WIA-Opinmine System in NTCIR-8 MOAT Evaluation

Lanjun Zhou
Department of SE&EM
The Chinese University of Hong Kong
Shatin, Hong Kong
ljzhou@se.cuhk.edu.hk

Yunqing Xia
Department of Computer Science and Technology
Tsinghua University
Beijing 100084, China
yqxia@tsinghua.edu.cn

Binyang Li
Department of SE&EM
The Chinese University of Hong Kong
Shatin, Hong Kong
byli@se.cuhk.edu.hk

Kam-Fai Wong
Department of SE&EM
The Chinese University of Hong Kong
Shatin, Hong Kong
kfwong@se.cuhk.edu.hk

ABSTRACT

This paper presents WIA-Opinmine system developed by CUHK_Tsinghua Web Information Analysis (WIA) Virtual Research Center for NTCIR-8 MOAT Task. The system is deemed special due to three facts. Firstly, the system is able to handle Simplified Chinese and Traditional Chinese at the same time. A tool is developed to convert Traditional Chinese into Simplified Chinese before opinion analysis. Secondly, a topic model based algorithm is found effective in relevance judgment. A co-clustering algorithm is incorporated in topic modeling. Thirdly, a ranking method is adopted to rank all holder (A0's) and target (A1's) candidates recognized by a semantic role labeling tool during which topic models for each topic are fully used for judging the importance of all candidates.

The NTCIR8 evaluation results as well as the post-NTCIR8 results show that our system could effectively recognize relevance sentences, opinionated sentences and polarities.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Text analysis

General Terms

Algorithms, Performance, Experimentation

Keywords

NTCIR MOAT, Opinion Mining, Term Extraction, Opinion Unit

1. INTRODUCTION

Opinion Mining (OM) nowadays becomes a very hot research topic. Aiming at identifying and analyzing opinions within texts, OM enhances many NLP applications such as information extraction, information retrieval, Questioning&Answering and text summarization. Generally speaking, research on OM are conducted on three levels, namely, document level [1], sentence level [2] and feature level [3].

Three approaches are adopted for opinion mining. Firstly, lexicon-based methods. Use sentiment lexicons and heuristic rules as major knowledge [4]. This approach typically faces the Out Of Vocabulary (OOV) problems. Secondly, supervised approaches are designed based on machine learning. As it is costly to annotate large amount of data, semi-supervised approaches are introduced to partially solve the problem [5]. High quality sentiment lexicons are still very important in these methods and classifiers are trained by utilizing linguistic features [6, 7, 8]. Thirdly, unsupervised approaches create a sentiment lexicon and use the lexicon to determine sentiment of given document or sentence. A typical unsupervised works are reported by Hatzivassiloglou and Wiebe [9] and Turney [10].

It is in NTCIR6 Opinion Analysis Pilot Task that Asian language opinion tasks are first introduced on Traditional Chinese (TC) and Japanese [11]. In NTCIR7 MOAT Task, Simplified Chinese (SC) is introduced [12]. In NTCIR-8 MOAT Monolingual Task[20], the following subtasks are defined:

(1) Opinionated judgment subtask (required)

(2) Relevance judgment subtask (optional)

(3) Opinion holder detection subtask (optional)

(4) Opinion target detection subtask (optional)

(5) Polarity judgment subtask (optional)

(6) Questioning&Answering subtask (optional)

CUHK_Tsinghua Web Information Analysis (WIA) Virtual Research Center WIA participated in subtasks (1)-(5) on both TC and SC sides.

Training data are crucial to WIA-Opinmine system. So we chose to use corpora in both SC and TC provided in NTCIR6 and NTCIR7. However, the two corpora are not used separately. A tool is developed to convert TC into SC, thus we finally obtained a bigger SC corpora. We argue the conversion is safe because formal news articles in both corpora do not significantly differ from each other in forming sentiment expressions. The special treatment on SC and TC is also reflected in opinion analysis, in which the tool is again used to convert TC text into SC before opinion analysis is conducted. As a result, both SC and TC texts could be processed by WIA-Opinmine system.

In addition, WIA-Opinmine system for NTCIR-8 is different with the system proposed in the previous NTCIR MOAT task [13, 14] in the following ways:

(1) Lexicons used in NTCIR-7 MOAT [14, 15] are refined

(2) A topic model based ranking model is used for relevance judgment instead of using a SVM classifier

(3) Polarity classification is regarded as a two-stage process. The first stage is recognition of opinionated sentences. Then the polarities will be judged.

