Overview of Multilingual Opinion Analysis Task at *NTCIR-8*

- A Step Toward Cross Lingual Opinion Analysis -

Yohei Seki
Toyohashi University of Technology
Aichi 441-8580, Japan
seki@tut.jp

Lun-Wei Ku
Naitonal Taiwan University
Taipei 10617, Taiwan
lwku@csie.ntu.edu.tw

Le Sun
Institute of Software, Chinese Academy of Sciences
Beijing 100080, P.R. China
sunle@iscas.cn

Hsin-Hsi Chen
Naitonal Taiwan University
Taipei 10617, Taiwan
hhchen@csie.ntu.edu.tw

Noriko Kando
National Institute of Informatics
Tokyo 101-8430, Japan
kando@nii.ac.jp

ABSTRACT

In this paper, we discuss the goal, task description, evaluation results, and the participants approaches for the Third Multilingual Opinion Analysis Task (MOAT) in the NTCIR-8 workshop[1]. We explored our task from our past experiences towards cross-lingual opinion analysis application. In order to solve this challenging problem, we believe that two solutions are required: (1) language-transfer approaches with semi-supervised techniques and (2) cross-lingual opinion question answering capabilities. To get closer to this goal, we created an opinion annotation corpora based on opinion Q&A in a common format across languages. Many teams participated in the subtasks for more than two language sides, and some teams also participated in the cross-lingual subtask. There were 56 result runs submitted from 16 participants, and half of the participants submitted the results in more than two language related tasks. We hope that the MOAT in NTCIR-8 will be a milestone for cross-lingual opinion analysis researches.

Categories and Subject Descriptors

H.3.1 [Information Storage and Retrieval]: linguistic processing

General Terms

Experimentation

Keywords

Multilingual Opinion Analysis, Sentiment Analysis, Polarity Classification, Opinion Holder Identification, Opinion Target Identification, and Crosslingual Opinion Question Answering

[1]http://research.nii.ac.jp/ntcir

NTCIR-8 Workshop Meeting, 2010, Tokyo, Japan.

1. INTRODUCTION

This paper describes the goal, task description, evaluation results, and discussions from the participants for the Third Multilingual Opinion Analysis Task (MOAT) in NTCIR-8 workshop. Opinion and sentiment analysis has recently been receiving a lot of attention in the natural language processing research community [6, 4, 10], also in the business application[2]. In TREC Blog track [5], an opinion finding task was conducted from 2007-2010. Opinion question & answering, and summarization tasks were conducted in TAC 2008 [3]. With the broad range of information sources available on the web, and the 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 the prevalent attitudes on the web. As a result, interest in automatically detecting sentences in which an opinion is expressed ([14] etc.), the polarity of the expression ([15] etc.), and the opinion holders ([2] etc.) & targets ([7] etc.) has been receiving more attention in the research community.

Based on these backgrounds, we started an opinion analysis task in NTCIR-6 as a pilot task [8] with four subtasks in three languages. In NTCIR-7 MOAT [9], we added an opinion target identification subtask and Simplified Chinese documents into the task definition. The document set was selected to focus on the IR task in NTCIR-6 and the question answering (ACLIA[3]) task in NTCIR-7. The number of MOAT participants from multilingual sides drastically increased from two in NTCIR-6 to nine in NTCIR-7.

Recently, many researchers have focused on a resource less approach for sentiment analysis. Blitzer et al. [1] proposed a domain adaptation approach for sentiment classification. Wan [12] solved the Chinese sentiment classification problem by using an English sentiment corpora on the Web. These researches can be categorized as semi-supervised approaches

[2]http://www.sentimentsymposium.com/
[3]http://aclia.lti.cs.cmu.edu/ntcir8

for opinion/sentiment analysis to solve the resource problem by using small labeled data and large unlabeled data. We recognize that solving the language resource problems in sentiment analysis for non-native researches are still important research questions. On the other hand, applications such as the EMM News Explorer[4] provides excellent services to catch the viewpoints from different countries. We also understand that providing different opinions from different countries gives us a chance for better worldwide communications.

In NTCIR-8, we extended our subtasks toward cross lingual opinion analysis oriented applications based on combinations of the element technologies in opinion analysis. There were 56 result runs submitted from 16 participants, and half the participants submitted the results in more than two languages related tasks, as listed in Table 8 in Section 2. Table 1 summarizes the MOAT progress in NTCIR-6, 7, and 8.

- We defined a new subtask: cross-lingual subtask to evaluate answer opinion extraction accuracy in the languages different from the question.
- We selected the document sets used in MOAT this time to focus on the opinion question answering task.
- We unified the annotation format across four languages using TSV& XML.
- We decreased the number of assessments compared to those in the previous MOAT, but increased the agreements between assessments from the strong opinion definition based on past annotation experiences.

This paper is organized as follows. In Section 2, we explain the task design in the NTCIR-8 Multilingual Opinion Analysis Task. Section 3 presents the evaluation results in (Traditional/Simplified) Chinese, Japanese, and English. Section 4 briefly discusses the system approaches taken by the participants. Finally, we present our conclusions in Section 5.

2. TASK DESIGN

2.1 Subtasks

2.1.1 *Conventional five subtasks: opinionated/relevance sentence judgment, polarity judgment, and opinion holder/target identification*

We continued five subtasks to evaluate the element technologies in opinion analysis. All the subtasks are prepared in four languages: English, Traditional Chinese, Simplified Chinese, and Japanese. The simple definitions are described as follows.

