Opinmine – Opinion Analysis System by CUHK for NTCIR-6 Pilot Task

Ruifeng Xu (1), Kam-Fai Wong and Yunqing Xia (2)
Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Shatin, Hong Kong, China
{rfxu, kfwong, yqxia}@se.cuhk.edu.hk

(1) *Rufeng Xu now works with Department of Computing, The Hong Kong Polytechnic University, China*
(2) *Yuqing Xia now works with Research Institute of Information Technology, Tsinghua University, China*

Abstract

This paper presents the CUHK opinion analysis system, namely Opinmine, for the NTCIR-6 pilot task. Opinmine comprises of three functional modules: (1) Preprocessing and Assignment Module (PAM) performs word segmentation, part-of-speech (POS) tagging and named entity recognition on the input Chinese text. It is based on lexicalized Hidden Markov Model and heuristic rules. (2) Knowledge Acquisition Module (KAM) applies unsupervised learning techniques to acquire different opinion knowledge including opinion operator, opinion indicator and opinion words from annotated data and Web data. (3) Sentence Analysis Module (SAM) analyzes each input sentence to determine whether it is opinionated. For each opinionated sentence, its opinion holders, opinion operators and opinion words are recognized and its polarity is determined. Furthermore, the relevance between the sentence and a topic are judged by based on sentence-topic and document-topic relevance. For lenient evaluation, the F1 performance of Opinmine in opinion extraction, polarity decision and relevance judgment are 0.635, 0.405 and 0.812, respectively; and for strict evaluation, the F_1 performances are 0.427, 0.296 and 0.616, respectively.

Keywords: *Opinion analysis, NTCIR, Opinion holder, Opinion operator*

1 Introduction

Automatic identification and analysis of opinions in running text have been the focus of research on information extraction in recent years in different application domains such as news articles and product reviews [Lu et al. 2005; Xu et al. 2007]. Various approaches have been proposed in subjectivity detection, semantic orientation detection, opinion identification and classification [Hatzivassiloglou et. al. 1997; Kim et. al. 2006; Lu et al. 2005]. These approaches were designed for different purposes and their performances are evaluated based on different datasets. For this reason, the evaluation is incomprehensive. This inevitably hinders research in opinion analysis.

NTCIR-6 provides a pilot task to evaluate and compare different approaches for multi-lingual opinion analysis [Sekiy et al. 2007]. The pilot task defines four evaluation categories:

1. Opinion sentence identification. Given an input text, extract all sentences, which carries an opinion.
2. Opinion holder identification. An opinion holder is the governor of an opinion. Typically, an opinion holder is a person, a country, an organization or a group, who expresses an opinion in a sentence. By grouping opinion holders with different stance in social and political topics, we can better understand the relationships among different people, countries, organizations…etc. [Kim et. al. 2006].
3. Determine the polarity of the opinionated sentence, i.e. to determine whether this sentence is positive, negative or neutral.
4. Judge the relevance between the sentence and the topic. This is very important to intelligent summarization as only relevant sentences are considered.

Notice that the first two categories are mandatory and the other two are optional.

CUHK has developed the Opinmine system and participated in the Chinese NTCIR-6 opinion analysis task. In this task, the opinion-annotated news corpus provided by National Taiwan University [Ku et al. 2005] is used. We took part in all of four evaluation categories. Since the size of the training dataset is limited, Opinmine acquires global opinion knowledge and local topic related knowledge from the Web. This knowledge facilitates opinion analysis. Opinmine comprises of three functional modules. The first module preprocesses the input text. It performs word

segmentation, POS tagging and named entity recognition over the text. The second module learns opinion-operator and opinion-indicator knowledge from a sentence. It involves recognition and normalization of negations and conjunctions. Such knowledge is topic independent. Furthermore, the knowledge of opinion word is learned based on a static opinion-word lexicon. Lexicon expansion is performed based on a synonym dictionary. Supervised techniques are applied to learning from annotated text, and unsupervised learning is applied to Web data. In this part of the module, global topic-related knowledge reflecting the polarity of an opinion word is obtained. The third module incorporates different sources of information to estimate the probabilities that a sentence is opinionated. And at the same time, the opinion holder and opinion polarity of the sentence are determined. This module also estimates the document-topic and sentence-topic relevance of each sentence. In lenient evaluation, the F_1 performances in opinion extraction, polarity decision and relevance judgment achieved by Opinmine are 0.635, 0.405 and 0.812, respectively; and the same in strict evaluation are 0.427, 0.296 and 0.616, respectively.