(4) Holder&target could be identified automatically based on SRL and ranking instead of using rules and patterns.

The rest of this paper is organized as follows. Section 2 gives the overview of WIA-Opinmine system proposed in NTCIR8-MOAT. Section 3 presents the methods for each module of our

system in detail including models for relevance judgment, opinionated sentence identification, polarity classification and holder&target recognition. Section 4 discusses the evaluation result. Section 5 gives the post-NTCIR8 experiments which prove the effectiveness of our proposed methods. Section 6 concludes this paper.

2. WIA-OPINMINE SYSTEM

2.1 Architecture and Workflow

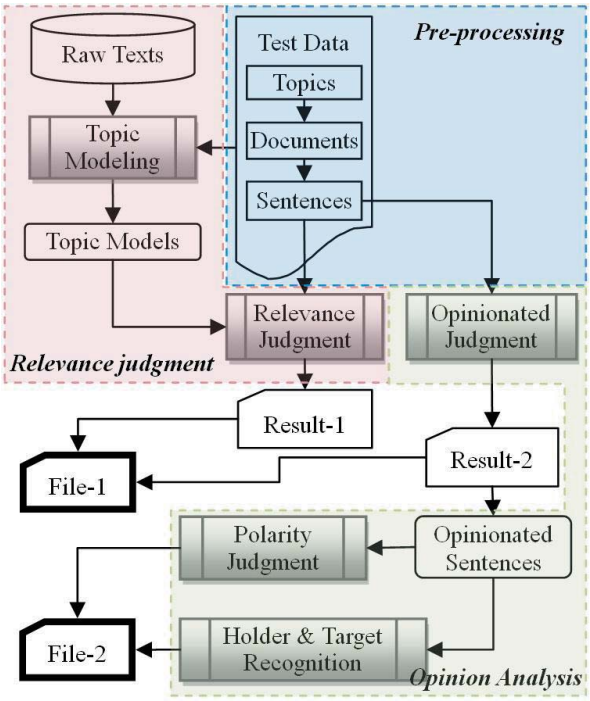

Figure 1. System architecture

It is shown in Figure 1 that the WIA-Opinmine system is comprised of three modules: (1) Pre-processing module reads all data including training data set, developing data set together with formal run data set and performs word segmentation, POS tagging, named entity recognition, dependency parsing and semantic role labeling. (2) Relevance judgment module builds topic models for each of the topics from formal run data and then ranks all sentences according to the similarity score between sentences and the corresponding topic model. We output top 60% of the ranked sentences of each topic as relevant according to the observation of training set. (3) Opinion analysis module analyzes each input sentences to determine whether it is opinionated and the polarity, holder&target of each non-factual sentences.

The NTCIR-8 MOAT tasks are achieved as follows:

Input: NTCIR MOAT Task formal run data (including STNO files, OTNO files, topic descriptions and raw texts)

Step 1: Building language models for each topic using topic descriptions and indexed raw texts (See section 2.2 for details).

Step 2: Performing word segmentation, POS tagging, named entity recognition, dependency parsing for each of the sentences in OTNO files and STNO files.

Step 3: Ranking all sentences according to their similarity score. The score is calculated between topic model and all sentences provided by STNO files. Top 60% sentences in each topic are marked as "relevant sentences"

Step 4: Using an opinionated classifier to judge all sentences provided by OTNO files. Opinionated sentences are preserved for polarity judgment and holder&target recognition.

Step 5: Judging the polarity of each sentence preserved in step 4 using a SVM classifier.

Step 6: Recognizing holder and target for each sentence in opinionated sentence set using SRL and patterns.

Output:

File-1: Containing results of relevance judgment and opinionated judgment.

File-2: Containing results of polarity and holder&target information for each opinionated sentence.

2.2 Development Data

The development data are necessary in topic modeling. We get full use of the raw texts provided by NTCIR-8 MOAT Task. The texts are news articles from Xinhua (Simplified Chinese, 2002-2005) and UDN (Traditional Chinese, 2002-2005). Information retrieval techniques are applied to find news texts according to a specific query q (see Section 3.1).

