Overview of Opinion Analysis Pilot Task at NTCIR-6

Yohei Seki†, David Kirk Evans‡, Lun-Wei Ku§,
Hsin-Hsi Chen§, Noriko Kando‡, Chin-Yew Lin¶
†Dept. of Information and Computer Sciences, Toyohashi University of Technology
Aichi 441-8580, Japan
seki@ics.tut.ac.jp
‡National Institute of Informatics
Tokyo 101-8430, Japan
{devans, kando}@nii.ac.jp
§Dept. of Computer Science and Information Engineering, National Taiwan University
Taipei 10617, Taiwan
{lwku, hhchen}@csie.ntu.edu.tw
¶Microsoft Research Asia
Beijing 100080, P.R. China
cyl@microsoft.com

Abstract

This paper describes an overview of the Opinion Analysis Pilot Task from 2006 to 2007 at the Sixth NTCIR Workshop. We created test collection for 32, 30, and 28 topics (11,907, 15,279, and 8,379 sentences) in Chinese, Japanese and English. Using this test collection, we conducted opinion extraction subtask. The subtask was defined from four perspectives: (a) opinionated sentence judgment, (b) opinion holder extraction, (c) relevance sentence judgment, and (d) polarity judgment. 21 run results were submitted by 14 participants with five results submitted by the organizers. We show the evaluation results of the groups participating in opinion extraction subtask.

Keywords: *Opinion Extraction, Opinion Holder, Relevance, Polarity, and NTCIR.*

1 Introduction

This paper describes an overview of *the Opinion Analysis Pilot Task* [5] from 2006 to 2007 at the Sixth NTCIR Workshop [4] (NTCIR-6 Opinion). This was the first effort to produce a multi-lingual test collection for evaluating opinion extraction at NTCIR.

Opinion and sentiment analysis has been receiving a lot of attention in the natural language processing research community recently [2, 9, 7]. 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 sentences in which an opinion is expressed ([12] etc.), the polarity of the expression ([13] etc.), targets, and opinion holders ([1] etc.) 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.

In the Sixth NTCIR Workshop, 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 test collection, task design, and evaluation results using the test collection across the Chinese, Japanese, and English data.

We believe that this pilot task 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.

This paper is organized as follows. In Section 2, we explain the task design for the *opinion extraction subtask*. Section 3, we briefly introduce the test collection used in *NTCIR-6 Opinion Analysis Pilot Task*. Section 4 presents the annotation methodology. Section 5 details the evaluation methodology used, and

explains the differences in the aproaches taken with examples. Section 6 describes participant system description. Section 7 presents evaluation results for the opinion extraction subtask in Chinese, Japanese and English. Finally, we present our conclusions in Section 8.

2 Task Design

2.1 Schedule

The time schedule for *the NTCIR-6 Opinion Analysis Pilot Task* is as follows.

2006-08-01	:	Start of Registration
2006-10-30	:	Registration Due
2006-11-21	:	Testing Sets Release (Chinese and Japanese *opinion extraction*)
2006-11-30	:	Submission of Results (Chinese and Japanese *opinion extraction*)
2006-12-11	:	Testing Sets Release (English *opinion extraction*)
2006-12-20	:	Submission of Results (English *opinion extraction*)
2007-02-08	:	Delivery of Evaluation Results
2007-03-08	:	Paper Due (for Proceedings)
2007-05-15	:	NTCIR Workshop-6 (Conference in Tokyo)

2.2 Participants

Results for *the opinion extraction subtask* have been collected. Five, three, and six teams participated in the Chinese, Japanese and English *opinion extraction* subtask. Two runs at most were accepted from each participant.

2.3 Opinion extraction subtask

Four Evaluation Categories

The opinion extraction subtask has four categories in evaluation, two of which are mandatory and two of which are optional. In Table 3, the two mandatory categories are to decide whether each sentence expresses an opinion or not. The two optional categories are whether the sentences are relevant to the set topic or not, and to decide the polarity of the opinionated sentences.

1. Opinionated sentences
   The opinionated sentences judgment 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.

2. Opinion holders
   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 assessors performed a kind of co-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.

3. Relevant sentences
   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 category evaluation, each sentence should be judged as either relevant (Y) or non-relevant (N) to the topic.

4. Opinion polarities
   Polarity is determined for each opinionated sentence, and for sentences where more than one opinion is expressed the assessors were instructed to determine the polarity of the 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.

Sample (Training) Data

Of 32, 30 topics in *NTCIR-6 Opinion Analysis Pilot Task* test collection, four topics were provided as a sample (training) data to participants in Chinese and Japanese. For English, only one topic was provided as a sample data because *MPQA opinion corpus* [11] was available for opinion extraction researchers in English.

Evaluation Metrics

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. In Chinese, Japanese, and English since all sentences were annotated by three assessors there is both a strict (all three assessors must have the same annotation) and a lenient standard (two of three assessors have the same annotation) for evaluation, both of which are being computed for all but the opinion holder evaluation, which require some manual judgment and will only be performed once for each participating group. Formal definition provided for evaluation is as follows.

1. Mandatory evaluation
   (a) Precision, Recall and F-measure of Opinion Holder using lenient gold standard.
   (b) Precision, Recall and F-measure of Opinion Holder using strict gold standard.

(c) Precision, Recall and F-measure of Opinion using lenient gold standard.

(d) Precision, Recall and F-measure of Opinion using strict gold standard.

2. Option 1 evaluation
   If Relevance information is provided, extra information will be reported including:
   (a) Precision, Recall and F-measure of Relevance using lenient gold standard.
   (b) Precision, Recall and F-measure of Relevance using strict gold standard.
3. Option 2 evaluation
   If Polarity information is provided, extra information will be reported including:
   (a) Precision, Recall and F-measure of Polarity using lenient gold standard.
   (b) Precision, Recall and F-measure of Polarity using strict gold standard.

3 Test collection

3.1 Document Sets

The test collection is based on *the NTCIR-3, 4, and 5 CLIR test collection* [6] documents and relevance judgments.