The rest of this paper is organized as follows. Section 2 presents the Preprocessing and Assignment Module (PAM), which involves word segmentation, pos-tagging and name entity recognition. Section 3 describes the Knowledge Acquisition Module (KAM) for learning opinion knowledge. Section 4 outlines the design and implement of the opinion Sentence Analysis Module (SAM) as well as the sentence-topic relevance estimation algorithm. Section 5 gives the evaluation results and finally, Section 6 concludes this paper.

2 PAM - Preprocessing and Assignment Module

The task of the first module is to segment Chinese sentences into words and assign each word a POS tag. This is an indispensable step in any Chinese sentence analysis application. Furthermore, named entities are recognized. These entities are candidates of opinion holders. The word segmentation algorithm proposed by [Lu et al. 2003] and the named entity recognizers by [Fu et al. 2006] are adopted in PAM. Further, they are trained using the Peking University People Daily corpus and the Sinica corpus, respectively.

Each tagged word in a sentence is represented as a lexical chunk, which consists of a sequence of lexicon words together with their lexical chunk tags. A lexical chunk tag follows the format T_1-T_2, where T_1 denotes the position pattern of a lexicon word in the segmented word and T_2 denotes the POS category of the segmented word. Here, a lexicon word refers to a word appeared in the dictionary. In general, a complete dictionary for Chinese language processing covers all possible Chinese characters. Any segmented word in a Chinese text is composed of one or several lexicon words. A position pattern refers to the location of a lexicon word within a segmented word. Generally, a lexicon word may take up one of the following position patterns after segmentation: (1) it is a segmented word by itself; (2) it occurs at the beginning of a segmented word; (3) it occurs in the middle of a segmented word; or (4) it occurs at the end of a segmented word.

Based on the above representations, Chinese word segmentation and POS tagging can be formulated as two new tasks: lexicon word segmentation and lexical chunking on a sequence of lexicon words. Word bigram language models and lexicalized hidden Markov Models (HMMs) are applied to the two tasks, respectively [Fu et. al. 2006]. The goal of lexicon word segmentation is to segment a sequence of Chinese characters into a meaningful sequence of lexicon words. According to a given lexicon, given a Chinese character string $C=c_1c_2...c_m$, there may be multiple candidate word sequences $\{W=w_1w_2..w_n\}$. Lexicon word bigram segmentation aims to find the most appropriate segmentation that maximizes the conditional probability $P(W|C)$, i.e.

$$\hat{W} = \arg\max_{W} P(W \mid C) \approx \arg\max_{W} \prod_{t=1}^{n} P(w_t \mid w_{t-1}) \quad (1)$$

where $P(w_i|w_{i-1})$ denotes the lexicon word bigram probability, which can be estimated from a segmented corpus using Maximum Likelihood Estimation (MLE). To resolve the issue of data sparseness in MLE, we apply linear interpolation technique to smoothen the estimated word bigram probabilities.

Lexical chunking involves assigning (i.e. tagging) each lexicon word in a sentence with an appropriate lexical chunk tag. Let $W=w_1w_2..w_n$ be a sequence of lexicon words for an input sentence, the task of lexical chunking is to find a sequence of lexical chunk tags that maximizes the conditional probability $P(W|C)$, namely,

$$\begin{aligned} \hat{T} &= \arg\max_{T} P(T \mid W) \\ &\approx \arg\max_{T} \prod_{i=1}^{n} \begin{pmatrix} P(w_i \mid w_{i-J} \cdots w_{i-1}, t_{i-I} \cdots t_{i-1} t_i) \\ \times P(t_i \mid w_{i-K} \cdots w_{i-1}, t_{i-L} \cdots t_{i-1}) \end{pmatrix} \end{aligned} \quad (2)$$