2.3 A Refined Opinion Lexicon

We continue using our NTCIR-7 opinion lexicon [14] in NTCIR-8 MOAT evaluations. However, it is observed in our study that the lexicon suffered from two problems:

(1) There were many contextual sentiment words and factual words (non-opinionated) in our lexicon.

(2) Some sentiment words were not generally used for opinions in news texts. i.e., "專家(expert)", "安全(safe)", "和平(peace)", "穩定(stability)".

Table 1. Lexicon list of WIA-Opinmine

Type	Lexicons
Sentiment Words	Positive sentiment words Negative sentiment words Contextual sentiment words
Degree Adverbs	Degree adverbs
Conjunctions	Coordinating conjunctions Subordinating Conjunctions Correlative Conjunctions
Other Words	Opinion indicators Opinion operators Negations

Extra work has been conducted to handle the above two problems. For problem (1), we invited and trained two human annotators to check all our lexicons manually and picked out all opinion words which were contextual dependently. A contextual opinion word w is defined as a word that appears in more than one elements of the sentence type set A.

A typical example of contextual words is "上升(going up)" which could be appear in at least 3 types of sentences. We did an

intersection of the words picked out by two annotators to maintain high precision of our contextual words lexicon.

$$A = \left\{ \begin{matrix} \text{POS\_SEN\_SET, NEG\_SEN\_SET,} \\ \text{NEU\_SEN\_SET, FACTUAL\_SEN\_SET} \end{matrix} \right\}$$

For problem (2), we calculated the *tf* value of all words in our opinion lexicon and remove the top 1.0% ranked words. We removed 426 words and all of the words mentioned in (2) were successfully discarded. Finally, we used the dictionaries shown in Table 1 in our system.

3. TASKS AND SOLUTIONS

3.1 Relevance Judgment

A topic model based algorithm is proposed for relevance sentence judgment in WIA-Opinmine system.

Generally speaking, the problem of relevance judgment at the sentence level could be viewed as retrieving sentences for a given topic. A simple way to solve this problem is to search topic words of each topic in the large scale raw texts (provided by the organizer) and then build topic models from retrieved documents. The sentences could be ranked according to the similarity between sentences and corresponding topic model. The topic model built in this way suffered following problems:

(1) Topic words are too general for some topics. e.g. "歐元 (Euro)" is a very general term. Thus, excessive number of documents would be retrieved.

(2) According to the observation of the training data, although the topic model may perform well in judging relevant documents, it performs poor on the sentence level.

(3) Documents of each topic in the test data could not cover every aspect of information. This implies that we need a more specific topic model for this subtask.

We improve the above method for building topic models by introducing additional information and utilizing coefficient of variation to enhance its performance on sentence level relevance judgment. Intuitively, we could extract topic information from the given documents to build topic models. To do so, a co-clustering based method is first applied to extract initial topic keywords from given documents [16]. For example, 15 terms (with weights) are extracted from the news articles on topic "歐元(Euro)". Query q is defined as:

$$q = term_1 \text{ OR } term_2 \text{ OR } \ldots\ldots \text{ OR } term_k$$

We search q then obtain the top 100 most relevant articles from the indexed development data (See section 2.2). Note that the duplicate articles are removed from the article set. In what follows, the topic model can be built with these articles. One observation is that the *tf* score of some named entities such as "歐洲(Europe)", "美國(USA)" are in the top area of *tf* score ranking list. We think these named entities are meaningless when building sentence-level topic model like "歐元(Euro)". We introduce coefficient of variation to "penalize" words whose occurrences from year 2002 to 2005 are very similar. Terms in our sentence-level topic model are weighted using the following equation:

$$w_i = tf_i \cdot cv_i$$

where cv_i denotes the coefficient of variation of $term_i$ and is defined as follows:

$$cv_i = \frac{1}{avg_i} \cdot \sqrt{\sum_{j=2002}^{2005} (tf_i^j - avg_i)^2 \cdot \frac{1}{4}}$$

where tf_i^j denotes the *tf* value of $term_i$ in year j and avg_i denotes the average *tf* value of $term_i$ within news reports in four years (i.e., from 2002 to 2005). Table 2 shows the top 20 words and their weights.