- It consists of Japanese data from 1998 to 2001 from the Yomiuri and Mainichi newspapers.
- The Chinese data contains data from 1998 to 2001 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 2001 with text from the Mainichi Daily News, Korea Times, and some data from Xinhua.

The test collection was created using about thirty queries over data from the *NTCIR Cross-Lingual Information Retrieval test collection* 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 judgments. For the Japanese and English portion of the test collection, 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-6 opinion analysis pilot task, please see Figure 1, which shows topic 010, "History Textbook Controversies, World War II".

Table 1 shows the number of topics, the number of documents, and the number of sentences for each language. The percentage of sentences that are opinionated and relevant are also computed for both the strict and lenient standards.

3.2 Topics

Table 2 lists the titles of all the topics in the data set. While only the English title is given, the topics and related meta-data as shown in Figure 1 have all been translated into each language.

4 Annotation

The NTCIR-6 Opinion Analysis Pilot Task extends previous work in opinion analysis [3, 8, 10] to a multilingual corpus. The initial category 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. Table 3 summarizes the annotation categories, which are all being performed for all three languages. All categories were annotated by three annotators in each language: Chinese, Japanese, and English. One sample topic was used for inter-coder session to improve the agreement between assessors.

4.1 Chinese Annotation Strategy

In the Chinese annotation effort, a pool of seven annotators were used to annotate the documents, with three annotators per document. Prior to annotation, the annotators underwent an hour-long orientation period where the purpose of the annotation was explained, and examples of sentences and their annotation were given. After the hour-long orientation session, the annotators were free to ask the annotation coordinator questions about specific sentences if they were unsure of the labelling, but no special care was taken to ensure consistency between the annotators in those cases.

4.2 Japanese Annotation Strategy

The Japanese data was annotated by three annotators, who were given basic instructions about the annotation task, and then annotated a sample topic. They held a meeting about six hours afterwards to discuss discrepancies with the explicit goal of trying to improve agreements between annotators. The general or common knowledge and future plans were not counted as opinions. The Japanese annotators agreed on a specific format for writing out opinion holder description strings. The three annotators were magazine or newspaper related editorials or translators.

4.3 English Annotation Strategy

Three annotators were used to mark the English data. One of the annotators was a journalist, another

Table 1. Test collection size at NTCIR-6 Opinion Analysis Pilot Task

Language	Topics	Documents	Sentences	Opinionated (Lenient / Strict)	Relevant
Chinese	32	843	11,907	62% / 25%	39% / 16%
Japanese	30	490	15,279	29% / 22%	64% / 49%
English	28	439	8,528	30% / 7%	69% / 37%

Table 2. Opinion Analysis Task Topic Titles

Number	Title
001	Time Warner, American Online (AOL), Merger, Impact
002	President of Peru, Alberto Fujimori, scandal, bribe
003	Kim Dae Jun, Kim Jong Il, Inter-Korea Summit
004	the US Secretary of Defense, William Sebastian Cohen, Beijing
005	G8 Okinawa Summit
006	Wen Ho Lee Case, classified information, national security
007	Ichiro, Rookie of the Year, Major League
008	Jennifer Capriati, tennis
009	EP-3 surveillance aircraft, F-8 fighter, aircraft collision
010	History Textbook Controversies, World War II
011	Tobacco business, accusation, compensation
012	Tiger Woods, sports star
013	”Chiutou” (Autumn Struggle), Appeal, Laborer, Protest, Taiwan
014	Expert, Opinion, International Monetary Fund (IMF), Asian countries
015	Find articles dealing with a teenage social problem
016	Divorce, Family Discord, Criticisms
017	China, Reaction, Taiwan, Diplomatic Relations
018	China, Stationing, Weapons, Taiwan
019	Animal Cloning Technique
020	Sexual Harassment, Lawsuits
021	Olympic, Bribe, Suspicion
022	North Korea, Daepodong, Asia, Response
023	Joining WTO
024	China Airlines Crash
025	Province-refining
026	Economic influence of the European monetary union
027	President Kim Dae-Jung’s policy toward Asia
028	Clinton scandals
029	War crimes lawsuits
030	Nuclear power protests
031	College Admission Policy
032	Counseling for Youths

Table 3. Four annotation categories at NTCIR-6 Opinion Analysis Pilot Task

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

<TOPIC>
<NUM>010<NUM>
<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>

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

was a translator, and the profession of the third annotator was unknown. Prior to annotation, there was a two hour meeting between the annotators and the English coordinator explaining the purpose of the annotation, and introducing them to the task with some sample annotations. Afterwards all three annotators annotated a sample topic, and later a four-hour meeting was held to discuss discrepancies and general approaches to annotation. By consensus, expression of common or general knowledge were not labeled as opinions, nor were statements from officials or companies about future plans or schedules.

4.4 Inter-annotator agreement

For the Japanese and English corpora all topics were annotated by the same three annotators, so it was possible to compute Cohen's Kappa for agreement over all topics between the annotators. Table 5 lists the pairwise agreement for annotators for the opinionated tagging subtasks.

Table 4 gives a summary of the Kappa agreement for annotators in each language. More specific agreement values for each language are given below. of determining whether a sentence contains opinionated language is open to individual interpretation regardless of the language.

In general, Japanese has the highest average agreement numbers for opinionated sentence detection. As the annotators for each language underwent different training and instruction, and come from different backgrounds, it is likely that much of the variation in agreement is not due to differences inherent in the language, but instead is due to differences in the annotators.

Table 4. Kappa summary

Language	Minimum	Maximum	Average
CH Opinionated	0.0537	0.4065	0.2328
JA Opinionated	0.5997	0.7681	0.6740
EN Opinionated	0.1704	0.4806	0.2947

Table 5. Pairwise Inter-annotator agreement using Cohen's Kappa for Japanese and English

Language	Annotator Pair	Task	Kappa
J	1-2	Opinionated	0.6541
J	1-3	Opinionated	0.5997
J	2-3	Opinionated	0.7681
E	1-2	Opinionated	0.4806
E	1-3	Opinionated	0.1704
E	2-3	Opinionated	0.2332

In Chinese, since there is a total set of seven annotators, not all topics were annotated by the same three annotators. It was thus not possible to compute agreement of three annotators over all topics, since the annotators change for each topic. Instead, for each topic the agreement between the three annotators was computed, and the average for each topic is shown in Table 6.

