion summarization summarizes opinions of articles by telling sentiment polarities, degrees and the correlated events. Moreover, opinion tracking monitors the changes of opinions over time.

Recently, several works dealt with opinion retrieval or opinion extraction. Wiebe, Wilson and Bell recognized opinionated documents (Wiebe *et. al.*, 2002). Pang, Lee, and Vaithyanathan classified documents by overall sentiment instead of topics (Pang *et. al.*, 2002). Dave's and Hu's researches (Dave *et. al.*, 2003; Hu and Liu, 2004) both focused on extracting opinions of reviews. Of course, the smallest unit of opinions is not a document. Riloff and Wiebe distinguished subjective sentences from objective ones (Riloff and Wiebe, 2003). Kim and Hovy proposed a sentiment classifier for English words and sentences, which utilized thesauri (Kim and Hovy, 2003).

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 (Pang *et. al.*, 2002). Both information retrieval (Dave *et. al.*, 2003) and information extraction (Cardie *et. al.*, 2003) technologies have also been explored. A statistical model was used for mining sentiment words too, but the experiment material was not described in detail (Takamura *et. al.*, 2005). The results for various metrics and heuristics varied depending on the testing situations.

To evaluate the performance of opinionated tasks, a set of annotations which integrate heterogeneous sources becomes indispensable. Ku, Wu, Lee and Chen (2005) constructed a corpus for opinion extraction. Opinions are always expressed towards a specific target. Therefore, relevance information is critical for opinionated tasks (Ku *et. al.*, 2005). To utilize resources and systems well developed for the research of information retrieval, compatible tags are necessary. However, the researches mentioned did not take this issue into consideration.

This paper studies the tagging format of corpora for opinion processing, the issues of inter-annotator agreement, and the effects of document sources. Opinionated annotations on TREC[1] and NTCIR[2] corpora are defined. Two sources of information are collected for the experiments, i.e., news and blog articles. The writing of the former is comparatively formal to that of the latter because blog articles, which express personal opinions of the writers, are often written in a casual style. These annotated resources are ready for opinionated tasks such as opinion extraction, opinion summarization, opinion tracking and opinionated question answering.

2. Annotation Format

Because opinions can be expressed in different granularities such as documents, sentences and words, an evaluation corpus should reflect such phenomena. Table 1 lists the annotation tags and their corresponding descriptions. Table 2 depicts the meanings of the attribute values. Every element has a pair of opening and closing tags as the XML language.

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

Table 1. Tag descriptions

<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 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 1. Tag descriptions (*Continued*)

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

Table 2. Abbreviations of attribute values and their meanings

Tag <OPINION_SEG> is especially useful in dealing with multi-perspective or opinion holder related issues. Consider an example shown as follows.

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

```
- <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 1. A tagging illustration of this example

Nested relations of opinion holders are critical to identify the owners of opinions, that is, multi-perspective issues. XML-like tags can easily represent nested relations and they are consistent with the tagging style of the famous TREC, CLEF and NTCIR information

retrieval evaluation corpora. If tagging with the above format, the evaluation corpora from the three worldwide IR forums can be reusable. A Chinese and an English tagging examples, selected from NTCIR-2 corpus, are illustrated in Figures 2 and 3, respectively.