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 an opinion question, 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 opinion questions. In the NTCIR-8 MOAT, only opinionated sentences were annotated for relevance.

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

5. Opinion targets
   The opinion targets were annotated in a similar manner to the opinion holders. Each opinion clause may have at least one opinion target.

Table 2 summarized the annotation values for the NTCIR-8 MOAT conventional subtasks.

2.1.2 *Cross-lingual Opinion Q&A Subtask*

The cross-lingual subtask is defined as the opinion question& answering task. Along with the questions in English, the answer opinions should be extracted in different languages. To keep it simple, the extraction unit is defined as sentences. The answer set is defined as the combination of the annotation in the conventional subtasks, as opinionatedness, polarity, and answerness should be matched against the definition in the question description.

2.2 Test Collection

In NTCIR-8 MOAT, we used news corpora in English, Simplified Chinese, Traditional Chinese, and Japanese. We changed the following points from the past MOAT.

- To use recently published corpora, we utilized the news corpus from 2002-2005.
- To analyze the sentiment expression in native English, we used the New York Times as the English side corpora.

The test collection size at NTCIR-8 MOAT is described in Table 3. The 20 opinion questions list used in NTCIR-8 MOAT is also listed in Table 4. The percentage of sentences that are opinionated and relevant or that of opinion clauses about the polarities are also computed, as listed in Tables 5 and 6.

2.3 Annotation

To improve the annotation consistency between assessors, we took on the two strategies:

[4]http://emm.newsbrief.eu/NewsBrief/clusteredition/en/latest.html

Table 1: MOAT Progress in NTCIR-6, 7, & 8

NTCIR	Language	Subtask	Unit	Focused Application	Corpora	Period
NTCIR-6	E,J,TC	opinionated relevance polarity holder	sentence	IR	Mainichi, Yomiuri, CIRB,Xinhua English, Hong Kong Standard etc.	1998-2001
NTCIR-7	+SC	+target	opinion clause	Q&A(ACLIA)	+Xinhua Chinese	-
NTCIR-8	-	+crosslingual	-	Opinion Q&A	+NYT, UDN	2002-2005

```
<TOPIC>
<NUM>N03</NUM>
<TITLE>Bali Island Terrorist Bombing</TITLE>
<QUESTION>What reasons were discussed about Bomb Terror in Bali Island in October, 2002?</QUESTION>
<POLARITY>Neutral</POLARITY>
<OPTYPE>reason controversy</OPTYPE>
<CONC>Bali Island Bombing Terrorism</CONC>
<PERIOD>2002-10</PERIOD>
</TOPIC>
```

Figure 1: Opinion question fields for sample topic N03

Table 2: Five subtasks for NTCIR-8 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 Clause
Opinion Holders	String, multiple	No	Opinion Clause
Opinion Targets	String, multiple	No	Opinion Clause

Table 3: Test collection size at NTCIR-8 MOAT

Language	Topics (Opinion Questions)			Documents			Sentences			Opinion Clauses		
	Sum	Sample	Test	Sum	Sample	Test	Sum	Sample	Test	Sum	Sample	Test
T-Chinese	21	1	20	787	12	775	9,684	160	9,524	N/A	N/A	N/A
Japanese	21	1	20	178	8	170	6,992	322	6,670	7,894	341	7,553
English	21	1	20	150	12	138	6,564	399	6,165	6,723	500	6,223
S-Chinese	20	1	19	410	25	385	4,720	228	4,492	4,744	232	4,512

Table 4: NTCIR-8 MOAT question lists

ID	Opinion Question	Notes
N01	What negative prospects were discussed about the Euro when it was introduced in January of 2002?	not used at SC side
(N03)	(What reasons were discussed about Bomb Terror in Bali Island in October, 2002?)	Sample
N04	What reasons have been given for the Space Shuttle Columbia accident in February, 2002?	
N05	What negative comments were discussed about Bush's decision to start Iraq war in March, 2003?	
N06	What negative prospects and opinions were discussed about SARS which started spreading in March, 2003?	
N07	What reasons are given for the blackout around North America in August, 2003?	
N08	What reasons and background information was discussed about the terrorist train bombing that happened in Madrid in March, 2004?	
N11	Why did supporters want to elect George W. Bush in the November 2004 American Presidential Election?	
N13	What positive comments were discussed to help the victims from earthquake and tsunami in Sumatera, Indonesia in December, 2004?	
N14	What objections are given for the US opposition to the Kyoto Protocol that was enacted in February 2005?	
N16	What reasons have been given for the anti-Japanese demonstrations that took place in April, 2005 in Peking and Shanghai in China?	
N17	In July 2005 there were terrorist bombings in London. What reasons and background were given, and what controversies were discussed?	
N18	What actions by President George Bush were criticized in response to Hurricane Katrina's August 2005 landing?	
N20	What negative opinions and discussion happened about the Bird Flu that started spreading in October, 2005?	
N24	Identify opinions that indicate that Arnold Schwarzenegger is a bad choice to be elected the new governor of California in the October 2003 election.	
N26	Find positive opinions about the reaction of Nuclear and Industrial Safety Agency officials to the Mihama nuclear powerplant accident in August 2004.	
N27	What were the advantages and disadvantages of the direct flight between Taiwan and Mainland China commercially?	
N32	What are good and bad approaches to losing weight?	
N36	What are complaints about XIX Olympic Winter Games that were held in and around Salt Lake City, Utah, United States in 2002?	
N39	What are the comments about China's first manned space flight which happened successfully in October 2003?	
N41	What negative comments were discussed when in April 2004 CBS made public pictures showing cruel U.S. military abuse of Iraqi prisoners of war?	

