Opinion Analysis across languages: An Overview of and Observations from the NTCIR6 Opinion Analysis Pilot Task

David Kirk Evans[1], Lun-Wei Ku[3], Yohei Seki[2], Hsin-Hsi Chen[3], and Noriko Kando[1]

[1] National Institute of Informatics, Tokyo, Japan, {devans,kando}@nii.ac.jp
[2] Dept. of Information and Computer Sciences, Toyohashi University of Technology, Japan seki@ics.tut.ac.jp
[3] Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan lwku@nlg.csie.ntu.edu.tw, hhchen@csie.ntu.edu.tw

Abstract. In this paper we introduce the NTCIR6 Opinion Analysis Pilot Task, information about the Chinese, Japanese, and English data, plans for future opinion analysis tasks at NTCIR, and a brief overview of the evaluation results. This pilot task is a sentence-level opinion identification and polarity detection task run over data from a comparable corpus in three languages: Chinese, English, and Japanese. We have manually annotated documents for this task in each language, producing what we believe to be the first multilingual opinion analysis data set over comparable data. Six participants submitted Chinese system results, three Japanese, and six English for this pilot task. We plan to release the data to the research community, and hope to spur further research into cross-lingual opinion analysis and its use in other NLP tasks. In particular, we look forward to researchers using this data to investigate cross-cultural perspective differences based on automatic sentiment analysis.

1 Introduction

Opinion and sentiment analysis has been receiving a lot of attention in the natural language processing research community recently. With the broad range of information sources available on the web, and rapid increase in the uptake of social community-oriented websites that foster user-generated content there has been further interest by both commercial and governmental parties in trying to automatically analyze and monitor the tide of prevalent attitudes on the web. As a result, interest in automatically detecting language in which an opinion is expressed, the polarity of the expression, targets, and opinion holders has been receiving more attention in the research community. Applications include tracking response to and opinions about commercial products, governmental policies, tracking blog entries for potential political scandals and so on.

The NII-NACSIS Test Collection for Information Retrieval (NTCIR) Workshops have been organized to improve the state of the art in Asian and Cross-

Lingual Information Retrieval, starting in 1999. [5, 6] In the Sixth NTCIR Workshop to be held in Tokyo, May 2007, a new pilot task for Opinion Analysis has been introduced. The pilot task has tracks in three languages: Chinese, English, and Japanese. In this paper, we present an overview of the corpus and evaluation results.

We believe that this corpus presents a unique opportunity to expand the study of opinionated text analysis across languages due to the comparable nature of the corpus. The documents have been carefully selected based on the manual relevance judgments assigned in a cross-lingual Information Retrieval task, ensuring a high quality corpus that is relevant in all three languages. There has been earlier work in creating annotated opinion corpora, for example, [13] describes a corpus tagged at the sentence level for subjectivity and Wiebe also distributes the well-known MPQA[4] corpus. There has also been work in collaborative filtering with the MovieLens corpus[5] and other review-oriented corpora. While there has been lots of research in English opinion analysis ([1, 2, 9, 14, 8, 15]) there has not been as much work in Chinese and Japanese.

Ku et al. [7] describe the construction of two Chinese corpora for opinion extraction, one based on news and one based on blog data, and also an algorithm for Chinese opinion identification at the document and sentence levels. They describe construction of a Chinese sentiment dictionary based on bootstrapping methods that also takes advantage of the ideographic nature of Chinese characters to predict polarity and strength of unknown words.

Seki et al. [10] conducted studies to build a Japanese multi-document summarizer depending on user-specified summary viewpoints. Once a set of documents is provided to the system, the user is presented with a list of topics discussed in the set and can select a topic of interest as well as the information type to focus on in the summaries, such as facts, opinions, or knowledge. The approach was then adapted to English and evaluated as part of the Document Understand Conference. [11] Kanayama et al. [3] re-cast the sentiment analysis problem into a machine translation framework, translating from free text to a more restricted set of sentiment units. They implemented systems for Japanese and English analysis based on two different transfer-based machine translations systems. Later work [4] automatically learns lexicons of polar clauses useful for domain-specific sentiment analysis.