Table 6. Inter-annotator agreement using Cohen's Kappa for Chinese

Topic	Opinionated	Topic	Opinionated
1	0.4009	17	0.1608
2	0.3772	18	0.2747
3	0.2327	19	0.3166
4	0.2210	20	0.0938
5	0.4065	21	0.2617
6	0.1046	22	0.1956
7	0.2355	23	0.3663
8	0.0706	24	0.1427
9	0.2254	25	0.3634
10	0.1228	26	0.2698
11	0.2367	27	0.3667
12	0.2351	28	0.1285
13	0.1942	29	0.3207
14	0.2714	30	0.0537
15	0.1344	31	0.2009
16	0.3119	32	0.1523

5 Evaluation Approach

In each language we tried to take a similar approach to evaluation, using precision, recall, and f-measure to report results. Each language had slight differences in how those measures were computed though. Details for each language are given below, but as a quick summary:

1. **Opinionated and Relevant**: Precision, recall, and f-measure was computed in the same manner for all languages.
2. **Polarity**: Three different approaches were used. We present results from all three approaches in each language in this overview paper.
3. **Opinion Holder**: The English and Japanese evaluations were similar, with a semi-automatic evaluation that relied on human judgments, and estimates the recall. The Chinese evaluators also used a semi-automatic approach but manually examined all instances where opinion holders were not exact string matches, and possibly skips some sentences similar to the polarity evaluation.

In the following sections, we will provide a description of the three evaluation approaches taken.The Chinese evaluation followed the LWK approach, the English evaluation followed the DKE approach, and the Japanese evaluation followed the YS approach.

5.1 LWK Evaluation Approach

5.1.1 Opinionated / Relevance

Under the strict evaluation, all three annotators must agree on the classification of the sentence to be counted as either an opinionated or relevant sentence. Under the lenient evaluation, two of the three annotators must agree on the classification of the sentence for it to be counted. Precision is computed as $\frac{\#systemcorrect}{\#systemproposed}$. Recall is computed as $\frac{\#systemcorrect}{\#sentences}$ where the number of sentences is either the number of opinionated or relevant sentences according to the strict or lenient criteria.

5.1.2 Polarity

The LWK approach evaluates only opinionated sentences that match the definitions for either the strict or lenient gold standards. For the strict standard, sentences on which all three annotators agree about the polarity, either all POS, all NEU, all NEG, or all "not opinionated". For the lenient evaluation two of the three annotators must agree that the sentence is opinionated to be included in the evaluation. All other sentences will not be included in the evaluation.

The polarity for the sentence is the polarity with the largest number of votes by the annotators. In cases where the polarity of the sentence is ambiguous, POS + NEU the gold standard is POS, for NEG + NEU the gold standard is NEG, for POS + NEG the gold standard is NEU, and for POS + NEU + NEG the gold standard is NEU.

5.1.3 Opinion Holder

The LWK evaluation approach for opinion holders is semi-automatic. All possible aliases of each opinion holder are generated manually first, for example, the names of holders with or without their titles. The results then are evaluated according to this information by keyword matching. At last, to ensure the correctness of the evaluation, every record which is different from all aliases of the correct holders is checked manually again.

Notice that if we are not sure the proposed answer is the same entity as the correct answer, it is treated as a wrong answer. For example, if the correct holder is "the president of America" but the participant reports "the president", there will not be a match. And also the resolution of the anaphor or the correference has not been evaluated yet, as we have mentioned earlier. That is, the holders of the sentence proposed by the participant should be the same as the form it appears in this sentence.

5.2 DKE Evaluation Approach

5.2.1 Opinionated / Relevance

The DKE evaluation approach for opinionated and relevant sentences is the same as described in Section 5.1.1.

5.2.2 Polarity

The general idea of the DKE approach is to weight system scores based on their agreement with the annotated data. The sentences that are evaluated for polarity are determined according to the lenient or strict standard: for lenient, only sentences in which at least two annotators have marked the sentence as opinionated are evaluated, for strict only sentences in which all three annotators marked the sentence as opinionated are evaluated.

The evaluation script creates a contingency table for the categories POS, NEU, NEG, and NONE, where NONE is category that is used when a sentence is not opinionated. For each sentence, the individual votes for each annotator are added to the appropriate cell based on the system's categorization. If, for example, the system assigned a sentence a polarity of NEG, and the annotators assign polarities of NONE (not an opinionated sentence), NEG, and NEU, the contingency table will be updated with 1 added to $t[GOLD_{NEG}][SYSTEM_{NEG}]$, $t[GOLD_{NEU}][SYSTEM_{NEG}]$, and $t[GOLD_{NONE}][SYSTEM_{NEG}]$. Precision and recall is then calculated in the normal way over the contingency table.

One advantage of this approach is that all of the annotations are taken into account, and the method scales well to any number of annotators. In addition, for sentences which are truly ambiguous for human annotators, the systems are partially rewarded based on how ambiguous a sentence is. For example, if one hundred annotators marked a sentence with 50 POS polarity annotations, and 50 NEU polarity annotations, a system that labels the sentence POS or NEU would benefit by agreeing with half of the annotators. Other schemes run the risk of declaring one of either POS or NEU to be correct, penalizing the system when the sentence is clearly difficult for humans to label one way or the other.

5.2.3 Opinion Holder

Opinion Holder evaluation under the DKE strategy used a perl script to implement a semi-automatic evaluation. For each document, an equivalence class is created for each opinion holder, and system opinion holders for a given sentence are matched using exact string matches to the opinion holders in the equivalence class. Matches are counted as correct opinion holders, if no matches are found then a human judge[1] is asked to determine if the system opinion holder matches ones of the opinion holders in the equivalence class for the sentence given the opinion holders in the equivalence class and the sentence text. If there is match, the system opinion holder is added to the equivalence class, otherwise it is marked as a known incorrect opinion holder.