Table 2. Top 20 terms for the topic model of "歐元"

Original	Original + Coefficient of Variation
歐元, 歐洲, 投資人, 銀行, 貨幣, 投資, 基金, 匯價, 資產, 市場, 美國, 價位, 主管, 升值, 指出, 經濟, 認為, 利率, 建議, 外匯	歐元, 匯價, 經濟, 羅尤, 基金, 升值, 銀行, 投資, 貨幣, 投資人, 價位, 復甦, 存款, 物價, 主管, 債券, 指出, 澳幣, 日圓, 央行

Refer to Table 2, significant improvement could be observed in the top 20 terms of topic model after introducing coefficient of variation in our method. Top 200 terms with their weight are preserved in our model and a simple cosine measure is applied to estimate the similarity between sentence S_i and topic model M.

$$Sim(S_i, M) = \frac{S_i \cdot M}{\|S_i\| \|M\|}$$

An observation of training data is about 60% of all sentences are relevant. Therefore, we simply output top ranked 60% sentences of each topic as relevant sentences in RUN-1. We found that many sentences with similarity scores over 0.8 are not belong to the top 60%. These sentences are marked as "relevant" together with the top 60% of sentences in RUN-2.

3.2 Opinionatedness Judgment

Experiments on NTCIR-6 and NTCIR-7 corpus show that our sentiment lexicon achieved 95.1% recall for the opinionated sentences. Further by using of opinion operators and opinion indicators, recall increased to 96.8%. Thus, the features adopted in the opinionated sentence classifier are mainly lexical. To boost the precision of our classifier, we use refined opinion lexicon (See section 2.3) and introduce some bi-gram features.

Table 3. Features adopted in the opinionated sentence classifier

Punctuation level features
The presence of direct quote punctuation "“" and "”" (SC)
The presence of direct quote punctuation "「" and "」" (TC)
The presence of other punctuations: "？" and "！"
Word-Level and entity-level features
The presence of known opinion operators

The percentage of known opinion word in sentence
Presence of a named entity
Presence of pronoun
Presence of known opinion indicators
Presence of known degree adverbs
Presence of known conjunctions
Bi-gram features
Named entities + opinion operators
Pronouns + opinion operators
Nouns or named entities + opinion words
Pronouns + opinion words
Opinion words (adjective) + opinion words(noun)
Degree adverbs + opinion words
Degree adverbs + opinion operators

Note: Opinion word comprises 4 types of words: Positive, Negative, Contextual and Neutral.

The features we adopted in this subtask are partly the same as the WIA-Opinmine in NTCIR-7. Consider Table 3, three types of features are adopted in the classifier.

These features are combined using a RBF kernel and a SVM classifier is trained leading to get a recall of over 80% with tolerable *F-score* on development set.

3.3 Polarity Judgment

Refer to Figure 1, opinionated sentences must be figured out as the input of polarity judgment classifier. Thus the recall of opinionated sentence classifier will directly affect the recall of polarity classifier on the test data. That's why we train our opinionated sentence classifier leading to a high recall. The result of 5-fold cross validation on training data shows that the precision of opinionated sentence classifier is about 60% while maintaining a recall higher than 80%. The result reveals the fact that there are still about 40% factual sentences in the input of polarity classifier.

In addition to the features shown in Table 3, we incorporate features of s-VSM(Sentiment Vector Space Model) [17] to enhance the performance of models only use lexicon and n-gram features. The principles of the s-VSM are listed as follows: (1) Only sentiment-related words are used to produce sentiment features for the s-VSM. (2) The sentiment words are appropriately disambiguated with the neighboring negations and modifiers. (3) Negations and modifiers are included in the s-VSM to reflect the functions of inversing, strengthening and weakening.

Sentiment unit is the appropriate element complying with the above principles. The notation for sentiment lexicon in s-VSM is as follows:

$$L = \{C, N, M\}; C = \{c_i\}, i = 1, ..., I$$
$$N = \{n_j\}, i = 1, ..., J$$
$$M = \{m_l\}, i = 1, ..., L$$

in which L represents the sentiment lexicon, C sentiment word set, N negation set and M modifier set. These words can be automatically extracted from our lexicon and each sentiment word is assigned a sentiment label, namely *strong* (positive and negative sentiment words) or *contextual* (contextual sentiment words) according to our lexical definition.