2 NTCIR6 Opinion Analysis Pilot Task

The NTCIR-6 Opinion Analysis Pilot Task extends previous work in opinion analysis to a multilingual corpus. The initial task focuses on a simplified sentence-level binary opinionated or not opinionated classification as opposed to more complicated contextual formulations, but we feel that starting with a simpler task will allow for wider participation from groups that may not have existing experience in opinion analysis.

[4] http://www.cs.pitt.edu/mpqa/

[5] http://www.grouplens.org/node/12

Analysis Task	Values	Req'd?
Opinionated Sentences	YES, NO	Yes
Opinion Holders	String, multiple	Yes
Relevant Sentences	YES, NO	No
Opinionated Polarities	POS, NEG, NEU	No

Table 1. Opinion Analysis task descriptions

The Opinion Analysis task has four subtasks, two of which are mandatory and two of which are optional. Table 1 summarizes the tasks, which are all being performed for all three languages. The two mandatory tasks are to decide whether each sentence expresses an opinion or not. For the Chinese data, all potential opinion holders are annotated whether the sentence in which the entity occurs is an opinionated sentence or not. In Japanese and English, opinion holders are only annotated for sentences that express an opinion, however, the opinion holder for a sentence can occur anywhere in the document. The annotators performed a kind of reference resolution by marking the opinion holder for the sentence, and if the opinion holder is an anaphoric reference noting the target of the anaphora. The opinionated sentences judgement is a binary decision, but in the case of opinion holders we allow for multiple opinion holders to be recorded for each sentence in the case that multiple opinions are expressed.

The two optional tasks are to decide the polarity of the opinionated sentences, and whether the sentences are relevant to the set topic or not. Each set contains documents that were found to be relevant to a particular topic, such as the one shown in Figure 1. For those participating in the relevance subtask each sentence should be judged as either relevant (Y) or non-relevant (N) to the topic. Polarity is determined for each opinionated sentence, and for sentences where more than one opinion is expressed the annotators were instructed to determine the polarity of the most main opinion expressed. In addition, the polarity is to be determined with respect to the set topic description if the sentence is relevant to the topic, and based on the attitude of the opinion if the sentence is not relevant to the topic.

Six teams participated in the Chinese opinion extraction subtask, six teams participated in the English opinion extraction subtask, and three teams participated in Japanese. Results for precision, recall, and F-measure will be presented for opinion detection and opinion holders, and optionally for sentence relevance and polarity for those participants that elected to submit results for those optional portions. Since all sentences were annotated by three annotators there is both a strict (all three annotators must have the same annotation) and a lenient standard for evaluation.

2.1 Corpus

The corpus is based on the NTCIR4 CLIR[6] documents and relevance judgments. It consists of Japanese data from 1998 to 1999 from the Yomiuri and Mainichi newspapers. The Chinese data contains data from 1998 to 1999 from the United Daily News, China Times, China Times Express, Commercial Times, China Daily News, Central and Daily News. The English data also covers from 1998 to 1999 with text from the Mainichi Daily News, Korea Times, and some data from Xinhua.

The corpus was created using about thirty queries over data from the NTCIR Cross-Lingual Information Retrieval corpus covering documents from 1998 to 2001. Document relevance for each set (query) had already been computed for the IR evaluation, so relevant documents for each language were selected based on the relevance judgements. For the Japanese and English portion of the corpus, a maximum of twenty documents were selected for each topic, while the Chinese portion might contain more than twenty documents for a topic. As an example of the topics in the NTCIR Opinion Analysis corpus, please see Figure 1, which shows topic 010, “History Textbook Controversies, World War II”.