The initial database of opinion holder equivalence classes is created by adding the opinion holders marked by the annotators. The database grows with each evaluated system, and after the first run for each system subsequent runs can be done automatically using the opinion holder database to match opinion holders.

Precision is computed as the number of correctly matched opinion holders divided by the number of offered opinion holders. Recall is only an approximation though: the evaluation script assumes one opinion holder for each opinionated sentence. While the specification allows for multiple opinion holders per sentence, only 3.5% of English annotations actually had more than one opinion holder annotated in the gold standard.

5.3 YS Evaluation Approach

5.3.1 Opinionated / Relevance

The YS evaluation approach for opinionated and relevant sentences is the same as described in Section 5.1.1. We provide the evaluation script in Perl on December, 2006 and participants could conduct post submission analysis using this script.

5.3.2 Polarity

The most important point of YS approach is consistency in evaluation strategies within four categories (opinionated sentence, relevance, polarity, and opinion holder). In polarity evaluation, the recall, precision, and F-value was computed based on leniently or strictly agreed results between assessors for positive, negative, and neutral values. Therefore, the evaluation results were slightly more strict than other two evaluation approaches. This evaluation script was also provided in Perl to participants on January, 2007.

5.3.3 Opinion Holder

Opinon holder evaluation strategy was also consistent with other three category evaluation strategies: they were evaluated based on leniently or strictly agreed opinion holders between assessors.

We only applied a sentence-based evaluation to evaluate the opinion holders. If multiple holders existed in one sentence, and the system detected one of them, then we regarded the system's extraction as valid.

In addition, we also applied a five-grade evaluation of the agreement between the system's and the assessor's detection, as follows. This strategy was useful to estimate the effectiveness of coreference resolution approach.

[1] For this evaluation, the co-author David Kirk Evans

1. Agreed semantically and strings were matched almost completely.
2. Agreed semantically and strings were matched partially, but a proper name was not detected.
3. Agreed semantically but strings were not matched.
4. Agreed partially in some aspect, but proper entity could not be specified.
5. Not agreed.

We counted the results using the above three grades for valid extractions and computed the precision, recall, and F-measure values. Opinion holder evaluation was conducted semiautomatically by combining perfect strict matching approach and manually conducted five graded estimation.

5.4 Comparison

Table 7 and Table 8 list a number of cases and the behavior of the three evaluation approaches.

Note that in the last example in Table 8 the Chinese score actually increases. This is due to the heuristic that says that under a lenient evaluation, a POS and NEU score by two annotators is treated as a POS sentence.

6 Participant System Descriptions

6.1 Chinese (in alphabetic order)

Five teams participated in Chinese side. The Chinese University of Hong Kong (CUHK) implemented the system with five modules based on knowledge learned from unsupervised web data. University of Sheffield (GATE) implemented SVM-based Chinese and English opinion extraction system based on sample four topics and MPQA corpus. Chinese Academy of Sciences (ISCAS) applied Conditional Random Filed (CRF) to find the opinion holders as a sequential labeling task. National Taiwan University (NTU) calculated polarity scores to decide the opinion polarities and strengths of words from composed characters. University of Maryland (UMCP) implemented the system based on sentiment lexicons and explored the effect of the lexicon size, etc.

6.2 English (in alphabetic order)

Six teams participated in English side. Cornell Univesity (Cornell) developed the system by using components and features from their previous work. University of Sheffield (GATE) used an SVM system to train a classifier over MPQA corpus and compare the differences between the MPQA corpus and the NTCIR-6 English corpus. Information and Communication University (ICU-IR) system was a hybrid machine-learning and rule-based system. They took a semi-supervised learning methods based on fourteen strong clue words and six seed rules. Illinois Institute of Technology (IIT) system uses a lexicon of words and phrases used to express appraisal attitudes. For opinion holders, they determine the subject or agent of the communication verb list and combine that with evidence from quote positions. National Institute of Informatics (NII) uses a machine-learning approach with shallow parsing to generate features used to train classifiers in the WEKA. Toyohashi University of Technology (TUT) system was based on SVM classifier trained over surface features and semantic primitives for predicates and subjects from a thesaurus.

6.3 Japanese (in alphabetic order)

Three teams participated in Japanese side. NEC Internet Systems Research Laboratory (NEC) took a SVM machine larning approach with four type features and related author and non-author opinion holder candidates to opinionated sentences. National Institute of Information and Communications Technology (NICT) implemented SVM-based opinion sentence classification and applied a pairwise classification with majority voting to polarity classification. Toyohashi University of Technology (TUT) implemented two-way opinion classification systems: an author and an authority opinion classificaition system crosslingualy in Japanese and English.

7 Evaluation Results

7.1 Chinese

Table 9 lists the evaluation results in Chinese opinion analysis based on lenient and strict standards. Though the CFP shows that the evaluation results of opinions and opinion holders together should be listed, they are separated evaluated because of the partial correct issue of opinion holders.

For opinion holder evaluation, we applied both the sentence-based evaluation and the holder-based evaluation, as shown in Table 12 and Table 13. In the sentence-based evaluation, because there may be multiple opinion holders in one opinion sentence, the number of Correct (With holder), Correct (Without holder), Partial Correct, Incorrect, Miss, False-alarm, precision, recall and f-measure are listed. The field ”Partial Correct” shows the number of sentences in which participants did not find all holders, while the field ”Incorrect” shows the number of sentences in which participants propose wrong holders. In the holder-based evaluation, the evaluation unit is one holder. The number of Correct, Incorrect, Miss, False-alarm, Proposed

Table 7. Comparison of Polarity Evaluation Approaches (Strict)

Annotation				System	Behavior
POS	NEU	NEG	NOT		
3	0	0	0	POS	LWK +, DKE +, YS +
2	0	1	0	POS	LWK sent. skipped, DKE -, YS -
0	0	0	3	POS	LKW -, DKE -, YS -
0	0	1	2	POS	LWK sent. skipped, DKE -, YS -

Table 8. Comparison of Polarity Evaluation Approaches (Lenient)

