Overview of Multilingual Opinion Analysis Task at NTCIR-7

Yohei Seki†, David Kirk Evans‡, Lun-Wei Ku§,
Le Sun¶, Hsin-Hsi Chen§, Noriko Kando‡
†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
¶Institute of Software, Chinese Academy of Sciences
Beijing 100080, P.R. China
sunle@iscas.cn

November 24, 2008

Abstract

This paper describes an overview of *the Multilingual Opinion Analysis Task* from 2007 to 2008 at the Seventh NTCIR Workshop. We created test collections of 22, 17, 17, 16 topics (7,163, 4,711, 6,174, and 5,301 sentences) in Japanese, English, Traditional Chinese, and Simplified Chinese. Using this test collection, we conducted five subtasks: (1) mandatory opinionated sentence judgment, and optional subtasks of (2) relevant sentence judgment, (3) polarity judgment, (4) opinion holder extraction, and (5) opinion target extraction. 32 results were submitted from 21 participants with five results submitted by the organizers. In this paper we present the task definition, the details of the test collection, the evaluation results of the groups that participated in this task, and their approach.

Keywords: Multilingual Opinion Analysis Task (MOAT), Opinionated Sentence, Opinion Holder, Opinion Target, Relevance, Polarity, and Opinion Expression Level Annotation and Evaluation.

1 Introduction

This paper describes an overview of *the Multilingual Opinion Analysis Task* from 2007 to 2008 at the Seventh NTCIR Workshop [4] (NTCIR-7 Opinion). This was the second effort to produce a multi-lingual test collection for evaluating opinion extraction at NTCIR, following the NTCIR-6 Opinion Analysis Pilot Task [9].

Opinion and sentiment analysis has been receiving a lot of attention in the natural language processing research community recently [6, 2, 10]. In TREC Blog track [5], opinion finding task was conducted in 2007 and 2008. 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 [11], 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 ([13] etc.), the polarity of the expression ([14] etc.), targets ([7] etc.), 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: (Traditional) Chinese, English, and Japanese. In the Seventh NTCIR Workshop, the work on opinion analysis was continued with the *Multilingual Opinion Analysis Task*, which has tracks in Chinese Simplified, Chinese Traditional, Japanese, and English. In this paper, we present an overview of the NTCIR-7 MOAT test collection, task design, and evaluation results using the test collection across the Chinese, Japanese, and English data.

This paper is organized as follows. In Section 2, we explain the task design. Section 3, we briefly introduce the test collection used in the *NTCIR-7 Multilingual Opinion Analysis Task*. Section 4 presents the annotation methodology. Section 5 details the evaluation methodology used, and explains the differences in the approaches taken with examples. Section 6 presents evaluation results in (Traditional/Simplified) Chinese, Japanese and English. Section 7 briefly discusses the system approaches taken by the participants. Finally, we present our conclusions in Section 8.

2 Task Design

2.1 Schedule

The time schedule for *the NTCIR-7 Multilingual Opinion Analysis Task* is as follows.

Table 1: NTCIR-7 MOAT schedule

Date	Event
2007-10-01	First call for participation
2008-08-01	Sample release (1 - 4 topics)
2008-09-01	Formal run topic release
2008-09-10	Japanese/Chinese (Traditional) formal run results due
2008-09-15	Chinese (Simplified) formal run results due
2008-09-26	English formal run results due
2008-10-20∼27	working paper submission
2008-11-15	Camera-ready paper Submission
2008-12-16∼19	NTCIR-7 Meeting

2.2 Participants

Table 2 shows the number of participants per language and across languages. There are a total of 32 results submitted across all languages.

2.3 Task definition

In the NTCIR-7 MOAT, opinion annotation is extended to the sub-sentence level. Annotators were instructed to annotate sentences in contiguous opinion expressions, with possibly multiple opinion expressions per sentence. Two of the annotation features, whether a sentence is opinionated or not, and whether the sentence is relevant to the topic, are evaluated at the sentence level. The other features, opinion holder, opinion target, and polarity, are evaluated at the opinion expression level.

Five evaluation subtasks

We set five subtasks in the evaluation, one of which is mandatory, and the rest of which are optional. In Table 6, the mandatory subtask is to decide whether each sentence expresses an opinion or not. The optional subtasks are to decide whether the sentences are relevant to the set topic or not, to decide the polarity of the opinionated sentences, to extract the opinion holder, and to extract the opinion target.

1. Opinionated sentences
   The opinionated sentences judgment is a binary decision for all sentences.

2. 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 subtask evaluation, each opinionated sentence should be judged as either relevant (Y) or non-relevant (N) to the topic, and non-opinionated sentences should be labeled Not Applicable (N/A). In the NTCIR-6 Opinion Analysis Pilot task, all sentences were annotated for relevance, but in the NTCIR-7 MOAT due to budgetary constraints, only opinionated sentences were annotated for relevance.

3. Opinion polarities
   Polarity is determined for each opinion expression. In addition, the polarity is to be determined with respect to the 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. The possible polarity values are positive (POS), negative (NEG), or neutral (NEU.)

4. Opinion holders
   Opinion holders are annotated for opinion expressions that express an opinion, however, the opinion holder for an opinion expression can occur anywhere in the document. The assessors performed a kind of co-reference resolution by marking the opinion holder for the opinion expression, if the opinion holder is an anaphoric reference noting the antecedent of the anaphora. Each opinion expression may have at least one opinion holder.

5. Opinion targets
   Opinion targets were annotated in a similar

Table 2: Number of participants

<table>
<tr><th colspan="2" rowspan="2">Language</th><th rowspan="2">Japanese</th><th rowspan="2">English</th><th colspan="2">Chinese</th></tr>
<tr><th>Trad.</th><th>Simp.</th></tr>
<tr><td colspan="2">Total</td><td>8</td><td>9</td><td>7</td><td>9</td></tr>
<tr><td rowspan="5">Multi-lingual Participants</td><td>J-E-TC-SC</td><td colspan="4">1</td></tr>
<tr><td>J-E-TC</td><td colspan="3">1</td><td></td></tr>
<tr><td>E-SC</td><td></td><td>1</td><td></td><td>1</td></tr>
<tr><td>E-J</td><td colspan="2">2</td><td colspan="2"></td></tr>
<tr><td>TC-SC</td><td colspan="2"></td><td colspan="2">4</td></tr>
</table>

manner to opinion holders. Each opinion expression may have at least one opinion target.

Sample (training) data

The NTCIR-6 Opinion Analysis Pilot Task (Japanese, Traditional Chinese, and English) was distributed as training data (approximately 32 topics per language.) Because the data is not the same between the NTCIR-6 and NTCIR-7 tasks, between 2 and 4 topics of sample data was distributed per language.

Evaluation metrics

Results for precision, recall, and F-measure will be presented for opinion detection, and for sentence relevance, polarity, opinion holders, and opinion targets 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 and target evaluation, which require some manual judgment and will only be performed once for each participating group. The formal definition is provided for the evaluation below.

1. Mandatory evaluation
   (a) Precision, Recall and F-measure of Opinion using lenient gold standard.
   (b) Precision, Recall and F-measure of Opinion using strict gold standard.
2. Optional evaluation
   For each optional subtask of the evaluation, polarity, relevance, opinion holders, and opinion targets, the following information will be reported:
   (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 Test collection

3.1 Document sets

The test collection is based on *the NTCIR-7 ACLIA Test Collection* which includes newspaper documents from 1998 to 2001. The corpus is a comparable corpus across the languages, with topics shared across languages when enough documents exist in the corpus.

- It consists of Japanese data from 1998 to 2001 from the Mainichi newspapers.
- The Traditional Chinese data contains data from 1998 to 2001 from the China Times, Commercial Times, China Times Express, Central Daily News, China Daily News, United Daily News, Economic Daily News, Min Sheng Daily, United Evening News, and Star News.
- The Simplified Chinese data contains documents from Xinhua News and Lianhe Zaobao from 1998 to 2001.
- The English data also covers from 1998 to 2001 with text from the Mainichi Daily News, Korea Times, Xinhua News, Hong Kong Standard, and the Straits Times.

The test collection was created using about twenty queries from the *NTCIR-7 ACLIA Task*. Relevant documents for each language were searched for using baseline IR systems, and then manually assessed for relevance. Each topic contains from 5 to 20 relevant documents. Unlike the NTCIR-6 Opinion Analysis Pilot Task, which has fairly verbose topic descriptions, the topics in the NTCIR-7 MOAT are short, and typically in the form of simple or complex questions.

As an example of the topics in the NTCIR-7 MOAT, please see Figure 1, which shows topic M01, “Tell me about regenerative medicine”.

Table 3 shows the number of topics, the number of documents, the number of sentences, and the number of opinion expressions for each language.

<TOPICDESCRIPTION><TOPIC>T01</TOPIC> <TITLE>Tell me about regenerative medicine.</TITLE> <NARRATIVE> I would like to know about regenerative medicine, and how effective it is. I would especially like to which diseases' patients it is effective in treating. </NARRATIVE> </TOPICDESCRIPTION>

Figure 1: Topic title and narrative fields for topic N01

Table 3: Test collection size at NTCIR-7 MOAT

Language	Topics			Documents			Sentences			Opinion Expressions (Sub-sentences/Clauses)		
	Sum	Sample	Test	Sum	Sample	Test	Sum	Sample	Test	Sum	Sample	Test
T-Chinese	17	3	14	246	58	188	6,174	1,509	4,655	N/A	N/A	4,657
Japanese	22	4	18	287	38	249	7,163	1,278	5,885	7,569	1,348	6,221
English	17	3	14	167	25	142	4,711	399	4,312	4,733	404	4,329
S-Chinese	16	2	14	271	19	252	5,301	242	4,877	7,523	570	6,953

The percentage of sentences that are opinionated and relevant or that of opinion expressions about polarities are also computed for both the strict and lenient standards, as shown in Table 4.

3.2 Topics

Table 5 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.