Table 5: Opinion percentage in NTCIR-8 MOAT test collection in English and Japanese

Topic	English						Japanese					
	Opinionated	Relevant (of Opinionated)	Answer (of Opinionated)	Polarity			Opinionated	Relevant (of Opinionated)	Answer (of Opinionated)	Polarity		
				POS	NEG	NEU				POS	NEG	NEU
N01	20.2	55	8.8	10.7	69.3	20	32.4	79.2	47.2	14.8	27.9	57.4
N03	25.2	83.7	14.6	19.7	45.1	35.2	25.2	59.3	21	14.9	33.8	51.4
N04	13.1	83	5.7	46.2	42.3	11.5	19.4	64.4	24.4	9.7	29.2	61.1
N05	21.4	92.9	25	15.4	73.1	11.5	35.6	57.6	23.1	6	32.1	62
N06	10	76.9	15.4	30.3	48.5	21.2	27	87.1	49.2	6.3	40.2	53.6
N07	11.5	91.5	53.2	17.1	58.5	24.4	29.5	40.3	37.5	1.6	20.3	78.1
N08	9.2	78.9	21.1	26.3	47.4	26.3	35.9	34.9	27.1	2.8	25.2	72
N11	20.7	76.7	9.3	48.6	51.4	0	35.3	45.3	19.5	18.2	20.4	61.3
N13	8.9	60.5	34.2	59.3	37	3.7	28.1	32.5	10.4	22	23.7	54.2
N14	21.3	95.6	26.7	13	73.9	13	34.4	28.6	12.7	5.7	39.6	54.7
N16	21.5	95.9	30.6	14	75.4	10.5	43.7	40.4	8	4.4	18	77.6
N17	15.8	92.5	55.2	22.6	54.8	22.6	30.6	50	18.8	2.7	15.5	81.8
N18	7.6	92.9	7.1	41.4	51.7	6.9	37.7	47.4	11.2	8.2	37.7	54.1
N20	12	97.6	64.3	14.3	71.4	14.3	37.5	40.2	22.5	3.7	28	68.2
N24	22.9	61.1	14.8	34.6	61.5	3.8	35.3	58.3	6.3	13	28.7	58.3
N26	17.9	84.2	0	25	58.3	16.7	33.3	25.4	5.1	4.7	21.9	73.4
N27	25.4	16.7	11.1	23.5	52.9	23.5	51.6	24.5	8.2	17.3	7.7	75
N32	12.3	91.7	66.7	30.6	62.9	6.5	27	66.4	40	10.6	19.5	69.9
N36	18	40.7	18.6	44.2	38.4	17.4	38.5	35	12	20.7	18.5	60.7
N39	21.2	79.2	37.5	60.9	21.7	17.4	31.4	28.1	25	5.1	9.1	85.9
N41	22.5	97.5	80	2.7	94.6	2.7	34.6	43.9	13.4	4	26.7	69.3
Macro Avg.	17.1	78.3	28.6	28.6	56.7	14.7	33.5	47.1	21.1	9.4	24.9	65.7
Micro Avg.	16.1	78.2	27.8	27.6	56.8	15.6	32.9	48.5	20.9	8.7	24.7	66.6

Table 6: Opinion percentage in NTCIR-8 MOAT test collection in Traditional/Simplified Chinese

Topic	Traditional Chinese						Simplified Chinese					
	Opinionated	Relevant (of Opinionated)	Answer (of Opinionated)	Polarity			Opinionated	Relevant (of Opinionated)	Answer (of Opinionated)	Polarity		
				POS	NEG	NEU				POS	NEG	NEU
N01	37.1	95.8	3.1	67.6	9.9	22.5	–					
N03	18.8	83.3	26.7	0	59.3	40.7	31	100	100	29.4	11.8	58.8
N04	22.9	63	21	20	27.1	52.9	16.9	100	100	14.6	9.8	75.6
N05	55.6	94.2	3.1	12.1	54.9	33	19.9	100	100	23.4	66	10.6
N06	20.3	98.8	14	3.8	26.9	69.2	22.7	94.7	94.7	22.6	59.7	17.7
N07	29.4	96.6	29.9	30.4	43.1	26.5	9.5	100	100	16.7	33.3	50
N08	31.8	100	32.4	9.6	57.7	32.7	12.8	100	100	3.2	41.9	54.8
N11	49.4	93.2	2.7	7.7	27.5	64.8	14.6	98	98	57.1	28.6	14.3
N13	15	66	4.3	26.5	61.8	11.8	19.6	100	100	62.9	14.3	22.9
N14	56.5	93.6	6.8	19	55	26	15.9	100	100	55.6	22.2	22.2
N16	19.4	61.1	11.1	0	35	65	20.6	93.8	92.5	20	12	68
N17	34.1	54.6	10	6.7	24	69.3	27.4	100	100	8.9	8.9	82.1
N18	29.7	62.9	14.5	14	86	0	24.5	100	100	2.8	25	72.2
N20	18.8	86.8	39.6	2	90	8	20.7	100	97.2	6.5	32.3	61.3
N24	21.9	61.3	2.7	15.6	6.3	78.1	20.5	100	100	21.4	21.4	57.1
N26	21	76.9	0	0	66.7	33.3	12.2	100	100	0	30	70
N27	57.6	69.4	3.9	36	27.6	36.4	26.9	98.9	98.9	52.8	12.5	34.7
N32	35.3	100	96.1	5.1	22.6	72.3	15.5	100	100	36.4	15.2	48.5
N36	20.3	100	20	33.3	42.6	24.1	21	100	100	18.2	36.4	45.5
N39	19	97.5	43.3	57.8	25	17.2	17.7	100	100	94.8	0	5.2
N41	47.7	100	17.8	9.9	77.5	12.7	14.9	96.4	94.6	0	80.4	19.6
Macro Avg.	31.5	83.6	19.2	18	44.1	37.9	19.2	99.1	98.8	27.4	28.1	44.6
Micro Avg.	34.9	86.6	18.9	20.4	38.8	41.2	19.6	98.6	98.3	30.3	27.4	42.3