Annotation				System	Behavior
POS	NEU	NEG	NOT		
3	0	0	0	POS	LWK +, DKE +, YS +
2	0	1	0	POS	LWK +, DKE + by $\frac{2}{3}$, YS +
0	0	0	3	POS	LWK -, DKE -, YS -
0	0	1	2	POS	LWK -, DKE -, YS -
1	0	2	0	POS	LWK -, DKE + by $\frac{1}{3}$, YS -
1	1	0	1	POS	LWK +, DKE + by $\frac{1}{3}$, YS P. down R. no change

Table 9. Chinese Opinion Analysis LWK Approach results

Group	L/S	Opinionated			Relevance			OpAndPolarity		
		P	R	F	P	R	F	P	R	F
CUHK	L	0.818	0.519	0.635	0.797	0.828	0.812	0.522	0.331	0.405
ISCAS	L	0.590	0.664	0.625	—	—	—	0.232	0.261	0.246
Gate-1	L	0.643	0.933	0.762	—	—	—	—	—	—
Gate-2	L	0.746	0.591	0.659	—	—	—	—	—	—
UMCP-1	L	0.645	0.974	0.776	0.683	0.516	0.588	0.292	0.441	0.351
UMCP-2	L	0.630	0.984	0.768	0.644	0.936	0.763	0.286	0.446	0.348
NTU	L	0.664	0.890	0.761	0.636	1.000	0.778	0.335	0.448	0.383
CUHK	S	0.341	0.575	0.428	0.468	0.900	0.616	0.197	0.596	0.296
ISCAS	S	0.221	0.662	0.331	—	—	—	0.059	0.314	0.099
Gate-1	S	0.253	0.979	0.402	—	—	—	—	—	—
Gate-2	S	0.330	0.696	0.448	—	—	—	—	—	—
UMCP-1	S	0.245	0.986	0.393	0.404	0.565	0.471	0.085	0.615	0.150
UMCP-2	S	0.239	0.993	0.385	0.354	0.953	0.516	0.081	0.604	0.143
NTU	S	0.258	0.921	0.404	0.343	1.000	0.511	0.104	0.662	0.180

Table 10. Chinese Opinion Analysis YS Approach results

Group	L/S	Opinionated			Relevance			Polarity		
		P	R	F	P	R	F	P	R	F
CUHK	L	0.819	0.520	0.636	0.797	0.828	0.813	0.480	0.431	0.454
NTU	L	0.630	0.890	0.738	0.603	1.000	0.752	0.269	0.537	0.358
UMCP-1	L	0.645	0.974	0.776	0.683	0.516	0.588	0.256	0.547	0.349
UMCP-2	L	0.630	0.984	0.768	0.644	0.936	0.763	0.248	0.548	0.341
ISCAS	L	0.590	0.664	0.625	—	—	—	0.170	0.271	0.209
GATE-1	L	0.643	0.933	0.761	—	—	—	—	—	—
GATE-2	L	0.747	0.591	0.660	—	—	—	—	—	—
CUHK	S	0.340	0.575	0.428	0.468	0.900	0.616	0.197	0.596	0.296
NTU	S	0.245	0.921	0.387	0.326	1.000	0.491	0.099	0.662	0.172
UMCP-1	S	0.245	0.987	0.393	0.404	0.565	0.471	0.086	0.615	0.150
UMCP-2	S	0.239	0.993	0.385	0.354	0.953	0.517	0.081	0.603	0.143
ISCAS	S	0.221	0.662	0.331	—	—	—	0.059	0.314	0.099
GATE-1	S	0.253	0.979	0.402	—	—	—	—	—	—
GATE-2	S	0.330	0.696	0.448	—	—	—	—	—	—

Table 11. Chinese Opinion Analysis DKE Approach results

Group	L/S	Opinionated			Relevance			Polarity		
		P	R	F	P	R	F	P	R	F
CUHK	L	0.819	0.520	0.636	0.797	0.828	0.813	0.480	0.431	0.454
NTU	L	0.667	0.888	0.762	0.636	1.000	0.777	0.286	0.538	0.374
UMCP-1	L	0.645	0.976	0.777	0.683	0.519	0.590	0.256	0.549	0.349
UMCP-2	L	0.630	0.986	0.769	0.644	0.943	0.765	0.248	0.549	0.341
ISCAS	L	0.590	0.664	0.625	—	—	—	0.170	0.271	0.209
GATE-1	L	0.643	0.933	0.761	—	—	—	—	—	—
GATE-2	L	0.747	0.591	0.660	—	—	—	—	—	—
CUHK	S	0.340	0.575	0.428	0.468	0.901	0.616	0.197	0.595	0.296
NTU	S	0.265	0.922	0.412	0.342	1.000	0.509	0.108	0.666	0.186
UMCP-1	S	0.245	0.988	0.393	0.404	0.570	0.473	0.086	0.615	0.150
UMCP-2	S	0.239	0.994	0.385	0.354	0.963	0.518	0.081	0.603	0.143
ISCAS	S	0.221	0.662	0.331	—	—	—	0.059	0.314	0.099
GATE-1	S	0.253	0.979	0.402	—	—	—	—	—	—
GATE-2	S	0.330	0.695	0.448	—	—	—	—	—	—

Holders, Correct Number (the total number of holders in the correct opinion sentences which participants proposed), precision, recall, and f-measure are listed. Opinion holders are only meaningful and extracted in opinion sentences. Therefore to avoid the propogate errors from the opinion sentence extraction, only the holders reported in correct opinion sentences proposed by participants are further evaluated.

7.2 English

Table 14 lists results using the lenient and strict standards. Of the nine submitted runs, six contained relevance information (four of the six groups) and seven contained polarity information (five of the six groups.) While there is no difference in the GATE runs reported in these results, the two runs took different strategies for opinion holder identification, but only the first priority run was evaluated for opinion holders. The polarity results differ slightly for the two TUT runs.

For the opinion holder analysis, the English co-organizer determined whether the system-predicted opinion holder matched one of the annotated opinion holders given the context of the sentence. The process was automated to some extent by looking for exact string matches, quite common with the -author- opinion holder, and memoization of previous human-made decisions.