Given a sentence, denoted as follows,

$$W = \{w_h\}, h = 1, ..., H$$

in which W denotes a set of words that appear in the sentence, the semantic lexicon is in turn used to locate sentiment units denoted as follows:

$$U = \{u_v\} = \{c_{i,v}, n_{j,v}, m_{l,v}\}$$
$$c_{i,v} \in W \cap C;\ n_{j,v} \in W \cap N;\ m_{l,v} \in W \cap M$$

We classify the sentiment units according to occurrence of sentiment words, negations and modifiers. If a sentiment word is mandatory for any sentiment unit, eight kinds of sentiment units are obtained. Let f_{PSW} denote count of positive sentiment words (PSW), f_{NSW} count of negative sentiment words (NSW), f_{NEG} count of negations (NEG) and f_{MOD} count of modifiers (MOD). Eight sentiment features are defined in Table 4.

Table 4. Definition of sentiment features.

f_i	Number of sentiment units satisfying …
f_1	$f_{PSW}>0, f_{NSW}=f_{NEG}=f_{MOD}=0$
f_2	$f_{PSW}=0, f_{NSW}>0, f_{NEG}=f_{MOD}=0$
f_3	$f_{PSW}>0, f_{NSW}=0, f_{NEG}>0, f_{MOD}=0$
f_4	$f_{PSW}=0, f_{NSW}>0, f_{NEG}>0, f_{MOD}=0$
f_5	$f_{PSW}>0, f_{NSW}=0, f_{NEG}=0, f_{MOD}>0$
f_6	$f_{PSW}=0, f_{NSW}>0, f_{NEG}=0, f_{MOD}>0$
f_7	$f_{PSW}>0, f_{NSW}=0, f_{NEG}>0, f_{MOD}>0$
f_8	$f_{PSW}=0, f_{NSW}>0, f_{NEG}>0, f_{MOD}>0$

Note: one sentiment unit contains only one sentiment word. f_{PSW} and f_{NSW} could not be positive at the same time because there is no overlap between PSW and NSW.

The performance of sentiment analysis system greatly degrades when neutral sentences are included in the experiments [18]. For this reason, we decided to extract some patterns as features to boost our classifier on neutral sentence classification. We select all neutral sentences and use PrefixSpan [21] to mine useful patterns while maintaining the sequence of words. Finally, top 20 patterns are chosen.

All features are combined using a RBF kernel and a SVM classifier is trained aiming to get best F-measure on development set.

3.4 Holder&Target Recognition

Different from NTCIR-7 Opinmine system, a totally automatic method is adopted to recognize opinion holders and targets.

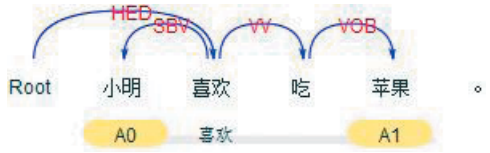

Figure 2. Dependency Parsing and SRL result for the sentence "小明喜欢吃苹果 (Xiao Ming likes eating apples)"

Both a dependency parser and a semantic role labeling (SRL)

tool (http://ir.hit.edu.cn/demo/ltp) are incorporated in our system to identify the semantic roles of each chunk based on verbs in a sentence. i.e., the parsing result of sentence "小明喜欢吃苹果 (Xiao Ming likes eating apples)" is shown in Figure 2. Refer to Figure 2, "小明(Xiao Ming)" is recognized as a A0, "苹果 (apples)" is recognized as a A1 and "喜欢(like)" is the sentiment verb connect A0 and A1. The holder and target of this sentence is "小明(Xiao Ming)" and "苹果(apples)" because there is only one candidate for A0's and A1's, respectively.

The meanings of A0's, A1's are different from one verb to another. i.e, the definition of A0 for "喜欢(like)" is "people described" and A1 is "entity A0 likes"; the definition of A0 for "会见(meet)" is "meeter" and A1 is "person met". In most conditions, A0 represents the subject of a verb and A1 represents the object. Then, we assume that in a sentence, holder should be one of the A0's and target should be one of the A1's. The problem becomes how to choose proper A0 and A1 for a sentence when more than one A0 or A1 exists. We propose a ranking method by using topic model (see section 3.1) and the position information. Given a sentence S with N words, we estimate the weight of argument A which belongs to verb V using the following equation:

$$score(A) = a_0 \cdot \frac{A \cdot M}{\|A\|\|M\|} + a_1 \cdot \log\frac{N}{ap} + a_2 \cdot \log\frac{N}{vp}$$

in which ap denotes the position of A, vp denotes the position of V, M denotes the topic model. We rank all noun and named entities if SRL could not find any A0's or A1's. In our experiment, (a_0, a_1, a_2) is estimated using linear regression with ordinary least square (OLS) method on training data. Note that the training data of holder&target recognition is only from NTCIR-7. After linear regression, the coefficients (a_0, a_1, a_2) are set to (0.5, 0.1, 0.05), respectively.