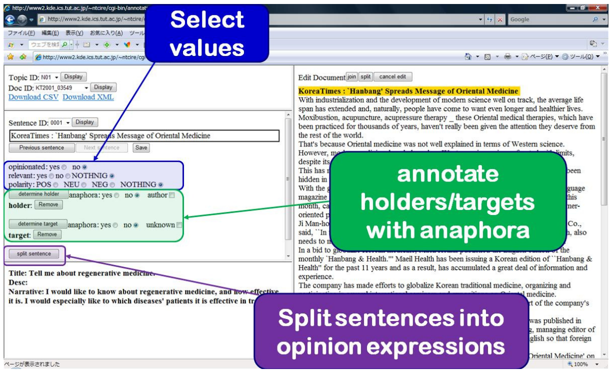
4 Annotation

The NTCIR-6 Opinion Analysis Pilot Task extends previous work in opinion analysis [3, 8, 12] 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.

In this task, we extended the annotation framework of polarities, opinion holders, and targets to a sub-sentence (opinion expression) level to improve the validness of the annotation if multiple opinion expressions are contained in one sentence. Table 6 summarizes the annotation subtasks, which are all being performed for all four languages. All subtasks were annotated by three annotators in each language: (Simplified or Traditional) Chinese, Japanese, and English. One sample topic was used for inter-coder session to improve the agreement between assessors.

4.1 Japanese annotation

The Japanese data was annotated by five annotators, and all topics were annotated by three assessors of them. They were given basic instructions about the annotation that was based on the same strategy in NTCIR-6: *i.e.*, the general or common knowledge and future plans were not counted as opinions, and so forth. Then, they annotated a sample topic (M04) and held a meeting about six hours afterwards to discuss discrepancies with the explicit goal of trying to improve agreements between annotators. The inter-annotator session was limited to one topic, and for the remaining twenty one topics, annotators worked independently. Even these independent results, for opinion and polarity annotation, the macro-averaged value of the kappa coefficient was above 0.70 and 0.60, which are considered high, as shown in Table 7.


Select values
annotate holders/targets with anaphora
Split sentences into opinion expressions
http://www2.kde.ics.tut.ac.jp/~ntcire/cgi-bin/annotat
http://www2.kde.ics.tut.ac.jp/~ntcire/
Google
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール
ウェブを検索
http://www2.kde.ics.tut.ac.jp/~ntcire/cgi
ページ(P) ツール(O)
Display
Display
Previous sentence
Next sentence
Save
determine holder
Remove
determine target
Remove
split sentence
join
split
cancel edit
ページが表示されました
100%

Topic ID: N01
Doc ID: KT2001_03549
Download CSV Download XML

Sentence ID: 0001
KoreaTimes : `Hanbang' Spreads Message of Oriental Medicine

opinionated: yes ○ no ●
relevant: yes ○ no ○ NOTHNIG ●
polarity: POS ○ NEU ○ NEG ○ NOTHING ●
anaphora: yes ○ no ● author □
holder:
anaphora: yes ○ no ● unknown □
target:

Title: Tell me about regenerative medicine.
Desc:
Narrative: I would like to know about regenerative medicine, and how effective it is. I would especially like to which diseases' patients it is effective in tr

Edit Document

KoreaTimes : `Hanbang' Spreads Message of Oriental Medicine
With industrialization and the development of modern science well on track, the average life span has extended and, naturally, people have come to want even longer and healthier lives.
Moxibustion, acupuncture, acupressure therapy _ these Oriental medical therapies, which have been practiced for thousands of years, haven't really been given the attention they deserve from the rest of the world.
That's because Oriental medicine was not well explained in terms of Western science.
However, m limits,
despite its
This has been
hidden in
With the guage
magazine this
month, ca mer-
oriented p
Ji Man-ho Co.,
said, ``In , also
needs to n
In a bid to g the
monthly `Hanbang & Health.'" Maeil Health has been issuing a Korean edition of ``Hanbang & Health'' for the past 11 years and as a result, has accumulated a great deal of information and experience.
The company has made efforts to globalize Korean traditional medicine, organizing and medicine.
rt of the company's
was published in
g, managing editor of
glish so that foreign
Oriental Medicine' on

Figure 2: Annotation tool at Japanese side

We developed a multilingual annotation tool to output CSV and XML formats, but this tool was used only at Japanese side because of time restriction. We also show the interface in Figure 2. With the radio buttons left above, we can select values such as opinionated or not in sentences or opinion expressions (sub-sentences/clauses), relevant, positive, neutral, negative, and so on. If you would like to annotate opinion holders or targets, you can use the buttons left middle and you can drag

Table 4: Opinion percentage in NTCIR-7 MOAT test collection

Language	Opinionated		Relevant (of Opinionated)		Polarity (POS/NEG/NEU)	
	Lenient	Strict	Lenient	Strict	Lenient	Strict
T-Chinese	46.8	44.3	82.72	90.16	34.1 / 40.3 / 25.6	33.2 / 41.2 / 25.6
Japanese	28.9	21.1	43.2	22.6	5.5 / 15.3 / 79.2	4.3 / 10.2 / 85.5
English	25.2	7.5	99.4	95.7	25.0 / 48.0 / 6.0	18.0 / 46.4 / 0.9
S-Chinese	38.3	18.4	95.1	88.7	30.7 / 25.8 / 43.5	30.9 / 6.5 / 62.6

Table 5: NTCIR-7 OMAT topic titles

ID	Title	Language ID			
		Japanese	English	Chinese	
				Traditional	Simplified
M00	Microsoft Anti-monopoly	N00			
M01	Regenerative medicine	N01	N01		N01
M02	American stance on depleted uranium bullets	N02	N02	N02	N02
M03	The impact of 911 terrorist attacks on America's economy	N03	N03	N03	N03
M04	HIV-tainted blood scandal	N04	N04		
M05	Cosovo civil war	N05	N05	N05	N05
M06	Incident with Nepal's ruling family (royalty)	N06	N06		N06
M07	Attacks toward Chinese Indonesian people	N07	N07	N07	N07
M08	Lawsuit American Government against Microsoft	N08	N08		N08
M09	Nuclear weapons tests	N09	N09	N09	N09
M10	Suriyah in the Middle East Peace Process.	N10	N10	N10	N10
M11	The relationship between AOL and Netscape	N11	N11	N11	N11
M12	El Nino	N12	N12	N12	N12
M13	The relationship between China and Russia	N13	N13		N13
M14	Greenhouse gasses	N14	N14	N14	N14
M15	The relationship between NATO and Poland	N15	N15	N15	N15
M16	Thailand in the Asian economic crisis	N16	N16	N16	N16
M17	Yasukuni Shrine	N17	T01		
M18	Chechin (Chechnia) civil war	N18		T96	
M19	Indonesian President Suharto	N19		N04	
M20	Nuclear missile abandonment of North Korea	N20		N13	
M21	Airplane crashes in Asia	N21		N08	
M22	The floods in the Mainland China			N01	
M23	The births of the cloned animals known to the world			N06	
M24	The responses of other countries to Lockerbie Air Disaster				N04

Table 6: Five annotation subtasks at NTCIR-7 Multilingual Opinion Analysis Task

Subtasks	Values	Req'd?	Annotation Unit
Opinionated Sentences	YES, NO	Yes	Sentence
Relevant Sentences	YES, NO	No	Sentence
Opinionated Polarities	POS, NEG, NEU	No	Opinion Expression
Opinion Holders	String, multiple	No	Opinion Expression
Opinion Targets	String, multiple	No	Opinion Expression

Table 7: Annotators agreements by topics at Japanese side

TopicID	Assessor			Sample /Test	Kappa Coefficient								
					Opinionatedness			Polarity			Relevance		
	1	2	3		a1-a2	a1-a3	a2-a3	a1-a2	a1-a3	a2-a3	a1-a2	a1-a3	a2-a3
M00	A	B	C	Sample	0.8029	0.7465	0.6619	0.6996	0.5777	0.5476	0.5432	0.3959	0.6100
M01	A	B	C	Sample	0.8707	0.8742	0.7511	0.7889	0.7448	0.6817	0.7795	0.8080	0.6805
M02	A	B	C	Sample	0.7931	0.8586	0.7416	0.6355	0.6330	0.6544	0.6440	0.7177	0.4971
M03	A	B	C	Test	0.7368	0.7276	0.6652	0.6812	0.6157	0.5390	0.6613	0.6695	0.6236
M04	A	B	C	Test	0.7837	0.8234	0.8202	0.6605	0.7327	0.6993	0.7171	0.7396	0.7366
M05	A	B	C	Test	0.6604	0.5969	0.4852	0.6350	0.6266	0.4443	0.6194	0.5394	0.4747
M06	A	B	C	Test	0.6542	0.6532	0.5932	0.6375	0.6339	0.6604	0.5699	0.6356	0.5568
M07	A	B	C	Test	0.7369	0.7143	0.6772	0.7113	0.6688	0.6276	0.5502	0.5887	0.4611
M08	A	B	C	Test	0.7056	0.6098	0.6773	0.6040	0.4648	0.5104	0.6457	0.6171	0.6233
M09	A	B	C	Test	0.7009	0.6238	0.6658	0.6700	0.5419	0.5815	0.6016	0.5209	0.5600
M10	A	B	C	Test	0.6663	0.7025	0.6747	0.6389	0.6625	0.6355	0.5440	0.7033	0.5537
M11	A	B	C	Test	0.6728	0.5875	0.6641	0.5859	0.4890	0.6421	0.6817	0.5666	0.6331
M12	E	B	C	Test	0.9296	0.8219	0.7896	0.8969	0.7488	0.7177	0.9029	0.7053	0.6777
M13	A	B	C	Test	0.6392	0.6874	0.6092	0.4987	0.5334	0.5538	0.6052	0.5843	0.5196
M14	E	B	C	Test	0.8314	0.8169	0.8212	0.7708	0.7235	0.7634	0.6899	0.5848	0.4645
M15	A	B	C	Test	0.6700	0.5707	0.7190	0.6423	0.4812	0.5999	0.6338	0.2903	0.3027
M16	A	B	C	Test	0.7457	0.6375	0.6782	0.7137	0.5333	0.5785	0.3402	0.6054	0.2705
M17	A	E	D	Sample	0.7711	0.7965	0.8504	0.6194	0.6380	0.6693	0.5108	0.4580	0.7129
M18	A	B	D	Test	0.7278	0.5422	0.6242	0.6587	0.4834	0.5279	0.7075	0.5744	0.6335
M19	A	B	D	Test	0.6477	0.6275	0.6790	0.6164	0.5532	0.5388	0.5035	0.4337	0.6054
M20	A	B	C	Test	0.7228	0.7798	0.8200	0.6902	0.7415	0.7669	0.3843	0.4525	0.7346
M21	A	B	D	Test	0.7025	0.6430	0.8058	0.7088	0.6259	0.6924	0.6719	0.5947	0.7471
Macro Avg						0.7135			0.6341			0.5905	
Micro Avg						0.7128			0.6380			0.5785	

opinion holder elements in the right pane. When you split the sentence into multiple opinion expressions, you can push the button left below.