Table 17 lists the precision, recall, and F-measure for both the lenient and strict evaluations of opinion holders. The script used to compute the results lists both precision and recall over all sentences — penalizing systems for suggesting opinion holders on non-opinionated sentences — and over only the sentences that are marked as opinionated according to the gold standard data. Table 17 lists results over all opinionated sentences to conform more closely with how the Chinese and Japanese evaluation was performed.Of the 6319 sentences marked with opinion holders, only 208 have more than one opinion holder, so I felt that this was a reasonable approximation.

7.3 Japanese

Table 18 lists the evaluation results of a Japanese opinion analysis based on lenient and strict standards.

- For opinionated sentence classification, NICT system performed best in precision and TUT performed best in recall.
- For opinion holder extraction, EHBN-2 best performed in precision and TUT performed best in recall.
- For relevance judgment, NICT-2 performed best in precision and NICT-1 performed best in recall.
- For polarity classification, NICT performed best in precision and TUT performed best in recall.

In summary, EHBN system got advantage in opinion holder extraction. NICT implemented balanced precision-focused system. TUT implemented recall-focused system and attained best F-values.

8 Discussions and Conclusions

8.1 Overview of Results in NTCIR-6

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

Table 12. Chinese Opinion Holders Analysis: Sentence-Based Results

Group	L/S	CRT-w	CRT-wo	P-CRT	InCRT	Miss	F-A	P	R	F
CUHK	L	1086	1070	189	84	81	319	0.647	0.754	0.697
ISCAS	L	665	1724	175	354	447	257	0.458	0.405	0.430
GATE-1	L	364	2685	100	345	1551	44	0.427	0.154	0.227
GATE-2	L	76	1554	5	112	1463	11	0.373	0.046	0.082
UMCP-1	L	1000	916	232	964	243	1955	0.241	0.410	0.303
UMCP-2	L	471	317	103	405	96	628	0.221	0.376	0.278
NTU	L	388	2564	57	120	1692	30	0.652	0.172	0.272
CUHK	S	550	371	81	41	29	106	0.707	0.785	0.744
ISCAS	S	293	544	84	157	188	89	0.470	0.406	0.436
GATE-1	S	165	933	47	171	677	11	0.419	0.156	0.227
GATE-2	S	42	617	3	66	694	3	0.368	0.052	0.091
UMCP-1	S	917	950	213	1051	257	1976	0.293	0.438	0.351
UMCP-2	S	441	327	95	442	97	631	0.274	0.410	0.329
NTU	S	179	863	27	53	753	12	0.661	0.177	0.279

Table 13. Chinese Opinion Holders Analysis: Holder-Based Results

Group	L/S	CRT	InCRT	Miss	F-A	P-H	CRT-NUM	P	R	F
CUHK	L	1375	92	1	386	1854	1476	0.742	0.932	0.826
ISCAS	L	871	422	0	396	1689	1958	0.516	0.445	0.478
GATE-1	L	475	363	0	66	904	2774	0.525	0.171	0.258
GATE-2	L	82	112	0	12	206	1943	0.398	0.042	0.076
UMCP-1	L	1232	964	0	1955	4151	2875	0.297	0.429	0.351
UMCP-2	L	1130	1051	0	1976	4157	2874	0.272	0.393	0.321
NTU	L	452	121	0	34	607	2672	0.745	0.169	0.276
CUHK	S	678	48	1	127	854	841	0.794	0.806	0.800
ISCAS	S	391	189	0	162	742	857	0.527	0.456	0.489
GATE-1	S	218	182	0	22	422	1244	0.517	0.175	0.262
GATE-2	S	46	66	0	4	116	952	0.397	0.048	0.086
UMCP-1	S	574	405	0	628	1607	1266	0.357	0.453	0.400
UMCP-2	S	536	442	0	631	1609	1266	0.333	0.423	0.373
NTU	S	209	53	0	13	275	1197	0.760	0.175	0.284

Table 14. English Opinion Analysis DKE Approach results

Group	L/S	Opinionated			Relevance			Polarity		
		P	R	F	P	R	F	P	R	F
IIT-1	L	0.325	0.588	0.419	—	—	—	0.120	0.287	0.169
IIT-2	L	0.259	0.854	0.397	—	—	—	0.086	0.376	0.140
TUT-1	L	0.310	0.575	0.403	0.392	0.597	0.473	0.088	0.215	0.125
TUT-2	L	0.310	0.575	0.403	0.392	0.597	0.473	0.094	0.230	0.134
Cornell†	L	0.317	0.651	0.427	—	—	—	0.073	0.197	0.107
NII	L	0.325	0.624	0.427	0.510	0.322	0.395	0.077	0.194	0.110
GATE-1	L	0.324	0.905	0.477	0.286	0.632	0.393	—	—	—
GATE-2	L	0.324	0.905	0.477	0.286	0.632	0.393	—	—	—
ICU-IR	L	0.396	0.524	0.451	0.409	0.263	0.320	0.151	0.264	0.192
IIT-1	S	0.070	0.578	0.125	—	—	—	0.027	0.322	0.049
IIT-2	S	0.056	0.840	0.105	—	—	—	0.016	0.359	0.031
TUT-1	S	0.065	0.553	0.117	0.171	0.605	0.266	0.016	0.195	0.029
TUT-2	S	0.065	0.553	0.117	0.171	0.605	0.266	0.019	0.229	0.034
Cornell†	S	0.069	0.662	0.125	—	—	—	0.010	0.135	0.018
NII	S	0.073	0.642	0.131	0.242	0.355	0.287	0.014	0.185	0.027
GATE-1	S	0.070	0.940	0.130	0.112	0.579	0.188	—	—	—
GATE-2	S	0.070	0.940	0.130	0.112	0.579	0.188	—	—	—
ICU-IR	S	0.102	0.616	0.175	0.177	0.266	0.213	0.034	0.301	0.061