4. Experiments

4.1 Evaluation Criteria

Five subtasks, including relevant sentence determination, opinionated sentence judgment, polarity classification, opinion holder and target recognition are evaluated. Among them, relevance sentence judgment, opinionated sentence judgment adopted the same metrics, i.e. Precision (P), Recall (R) and F-measure(F) [12].

$$P = \frac{\#system\_correct}{\#system\_proposed}$$

$$R = \frac{\#system\_correct}{\#gold\_answer}$$

$$F = \frac{2 \cdot P \cdot R}{P + R}$$

For the polarity determination in NTCIR-8, lenient recall-based criteria are adopted. The recall-based precision (R_P), recall-based Recall (R_R) and recall-based F are defined as:

$$R\_P = \frac{\#system\_correct(polarity = POS, NEU, NEG)}{\#system\_proposed(opinionated = Y)}$$

$$R\_P = \frac{\#system\_correct(polarity = POS, NEU, NEG)}{\#gold(opinionated = Y)}$$

$$R\_F = \frac{2 \cdot R\_P \cdot R\_R}{R\_P + R\_R}$$

The evaluation on recognition of opinion holder and opinion target adopts the metric similar to polarity judgment.

Two annotators were induced for labeling each sentence in NTCIR-8 while there were three annotators in NTCIR-6 and NTCIR-7. Accordingly, only the lenient way is adopted in the evaluation of NTCIR-8.

4.2 NTCIR-8 EXPERIMENTS

Relevance Judgment

Performance of WIA-Opinmine in relevance judgment on both TC and SC are given in Table 5.

Table 5. Evaluation result of relevance judgment

RUN-ID		TC	SC
WIA RUN-1	**P**	89.44	97.74
	R	58.04	58.33
	F	70.40	73.19
WIA RUN-2	**P**	89.46	98.22
	R	58.74	59.17
	F	70.92	73.85
Best*	**P**	86.35	97.78
	R	93.56	59.64
	F	89.81	74.09

Best*: The best result in NTCIR-8 MOAT Evaluation

The result of RUN-2 outperforms RUN-1 slightly on both TC and SC. Our method achieves almost 0.60 of recall when we set the threshold to 60%. This reflects that our topic model accurately ranks sentences considering relevant. Based on such an observation, it can be safely claimed that our system may perform better in relevance judgment if the threshold is enlarged from 60% to 90%. This claim has been justified in our post-NTCIR8 experiments (see Section 5.1).

Opinionatendness Judgment

Secondly, the performance of opinionated sentence judgment is evaluated and the results are listed in Table 6.

Table 6. Evaluation result of opinionated sentence judgment

RUN-ID		TC	SC
WIA RUN-1&2	**P**	53.39	29.2
	R	83.68	95.9
	F	65.19	44.77
Best*	**P**	56.37	41.34
	R	85.71	83.35
	F	68.01	55.27

Best*: The best result in NTCIR-8 MOAT Evaluation

The performance of opinionated sentence judgment between TC and SC are very different. Our model achieved a similar result on TC as the performance on the development data. But the same

model performed poor on SC. We use the same lexicons, features and classifier for both of TC and SC. This result reveals the fact that the annotators of TC and SC had different criteria to annotate the sentences. And the criteria of annotation of opinionated sentence for SC may be different with the ones used in NTCIR-7 because our model achieved a much better result on development data. Results on SC of other teams are similar to ours. Recall that our model is leading to a high recall; this means our model could achieve a better *F-score* if parameters are tuned to get the best *F-score* (See 5.2).

Table 7. Evaluation result of polarity classification

RUN-ID		TC	SC
WIA RUN-1	**P**	50.65	50.72
	R	41.11	46.57
	F	45.38	48.56
WIA RUN-2	**P**	50.63	51.18
	R	40.42	45.91
	F	44.95	48.40
Best*	**P**	76.48	67.39
	R	53.03	52.90
	F	62.63	59.27

Best*: The best result in NTCIR-8 MOAT Evaluation

Polarity Judgment

Thirdly, the performance on polarity classification is evaluated. Refer to Table 7, the difference between RUN-1 and RUN-2 are different parameters of the SVM classifier. Our performance ranked top 2 on both TC and SC. CTL got a surprisingly high *F*-measure of 59.27 on SC and 62.63 on TC. Our model performs better on SC using the same model. One possible reason is that our word segmentation tool for TC is not as good as the one for SC.