<TOPIC> <NUM>010<NUM> <SLANG>CH<SLANG> <TLANG>ENG<TLANG>
<TITLE>History Textbook Controversies, World War II</TITLE>
<DESC>Find reports on the controversial history textbook about the Second World War approved by the Japanese Ministry of Education.</DESC>
<NARR> <BACK>The Japanese Ministry of Education approved a controversial high school history textbook that allegedly glosses over Japan’s atrocities during World War Two such as the Nanjing Massacre, the use of millions of Asia women as ”comfort women” and the history of the annexations and colonization before the war. It was condemned by other Asian nations and Japan was asked to revise this textbook.</BACK>
<REL>Reports on the fact that the Japanese Ministry of Education approved the history textbook or its content are relevant. Reports on reflections or reactions to this issue around the world are partially relevant. Content on victims, ”comfort women”, or Nanjing Massacre or other wars and colonization are irrelevant. Reports on the reflections and reactions of the Japanese government and people are also irrelevant.</REL>
</NARR>
<CONC>Ministry of Education, Japan, Junichiro Koizumi, textbook, comfort women, sexual slavery, Nanjing Massacre, annexation, colonization, protest, right-wing group, Lee Den Hui</CONC> </TOPIC>

Fig. 1. Topic title, description, and relevance fields for set 010

Table 2 shows the number of topics, documents, and sentences for each language, as well as the percentage of opinionated and relevant sentences. The Chinese corpus creation was started in advance of the Japanese and English

[6] `http://research.nii.ac.jp/ntcir/permission/ntcir-4/perm-en-CLIR.html`

Language	Topics	Documents	Sentences	Opinionated (Lenient / Strict)	Relevant
Chinese	32	843	8,546	62% / 25%	39% / 16%
English	28	439	8,528	30% / 7%	69% / 37%
Japanese	30	490	12,525	29% / 22%	64% / 49%

Table 2. General information about NTCIR6 Opinion Analysis Corpus

sides of the corpus, subsequently a larger number of documents was annotated whereas the English and Japanese sides of the corpus limit each topic to twenty documents.

2.2 Annotator Agreement

For English and Japanese, where three annotators were used to annotate all topics, we have computed inter-annotator agreement using Cohen's Kappa. For complete details on inter-annotator agreement, please see [12]. Table 3 shows the minimum and maximum Kappa scores between annotator pairs, as well as the average. One of the annotators in English consistently did not agree with the other two annotators, significantly lowering overall agreement scores. For Chinese, Kappa scores are computed for each topic, with the minimum, maximum, and average reported here over all 31 topics. The average Chinese Kappa agreement scores are similar to the average English scores, although the Chinese annotators are more consistent in polarity tagging. The Japanese annotators overall are much more consistent than either the Chinese or English annotators.

Language	Minimum	Maximum	Average
Chinese Opinionated	0.0537	0.4065	0.2328
Chinese Relevant	0.0441	0.6827	0.2885
Chinese Polarity	0.1605	0.8989	0.4733
English Opinionated	0.1704	0.4806	0.2947
English Relevant	0.0618	0.5298	0.3719
English Polarity	0.2039	0.5457	0.3380
Japanese Opinionated	0.5997	0.7681	0.6740
Japanese Relevant	0.6966	0.8394	0.7512
Japanese Polarity	0.6367	0.7875	0.7054

Table 3. Inter-annotator agreement Kappa summary

3 Evaluation

For a detailed description of the evaluation approach and methodology, please see [12]. Table 4 presents the results for Chinese, English, and Japanese opinion analysis under the lenient evaluation metric, where two of the three annotators must agree for a value to be included in the gold standard. The results from the strict evaluation have been omitted in the interest of brevity.