```
- <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」為最高原則,希望外界不要過於
    <SENTIMENT_KW TYPE="NEG">焦慮</SENTIMENT_KW>
    。
  </OPINION_SEG>
</SEN_ATTITUDE>
```

Figure 2. 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 3. Civil ID card example in English

These two figures show a passage opinion for topic ZH021 of NTCIR-2 in Chinese and in English shown in Table 3. It contains an opinion keyword "表示" (point out) and a negation "不要" (**should not**) which modifies a non-supportive keyword "焦慮" (concerned). The negation reverses the sentiment polarity from negative to positive. The opinion holder is "李雪津" (Hsuehchin Li).

3. Source of Documents

To compare the characteristics of different information sources, news articles in TREC 2003 novelty track and NTCIR-2 are adopted, and blog articles are selected from the web. In novelty track (Soboroff and Harman, 2003), there are 50 document sets, and each set has 25 documents. All documents in the same set are relevant to one topic. Total 22 topics are opinionated. Chen and Chen (2002) developed a test collection CIRB010 for Chinese information retrieval in NTCIR-2. It consists of 50 topics and 6 of them are opinionated topics. Opinionated topics for annotation are shown in Table 3.

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 3. Opinionated Topics in CIRB010

Total 192 relevant documents of the 6 opinionated topics are annotated.

Blog is a new rising community, and articles inside express many personal opinions. It is selected as the third source. Documents selected from these three sources are in different languages, so they are useful for studying cross-lingual opinionated issues. Besides, they are from the public media and the web. Opinions from different social classes can be compared. To study the effects of different sources, articles of the same topic are used. Documents of the opinionated topic, Set 2 ("clone Dolly sheep"), in TREC corpus and documents of an additional topic "animal cloning" of NTCIR-3 are selected as the experimental material. For blogs, we retrieve documents from blog portals by the query "animal cloning". Numbers of documents related to "animal cloning" are listed in Table 4.

Source	TREC	NTCIR	BLOG
Quantity	25	17	20

Table 4. Quantities of documents for opinion summarization

4. Inter-Annotator Agreement

The agreement of annotations is analyzed to study the characteristics of opinions. Two Chinese materials, i.e., NTCIR and BLOG, are annotated for the analyses of the inter-annotator agreement. The metric of the inter-annotator agreement between annotators A and B is shown is Formula 1.

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

The agreements at word, sentence and document levels are listed in Tables 5, 6, and 7, 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 5. Agreements of annotators at word level

Annotators	A vs. B	B vs. C	C vs. A	Ave
Percentage	73.06%	68.52%	59.67%	67.11%
All agree	52.19%			

Table 6. Agreements of annotators at sentence level

Annotators	A vs. B	B vs. C	C vs. A	Ave
Percentage	73.57%	68.86%	60.44%	67.62%
All agree	52.86%			

Table 7. Agreements of annotators at document level

Agreements of data from news and blogs are listed in Table 8 for comparison.

Source	NTCIR		BLOG	
Level	Sentence	Document	Sentence	Document
Average agreements of two annotators	53.33%	41.18%	73.85%	64.71%
All agree	33.33%	17.65%	61.40%	41.18%

Table 8. Agreements in different information sources

Table 8 shows that annotations of news articles have lower agreement rates than annotations of web blogs. This is because blog articles may use simpler words and are easier to understand by human annotators than news articles.

From the analyses of inter-annotator agreement, we find that the agreement drops fast when the number of annotators increases. It is less possible to have consistent annotations when more annotators are involved. Here we adopt voting (i.e., the majority) to create the gold standard for evaluation. If the annotations of one instance are all different, this instance is dropped. Table 9 summarizes the statistics of the annotated testing data.

	Positive	Neutral	Negative	Non-opinionated	Total
Word	256	27	243	312	838
Sentence	48	3	93	432	576
Document	7	2	11	14	34

Table 9. Summary of testing data

Table 10 shows the annotation results of three annotators comparing to the gold standard. On average, an annotator can "monitor" the opinions of the whole to around 80.14%. This value can be considered as a reference when evaluating the performance of algorithms. Because the decision of opinion polarities depends much on human perspectives, the information entropy of testing data should also be taken into consideration when comparing system performance.

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%

(a) Word level

Annotators	A	B	C	Average
Recall	94.44%	38.89%	90.97%	74.77%
Precision	71.20%	74.67%	50.19%	65.35%
f-measure	81.19%	51.14%	64.69%	65.67%

(b) Sentence level

Annotators	A	B	C	Average
Recall	100%	50%	85%	78.33%
Precision	71.43%	71.43%	65.38%	69.41%
f-measure	83.33%	58.82%	73.91%	72.02%

(c) Document level

Table 10. Annotators' performance referring to gold standard

5. Applications

The gold standard may be used to the opinionated tasks, such as sentiment word mining, opinionated sentence extraction, opinionated document extraction, opinion summarization, opinion tracking, and opinionated question answering. For opinion summarization, tracking and question answering, not only English and Chinese opinionated reports, but also news and Blog opinionated reports are generated for comparison.

For the applications, a Chinese sentiment dictionary is built. Two sets of sentiment words are selected, including General Inquirer [3] (abbreviated as GI) and Chinese Network Sentiment Dictionary[4] (abbreviated as CNSD). The former is in English and translated into Chinese. The latter, whose sentiment words are collected from the Internet, is in Chinese. Table 11 shows the statistics of the revised dictionaries. Words from these two resources form the "seed vocabulary" in our dictionary.

Dictionary	Positive	Negative
GI	2,333	5,830
CNSD	431	1,948
Total	2,764	7,778

Table 11. Qualified seeds

This dictionary, named as NTU sentiment dictionary (NTUSD), provides positive and negative words revised by human. It can serve as a basis for opinionated tasks. Experimental resources and tools in this paper are available at:

http://nlg18.csie.ntu.edu.tw:8080/opinion/index.html.

6. Conclusion

A set of tags to describe the basic building blocks of opinionated documents is defined in this paper. Experiment materials are developed and then the tags are applied on these materials by annotators. The average agreements of annotators at word, sentence and document level are 68.62%, 67.11% and 67.62%, respectively. The performance of one single annotator achieves 80.14%, 65.67% and 72.02% at word, sentence and document level, respectively. An annotator cannot monitor the opinions of the whole to 100% degree because opinionated issues concern human perspective.

---

[3] http://www.wjh.harvard.edu/~inquirer/
[4] http://134.208.10.186/WBB/EMOTION_KEYWORD/Atx_emtwordP.htm

References

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

Chen, K.-H. and Chen, H.-H. (2002) Cross-language Chinese text retrieval in NTCIR workshop – towards cross-language multilingual text retrieval. *ACM SIGIR Forum*, **35**(2), pages 12-19.

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

Hu, Minqing and Liu, Bing. (2004) Mining and summarizing customer reviews. *SIGKDD 2004*, pages 168-177.

Kim, S.-M. and Hovy, E. (2004). Determining the sentiment of opinions. *Proceedings of the 20th International Conference on Computational Linguistics*, pages 1367-1373.

Ku, L.-W., Lee, L.-Y., Wu, T.-H. and Chen, H.-H. (2005). Major topic detection and its application to opinion summarization. *SIGIR 2005*, pages 627-628.

Ku, L.-W., Liang, Y.-T. and Chen, H.-H. (2006). Opinion extraction, summarization and tracking in news and blog Corpora." *Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*, AAAI Technical Report.

Ku, L.-W., Wu, T.-H., Lee, L.-Y., and Chen, H.-H. (2005). Construction of an evaluation corpus for opinion extraction. *Proceedings of the 5th NTCIR Workshop Meeting*, pages 513-520.

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

Riloff, E. and Wiebe, J. (2003). Learning extraction patterns for subjective expressions. *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 105-112.

Soboroff, I. and Harman, D. (2003). Overview of the TREC 2003 novelty track. *The Twelfth Text REtrieval Conference,* National Institute of Standards and Technology, pages 38-53.

Takamura, H., Inui, T. and Okumura, M. (2005). Extracting semantic orientations of words using spin model. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 133-140.

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*.