4.2 Traditional Chinese annotation

In NTCIR-6, we defined that one sentence should end with a period, so that we could segment sentences automatically. However, we found the length of many sentences was extremely long, and these sentences were very complex in annotating their opinionated features. Therefore, in NTCIR-7, we segmented sentences into their opinion units as possible as we could before annotations. In the end, we only found two sentences with more than one opinion clauses in testing data. Therefore, we evaluated results at sentence level since only two extra opinion clauses were found.

After the sentence segmentation, a pool of ten annotators were used to annotate the documents, with three annotators per topic. Prior to annotation, the annotators underwent a two-hour-long orientation period where the purpose of the annotation was explained, and examples of sentences and their annotations were given. After this orientation session, the annotators started their practice session for three hours. They were free to ask the annotation coordinator questions about specific sentences if they were unsure of the labeling, and they can discuss with each other in this session to ensure consistency between the annotators in those cases. After the practice session they started to annotate officially, and they were not allowed to discuss their labeling with each other in this phase. They could still ask the annotation coordinator if they had any questions.

For annotations of the traditional Chinese corpus, two stages were performed. The first stage, as mentioned in the last paragraph, we had three annotators annotate each topic. In the second stage, we calculated the pairwise kappa of opinionated tags among them. If any kappa below 0.3 appeared, we asked the forth annotators to annotate this topic. And if any kappa below 0.3 still appeared among three sets of data with highest kappa values, we had the fifth annotator generate new data again. Finally, we selected three sets of data with highest pairwise kappa values as the final data to generate the gold standard.

For each topic the agreement between the three annotators was computed, and the macro average (over topics) and the micro average (over sentences) are shown in Table 8. Because the corpus is annotated by ten annotators, the column annotator n (n=1,2,3) does not necessary denote the same person. Therefore, the values of macro and micro average are calculated by the average of three annotators. Though agreements vary from topics, we still can see worse agreements for the relevance tags from statistics. Note that we reannotated according to the pairwise kappa values of opinionated tags, and it may be one of the reason that we had higher agreements among opinionated and polarity tags. We also found that the reason why we had low relevance agreements is because some annotators gave many more Y relevance tags than the others instead of real contradictions. In other words, annotators might have problems in judging which sentences are relevant by only reading topic descriptions. This might be due to the shorter descriptions of topics we adopted from the ACLIA cluster, which contains multilingual QA tasks.

4.3 Simplified Chinese annotation

The simplified Chinese data was annotated by twelve annotators, and all topics were annotated by three of them. Prior to formal annotation, they were given a basic instruction note about the purpose of the annotation and annotation rules based on English 2008 Opinion Annotator Sample Instructions by David and Lun-Wei Ku's OAT (Opinion Annotation Tool) 3.0 Manual. We use this tool for our annotation work. Firstly, four annotators are selected to do the sample annotation. They annotated two sample topics and held a meeting about 4 hours to discuss the discrepancies and the ambiguous statements in the instruction rules. Then all the annotators held another meeting about 4 hours to learn how to do the annotation work based on the finished sample date and the sample in OTA3.0 manual. Four teams are set up and the person who done the sample annotation are set as the leader of the four-person team. The annotators can ask any questions to the team leader about the specific sentences if they were unsure of the labeling. Three annotators will annotate each sentence in the data set for opinion features. When more than one opinion is present in a sentence, the annotators will separate the sentence into separate opinion clauses, which are then annotated for the opinion features. There are also sentence-level features that only pertain to the sentence as a whole.

After all annotation data finished, we check them by calculating the pairwise kappa for relevance and polarity tags. If the kappa value of one topic is obviously below than others, we ask the team leader to check this topic again in case any misunderstanding by one of the annotator in his team. Finally, we calculated the agreement between the three annotators, and the macro average (over topics) and the micro average (over sentences) are shown in Table 9. Because the corpus is annotated by twelve annotators, the column annotator a1, a2, a3 does not necessary denote the same person.

4.4 English annotation

The English data was annotated using a pool of six annotators. The annotators underwent an initial two hour training session, then annotated a training topic over the span of a week. The annotators then took part in a second two hour meeting where we discussed examples from the training topic that were contentious, and strategies for improving consistency between annotators. After the two training meetings, annotators began work on the sample data and formal run data. The annotators used the same tool that was used for the Siplified and Traditional Chinese annotation, Lun-Wei Ku's Opinion Annotation Tool version 3.0.

After the annotation for the topics was completed, for topics with low kappa agreement the annotators were asked to review their annotation and confirm their annotation. The kappa results are shown in Table 10.

5 Evaluation Approach

5.1 Traditional Chinese evaluation

5.1.1 Opinionated / relevance

Relevance, opinion extraction, polarity detection, holder and target identifications were evaluated at traditional Chinese side. Among them, relevance and opinion extraction adopted the same metric, while metrics for polarity detection, holder and target identification are similar.

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 (P) is computed as

$$\frac{\#\text{system\_correct}}{\#\text{system\_proposed}}.$$

Recall (R) is computed as

$$\frac{\#\text{system\_correct}}{\#\text{gold}}.$$

And the F-measure (F) is computed as

$$\frac{2 \times \mathrm{P} \times \mathrm{R}}{\mathrm{P} + \mathrm{R}}.$$

where the number of sentences is either the number of opinionated or relevant sentences according to the strict or lenient criteria.

Table 8: Annotators agreements by topics at traditional Chinese side

TopicID	Assessor			Sample	Kappa Coefficient								
(TrC/Multi				/Test	Opinionatedness			Polarity			Relevance		
-lingual)	1	2	3		a1-a2	a1-a3	a2-a3	a1-a2	a1-a3	a2-a3	a1-a2	a1-a3	a2-a3
N03/M03	A	B	C	Test	0.4958	0.7276	0.4968	0.7162	0.7830	0.5339	0.5228	0.1810	0.2648
N04/M19	A	B	C	Test	0.3120	0.2331	0.5824	0.5657	0.6360	0.8691	0.0930	0.2583	0.1800
N05/M05	A	B	C	Test	0.4145	0.4796	0.4053	0.8207	0.7408	0.8530	0.2170	0.1708	0.0489
N06/M23	A	B	C	Test	0.3429	0.3742	0.7005	0.8042	0.9506	0.8504	0.3225	0.4890	0.5028
N07/M07	A	B	C	Test	0.4176	0.5129	0.4520	0.5499	0.5927	0.5797	0.5823	0.7100	0.7221
N08/M21	A	B	C	Test	0.3880	0.3760	0.5200	0.8825	0.8446	0.9674	0.5645	0.6005	0.6960
N09/M09	A	B	C	Test	0.3307	0.5168	0.3535	0.7772	0.7338	0.8307	0.2027	0.4923	0.4162
N10/M10	A	B	C	Test	0.6204	0.5814	0.5863	0.7500	0.8238	0.7439	0.2496	0.2355	0.5224
N11/M11	A	B	C	Test	0.2915	0.3863	0.5762	0.8880	0.9566	0.9109	0.0992	0.2707	0.3729
N12/M12	E	B	C	Test	0.5382	0.4147	0.4176	0.9603	0.6497	0.6551	0.3018	0.1626	0.2069
N13/M20	A	B	C	Test	0.4994	0.3710	0.3624	0.3624	0.7635	0.8312	0.3624	0.3624	0.6799
N14/M14	E	B	C	Test	0.4580	0.5004	0.4885	0.7912	0.9330	0.7313	0.3493	0.0908	0.3497
N15/M15	A	B	C	Test	0.4178	0.5343	0.5837	0.8577	0.7936	0.7427	0.0723	0.0313	0.5572
N16/M16	A	B	C	Test	0.3926	0.4873	0.2987	0.7818	0.7531	0.8168	0.0098	0.3973	0.0595
Macro Avg					0.4581			0.7709			0.3329		
Micro Avg					0.4706			0.7699			0.3376		