Table 7: κ coefficient at NTCIR-8 MOAT

Attribute	English	Japanese	Chinese	
			Simplified	Traditional
Opinionated	0.7309	0.7174	0.9722	0.4568
Polarity	0.7069	0.6330	0.8901	0.3521
Relevance	0.6907	0.6199	0.9716	0.4003
Answer	0.6607	0.5580	0.9712	0.2901

1. By using an online opinion annotation tool, the assessment manager and assessors shared the annotation data when the annotation was ongoing. The annotation tool output the XML & TSV formats, and we could share several scripts across languages based on the common formats. The annotation tool is shown in Figure 2.

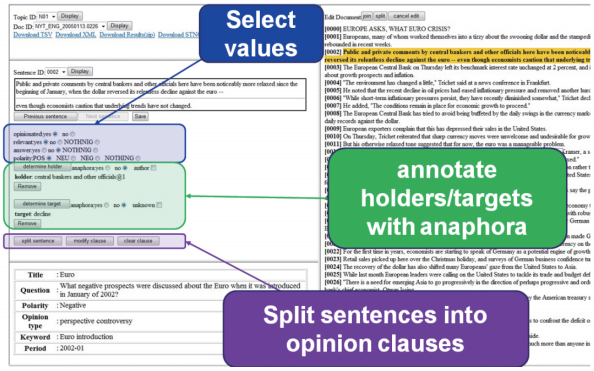

Figure 2: Online annotation tool

2. We provide a strong definition on opinion annotation from the past two MOAT experiences. The details for the opinion annotation scheme are described in Appendix A.

3. One sample topic (N03) was used for the inter-coder session to improve the agreement between assessors.

Based on these, the κ coefficient between assessors drastically improved from the past NTCIR MOAT, as shown in Table 7.

2.4 Evaluation Metrics

Results for precision, recall, and F-measure will be presented for opinion detection, and for sentence relevance, polarity, opinion holders, opinion targets, and cross-lingual subtasks for those participants that elected to submit results for those optional portions.

2.4.1 *Opinionated sentence*

For opinionated sentence evaluation, we took the same approach across all languages. We prepared the correct answer set of sentences that two of two assessors agreed were opinionated. Then, the precision, recall, and F-measure are defined as follows:

$$\begin{aligned}
\text{Precision(P)} &= \frac{\#\text{system_correct(opn = Y)}}{\#\text{system_proposed(opn = Y)}}. \\
\text{Recall(R)} &= \frac{\#\text{system_correct(opn = Y)}}{\#\text{assessors_agreeed(opn = Y)}}. \\
\text{F_measure(F)} &= \frac{2 \times \text{P} \times \text{R}}{\text{P} + \text{R}}.
\end{aligned}$$

2.4.2 *Relevance*

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

$$\begin{aligned}
\text{Precision(P)} &= \frac{\#\begin{pmatrix} \textit{system_correct}(rel = Y)\& \\ \textit{system_proposed}(opn = Y) \end{pmatrix}}{\#\begin{pmatrix} \textit{system_proposed}(rel = Y)\& \\ \textit{system_proposed}(opn = Y)\& \\ \textit{assessors_agreed}(opn = Y) \end{pmatrix}}. \\
\text{Recall(R)} &= \frac{\#\begin{pmatrix} \textit{system_correct}(rel = Y)\& \\ \textit{system_proposed}(opn = Y) \end{pmatrix}}{\#\text{assessors_agreed(rel = Y)}}. \\
\text{F_measure(F)} &= \frac{2 \times \text{P} \times \text{R}}{\text{P} + \text{R}}.
\end{aligned}$$

The reason that the precision was computed as above was 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 the opinionated sentences.

- The numerators of the precision and recall were counted for the opinionated sentences the system proposed, to exclude the non-opinionated cases results.
- The denominator of the precision was counted for the opinionated sentences that the assessors agreed upon. With the pure precision, the participants who submitted the results for all sentences will be at a disadvantage, because their correct answers for the non-opinionated sentences are not counted with the former reason. Based on this discussion, we decided to define the precision as described here.

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

2.4.3 *Polarity*

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

For the agreement estimation between assessors, we implemented the YS method (= to use in the answer set of polarity as POS/NEG/NEU sets that the assessors agreed with) and DKE method (= weigh the answer set according to the number of agreed assessors) in the evaluation script, that were defined in the NTCIR-6 workshop [8]. We set the YS method as a default function and the DKE method as an

optional function. We also implemented the LWK method (= majority voting) as another optional function.

2.4.4 Opinion Holder and Target

The Opinion Holder and Target evaluation 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 the system opinion holders or targets for a given sentence are matched using the 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 is asked to determine if the system answer matches the ones of the opinion holders or targets in the equivalence class for the sentence. If there is a 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 the opinion holder and target equivalence classes is created by adding the 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 the opinion holders.