Group	L	Opinionated			Holder			Relevance			Polarity		
		P	R	F	P	R	F	P	R	F	P	R	F
CHUK	C	0.818	0.519	0.635	0.647	0.754	0.697	0.797	0.828	0.812	0.522	0.331	0.405
ISCAS	C	0.590	0.664	0.625	0.458	0.405	0.430	—	—	—	0.232	0.261	0.246
Gate-1	C	0.643	0.933	0.762	0.427	0.154	0.227	—	—	—	—	—	—
Gate-2	C	0.746	0.591	0.659	0.373	0.046	0.082	—	—	—	—	—	—
UMCP-1	C	0.645	0.974	0.776	0.241	0.410	0.303	0.683	0.516	0.588	0.292	0.441	0.351
UMCP-2	C	0.630	0.984	0.768	0.221	0.376	0.278	0.644	0.936	0.763	0.286	0.446	0.348
NTU	C	0.664	0.890	0.761	0.652	0.172	0.272	0.636	1.000	0.778	0.335	0.448	0.383
IIT-1	E	0.325	0.588	0.419	0.198	0.409	0.266	—	—	—	0.120	0.287	0.169
IIT-2	E	0.259	0.854	0.397	—	—	—	—	—	—	0.086	0.376	0.140
TUT-1	E	0.310	0.575	0.403	0.117	0.218	0.153	0.392	0.597	0.473	0.088	0.215	0.125
TUT-2	E	0.310	0.575	0.403	—	—	—	0.392	0.597	0.473	0.094	0.230	0.134
Cornell†	E	0.317	0.651	0.427	0.163	0.346	0.222	—	—	—	0.073	0.197	0.107
NII	E	0.325	0.624	0.427	0.066	0.166	0.094	0.510	0.322	0.395	0.077	0.194	0.110
GATE-1	E	0.324	0.905	0.477	0.121	0.349	0.180	0.286	0.632	0.393	—	—	—
GATE-2	E	0.324	0.905	0.477	—	—	—	0.286	0.632	0.393	—	—	—
ICU-KR	E	0.396	0.524	0.451	0.303	0.404	0.346	0.409	0.263	0.320	0.151	0.264	0.192
EHBN-1	J	0.531	0.453	0.489	0.138	0.085	0.105	—	—	—	—	—	—
EHBN-2	J	0.531	0.453	0.489	0.314	0.097	0.149	—	—	—	—	—	—
NICT-1	J	0.671	0.315	0.429	0.238	0.102	0.143	0.598	0.669	0.632	0.299	0.149	0.199
NICT-2	J	0.671	0.315	0.429	0.238	0.102	0.143	0.644	0.417	0.506	0.299	0.149	0.199
TUT	J	0.552	0.609	0.579	0.226	0.224	0.225	0.630	0.646	0.638	0.274	0.322	0.296

Table 4. Chinese, English, and Japanese Opinion Analysis Lentient results

Performance across languages varies greatly, and due to both corpora and annotator differences are difficult to compare directly. In this pilot task, each language was evaluated independently, and actually different formulations for precision and recall were used under each language. The task overview paper presents the differences between the evaluation approaches, and also presents evaluations for each language using each approach, but the numbers reported here are the official results. Opinion Holder evaluation for English was performed semi-automatically, but due to the manual effort involved only the first priority run from each participant was evaluated. The Chinese and Japanese evaluation also used semi-automatic approaches to opinion holder evaluation, but were able to evaluate all submitted runs.

Of the groups that participated, one group (GATE) participated in both the Chinese and English task, and one group (TUT) participated in both the English and Japanese task. Despite using similar approaches, their results differ in each language in part due to the difference in annotation between the languages. An interesting question for future work is whether these differences stem more from annotator training, differences in the documents that make up the corpus, or cultural and language differences.

4 Future Work

The NTCIR Opinion Analysis Pilot task is in the first year of operation, and has started with a fairly simple task in three languages. We have proposed multiple evaluation approaches and held a workshop in May with participants discussing the evaluation results and both positive and negative experiences with this cross-lingual evaluation. We hope to foster more research into multi-lingual aspects of sentiment analysis and hope to see more sites participate in analysis for multiple languages. The next section presents the roadmap for future NTCIR Opinion Analysis Tasks.