Table 9: Annotators agreements by topics at Simplified Chinese side

TopicID	Assessor			Sample	Kappa Coefficient								
(SiC/Multi				/Test	Opinionatedness			Polarity			Relevance		
-lingual)	1	2	3		a1-a2	a1-a3	a2-a3	a1-a2	a1-a3	a2-a3	a1-a2	a1-a3	a2-a3
N03/M03	A	B	C	Test	0.3862	0.3890	0.5455	0.2937	0.3119	0.5577	0.7786	0.7368	0.8710
N04/M24	A	B	C	Test	0.3975	0.4567	0.9222	0.3067	0.3225	0.7603	0.3926	0.3926	1.0000
N05/M05	A	B	C	Test	0.5209	0.4295	0.7442	0.3670	0.3932	0.5682	1.0000	1.0000	1.0000
N06/M06	A	B	C	Test	0.3847	0.3392	0.8925	0.1890	0.1935	0.8081	0.4956	0.6647	0.6647
N07/M07	A	B	C	Test	0.6045	0.5050	0.6577	0.3782	0.4166	0.4433	0.6946	0.0534	0.0139
N08/M08	A	B	C	Test	0.4286	0.4592	0.6936	0.3151	0.4551	0.5562	0.5203	0.6299	0.7668
N09/M09	A	B	C	Test	0.2518	0.3163	0.4856	0.3119	0.3864	0.4649	0.5093	0.5879	0.6501
N10/M10	A	B	C	Test	0.4913	0.4736	0.5072	0.3798	0.4593	0.4382	0.3958	0.0401	0.1773
N11/M11	A	B	C	Test	0.6244	0.6148	0.4436	0.5119	0.4985	0.3908	0.7898	1.0000	0.7898
N12/M12	A	B	C	Test	0.2906	0.1628	0.1058	0.1997	0.1748	0.0952	0.1261	0.3130	0.3225
N13/M13	A	B	C	Test	0.1297	0.3718	0.1738	0.1044	0.3369	0.1791	0.3243	0.6422	0.1898
N14/M14	A	B	C	Test	0.3711	0.4435	0.4932	0.3283	0.3936	0.4213	0.6157	0.6157	0.5272
N15/M15	A	B	C	Test	0.2173	0.5064	0.3147	0.2089	0.4390	0.2602	0.9347	0.9321	0.9538
N16/M16	A	B	C	Test	0.1139	0.4753	0.1835	0.1197	0.3733	0.1497	0.9316	1.0000	0.9316
Macro Avg					0.4362			0.3634			0.6185		
Micro Avg					0.3862			0.3314			0.6030		

Table 10: Annotators agreements by topics for English data

TopicID	Assessor			Sample	Kappa Coefficient								
				/Test	Opinionatedness			Polarity			Relevance		
	1	2	3		a1-a2	a1-a3	a2-a3	a1-a2	a1-a3	a2-a3	a1-a2	a1-a3	a2-a3
N03	A	B	C	Test	0.3675	0.3066	0.3430	0.3944	0.3405	0.3526	0.5684	0.5017	0.5561
N04	A	B	C	Test	0.3850	0.2940	0.3850	0.3108	0.2226	0.3108	0.5665	0.4556	0.5665
N05	A	B	C	Test	0.2483	0.2637	0.2203	0.2581	0.1868	0.2340	0.4469	0.3951	0.4469
N06	A	B	C	Test	0.1132	0.1189	0.0868	0.0517	0.0735	0.0372	0.4215	0.3491	0.4317
N07	A	B	C	Test	0.1782	0.2101	0.1784	0.1291	0.1205	0.1219	0.2427	0.2353	0.2470
N08	D	E	F	Test	0.3846	0.3790	0.3704	0.4562	0.4193	0.4409	0.3534	0.3577	0.3314
N09	A	B	C	Test	0.4218	0.4136	0.4081	0.1641	0.1349	0.1680	0.2913	0.2410	0.2913
N10	D	E	F	Test	0.4555	0.4292	0.4405	0.3837	0.3979	0.3837	0.5297	0.4641	0.5186
N11	A	B	C	Test	0.2571	0.2839	0.2897	0.0098	0.0177	0.0114	0.1167	0.0970	0.1206
N12	D	E	F	Test	-0.0101	-0.0080	-0.0101	-0.0111	-0.0065	-0.0111	0.1207	0.0934	0.1207
N13	D	E	F	Test	0.3047	0.2996	0.2838	0.3725	0.3004	0.3543	0.4082	0.3353	0.4034
N14	D	E	F	Test	0.0549	0.0171	0.0549	0.1242	0.1212	0.1293	0.3457	0.5894	0.3085
N15	A	B	C	Test	0.3516	0.4117	0.3516	0.1251	0.1938	0.1251	0.4188	0.4276	0.4188
N16	D	E	F	Test	0.0706	0.0831	0.0525	0.0860	0.0820	0.0834	0.1331	0.1247	0.1346
Macro Avg					0.2369			0.1958			0.3459		
Micro Avg					0.2343			0.2156			0.3282		

5.1.2 Polarity

The LWK approach gives two evaluation results. The traditional metric, which is the same as that used in NTCIR6, evaluates opinionated sentences correctly reported by participants' systems under the opinionated strict or lenient gold standards. Under the traditional metric, a Set Precision (S-P) is computed as

$$\frac{\#\text{system\_correct}(\text{polar} = \text{POS}, \text{NEU}, \text{NEG})}{\#\text{system\_correct}(\text{opn} = \text{Y})}.$$

The other recall based metric evaluates all opinionated sentences under the opinionated strict or lenient gold standards.

Under this recall based metric, Precision is computed as

$$\frac{\#\text{system\_correct}(\text{polar} = \text{POS}, \text{NEU}, \text{NEG})}{\#\text{system\_proposed}(\text{opn} = \text{Y})}.$$

Recall is computed as

$$\frac{\#\text{system\_correct}(\text{polar} = \text{POS}, \text{NEU}, \text{NEG})}{\#\text{gold}(\text{opn} = \text{Y})}.$$

And the F-measure is computed as

$$\frac{2 \times \text{P} \times \text{R}}{\text{P} + \text{R}}.$$

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/targets

The evaluation at traditional Chinese for opinion holders / targets 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. To ensure the correctness of the evaluation, every record which is different from all aliases of the correct holders is checked manually again. If any correct answer is found, it is added to the gold standard. At last, all runs are evaluated according to the final gold standard by keyword matching and their precision, recall and F-measure are calculated.

The traditional metric (Holder-T, Target-T) and the recall based metric (Holder-RB, Target-RB) are both adopted for evaluation again. ¡ They evaluate opinionated sentences correctly reported by participants' systems under the opinionated strict or lenient gold standards. In these opinion sentences, all holders/targets annotated by three annotators are viewed as correct answers. Participants can report anyone of them to get the score. Under this traditional metric, Precision is computed as

$$\frac{\#\text{system\_correct}(\text{polar} = \text{POS}, \text{NEU}, \text{NEG})}{\#\text{system\_correct}(\text{opn} = \text{Y})}.$$

Recall is computed as

$$\frac{\#\text{system\_correct}(\text{polar} = \text{POS}, \text{NEU}, \text{NEG})}{\#\text{system\_correct}(\text{opn} = \text{Y})}.$$

And the F-measure is computed as

$$\frac{2 \times \text{P} \times \text{R}}{\text{P} + \text{R}}.$$

Under the recall based metric, Precision is computed as

$$\frac{\#\text{system\_correct}(\text{polar} = \text{POS}, \text{NEU}, \text{NEG})}{\#\text{system\_proposed}(\text{opn} = \text{Y})}.$$

Recall is computed as

$$\frac{\#\text{system\_correct}(\text{polar} = \text{POS}, \text{NEU}, \text{NEG})}{\#\text{gold}(\text{opn} = \text{Y})}.$$

And the F-measure is computed as

$$\frac{2 \times \text{P} \times \text{R}}{\text{P} + \text{R}}.$$

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 proposed answer without the resolution of the anaphor or the coreference is treated as correct. Of course, the resolved form of an anaphor or a coreference is correct.

5.2 Simplified Chinese evaluation

The evaluation approach we used is almost the same as in the traditional Chinese side. However, when calculating the value of #system-correct holder or #system-correct target, we use the automatically matching method for tight schedule. In the lenient metric, for each sentence the holders/targets of the three annotators tagged will be combined, if the system proposed holder/target can match one of them, it will be considered correct. In the strict metric, the holder/target of the three annotators must be identical, only when the system proposed holder/target can match the annotators' holder/target, it will be considered correct. This is one of the reasons that the holder and target performances are relative lower than other languages.

5.3 English evaluation

5.3.1 Opinionated sentence

The English evaluation takes the same lenient and strict definitions as the other languages, where two annotators must agree for the lenient case, and all three for the strict case. Opinionated sentence precision, recall, and F-measure is defined as in Section 5.4.1.

5.3.2 Relevance

For relevance evaluation, a contingency table for the categories YES, NO, NA, and NONE, where NONE is the category used when annotators do not agree on the relevance of the sentence. NA is used when a sentence is not opinioated. Precision and Recall are computed as given below.

$$\begin{aligned}\text{Precision(P)} &= \frac{\#\text{system\_correct}(\text{rel} = \text{YES}|\text{NO}|\text{NA})}{\#\text{system\_proposed}(\text{rel} = \text{YES}|\text{NO}|\text{NA})}.\\ \text{Recall(R)} &= \frac{\#\text{system\_correct}(\text{rel} = \text{YES})}{\#\text{assessors\_agreeed}(\text{rel} = \text{YES})}.\end{aligned}$$

Unfortunately, for English while relevance was defined only to be annotated for Opinionated sentences, many annotators also annotated non-opinionated sentences, so the denominator for recall is larger than the set of all opinionated sentences.

5.3.3 Polarity

In the NTCIR-6 evaluation polarity was weighted according to the annotator scores, but this year the English evaluation is more similar to the Chinese and Japanese evaluations: two or more annotators must agree in the lenient case for the annotation to be used, or all three annotators for the strict evaluation. 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 or not enough annotators agree on the polarity.

$$\begin{aligned}\text{Precision(P)} &= \frac{\#\text{system\_correct}(\text{pol} = \text{POS}|\text{NEG}|\text{NEU})}{\#\text{system\_proposed}(\text{pol} = \text{POS}|\text{NEG}|\text{NEU})}.\\ \text{Recall(R)} &= \frac{\#\text{system\_correct}(\text{pol} = \text{POS}|\text{NEG}|\text{NEU})}{\#\text{assessors\_agreeed}(\text{pol} = \text{POS}|\text{NEG}|\text{NEU})}.\\ \text{F\_measure(F)} &= \frac{2 \times \text{P} \times \text{R}}{\text{P} + \text{R}}.\end{aligned}$$

5.3.4 Opinion Holder and Target

Opinion Holder and Target 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 or target, and system opinion holders or targets for a given sentence are matched using exact string matches to the opinion holders or targets in the equivalence class. Exact matches are counted as correct, and if no matches are found then a human judge[1] is asked to determine if the system answer matches ones of the opinion holders or targets in the equivalence class for the sentence. If there is match, the system opinion holder or target is added to the equivalence class, otherwise it is marked as a known incorrect opinion holder or target.