The 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.

In the Traditional Chinese side, a partially correct judgment was conducted and weighted as 1/2 a correct judgment. Non-agreed upon evaluation results (explained in Section 2.4.5) were also provided to the TC side participants, but this is a minor change.

2.4.5 Cross-lingual Opinion Q&A Subtask

We evaluated the cross-lingual subtask as the precision, recall, and f-value to estimate the answer extraction capability for opinion questions. The answer set from the human assessments were defined as the sentence ID group from the four languages based on the following conditions: opinionatedness:"Y"; answerness:"Y"; and the polarity is matched with the opinion question description if specified within it.

In NTCIR-8 MOAT, human assessment was conducted by two assessors in each language. In the cross-lingual subtask, agreement should be conducted based on the opinionatedness, polarity, and answerness. The number of agreements seemed slightly low compared to the mono-lingual subtasks, and two results are provided in this subtask.

- Agreed upon results: evaluation based on the intersection (agreed) annotation between two assessors.
- Non-agreed upon results: evaluation based on the union annotation from two assessors.

The term "agreed" means the agreement between assessors, not the agreement between the system and the annotation.

2.5 Participants

In NTCIR-8 MOAT, 16 teams participated in the task with 56 submission runs. The team ID is listed in Table 8. Half the teams submitted the results in more than two language related tasks. We maximally accepted three runs for each subtask.

Table 8: Participants list at NTCIR-8 MOAT (alphabetical order)

TeamID	Affiliation	# of Submission Runs				
		EN	SC	TC	JA	CL
BUPT	Beijing University of Posts and Telecommunications		2			
CTL	City University of Hong Kong		1	1		
CityUHK	City University of Hong Kong			3		
cyut	Chaoyang University of Technology			3		
IISR	Yuan Ze University				3	
KAIST	Korea Advanced Institute of Science and Technology	2				
KLELAB	Pohang University of Science and Technology	3		3		
NECLE	NEC Laboratories China	2	2			
NTU	National Taiwan University	2		2		2
OPAL	University of Alicante	3				3
PKUTM	Peking University		3			
PolyU	The Hong Kong Polytechnic University	3	1			
SICS	Swedish Institute of Computer Science	1				
TUT	Toyohashi University of Technology				3	
UNINE	University of Neuchâtel	2		1	1	
WIA	The Chinese University of Hong Kong		2	2		
# of teams		8	6	7	3	2
# of runs		18	11	15	7	5

3. EVALUATION RESULTS

3.1 Opinion Analysis Subtasks Results

Tables 9-12 lists the evaluation results of the opinionated, relevance, polarity, opinion holder, and opinion target analysis for English, Simplified/Traditional Chinese, and Japanese[5]. In order to determine whether there were any statistically significant differences between the runs, we performed an analysis of variance (ANOVA) on the F-values in the opinoinated sentence judgment subtask, followed by a multiple comparison test according to Turkey's significant difference criterion to determine which pairs of runs are significantly different at the 95% confidence level, as listed in Tables 9-12. Note that we ranked all the runs according to the macro-averaged values of the F-values over all the topics.

[5]Note that the recall/F-values evaluation is not provided in the TC side. To have consistency with the holder/target evaluation strategy in other languages, we provided the agreed upon evaluation results as the official evaluation in this paper.

Table 9: English opinionated/relevance/polarity/holder/target evaluation results

Group	Run ID	Opinionated			Relevance			Polarity			Holder			Target			Significance							
		P	R	F	P	R	F	P	R	F	P	R	F	P	R	F								
UNINE	1	29.44	62.84	40.1	83.68	32.74	47.07	50.29	29.58	37.25				–			A							
NECLC	bsf	26.5	58.74	36.52						–							A	B						
NECLC	bs1	21.79	78.84	34.14						–							A	B	C					
NECLC	bs0	25.85	58.11	35.78						–							A	B	C	D				
UNINE	2	19.32	81.79	31.26	84.39	36.01	50.48	48.35	37.8	42.43				–				B	C	D				
KLELAB	3	19.68	68	30.53				–			43.4	27.8	33.9		–			B	C	D	E			
KAIST	2	19	65.26	29.43				–				–		23.1	34.6	27.7			C	D	E			
KLELAB	2	17.9	82	29.39				–			41.1	31.7	35.8		–				C	D	E			
KAIST	1	18.88	64.84	29.24				–				–		23.1	34.6	27.7			C	D	E			
NTU	2	17.02	93.68	28.81	77.75	94.02	85.11	49.22	46.23	47.68				–						D	E	F		
NTU	1	16.8	95.47	28.57	77.95	96.06	86.06	52.17	49.94	51.03				–						D	E	F		
KLELAB	1	16.82	95.37	28.6				–			39.7	35.6	37.5		–					D	E	F		
OPAL	1	17.99	45.16	25.73	82.05	47.83	60.43	38.13	12.82	19.19				–							E	F	G	
OPAL	2	19.44	44	26.97	82.61	5.16	9.71	50.93	12.26	19.76				–								F	G	
OPAL	3	19.44	44	26.97	76.32	3.94	7.49					–										F	G	
PolyU	1	24.58	21.47	22.92						–													G	
SICS	1	13.87	31.37	19.24						–													G	H
PolyU	2	19.51	13.37	15.87						–														H

Table 10: Simplified Chinese opinionated/relevance/polarity/holder/target evaluation results