Table 15. English Opinion Analysis LWK Approach results

Group	L/S	Opinionated			Relevance			Polarity		
		P	R	F	P	R	F	P	R	F
IIT-1	L	0.326	0.585	0.419	—	—	—	0.136	0.238	0.173
IIT-2	L	0.260	0.844	0.397	—	—	—	0.108	0.343	0.164
TUT-1	L	0.311	0.572	0.402	0.395	0.595	0.475	0.129	0.232	0.166
TUT-2	L	0.311	0.572	0.402	0.395	0.595	0.475	0.125	0.226	0.161
Cornell†	L	0.326	0.524	0.402	—	—	—	0.128	0.200	0.156
NII	L	0.327	0.625	0.429	0.511	0.321	0.395	0.122	0.228	0.159
GATE-1	L	0.324	0.821	0.465	0.291	0.609	0.394	—	—	—
GATE-2	L	0.324	0.821	0.465	0.291	0.609	0.394	—	—	—
ICU-IR	L	0.397	0.532	0.454	0.408	0.262	0.319	0.189	0.247	0.214
IIT-1	S	0.073	0.579	0.129	—	—	—	0.028	0.321	0.051
IIT-2	S	0.058	0.832	0.108	—	—	—	0.017	0.348	0.032
TUT-1	S	0.067	0.551	0.120	0.173	0.603	0.268	0.016	0.195	0.030
TUT-2	S	0.067	0.551	0.120	0.173	0.603	0.268	0.019	0.225	0.035
Cornell†	S	0.072	0.516	0.127	—	—	—	0.010	0.106	0.019
NII	S	0.075	0.638	0.135	0.242	0.353	0.288	0.015	0.181	0.027
GATE-1	S	0.071	0.804	0.131	0.115	0.558	0.191	—	—	—
GATE-2	S	0.071	0.804	0.131	0.115	0.558	0.191	—	—	—
ICU-IR	S	0.103	0.615	0.177	0.178	0.265	0.213	0.035	0.300	0.062

Table 16. English Opinion Analysis YS Approach results

Group	L/S	Opinionated			Relevance			Polarity		
		P	R	F	P	R	F	P	R	F
IIT-1	L	0.326	0.583	0.418	—	—	—	0.120	0.284	0.169
IIT-2	L	0.260	0.842	0.397	—	—	—	0.086	0.370	0.140
TUT-1	L	0.311	0.571	0.403	0.393	0.598	0.474	0.088	0.214	0.125
TUT-2	L	0.311	0.571	0.403	0.393	0.598	0.474	0.095	0.229	0.134
Cornell†	L	0.317	0.500	0.388	—	—	—	0.073	0.152	0.098
NII	L	0.326	0.619	0.427	0.512	0.322	0.395	0.077	0.193	0.110
GATE-1	L	0.327	0.792	0.463	0.287	0.593	0.387	—	—	—
GATE-2	L	0.325	0.813	0.464	0.286	0.612	0.390	—	—	—
ICU-IR	L	0.392	0.493	0.437	0.409	0.261	0.318	0.149	0.247	0.186
IIT-1	S	0.070	0.578	0.126	—	—	—	0.027	0.321	0.049
IIT-2	S	0.056	0.835	0.105	—	—	—	0.016	0.355	0.031
TUT-1	S	0.066	0.555	0.118	0.171	0.605	0.267	0.016	0.194	0.029
TUT-2	S	0.066	0.555	0.118	0.171	0.605	0.267	0.019	0.229	0.034
Cornell†	S	0.069	0.499	0.121	—	—	—	0.001	0.102	0.018
NII	S	0.074	0.641	0.132	0.242	0.353	0.287	0.014	0.184	0.027
GATE-1	S	0.071	0.788	0.130	0.113	0.541	0.186	—	—	—
GATE-2	S	0.070	0.804	0.129	0.113	0.561	0.188	—	—	—
ICU-IR	S	0.100	0.576	0.170	0.178	0.263	0.212	0.032	0.270	0.057

Table 17. English Opinion Holders Analysis results

Group	Lenient			Strict		
	P	R	F	P	R	F
IIT-1	0.198	0.409	0.266	0.054	0.461	0.097
TUT-1	0.117	0.218	0.153	0.029	0.241	0.051
Cornell†	0.163	0.346	0.222	0.041	0.392	0.074
NII	0.066	0.166	0.094	0.018	0.169	0.032
GATE-1	0.121	0.349	0.180	0.029	0.398	0.055
ICU-IR	0.303	0.404	0.346	0.085	0.515	0.146

Table 18. Japanese Opinion Analysis YS Approach results

Group	L/S	Opinionated			Holder (S/A/B/C/D/OE/LE)			Relevance			Polarity		
		P	R	F	P	R	F	P	R	F	P	R	F
EHBN-1	L	0.531	0.453	0.489	0.138	0.085	0.105 (224/46/6/34/806/880/2129)	-	-	-	-	-	-
EHBN-2	L	0.531	0.453	0.489	0.314	0.097	0.149 (236/39/41/77/321/293/2531)	-	-	-	-	-	-
NICT-1	L	0.671	0.315	0.429	0.238	0.102	0.143 (86/0/246/224/378/462/2311)	0.598	0.669	0.632	0.299	0.149	0.199
NICT-2	L	0.671	0.315	0.429	0.238	0.102	0.143 (86/0/246/224/378/462/2311)	0.644	0.417	0.506	0.299	0.149	0.199
TUT	L	0.552	0.609	0.579	0.226	0.224	0.225 (472/137/118/134/1006/1354/1378)	0.630	0.646	0.638	0.274	0.322	0.296
EHBN-1	S	0.414	0.479	0.444	0.079	0.094	0.086 (128/28/2/22/405/1411/1095)	-	-	-	-	-	-
EHBN-2	S	0.414	0.479	0.444	0.183	0.110	0.137 (130/25/29/31/166/626/1299)	-	-	-	-	-	-
NICT-1	S	0.546	0.348	0.425	0.133	0.110	0.120 (73/0/112/104/214/893/1177)	0.470	0.693	0.560	0.168	0.150	0.158
NICT-2	S	0.546	0.348	0.425	0.133	0.110	0.120 (73/0/112/104/214/893/1177)	0.525	0.446	0.482	0.168	0.150	0.158
TUT	S	0.414	0.620	0.497	0.131	0.251	0.172 (292/68/61/63/501/2236/695)	0.505	0.681	0.580	0.161	0.339	0.218