The initial database of opinion holder and target equivalence classes is created by adding the

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

opinion holders and targets 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 and target database to match opinion holders.

Precision is computed as the number of correctly matched opinion holders or targets divided by the number of offered opinion holders or targets. The denominator for Recall is computed by assuming one opinion holder and one target for each opinion unit.

5.4 Japanese evaluation

5.4.1 Opinionated sentence

For opinionated sentence evaluation, we took the same approach with other languages. For lenient standards, we prepared the correct answer set of sentences that two of three assessors agreed them as opinionated. For strict standards, we prepared the answer set of sentences that three of three assessors agreed them as opinionated. Then, the precision, recall, and F-measure are defined as follows:

$$
\begin{aligned}
\text{Precision(P)} &= \frac{\text{\#system\_correct(opn = Y)}}{\text{\#system\_proposed(opn = Y)}}. \\
\text{Recall(R)} &= \frac{\text{\#system\_correct(opn = Y)}}{\text{\#assessors\_agreeed(opn = Y)}}. \\
\text{F\_measure(F)} &= \frac{2 \times \text{P} \times \text{R}}{\text{P} + \text{R}}.
\end{aligned}
$$

5.4.2 Relevance

For relevance judgment, we only annotated relevant information for opinionated sentences. Then, the precision, recall, and F-measure are defined based on opinionated sentences as follows:

$$
\begin{aligned}
\text{Precision(P)} &= \frac{\# \begin{pmatrix} system\_correct(rel = Y)\& \\ system\_proposed(opn = Y) \end{pmatrix}}{\# \begin{pmatrix} system\_proposed(rel = Y)\& \\ system\_proposed(opn = Y)\& \\ assessors\_agreed(opn = Y) \end{pmatrix}}. \\
\text{Recall(R)} &= \frac{\# \begin{pmatrix} system\_correct(rel = Y)\& \\ system\_proposed(opn = Y) \end{pmatrix}}{\text{\#assessors\_agreed(rel = Y)}}. \\
\text{F\_measure(F)} &= \frac{2 \times \text{P} \times \text{R}}{\text{P} + \text{R}}.
\end{aligned}
$$

The reason that the precision at Japanese side was computed as above is to treat the different number of submitted sentences from all the participants equally: some participants submitted the results for all sentences, while the other participants submitted the results only for opinionated sentences.

- The numerators of the precision and recall was counted on the opinionated sentences the system proposed, to exclude N/A cases, as defined in Section 2.3.

- The denominator of the precision is counted on the opinionated sentences that the assessors agreed. This corresponds to the *set precision* metric at the other language sides. However, with the pure precision, the participants who submitted the results for all sentences will be disadvantage, because their correct answers in the non-opinionated sentence are not counted with the former reason. Based on this discussion, we decided to define *set precision* as the precision metric at Japanese side.

The participants can also evaluate the results for whole sentences by using the evaluation script we provided with the option '-w'. In this case, the denominator of the precision is counted on all sentences, and three metrics are computed based on the same approach with the opinionated sentence judgment case.

5.4.3 Polarity

For polarity judgment, we also took the same approach with the relevance judgment, based on the same reason. In this subtask, however, the evaluation was conducted at the opinion expression (subsentence/clause) level, not at the sentence level.

For the agreement estimation between assessors, we implemented *YS method* (= to use in the answer set of polarity as POS/NEG/NEU sets that the assessors agreed strictly or leniently) and *DKE method* (= in strict case, the same approach with *YS*; in lenient case, weight the answer set according to the number of agreed assessors) in the evaluation script, that were defined in NTCIR-6 workshop [9]. We set *YS method* as a default function and *DKE method* as an optional function. We also implemented *LWK method* (= majority voting) as another optional function, that was explained in Section 5.1.2.

5.4.4 Opinion holder & target

Unfortunately, we did not receive the submissions of opinion holder/target subtasks from the participants at Japanese side (except the organizer) this time.

6 Evaluation Results

6.1 Simplified Chinese

Table 11 shows the evaluation results of opinionated, relevance, and polarity based on the lenient and strict gold standards. Table 12 shows the evaluation results of holder and target based on the lenient and strict standards we described in 5.2.

6.2 English

Table 13 lists the evaluation results of the opinionated, relevance, and polarity analysis for English based on lenient and strict standards.

6.3 Japanese

Table 15 lists the evaluation results of opinionated, relevance, and polarity analysis at Japanese side based on lenient and strict standards.

6.4 Traditional Chinese

Table 16 lists the evaluation results for relevance, opinion extraction and polarity detection based on lenient and strict gold standards. The results of relevance, opinion extraction and polarity detection are evaluated automatically while holder and target identifications manually, so they are separated evaluated and listed in two tables.

For evaluation of opinion holder and target identifications, we applied both the sentence-based traditional evaluation (Holder-T) and the recall-based evaluation (Holder-RB), as shown in Table 17. In both evaluation metrics, precision, recall and F-measure are listed, but their definitions, introduced in LWK evaluation approach, are different. Because the denominators in the formulae for calculating the precision and recall in the traditional evaluation are the same, the values of precision, recall and f-measure are the same. Note that the traditional metric may have preference in a high precision opinion system, while the performance of the opinion extraction task may also influence the performance evaluated by the recall-based metric.

7 Discussions on Participant System

7.1 English

Nine groups submitted runs for the English evaluation.

The Sussex group use very few language-dependent features, and have an interesting segmentation method for Japanese and Chinese. Their relevance system is unsupervised and compares word ranks across topics to make a relevance decision, while they used the training data to collect a list of opinionated terms, then manually selected a small number of terms and an automatically expanded set of related terms as features for their opinionated judgement system.

The North Eastern University group's submission for English took advantage of an existing sentiment lexicon with a rule-based system and compared that to a Naive Bayes system trained over the NTCIR-6 Opinion Pilot Task data and MPQA data.

The TUT submission separates opinions into two classes: those from the author's point of view, and those from an authority's point of view. They use a χ^2 test to identify syntactic and semantic features from the NTCIR-6 and MPQA corpora, and compare SVM voting and Multi-label classification approaches. They participated in the opinionated, relevance judgement, polarity classification, and opinion holder identification sub-tasks.

The KLE group uses data from SentiWordNet and Levine's Verb Classes to identify opinionated features, and weight them using a BM25 algorithm and the NTCIR-6 corpus to estimate some smoothing values for the opinionated component. They determine polarity based on the weights of the features from the sentiment lexicons, and they use features from a dependency parse to determine the opinion holder.

The University of Neuchâel group takes a largely language-independent approach to the task, and use a statistical method to identify tokens (in the English case words, but unigrams and bigrams in Japanese and Chinese) that are useful for opinionated and polarity detection.

The MIRACLE team use a machine learning approach for the English and Japanese opinionated, polarity, and relevance tasks. They use seed terms from an existing lexicon that have been annotated for polarity as features and a KNN clustering approach for opinionated and polarity classification. Relevance uses a distance metric between expanded semantic space vectors for each of the topics and the sentences, and data from NTCIR-6 for parameter optimization.

The UKP group use a sequential SVM to label words, and compute the opinion holder, opinionated, and polarity tasks simultaneously over the learned tags. They use both lexical and syntactic features and existing sentiment lexicons, and investigate domain adaptation using a structural correspondence learning approach.

Table 11: Simplified Chinese opinionated/relevance/polarity analysis results

Group	RunID	L/S	Opinionated			Relevance			Polarity	Recall-based Polarity		
			P	R	F	P	R	F	S-P	P	R	F
BUPT	1	L	0.604	0.3991	0.4807	N/A			N/A			
ICLPKU	1	L	0.4803	0.8004	0.6003	0.9775	0.6559	0.785	0.4505	0.2164	0.3606	0.2705
ICLPKU	2	L	0.4487	0.7983	0.5745	0.9775	0.6559	0.785	0.4524	0.203	0.3612	0.2599
NEUNLP	1	L	0.4721	0.7116	0.5676	N/A			N/A			
NLCL	1	L	0.4425	0.3991	0.4197	0.963	0.3258	0.4869	N/A			
NLCL	2	L	0.4822	0.3686	0.4178	0.9752	0.2799	0.4349	N/A			
NLCL	3	L	0.4316	0.6988	0.5336	0.9714	0.585	0.7302	N/A			
NLPR	1	L	0.5822	0.7753	0.665	N/A			N/A			
NLPR	2	L	0.588	0.4842	0.5311	N/A			N/A			
NLPR	3	L	0.4551	0.5725	0.5071	N/A			N/A			
NLPR	4	L	0.5769	0.5639	0.5703	N/A			N/A			
NTU	1	L	0.5939	0.6089	0.6013	0.9656	0.7693	0.8564	0.4956	0.2944	0.3018	0.298
NTU	2	L	0.5956	0.6067	0.6011	0.9796	0.5798	0.7284	0.5079	0.3025	0.3082	0.3053
NTU	3	L	0.5956	0.6067	0.6011	0.9767	0.5796	0.7275	0.5159	0.3072	0.313	0.3101
TTRD	1	L	0.412	0.9636	0.5772	0.9507	0.6981	0.8051	0.4348	0.1791	0.4189	0.251
TTRD	2	L	0.4456	0.756	0.5607	0.968	0.7363	0.8364	0.4947	0.2204	0.374	0.2774
WIA **	1	L	0.5862	0.8208	0.6839	0.994	0.5032	0.6682	0.7419	0.4348	0.6089	0.5074
ISCAS*	1	L	0.4649	0.7442	0.5723	0.9703	0.9288	0.9491	N/A			
BUPT	1	S	0.6312	0.4421	0.52	N/A			N/A			
ICLPKU	1	S	0.4486	0.8207	0.5801	0.9845	0.6743	0.8004	0.2836	0.1272	0.2327	0.1645
ICLPKU	2	S	0.3984	0.8252	0.5373	0.9845	0.6743	0.8004	0.2807	0.1118	0.2316	0.1508
NEUNLP	1	S	0.4358	0.7339	0.5469	N/A			N/A			
NLCL	1	S	0.3857	0.402	0.3937	0.9736	0.3326	0.4959	N/A			
NLCL	2	S	0.4425	0.3898	0.4144	0.9848	0.2846	0.4415	N/A			
NLCL	3	S	0.3667	0.706	0.4827	0.9827	0.5897	0.7371	N/A			
NLPR	1	S	0.6096	0.892	0.724	N/A			N/A			
NLPR	2	S	0.6129	0.5501	0.5798	N/A			N/A			
NLPR	3	S	0.4197	0.637	0.506	N/A			N/A			
NLPR	4	S	0.5973	0.6459	0.6207	N/A			N/A			
NTU	1	S	0.6314	0.7517	0.6863	0.9748	0.7859	0.8702	0.3378	0.2133	0.2539	0.2318
NTU	2	S	0.6343	0.7494	0.6871	0.9878	0.5969	0.7441	0.3611	0.229	0.2706	0.2481
NTU	3	S	0.6343	0.7494	0.6871	0.9866	0.5943	0.7418	0.373	0.2366	0.2795	0.2563
TTRD	1	S	0.3481	0.9699	0.5124	0.9631	0.7006	0.8112	0.2882	0.1003	0.2795	0.1476
TTRD	2	S	0.3958	0.755	0.5193	0.9759	0.7487	0.8474	0.3923	0.1553	0.2962	0.2038
WIA **	1	S	0.6098	0.8964	0.7259	0.9969	0.524	0.687	0.5329	0.3250	0.4777	0.3868
ISCAS*	1	S	0.4271	0.8118	0.5597	0.9828	0.9369	0.9593	N/A			