Group	Run ID	Opinionated			Relevance			Polarity			Holder			Target			Significance				
		P	R	F	P	R	F	P	R	F	P	R	F	P	R	F					
PKUTM	2	41.34	83.35	55.27				–			89.6	73.2	80.5	55.4	43.1	48.5	A				
PKUTM	1	37.21	83.7	51.52				–			89.2	73.6	80.6	55.0	43.4	48.5	A	B			
PKUTM	3	34.05	90.62	49.5				–			87.7	79.2	83.2	54.8	47.3	50.8	A	B			
BUPT	2	35.02	87.81	50.07	97.96	51.55	67.55	58.13	31.13	40.55	89.4	49.6	63.8		–			B			
CTL	1	36.46	78.9	49.87		–		67.39	52.9	59.27	95.3	73.2	82.8	73.5	56.4	63.8		B	C		
BUPT	1	38.43	67.53	48.98	98.23	39.64	56.49	60.41	27.18	37.49	92.9	47.3	62.7		–			B	C		
WIA	1	29.2	95.9	44.77	98.2	58.33	73.19	50.72	46.57	48.56	85.5	76.8	80.9	36.9	33.0	34.9			C	D	
WIA	2	29.2	95.9	44.77	98.22	59.17	73.85	51.18	45.91	48.4	85.3	74.5	79.5	37.0	32.2	34.4			C	D	
NECLC	bsf	28.66	74.56	41.4						–										D	
NECLC	bs0	28.53	72.22	40.9						–										D	
NECLC	bs1	24.39	92.38	38.59						–										D	
PolyU	1	21.52	59.91	31.67		–		68.95	31.93	43.65			–								E

3.2 Cross-lingual Opinion Q&A Subtask Results

Table 13 lists the evaluation results of the cross-lingual subtask to evaluate the precision, recall, and F-values for the answer opinion extraction. We performed an analysis of variance (ANOVA) on the F-values in the cross-lingual answer opinion extraction subtask each for agreed and non-agreed standards between two assessments, followed by a multiple comparison test according to Turkey's significant difference criterion to determine which pairs of runs are significantly different at the 95% confidence level.

4. DISCUSSION

From the experience in NTCIR-8 MOAT, we discuss effective approach and evaluation strategy towards better understanding for sentiment analysis technology.

4.1 Effective Approach

The participants who attained better evaluation results in NTCIR-8 MOAT based on the smart technologies: (1) feature filtering technology, (2) effective machine learning methodology, and (3) lexicon resources. The approaches taken in the top team of opinion judgment subtask are summarized in Table 14.

As shown in Table 14, UNINE combined corpus-based statistical significance filtering approach (Z-score) with minimizing nautral language exception and ambiguous errors using SentiWordNet. PKUTM used In-house opinion word list based on HowNet, NTU lexicon, and Jun Li's opinion word list. CTL and PKUTM set several types of features: punctuation, word and entities, lexical collocation, and subjective clues at sentence/paragraph/document levels. CityUHK combined NTU, LCPW, LCNW, and CityU's in-house polar word/phrase list into polar item lexicon and also prepared reporting verbs, and filtered the noisy terms out using training data to adjust them to the news domain. They also implemented voting ensemble scheme based on SVM, ME, and supervised lexicon.

On the other hand, NECLC team took the self learning approach based on ME model, while BUPT team tried semi-supervised approach with TSVM. NECLC achieved satisfactory results in English, but not so highly ranked in Traditional Chinese.

For opinion holder & target identification, WIA proposed ranking model based on the dependency parser and semantic role labeling information. KAIST proposed to use document theme to identify the opinion target. CTL, PKUTM and CityUHK developed heuristic rules, and the latter approach was based on dependency parsing. PKUTM, KLELAB, and BUPT used CRF model for opinion holder identification. CTL incorporates sentence similarity approach simplified with Chinese synonym dictionary (TongYiCi CiLin), etc. into feature based approach, and achieved satisfactory results in SC&TC languages. This similarity approach is mainly applied to opinion holder & target identification subtasks.

Cross-lingual opinion question & answering is still a challenging problem. OPAL team took the triangulation approach to create Traditional Chinese resource from English & Spanish resources. Their results suggested that the extensive filtering approach should increase the precision, but decrease the recall a lot, and the language model based approach might be necessary to better account for language variability.

4.2 Insights from Evaluations

The difference of evaluation results by each language depend on the difference of opinion percentage listed in Tables 5 and 6. In Traditional Chinese and Japanese, evaluation results tend to be higher values than English and Simplified Chinese, just as the opinion percentage is.

The opinion definitions tend to differ according to the assessors, so the strong definition must be needed for consistent annotation. In the inconsistent case, the evaluation results should be unstable. In such a case, evaluation measure should be fixed according to the task goal. For example, loose agreement evaluation measure will be effective if the task goal is to extract the relevant opinion as much as possible.

5. CONCLUSION

In this paper, we presented our experiences in Multilingual Opinion Analysis Task (MOAT) in NTCIR workshop. In this task, many participants proposed and challenged a new opinion extraction technology. The smart feature filter and machine learning technologies with rich lexicon should be successful keys for opinion analysis. Towards cross-lingual opinion analysis, we should be careful for the language variability to create resources.

6. ACKNOWLEDGMENTS

We greatly appreciate the efforts of all the participants in the Multilingual Opinion Analysis Task at the Eighth NTCIR Workshop. We also appreciate NTCIR secretaries Ms. Miho Sugimoto, Ms. Ruiko Homma, and Ms. Fumiko Koiwahara deeply for their sincere supports to the task management.