Table 8. Evaluation result of Holder&Target recognition (SC)

RUN-ID		SC	
		Holder	Target
WIA RUN-1	**P**	85.5	36.9
	R	76.8	33.0
	F	80.9	34.9
WIA RUN-2	**P**	85.3	37.0
	R	74.5	32.2
	F	79.5	34.4
Best*	**P**	87.7	73.5
	R	79.2	56.4
	F	83.2	63.8

Best*: The best result in NTCIR-8 MOAT Evaluation

Holder&Target recognition

Finally, the performance of holder and target recognition is evaluated. The result could be found in Table 8 and Table 9. Our method performed poorly on target recognition. After processing error-analysis of the text data manually, we found the following reasons:

(1) The dependency parser and semantic role labeling tool we adopt performs poorly on long sentences (more than one verb or contains commas).

(2) Our named entity recognizer performs poorly and we did not integrate the weight of named entities in the formula of ranking candidate A0's and A1's.

(3) The parameters of our ranking method are chosen in ad-hoc manner. More corpora [19] could be used in the tuning of parameters together with training data.

(4) Our ranking model performs well in holder recognition but poorly on target recognition. Maybe different parameters are needed for target ranking.

Table 9. Evaluation result (Precision) of Holder&Target recognition (TC)

RUN-ID		TC	
		Holder	Target
WIA RUN-1	Strict	62.1	28.3
	Lenient	51.3	23.3
WIA RUN-2	Strict	60.5	24.6
	Lenient	49.6	19.6
Best*	Strict	84.9	54.4
	Lenient	72.0	45.7

Best*: The best result in NTCIR-8 MOAT Evaluation

5. Post-NTCIR-8 Experiments

5.1 Relevance Judgment

In this section, we introduce our Post-NTCIR8 experiments to reveal that after parameter tuning, our method could achieve some significant improvements. Note that we only change the threshold or parameters of our methods in these experiments. And we only use newest officially released evaluation tools to get the post-NTCIR-8 results.

Refer to Table 5, the precision of our relevant judgment system is satisfaction but the recall is low. We increase our threshold from 60% to 90%. Consider Table 9, significant improvement has been achieved after the threshold was set to 90%. The *F-score* increases for about 12% on TC and 20% on SC. We achieve a surprising result on SC by achieving 93.37% of *F-score*.

Table 10. Post-NTCIR8 experimental result of relevance sentence judgment

Threshold		Post-TC	Post-SC
80%	**P**	87.76	98.14
	R	71.95	79.37
	F	79.07	87.76
90%	**P**	**87.39**	**98.17**
	R	78.31	**89.01**
	F	82.60	**93.37**
Best*	**P**	86.35	97.78
	R	**93.56**	59.64
	F	**89.81**	74.09

Best*: The best result in NTCIR-8 MOAT Evaluation

5.2 Opinionatedness Judgment

Refer to Table 6, our model achieved the recall of 95.27% on SC but the precision was poor. That's because our model focused on get a better recall while the *F-score* are tolerable (see Section 3.2) on development set. To objectively evaluate our model, the

parameters of SVM classifier are tuned to get better precisions on development data tending to get the best *F-score* on test data officially released. The experimental result on SC is shown in the following table. Refer to Table 11, our Post-NTCIR8 result is slightly better than the best result in NTCIR-8 evaluation, this shows the effectiveness of our model.

Table 11. Post-NTCIR8 experimental result of opinionated sentence judgment (SC)

	Best*	Post-SC
P	41.34	**47.56**
R	**83.35**	68.05
F	55.27	**55.99**

Best*: The best result in NTCIR-8 Evaluation

6. CONCLUSIONS AND FUTURE WORK

In this paper, we present a framework for NTCIR-8 MOAT monolingual tasks. All of our methods were designed regardless of language and all modules are built automatically without human effort. Our topic model based method is proved to be very effective on relevance judgment subtask. Owning to the limit training data, we combine all training data from both SC and TC and train general models for opinionated judgment and polarity classification. The experimental result shows that we achieve top 2 of performance on both SC and TC. But there are still much work to do on target recognition. The future work will be focused on two directions: (1) introducing discourse information in opinionated and polarity judgment such as sentence-level, paragraph-level and document-level features; (2) Boosting the performance of holder and target recognition.