* = organizer

** = late submission

Table 12: Simplified Chinese opinion holder/target analysis results

Group	Run	L/S	Holder-T			Holder-RB			Target-T			Target-RB		
	ID	/S	P	R	F	P	R	F	P	R	F	P	R	F
ICLPKU	1	L	0.4124	0.4124	0.4124	0.1981	0.3301	0.2476	0.0033	0.0033	0.0033	0.0016	0.0027	0.0020
ICLPKU	2	L	0.4095	0.4095	0.4095	0.1838	0.3269	0.2353	0.0034	0.0034	0.0034	0.0015	0.0027	0.0019
TTRD	1	L	0.1129	0.1129	0.1129	0.0464	0.1081	0.0649	0.0151	0.0151	0.0151	0.0062	0.0144	0.0087
TTRD	2	L	0.1270	0.1270	0.1270	0.0566	0.0958	0.0711	0.0546	0.0546	0.0546	0.0243	0.0412	0.0306
NLPR	1	L	0.4286	0.4286	0.4286	0.2495	0.3323	0.2850	N/A			N/A		
NLPR	2	L	0.4497	0.4497	0.4497	0.2645	0.2178	0.2389	N/A			N/A		
NLPR	3	L	0.4037	0.4037	0.4037	0.1838	0.2311	0.2047	N/A			N/A		
NLPR	4	L	0.4298	0.4298	0.4298	0.2479	0.2424	0.2451	N/A			N/A		
NTU	1	L	0.2909	0.2909	0.2909	0.1728	0.1771	0.1749	N/A			N/A		
NTU	2	L	0.0397	0.0397	0.0397	0.0236	0.0241	0.0239	N/A			N/A		
NTU	3	L	0.1587	0.1587	0.1587	0.0945	0.0963	0.0954	N/A			N/A		
WIA**	1	L	0.6656	0.6656	0.6656	0.3901	0.5463	0.4552	0.4505	0.4505	0.4505	0.2640	0.3697	0.3081
ICLPKU	1	S	0.4104	0.4104	0.4104	0.0937	0.3216	0.1451	0	0	0	0	0	0
ICLPKU	2	S	0.4275	0.4275	0.4275	0.0829	0.3363	0.1329	0	0	0	0	0	0
TTRD	1	S	0.1719	0.1719	0.1719	0.0282	0.1608	0.0480	0	0	0	0	0	0
TTRD	2	S	0.1125	0.1125	0.1125	0.0218	0.0819	0.0345	0.0625	0.0625	0.0625	0.0061	0.0479	0.0108
NLPR	1	S	0.4759	0.4759	0.4759	0.1719	0.4035	0.2410	N/A			N/A		
NLPR	2	S	0.4821	0.4821	0.4821	0.1688	0.2368	0.1971	N/A			N/A		
NLPR	3	S	0.4689	0.4689	0.4689	0.0980	0.2866	0.1461	N/A			N/A		
NLPR	4	S	0.4715	0.4715	0.4715	0.1558	0.2661	0.1965	N/A			N/A		
NTU	1	S	0.3839	0.3839	0.3839	0.1392	0.2515	0.1792	N/A			N/A		
NTU	2	S	0.0268	0.0268	0.0268	0.0098	0.0175	0.0126	N/A			N/A		
NTU	3	S	0.1652	0.1652	0.1652	0.0605	0.1082	0.0776	N/A			N/A		
WIA**	1	S	0.7574	0.7574	0.7574	0.2817	0.6754	0.3976	0.5185	0.5185	0.5185	0.1077	0.4795	0.1759

** = late submission

Table 13: English opinionated/relevance/polarity analysis results

Group	RunID	L/S	Opinionated			Relevance			Polarity		
			P	R	F	P	R	F	P	R	F
ICU	1	L	0.2435	0.3687	0.2933	0.2758	0.3648	0.3141	N/A		
ICU	2	L	0.2435	0.3687	0.2933	0.2757	0.3648	0.3141	N/A		
kle	1	L	0.3529	0.7272	0.4752	N/A			0.2586	0.4301	0.3230
kle	2	L	0.3751	0.5410	0.4430	N/A			0.2608	0.3159	0.2857
kle	3	L	0.2736	0.9327	0.4231	N/A			0.2381	0.5536	0.3330
MIRACLE	1	L	0.5952	0.0116	0.0227	0.3741	0.3189	0.3444	N/A		
NEUNLP	1	L	0.3522	0.7788	0.4851	N/A			N/A		
NEUNLP	2	L	0.2952	0.8986	0.4444	N/A			N/A		
NLCL	1	L	0.3780	0.1014	0.1599	0.1296	0.0685	0.0896	N/A		
NLCL	2	L	0.3384	0.1862	0.2402	0.1746	0.1438	0.1577	N/A		
NLCL	3	L	0.2766	0.8461	0.4169	0.4817	0.6887	0.5669	N/A		
sics	1	L	0.4192	0.6101	0.4970	N/A			0.1838	0.2413	0.2087
sics	2	L	0.2847	0.8452	0.4259	N/A			0.0278	0.0594	0.0379
sics	3	L	0.2916	0.9235	0.4433	N/A			0.0309	0.0699	0.0428
TUT	1	L	0.3185	0.4092	0.3582	0.2092	0.1755	0.1909	0.1943	0.1830	0.1885
TUT	2	L	0.3282	0.2562	0.2878	0.1647	0.1136	0.1344	0.1896	0.1142	0.1425
TUT	3	L	0.3185	0.4092	0.3582	0.2092	0.1755	0.1909	0.1621	0.1527	0.1573
UKP07	1	L	0.3305	0.9060	0.4844	N/A			0.2028	0.4394	0.2775
UKP07	2	L	0.3305	0.9060	0.4844	N/A			0.2001	0.4336	0.2738
UniNe	1	L	0.3322	0.6995	0.4504	0.4170	0.5992	0.4918	0.2279	0.3671	0.2812
UniNe	2	L	0.3774	0.5760	0.4560	0.3423	0.4539	0.3903	0.2457	0.3193	0.2777
UniNe	3	L	0.3829	0.5530	0.4525	0.3305	0.4325	0.3747	0.2502	0.3100	0.2769
ICU	1	S	0.0743	0.3777	0.1241	0.0981	0.3797	0.1559	N/A		
ICU	2	S	0.0743	0.3777	0.1241	0.0981	0.3797	0.1559	N/A		
kle	1	S	0.1109	0.7678	0.1938	N/A			0.0687	0.4645	0.1197
kle	2	S	0.1195	0.5789	0.1981	N/A			0.0683	0.3365	0.1136
kle	3	S	0.0808	0.9257	0.1487	N/A			0.0657	0.6209	0.1188
MIRACLE	1	S	0.2857	0.0116	0.0222	0.0853	0.3040	0.1333	N/A		
NEUNLP	1	S	0.1105	0.8204	0.1947	N/A			N/A		
NEUNLP	2	S	0.0881	0.9009	0.1605	N/A			N/A		
NLCL	1	S	0.1168	0.1053	0.1107	0.0526	0.0847	0.0649	N/A		
NLCL	2	S	0.1089	0.2012	0.1413	0.0744	0.1876	0.1066	N/A		
NLCL	3	S	0.0838	0.8607	0.1527	0.1644	0.7266	0.2682	N/A		
sics	1	S	0.1336	0.6533	0.2219	N/A			0.0524	0.2796	0.0883
sics	2	S	0.0832	0.8297	0.1512	N/A			0.0005	0.0047	0.0010
sics	3	S	0.0879	0.9350	0.1607	N/A			0.0010	0.0095	0.0019
TUT	1	S	0.0961	0.4149	0.1561	0.0740	0.1853	0.1057	0.0569	0.2180	0.0903
TUT	2	S	0.1039	0.2724	0.1504	0.0615	0.1220	0.0817	0.0484	0.1185	0.0687
TUT	3	S	0.0961	0.4149	0.1561	0.0740	0.1853	0.1057	0.0359	0.1374	0.0569
UKP07	1	S	0.1026	0.9443	0.1850	N/A			0.0570	0.5024	0.1024
UKP07	2	S	0.1026	0.9443	0.1850	N/A			0.0576	0.5071	0.1034
UniNe	1	S	0.1050	0.7430	0.1840	0.1614	0.6768	0.2607	0.0637	0.4171	0.1105
UniNe	2	S	0.1196	0.6130	0.2001	0.1431	0.5627	0.2282	0.0687	0.3602	0.1146
UniNe	3	S	0.1225	0.5944	0.2032	0.1382	0.5367	0.2198	0.0668	0.3365	0.1115