7. REFERENCES

[1] J. Blitzer, M. Dredze, and F. Pereira. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL), pages 440–447, 2007.

[2] 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.

[3] H. T. Dang. Overview of the TAC 2008 Opinion Question Answering and Summarization Tasks. In Text Analysis Conference (TAC 2008) Workshop Notebook Papers, 2008. [cited 2010-04-01]. Available from: <http://www.nist.gov/tac/tracks/2008/index.html>.

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

Table 11: Traditional Chinese opinionated/relevance/polarity/holder/target evaluation results

Group	Run ID	Opinionated			Relevance			Polarity			Holder	Target	Significance			
		P	R	F	P	R	F	P	R	F	P	P				
CTL	1	65.14	68.79	66.92	–			76.5	53.06	62.66	84.9	54.4	A			
CityUHK	2	56.39	85.71	68.03	–			44.14	38.5	41.13	72.1	48.5	A			
CityUHK	1	50.92	91.98	65.55	–			45.17	41.93	43.49	70.0	25.9	A			
CityUHK	3	50.92	91.98	65.55	–			45.17	41.93	43.49	68.1	23.3	A			
WIA	1	53.41	83.68	65.2	89.44	58.04	70.4	50.68	41.14	45.41	62.1	28.3	A			
WIA	2	53.41	83.68	65.2	89.46	58.74	70.92	50.66	40.45	44.98	60.5	24.6	A			
KLELAB	3	44.51	87.92	59.1	–									B		
KLELAB	1	41.98	94.94	58.22	–						29.6*	–		B		
KLELAB	2	41.98	94.94	58.22	–						26.2*	–		B		
NTU	2	41.85	92.22	57.57	86.44	92.06	89.16	44.35	41.19	42.71	–			B		
NTU	1	41.41	93.82	57.46	86.35	93.57	89.82	45.57	42.83	44.16	–			B		
cyut	1	42.71	87.74	57.45	–			40.49	35.6	37.89	–			B		
cyut	2	41.13	82.41	54.87	–			31.26	25.95	28.36	–			B	C	
UNINE	1	52.37	48.47	50.34	86.2	48.25	61.87	47.01	23.27	31.13	–				C	D
cyut	3	47.55	43.99	45.7	–			36.68	16.19	22.46	–					D

*: late submission (unofficial run)

Table 12: Japanese opinionated/relevance/polarity evaluation results

Group	Run ID	Opinionated			Relevance			Polarity			Significance			
		P	R	F	P	R	F	P	R	F				
TUT	1	66.83	61.73	64.18	49.74	26.83	34.86	54.13	27.43	36.41	A			
TUT	3	68.04	55.32	61.02	56.89	30.21	39.46	65.29	29.95	41.06	A	B		
IISR	3	67.86	51.53	58.58	–							B	C	
TUT	2	68.69	50.99	58.53	55.84	26.92	36.33	64.57	26.8	37.88		B	C	
IISR	1	67.3	49.86	57.28	–							B	C	
IISR	2	67.74	47.65	55.95	–								C	
UNINE	1	63.3	28.56	39.36	48.18	28.61	35.9	42.8	8.95	14.8				D

Table 13: Cross-lingual answer opinion extraction evaluation results

Group	Run ID	lang	evaluation type	Answer Extraction			Significance		
				P	R	F			
NTU	2	TC	Agree	7.8	38.46	9.38	–		
NTU	1	EN	Agree	5.63	45.81	7.65	–		
NTU	2	EN	Agree	4.87	39.99	7.35	–		
OPAL	1	TC	Agree	3.54	56.23	6.34	–		
OPAL	3	TC	Agree	3.42	72.13	6.32	–		
NTU	1	TC	Agree	5.54	39.07	5.99	–		
OPAL	2	TC	Agree	3.35	42.75	5.78	–		
OPAL	3	TC	Non-Agree	15.02	77.68	23.55	A		
NTU	2	TC	Non-Agree	24.59	41.11	23.41	A		
OPAL	1	TC	Non-Agree	14.62	60.47	21.36	A		
OPAL	2	TC	Non-Agree	14.64	49.73	19.57	A		
NTU	1	TC	Non-Agree	20.09	41.19	19.26	A	B	
NTU	2	EN	Non-Agree	8.46	37.58	11.35		B	C
NTU	1	EN	Non-Agree	8.61	41.55	10.86			C

Table 14: The approaches taken in the top team of opinion judgment subtask

TeamID	Lang	Feature Filtering	Machine Learining	Lexicon Resource
UNINE	EN	Z score	logistic regression	SentiWordNet
PKUTM	SC	Iterative classifier	SVM (better than NB, ME, DT)	In House, NTU, and Jun Li's lexicon
CityUHK	TC	Supervised Lexicon	Ensemble	NTUSD, LCPW, LCNW, CPWP, SKPI

[5] National Institute of Standards and Technology. TREC (Text Retrieval Conference) 2006-2009: BLOG Track [online]. In TREC-BLOG Information Retrieval Wiki, 2009. [cited 2010-04-01]. 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, 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, pages 265–278, NII, Japan, May 2007.

[9] Y. Seki, D. K. Evans, L. W. Ku, L. Sun, H. H. Chen, and N. Kando. Overview of Multilingual Opinion Analysis Task at NTCIR-7. In Proc. of the Seventh NTCIR Workshop Meeting on Evaluation of Information Access Technologies, pages 185–203, NII, Japan, December 2008.

[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, San Jose, California, March 2009. [cited 2008-10-10]. Available from: <http://www.icwsm.org/2009/index.shtml>.