7. ACKNOWLEDGEMENTS

This work is partially supported by the Innovation and Technology Fund of Hong Kong SAR (No. ITS/182/08) and National 863 program (No. 2009AA01Z150).

8. REFERENCES

[1] B. Pang, L.L. Lee and S. Vaithyanathan, Thumbs up? Sentiment Classification Using Machine Learning Techniques. In EMNLP'02, pp.79-86, 2002

[2] E. Riloff, J. Wiebe and T. Wilson, "Learning Subjective Nouns Using Extraction Pattern Bootstrapping." In CoNLL'03, pp.25-32, 2003

[3] M. Hu and B. Liu., Mining and summarizing customer reviews, In SIGKDD'04, pp.168-177, 2004

[4] L.W. Ku, T.H. Wu, L.Y. Lee, H.H. Chen, Construction of an Evaluation Corpus for Opinion Extraction, In NTCIR-5, pp.513-520, Japan, 2005

[5] S. Dasgupta and V. Ng. Mine the Easy, Classify the Hard: A Semi-Supervised Approach to Automatic Sentiment Classification. In ACL'09, pp. 701–709, 2009.

[6] B. Pang and L.L. Lee, A Sentiment Education: Sentiment Analysis Using Subjectivity Summariza-tion based on Minimum Cuts. In ACL'04, pp. 271-278, 2004

[7] W.H. Lin, T. Wilson, J. Wiebe and A. Hauptmann, Which Side are You on? Identifying Perspectives at the Document and Sentence Level, In CoNLL'06, pp.109-116, 2006

[8] E. Riloff, S. Patwardhan and J. Wiebe, Feature Subsumpti-on for Opinion Analysis, In EMNLP'06, 2006

[9] V. Hatzivassiloglou and J. Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. In COLING'00, 2000.

[10] P. Turney. Thumbs up or thumbs down? Semantic orientateon applied to unsupervised classification of reviews. In ACL'02, pp. 417–424, 2002.

[11] 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 NTCIR-6 Workshop Meeting, pp.265-278, 2007.

[12] Y. Seki, D.K.Evans and L.W. Ku, Overview of Multilingual Opinion Analysis Task at NTCIR-7, In NTCIR-7 Workshop Meeting, 2008

[13] R.F. Xu, K.F. Wong and Y.Q. Xia, Opinmine – Opinion Analysis System by CUHK for NTCIR-6 Pilot Task. In NTCIR-6 Workshop, Japan, pp.350-357, 2007

[14] R.F. Xu, K.F. Wong and Y.Q. Xia, Coarse-Fine Opinion Mining–WIA in NTCIR-7 MOAT Task. In NTCIR-7 Workshop, Japan, pp.307-313, 2008

[15] R.F. Xu, K.F. Wong, Q. Lu, Y. Xia, W. Li, Learning Knowledge from Relevant Webpage for Opinion Analysis. In WI-IAT'08, Vol 1, 2008

[16] Y. Xia, Y. Zhang, W. Su and J. Yao. 2010. Co-Clustering Sentences and Terms for Generic Multi-Document Summarization. *Submitted to ACL2010*.

[17] Y. Xia, L. Wang, K.F. Wong, and M. Xu, Sentiment Vector Space Model for Lyric-based Song Sentiment Classification, In ACL08 Short Papers, pp.133-136.

[18] T. Wilson, J. Wiebe, and P. Hoffmann, Recognizing Contextual Polarity an exploration of features for phrase-level sentiment analysis, Computational Linguistics, Vol.35, pp.399-433, MIT Press, 2009

[19] R.F. Xu, Y. Xia, K.F. Wong, W.J. Li, Opinion Annotation in On-line Chinese Product Reviews, In LREC'08, Marrakesh, Morocco, May 26-30, 2008.

[20] Y. Seki, L.W. Ku, L. Sun, H. Chen and N. Kando, Overview of Multilingual Opinion Analysis Task at NTCIR-8, In NTCIR-8 Workshop Meeting, 2010

[21] J. Pei et al., PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth, In Proceedings of ICDE'01, pp.215-226, 2001