Table 14: English Opinion Holder and Opinion Target results

		Holder			Target		
Group	Type	P	R	F	P	R	F
ICU	L	NA			0.1059	0.1761	0.1324
ICU	S	NA			0.0374	0.1793	0.0618
kle	L	0.4000	0.5076	0.4474	NA		
kle	S	0.1333	0.5322	0.2132	NA		
TUT	L	0.3923	0.2833	0.3290	NA		
TUT	S	0.1250	0.2829	0.1735	NA		

Table 15: Japanese opinionated/relevance/polarity analysis results

Group	RunID	L/S	Opinionated			Relevance			Polarity		
			P	R	F	P	R	F	P	R	F
EHBN	1	L	0.4921	0.7313	0.5883	0.4819	0.6354	0.5481	N/A		
HCU	1	L	0.619	0.5138	0.5615	N/A			N/A		
HCU	2	L	0.7754	0.2111	0.3319	N/A			N/A		
MIRAC	1	L	0.316	0.0894	0.1394	0.4545	0.0816	0.1384	0.2465	0.0183	0.0341
NAK	1	L	0.8115	0.3416	0.4808	N/A			0.4922	0.1801	0.2637
NAK	2	L	0.7886	0.3092	0.4442	N/A			0.4977	0.167	0.2501
NAK	3	L	0.7813	0.3633	0.496	N/A			0.4934	0.1936	0.2781
NLCL	1	L	0.4255	0.2234	0.293	0.5367	0.1891	0.2797	N/A		
TAK	1	L	0.5191	0.2798	0.3636	N/A			0.4638	0.1138	0.1828
TUT*	1	L	0.6742	0.562	0.613	0.5527	0.2925	0.3825	0.4596	0.214	0.292
TUT*	2	L	0.6742	0.562	0.613	0.5527	0.2925	0.3825	0.4283	0.1994	0.2721
UniNe	1	L	0.5363	0.1999	0.2912	0.4147	0.1918	0.2623	0.3251	0.0548	0.0938
EHBN	1	S	0.3738	0.7627	0.5017	0.2808	0.7321	0.4059	N/A		
HCU	1	S	0.4894	0.5577	0.5213	N/A			N/A		
HCU	2	S	0.6544	0.2446	0.3561	N/A			N/A		
MIRAC	1	S	0.2412	0.0936	0.1349	0.2233	0.0821	0.1201	0.2394	0.0166	0.031
NAK	1	S	0.6885	0.3979	0.5043	N/A			0.4933	0.2154	0.2999
NAK	2	S	0.6612	0.3559	0.4627	N/A			0.5062	0.1998	0.2865
NAK	3	S	0.6574	0.4197	0.5123	N/A			0.5022	0.2232	0.309
NLCL	1	S	0.3135	0.226	0.2627	0.301	0.2107	0.2479	N/A		
TAK	1	S	0.4166	0.3083	0.3544	N/A			0.5172	0.1316	0.2098
TUT*	1	S	0.5416	0.6199	0.5781	0.3062	0.3357	0.3203	0.4806	0.2417	0.3216
TUT*	2	S	0.5416	0.6199	0.5781	0.3062	0.3357	0.3203	0.4535	0.2281	0.3035
UniNe	1	S	0.4164	0.2131	0.2819	0.1553	0.1464	0.1507	0.2914	0.0497	0.0849

* = organizer

Table 16: Traditional Chinese opinionated/relevance/polarity analysis results

Group	RunID	L/S	Opinionated			Relevance			Polarity	Recall-based Polarity		
			P	R	F	P	R	F	S-P	P	R	F
WIA	1	L	0.7298	0.5211	0.6080	0.9949	0.5306	0.6921	0.6931	0.5058	0.3611	0.4214
CityUHK	1	L	0.6601	0.8446	0.7411	N/A			0.5361	0.3539	0.4528	0.3973
CityUHK	2	L	0.7432	0.6526	0.6950	N/A			0.5197	0.3862	0.3391	0.3612
CityUHK	3	L	0.6520	0.8698	0.7453	N/A			0.5053	0.3294	0.4395	0.3766
iclpku	1	L	0.7015	0.6279	0.6626	0.9943	0.6768	0.8054	0.4810	0.3374	0.3020	0.3187
iclpku	2	L	0.5812	0.7383	0.6504	0.9943	0.6768	0.8054	0.4513	0.2623	0.3332	0.2935
NLCL	1	L	0.5358	0.2676	0.3570	0.9240	0.1801	0.3015	N/A			
NLCL	2	L	0.4760	0.7415	0.5798	0.9283	0.4846	0.6368	N/A			
NLCL	3	L	0.4944	0.5064	0.5003	0.9298	0.3407	0.4987	N/A			
NTU*	1	L	0.5648	0.8969	0.6931	0.9615	0.7103	0.8170	0.4875	0.2753	0.4372	0.3379
NTU*	2	L	0.5575	0.8868	0.6846	0.9804	0.6448	0.7780	0.4796	0.2674	0.4253	0.3283
NTU*	3	L	0.5575	0.8868	0.6846	0.9807	0.6324	0.7689	0.4811	0.2682	0.4267	0.3294
TTRD**	1	L	0.5110	0.9345	0.6607	0.9673	0.8413	0.8999	0.3747	0.1915	0.3501	0.2476
TTRD	2	L	0.5664	0.6622	0.6106	0.9660	0.8967	0.9300	0.4651	0.2634	0.3080	0.2840
UniNe	1	L	0.5428	0.9267	0.6846	0.9614	0.8456	0.8998	0.4293	0.2330	0.3978	0.2939
WIA	1	S	0.8520	0.6003	0.7043	0.9788	0.4061	0.5740	0.7003	0.5966	0.4204	0.4932
CityUHK	1	S	0.8364	0.9037	0.8687	N/A			0.5463	0.4569	0.4936	0.4746
CityUHK	2	S	0.9003	0.7333	0.8082	N/A			0.5288	0.4761	0.3877	0.4274
CityUHK	3	S	0.8178	0.9220	0.8668	N/A			0.5259	0.4301	0.4849	0.4558
iclpku	1	S	0.8567	0.6998	0.7704	0.9530	0.5626	0.7075	0.5085	0.4357	0.3559	0.3918
iclpku	2	S	0.7423	0.7866	0.7638	0.9530	0.5626	0.7075	0.4909	0.3644	0.3861	0.3750
NLCL	1	S	0.6259	0.2930	0.3991	0.8487	0.1454	0.2482	N/A			
NLCL	2	S	0.5830	0.7412	0.6526	0.8573	0.4110	0.5556	N/A			
NLCL	3	S	0.6005	0.5255	0.5605	0.8640	0.2863	0.4301	N/A			
NTU*	1	S	0.7076	0.9307	0.8040	0.8849	0.6437	0.7453	0.4979	0.3523	0.4634	0.4003
NTU*	2	S	0.6978	0.9172	0.7926	0.9199	0.5890	0.7182	0.4809	0.3356	0.4411	0.3811
NTU*	3	S	0.6978	0.9172	0.7926	0.9123	0.5849	0.7128	0.4844	0.3380	0.4443	0.3839
TTRD**	1	S	0.6452	0.9395	0.7650	0.8992	0.8044	0.8491	0.3924	0.2531	0.3686	0.3002
TTRD	2	S	0.7384	0.6744	0.7050	0.8885	0.8676	0.8779	0.4758	0.3514	0.3209	0.3354
UniNe	1	S	0.6921	0.9379	0.7965	0.8746	0.8443	0.8592	0.4431	0.3067	0.4156	0.3529

* = organizer
** = late submission

Table 17: Traditional Chinese opinion holder/target analysis results

Group	Run	L	Holder-T			Holder-RB			Target-T			Target-RB		
	ID	/S	P	R	F	P	R	F	P	R	F	P	R	F
WIA	1	L	0.8254	0.8254	0.8254	0.2992	0.4305	0.3531	0.6058	0.6058	0.6058	0.2196	0.3160	0.2591
iclpku	1	L	0.5872	0.5872	0.5872	0.2051	0.3684	0.2635	0.0213	0.0213	0.0213	0.0074	0.0133	0.0095
iclpku	2	L	0.5988	0.5988	0.5988	0.1733	0.4420	0.2490	0.0199	0.0199	0.0199	0.0058	0.0147	0.0083
NTU*	1	L	0.5028	0.5028	0.5028	0.1414	0.4508	0.2153	N/A			N/A		
NTU*	2	L	0.4587	0.4587	0.4587	0.1273	0.4066	0.1939	N/A			N/A		
NTU*	3	L	0.3191	0.3191	0.3191	0.0886	0.2829	0.1349	N/A			N/A		
TTRD**	1	L	0.5645	0.5645	0.5645	0.1443	0.5373	0.2275	0.0358	0.0358	0.0358	0.0091	0.0340	0.0144
TTRD	2	L	0.5947	0.5947	0.5947	0.1678	0.4002	0.2365	0.1066	0.1066	0.1066	0.0301	0.0718	0.0424
WIA	1	S	0.8238	0.8238	0.8238	0.1988	0.4952	0.2838	0.6331	0.6331	0.6331	0.1528	0.3806	0.2181
iclpku	1	S	0.5797	0.5797	0.5797	0.1303	0.4053	0.1972	0.0228	0.0228	0.0228	0.0051	0.0159	0.0077
iclpku	2	S	0.5816	0.5816	0.5816	0.1035	0.4570	0.1688	0.0223	0.0223	0.0223	0.0040	0.0175	0.0065
NTU*	1	S	0.4825	0.4825	0.4825	0.0814	0.4490	0.1378	N/A			N/A		
NTU*	2	S	0.4358	0.4358	0.4358	0.0723	0.3997	0.1225	N/A			N/A		
NTU*	3	S	0.2969	0.2969	0.2969	0.0493	0.2723	0.0834	N/A			N/A		
TTRD**	1	S	0.5496	0.5496	0.5496	0.0822	0.5295	0.1423	0.0372	0.0372	0.0372	0.0056	0.0358	0.0096
TTRD	2	S	0.5840	0.5840	0.5840	0.0972	0.4013	0.1565	0.1356	0.1356	0.1356	0.0226	0.0932	0.0363

* = organizer
** = late submission

The ICU group participated in the opinionated, relevance, and target identification tasks. They use the KLDivergence statistic to identify words that are discriminative for opinionated or non-opinionated sentences, and use tri-gram language models for each topic with web snippets used for query expansion for relevance. They use a syntactic parse and distance from potential opinion targets in the parse tree with a machine learning framework over the parse constituents to identify targets.