4.1 NTCIR OAT Roadmap

We plan to conduct the Opinion Analysis Task again in NTCIR-7 and NTCIR-8. The NTCIR meetings are held every year and a half. For NTCIR-7 we plan to add a new genre to the task, reviews, in addition to the news genre used in NTCIR-6. We are currently exploring using review web sites as a source of data. NTCIR-7 and 8 will both continue to use Chinese, English, and Japanese, and while no further languages are slated for addition at this time, Korean is a possible candidate since relevance judgments for some of the topic already exist. NTCIR-7 will also add a strength of opinion and stakeholder evaluation in addition to the subjectivity, polarity, and opinion holder evaluation performed in NTCIR-6. NTCIR-8 will add a temporal evaluation, and possibly expand to clause-level subjectivity.

5 Conclusions

In this paper we have presented the NTCIR Opinion Analysis Pilot Task, the corpus used in the workshop, and an overview of the evaluation results. We look forward to future iterations of the NTCIR Opinion Analysis Task which will add new genres to the evaluation, and add further features for extraction.

References

1. V. Hatzivassiloglou and K. R. McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 174–181, Morristown, NJ, USA, 1997. Association for Computational Linguistics.
2. V. Hatzivassiloglou and J. M. Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th International Conference on Computational Linguistics*, 2000.

3. H. Kanayama and T. Nasukawa. Deeper sentiment analysis using machine translation technology. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, pages 494–500, 2004.
4. H. Kanayama and T. Nasukawa. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 355–363, Sydney, Australia, July 2006. Association for Computational Linguistics.
5. N. Kando, K. Kuriyama, T. Nozue, K. Eguchi, H. Karo, S. Hidaka, and J. Adachi. The ntcir workshop: the first evaluation workshop on japanese text retrieval and cross-lingual information retrieval. In *Proceedings of the 4th International Workshop on Information Retrieval with Asian Languages (1RAL'99)*, 1999.
6. K. Kishida, K. hua Chen, S. Lee, K. Kuriyama, N. Kando, H.-H. Chen, and S. H. Myaeng. Overview of clir task at the fifth ntcir workshop. In *Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, Tokyo, Japan, December 2005. National Institute of Informatics.
7. L.-W. Ku, Y.-T. Liang, and H.-H. Chen. Opinion extraction, summarization and tracking in news and blog corpora. In *Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs, AAAI Technical Report*, 2006.
8. B. Pang and L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*, pages 115–124, 2005.
9. E. Riloff and J. Wiebe. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing (EMNLP 2003)*, pages 105–112, Sapporo, Japan, July 2003.
10. Y. Seki, K. Eguchi, and N. Kando. Multi-document viewpoint summarization focused on facts, opinion and knowledge. In J. G. Shanahan, Y. Qu, and J. Wiebe, editors, *Computing Attitude and Affect in Text: Theories and Applications*, chapter 24, pages 317–336. Springer, Dordrecht, The Netherlands, December 2005.
11. Y. Seki, K. Eguchi, N. Kando, and M. Aono. Opinion-focused Summarization and its Analysis at DUC 2006. In *Proc. of the Document Understanding Conf. Wksp. 2005 (DUC 2006) at the Human Language Technology Conf. - North American chapter of the Association for Computational Linguistics (HLT-NAACL 2006)*, pages 122–130, New York Marriott, June 2006.
12. Y. Seki, D. K. Evans, L.-W. Ku, H.-H. Chen, N. Kando, and C.-Y. Lin. Overview of opinion analysis pilot task at ntcir-6. In *Proceedings of the Sixth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access.* National Institute of Informatics, May 2007.
13. J. Wiebe, R. Bruce, and T. O'Hara. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th Association of Computational Linguistics*, pages 246–253, 1999.
14. J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *The Third IEEE International Conference on Data Mining*, pages 427–343, November 2003.
15. H. Yu and V. Hatzivassiloglou. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, volume 10, pages 129–136, Morristown, NJ, USA, 2003. Association for Computational Linguistics.