S/A/B/C/D = Five graded evaluation
OE = Over Estimation
LE = Lack of Estimation

Table 19. Japanese Opinion Analysis DKE Approach results

Group	L/S	Opinionated			Relevance			Polarity		
		P	R	F	P	R	F	P	R	F
EHBN-1	L	0.531	0.453	0.488	-	-	-	-	-	-
EHBN-2	L	0.531	0.452	0.488	-	-	-	-	-	-
NICT-1	L	0.671	0.315	0.429	0.598	0.669	0.632	0.298	0.149	0.199
NICT-2	L	0.671	0.315	0.429	0.644	0.417	0.506	0.298	0.149	0.199
TUT-1	L	0.552	0.609	0.589	0.630	0.645	0.638	0.274	0.322	0.296
EHBN-1	S	0.414	0.479	0.444	-	-	-	-	-	-
EHBN-2	S	0.414	0.479	0.444	-	-	-	-	-	-
NICT-1	S	0.545	0.348	0.425	0.470	0.693	0.560	0.168	0.150	0.158
NICT-2	S	0.545	0.348	0.425	0.525	0.446	0.482	0.168	0.150	0.158
TUT-1	S	0.414	0.620	0.497	0.505	0.681	0.580	0.161	0.339	0.218

Table 20. Japanese Opinion Analysis LWK Approach results

Group	L/S	Opinionated			Relevance			Polarity		
		P	R	F	P	R	F	P	R	F
EHBN-1	L	0.531	0.453	0.489	-	-	-	-	-	-
EHBN-2	L	0.531	0.452	0.489	-	-	-	-	-	-
NICT-1	L	0.669	0.313	0.426	0.596	0.666	0.629	0.308	0.140	0.192
NICT-2	L	0.669	0.313	0.426	0.644	0.420	0.509	0.308	0.140	0.192
TUT-1	L	0.550	0.614	0.580	0.628	0.646	0.637	0.287	0.311	0.298
EHBN-1	S	0.412	0.476	0.442	-	-	-	-	-	-
EHBN-2	S	0.412	0.476	0.442	-	-	-	-	-	-
NICT-1	S	0.542	0.343	0.420	0.475	0.690	0.563	0.165	0.143	0.154
NICT-2	S	0.542	0.343	0.420	0.527	0.446	0.483	0.165	0.143	0.154
TUT-1	S	0.411	0.621	0.495	0.510	0.680	0.583	0.160	0.331	0.216

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.

8.2 Directions for NTCIR-7 Opinion Analysis Task

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.

Acknowledgements

We greatly appreciate the efforts of all the participants in *the Opinion Analysis Pilot Task* at the Sixth NTCIR Workshop. We also greatly appreciate Prof. Janyce Wiebe at the University of Pittsburgh for her advisory comments.

References

[1] Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan. Identifying sources of opinions with conditional random fields and extraction patterns. In *Proc. of the 2005 Human Language Technology Conf. and Conf. on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, Vancouver, B. C., 2005.

[2] M. Gamon and A. Aue. *Proc. of Wksp. on Sentiment and Subjectivity in Text at the 21th Int'l Conf. on Computational Linguistics / the 44th Ann. Meeting of the Assoc. for Computational Linguistics (COLING/ACL 2006)*. The Association for Computational Linguistcs, Sydney, Austraria, July 2006.

[3] L. W. Ku, T. H. Wu, L. Y. Lee, and H. H. Chen. Construction of an Evaluation Corpus for Opinion Extraction. In *Proc. of the Fifth NTCIR Wksp. on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access*, pages 513–520, December 2005.

[4] National Institute of Informatics. NTCIR (NII-NACSIS Test Collection for IR Systems) Project [online]. In *NTCIR (NII-NACSIS Test Collection for IR Systems) Project website*, 1998-2007. [cited 2007-01-26]. Available from: <http://research.nii.ac.jp/ntcir/>.

[5] National Institute of Informatics. NTCIR-6 Opinion Analysis Pilot Task [online]. In *NTCIR website*, 2006. [cited 2007-1-26]. Available from: <http://research.nii.ac.jp/ntcir/ntcir-ws6/opinion/index-en.html>.

[6] National Institute of Informatics. NTCIR CLIR Task [online]. In *NTCIR*, 2006. [cited 2007-1-26]. Available from: <http://homepage3.nifty.com/kz_401/>.

[7] National Institute of Standars and Technology. TREC (Text REtrieval Conference) 2006-2007: BLOG Track [online]. In *TREC website*, 2006. [cited 2007-1-26]. Available from: <http://trec.nist.gov/tracks.html>.

[8] 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: Theory and Applications*, volume 20 of *The Information Retrieval Series*, chapter 24, pages 317–336. Springer-Verlag, New York, December 2005.

[9] J. G. Shanahan, Y. Qu, and J. Wiebe. *Computing Attitude and Affect in Text: Theory and Applications*, volume 20 of *The Information Retrieval Series*. Springer-Verlag, New York, December 2005.

[10] J. Wiebe, T. Wilson, and C. Cardie. Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation*, 39(2-3):165–210, 2005.

[11] J. M. Wiebe, E. Breck, C. Buckley, C. Cardie, P. Davis, B. Fraser, D. Litman, D. Pierce, E. Riloff, and T. Wilson. MPQA: Multi-Perspective Question Answering Opinion Corpus Version 1.2, 2006. [cited 2007-1-26]. Available from: <http://www.cs.pitt.edu/mpqa/databaserelease/>.

[12] J. M. Wiebe, T. Wilson, R. F. Bruce, M. Bell, and M. Martin. Learning subjective language. *Computational Linguistics*, 30(3):277–308, 2004.

[13] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proc. of the 2005 Human Language Technology Conf. and Conf. on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, Vancouver, B. C., 2005.