The SICS team use only general linguistic information and dependency parsing to extract three different types of features from the NTCIR-6 data for an SVM-based approach. They present an interesting investigation of feature selection for the opinionated sentence detection task.

7.2 Japanese

Eight teams participated in the task at Japanese side. We introduced their systems by dividing them into two parts: (1) four multilingual participants by focusing on language portable approach and (2) four Japanese native participants by focusing on opinion clues from Japanese native viewpoints.

Four teams of Japanese side participants challenged at multilingual sides with language portable approach. The *Sussex University* (NLCL) team did not use preliminary word segmentation in Japanese and Chinese. They used the same routine for finding basic lexical units in all languages. They extracted opinion bearing terms using χ^2 score. Their results improved with the manual list plus all associated words. *University of Neuchâtel* (UniNe) team utilized bigram indexing scheme for Japanese and suggested using a statistical method (Z score) to identify the terms that adequately characterize subsets of the corpus belonging to positive, negative, neutral or non opinionated subsets. *MIRACLE* team is a research consortium formed by research groups of three different universities in Madrid. They implemented a semantic knowledge base using machine-translated dictionary. They provided some interesting approaches for non-native speakers to implement a multilingual opinion extraction system. A Japanese side organizer also submitted the results as *Toyohashi University of Technology* (TUT) team in Japanese and English. The feature selection was based on statistical χ-square test and implemented polarity classification systems using multi-label classification methods.

Other four Japanese participants focused on Japanese oriented clues to extract opinion or classify polarity. *NEC* (EHBN) focused on consecutive property of opinionated or relevant sentences in Japanese and proposed *Sliding Window Framework* and proved to obtain high recall. *Hiroshima City University* (HCU) implemented their opinion extraction system with 760,000 sentence-final expressions collected from their newspaper corpus and provided the detailed analysis for their results. *Keio University* (NAK) also focused on auxiliary verbs and utilized some other features such as character types to extract opinions and classify the polarity. *Sokendai University* (TAK) implemented a polarity classification system using a small number of signpost expressions.

7.3 Traditional Chinese

Seven teams participated in the task at traditional Chinese side. In this section, we introduced only participants who participated in both the traditional and simplified Chinese tasks, or only in the traditional Chinese task, in alphabetic order. Descriptions of other multilingual participants' systems (NLCL and UniNe) are already provided by the Japanese side.

CityUHK: Language Information Sciences Center, City University of Hong Kong In their system, supervised approaches and ensemble techniques have been used and compared in our participating system. Two kinds of supervised approaches were employed: 1) the supervised lexicon-based approach, 2) machine learning approaches, and ensemble techniques were also used to combine the results given by different approaches. Three classifiers, the supervised lexicon-based classifier, SVM classifier, and Bayes classifier, are adopted in their system.

iclpku: Institute of Computational Linguistic, Peking University In their system, maximum entropy model is used to predict the polarity class. A rule-based pattern matching scheme is devised to find topic-relevant sentence. For the subtask of detecting holders and targets, the CRF model is adopted.

NTUCopeOpi: National Taiwan University They adopted their opinion analysis system CopeOpi to analyze opinionated information in NTCIR-7 MOAT tasks document collections. For opinion extraction task, their algorithm was based on the bag-of-character methods proposed in NTCIR-6 and considered morphological structures of Chinese words to extract opinion words correctly. How distant an opinion word is to the end of the sentence was also considered to adjust its opinion weight. For the relevance judgment task, the distance of two sentences were also considered as a factor in their weighting formula.

TTRD: Tornado Technologies Co. Their method for opinion analysis tasks involves two different approaches: (1) the machine learning-based prototype system (on the basis of support vector machines (SVMs)) and (2) stochastic estimation of the character-level of words. For relevance judgment, they adopt lemur as their language model and Tornado Search 5.0 to perform the second retrieval. Two-stage Dirichlet smoothing strategy is then applied.

7.4 Simplified Chinese

Nine groups submitted runs for the simplified Chinese evaluation. It's a pity that one group give up to submit the technical paper. In this section, we introduced the system who participated in the simplified Chinese task and the system are not described by others in alphabetic order. Descriptions of the multilingual participant's system (NLCL) are already provided by the Japanese side. Description of other systems who participated both the traditional and simplified Chinese tasks (iclpku, NTUCopeOpi, TTRD) are already provided by the traditional Chinese side.

ISCAS: Institute of Software, Chinese Academy of Sciences For identifying the opinionated sentences, in their system an EM algorithm is used to extract the sentiment words based on the sentimental dictionary, and then an iterative algorithm is used to estimate the score of the sentiment words and the sentences. In relevant sentences detection sub-task, pseudo feedback and query extension methods are used based on the traditional IR model.

NEU: Northeastern University Their system adopts a sentiment lexicon-based (SLB) approach to identifying opinionated sentences. Their Chinese sentiment dictionary is extracted from the famous Chinese concept lexicon Hownet.

NLPR: Institute of Automation, Chinese Academy of Sciences In their system the domain adaptation technique is used for identifying the subjective sentences and the data in NTCIR6 is used for training subjective classifier. In order to extract the opinion holder, they use the CRF model, which is combined with manual designed heuristics rules. The features in their CRF model include part-of-speech features, semantic class features, contextual features, and dependency features through parsing analysis.

WIA: Chinese University of Hong Kong They adopt a multi-pass coarse-fine analysis strategy for detecting opinionated sentences. A base classifier firstly coarsely estimates the opinion of sentences and the document. The obtained document-level and sentence-level opinions are then incorporated in a complex classifier to analyze the opinion of sentences again to generate refined sentence and document opinions. The updated opinion features are feed back to the classifier to further refine the opinion analysis. Such circles terminate until the analysis results converge. Similar strategy is adopted to sentence-topic relevance estimation. Furthermore, the mutual reinforcement between the analysis of sentence relevance and sentence opinion are integrated in one framework.

8 Conclusions

8.1 Overview of results in NTCIR-7

In *NTCIR-7 MOAT*, we have several new challenging points different from the first *NTCIR-6 OAT* as follows:

1. The participants could use *NTCIR-6 OAT corpus*: large size test collection with detailed annotation appropriate for training use.

2. The task focused on not only sentence-level annotation but also opinion expression (subsentence/clause) level annotation.

3. The number of teams who participated at multilingual sides with language portable approaches increased (two teams ⇒ eight teams).

4. A new target language (Simplified Chinese) and a new subtask (opinion target detection) were added to the scope of *MOAT*.

Basically, the evaluation results of the participants improved a lot based on the first point and the mature of the opinion analysis technology. We also succeeded in opinion expression (subsentence/clause) level annotation using annotation tools and attained high κ coefficient for the assessors' agreements. For the third and fourth points, many pariticpants challenged their new exciting approach this year and provided new precious insights.

8.2 Directions for next challenge

Instead of sudden change of the target document genre, 21 teams participated in the MOAT this year. We greatly appreciate their efforts and try to advance the community for the next challenge. We took the legally safe approach this year with considering the present state of the research community, but we also understood the value of opinion analysis of the user generated contents such as blogs deeply. We struggle to solve the problem and hope to collaborate with participants to break out a new challenge.

Acknowledgements

We greatly appreciate the efforts of all the participants in *the Multilingual Opinion Analysis Task* at the Seventh NTCIR Workshop.

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 Test Collection for IR Systems) Project [online]. In *NTCIR (NII Test Collection for IR Systems) Project website*, 1998-2008. [cited 2008-10-10]. Available from: <http://research.nii.ac.jp/ntcir/>.

[5] National Institute of Standars and Technology. TREC (Text REtrieval Conference) 2006-2008: BLOG Track [online]. In *TREC-BLOG Information Retrieval Wiki*, 2008. [cited 2008-10-10]. Available from: <http://ir.dcs.gla.ac.uk/wiki/TREC-BLOG>.

[6] B. Pang and L. Lee. *Opinion Mining and Sentiment Analysis*. Now Publishers Inc, July 2008.

[7] J. Ruppenhofer, S. Somasundaran, and J. Wiebe. Finding the Sources and Targets of Subjective Expressions. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008.

[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] 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 *Proc. of the Sixth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access*, pages 265–278, NII, Japan, May 2007.

[10] 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.

[11] The Association for the Advancement of Artificial Intelligence. International Conference on Weblogs and Social Media. In *Proc. of the third Int'l AAAI Conference on Weblogs and Social Media (forthcoming)*, San Jose, California, March 2009. [cited 2008-10-10]. Available from: <http://www.icwsm.org/2009/index.shtml>.

[12] 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.

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

[14] 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.