[12] X. Wan. Co-Training for Cross-Lingual Sentiment Classification. In Proceedings of the 47th Annual Meeting of the Association of Computational Linguistics (ACL), pages 235–243, Suntec, Singapore, 2009.

[13] 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.

[14] J. M. Wiebe, T. Wilson, R. F. Bruce, M. Bell, and M. Martin. Learning Subjective Language. Computational Linguistics, 30(3):277–308, 2004.

[15] 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.

APPENDIX

A. OPINION ANNOTATION SCHEME

A.1 Opinion Annotation

In [13], they annotate three types of private state expressions:

1. explicit mention of private states
2. speech events expressing private states
3. expressive subjective elements

In our task, we do not need to distinguish between the three types of private state expressions, but the guidelines for each type are a good indicator for what should be annotated. Examples of the three types of expressions are as follows.

1. Explicit mentions of private states by a person, nation, or organization:
   - Psychologists argue that teenagers are not old enough to make permanent decisions about changing their bodies, because their sexual identity is still in flux.
   - The U.S. fears a spill-over.
2. Speech events expressing private states by an agent:
   - Ito said the government must concentrate now on reviving its fragile financial health so as to restore international confidence in Japan.
   - "The report is full of absurdities," Xirao-Nima said.

   In this work, the term speech event is used to refer to any speaking or writing event. A speech event has a writer or speaker as well as a target, which is whatever is written or said.
3. Expressive subjective elements:
   - Japan must seem to be a country full of antiquated rules.
   - The time has come, gentlemen, for Sharon, the assassin, to realize that injustice cannot last long.

   The private states in these sentences are expressed entirely by the words and the style of language that is used. In the latter example, although the writer does not explicitly say that he hates Sharon, his choice of words clearly demonstrates a negative attitude toward him. As used in these sentences, the phrases "The time has come," "gentlemen," "the assassin," and "injustice cannot last long," are all expressive subjective elements. Expressive subjective elements are used by people to express their frustration, anger, wonder, positive sentiment, mirth, etc., without explicitly stating that they are frustrated, angry, etc. Sarcasm and irony often involve expressive subjective elements.

   In addition, note that there are many expressions like "the poor mouse function often braked" or "the price of this famous musical is extremely expensive." We regard them as expressive subjective elements, and therefore, as opinionated.

A.2 Insubstantial and Hypothetical cases

In general, the following items should not be considered opinions:

- Cases in which the subjective language is considered "insubstantial" by Wiebe et al. [13] should not be considered opinionated. (See section 3.6 in their paper.)
- Statements that might be controversial in some context, but that are expressed as fact within the context of the document. For example, in a science article a statement like "Humans evolved from lower-order life forms over millions of years." should not be considered an opinion. In an article about religious views on science, where the topic of evolution is presented as disputed and controversial, it may be considered an opinion.

- Statements that are generally accepted to be true and "common sense" should not be considered opinions.
- Expressions of general plans and schedules should not be considered opinionated. For example, "The building will be completed by the deadline" should not be labelled opinionated, unless it is a disputed topic in the context of the article (someone else counters that the building will not be completed by the deadline.)
- Indirect hearsay or rumors from anonymous people, which could not specify the opinion holder, should not be considered opinionated.
- Declaration or assertion by an organization should not be considered opinionated.
- The other people, health, or weather descriptions should not be considered opinionated.
- The advertising copy or a general phrase used in daily life should not be considered opinionated.

A.3 Opinion Holder Annotation

- Identify the agent (person, object, government, group, etc.) that is expressing the subjective content. The author or the elements in the sentences should be selected as holders.
- The opinion holder is judged by using the following three opinion holder types:
  1. Person or organization that expresses private states
  2. Speech events agents
  3. Agent that implicitly expresses subjective information
- If the opinion holder element appeared in the previous text, that is, the antecedent element is referenced for the opinion holder in the sentence; you should check the anaphora checkbox and select the antecedent and the anaphora expression.
- The anaphora expression is not only a demonstrative pronoun, but also a zero pronoun.
- You should attach the affiliation or country name directly modifying the opinion holder to the opinion holder element.
- You should not attach the relative or modifying expression to the affiliation or country name.
- The quantifiers phrase modifying opinion holder should be attached.
- When the opinion holder is different according to each clause, the opinion holder of the main clause should be the holder of the sentence.

A.4 Opinion Target Annotation

- Identify the string from the text that corresponds to the object that the opinion is about. The opinion target type should not need to be checked.
- The anaphora information should be checked in the same way as the opinion holder.
- The target element should be checked using the following priority:
  1. The element appeared in the sentence.
  2. The element in the previous sentence.
  3. The element in the text that is valid in the context.
  4. The title/product element that is topicalized in the text.
- The element should not be too lengthy. The element could be edited to be short.
- Words that do not appear in the text should not be used.

A.5 Polarity Annotation

- The polarity was judged based on the opinionated sentences. This relationship was automatically checked in the tool.
- The polarity is determined by taking into account the topics and context considering not on surface terms. Therefore, the same term could be judged as positive or negative according to the context of the usage.

A.6 Split multiple opinion clausal units in one sentence

- If one sentence contains several opinions, you should split it into opinion clausal units. This splitting point must be agreed upon between assessors based on the inter-annotator session.
- The opinion clausal unit is split based on the opinion segmentation points, not on the grammatical segmentation points.
- If one sentence contains several opinions, the polarity of the sentence should be determined based on the polarity of the main clause.