Equation (2) gives a general form of the uniformly lexicalized HMMs for Chinese lexical chunking, where the following hypotheses are made: the appearance of current word w_i is assumed to depend not only on the current tag t_i and the previous $I(1 \le I \le i-1)$ tags $t_{i-I}...t_{i-1}$ but also the previous $J(1 \le J \le i-1)$ words $w_{i-j}...w_{i-1}$; The assignment of current tag t_i is supposed to depend on both its previous $K(1 \le K \le i-1)$ words $w_{i-k}\ldots w_{i-1}$ and L$(1 \le L \le i-1)$ tags $t_{i-L}\ldots t_{i-1}$. To account for data sparseness, we set I=0 and $J=K=L$=1. With the MLE technique, the uniformly lexicalized HMMs are trained. To address the problem of zero probabilities in MLE, the linear interpolation technique is employed to smoothen lexicalized parameters with relevant non-lexicalized probabilities.

Based on word segmentation and POS tagging results, named entity recognition are performed. In fact at this stage, some of the named entities has already been identified and assigned with POS tags, e.g. /nr for person name etc. The *personal name recognition component* is used to recognize unknown personal names. 《姓氏人名用字分析統計》[CSSA 2002] is referenced. It records 737 Chinese family names; out of which, 8 entails 2 characters. A personal name is derived out of 3,345 characters, and the distribution is quite even. This component first identifies all possible character combinations used to form names; and puts them in a candidate list. Each candidate will then be processed using a set of rules. Any candidates that violate a rule will be removed. Each of the remainders will then be rated. Only candidates with the high ratings are retained. The algorithm produces a family-name list, a first-given-name list, a second-given-name list, a prefix list, and a suffix list. Affixes are words that usually appear right before or after a Chinese name, such as 先生(*Mr.*), 女士(*Mrs.*) and 總統(*President*). Each character in the family-name list, first-given-name list and second-given-name list is also accompanied by statistical information. Although segmentation of foreign names is not straightforward, it is relatively simple because, with the exception of Japanese names, characters used in non-Chinese names can be identified easily. Moreover, some of those characters can only be used either at the beginning or at the end, while others are only used in the middle. Thus, the algorithm uses several character lists that are commonly used in non-Chinese names.

The *place name recognition component* primarily uses a place dictionary to identify place names. There are two reasons for this: first, it is difficult to collect reputable statistical data for place names; second, segmentation based on place dictionary produces acceptable results. Our algorithm only uses a suffix table as supplement. Similar to the place name component, the recognition of organization names is also heavily dependent on a dictionary plus company indicators such as 公司(*company*)，有限公司(*co. Ltd.*).

3 KAM – Knowledge Acquisition Module

3.1 Knowledge of Opinion Operators and Opinion Indicators

People express their opinions in some conventional patterns, especially in formal news text. A complete opinion typically consists of the following components.

1. An **opinion holder** is the governor of an opinion and normally represents a person, a countries or an organization.
2. An **opinion object** is the target of the opinion. It may be a person, a country, an organization, an event, a product or a service.
3. An **opinion word** expresses the polarity, i.e. favorable or unfavorable, of an opinion.
4. An **opinion operator** is the verb indicating an opinion event.
5. An **opinion indicator** is the word indicating the orientation of an opinion or the orientation trend of multiple opinions [Ku et al. 2005].

Consider the following example:

李博士/nr 指出/v，目前的/a 策略/n 可以/v 有效的/ad 控制/v 風險/n，/w 但是/adv，問題/n 依然/adv 存在/v。/w

(Dr. Li pointed out that current strategy reduce the risks effectively, however, the problems are still existed.)

In this opinionated sentence, the opinion holder is 李博士/nr (*Dr. Li*), the opinion object is 目前的/a 策略/n (*current strategy*), the opinion operator is 指出/v (*point out*) and the opinion word is 有效的/ad (*effectively*). Also, the word 但是/adv (*however*) serves as an opinion indicator reflecting that the polarity of the following clause is negative to the preceding clause. [Xu et al. 2007] observed that the use of opinion words was more versatile and the use of opinion operator and opinion indicators was more regular. Thus, we first learn the knowledge of opinion operators and opinion indicators.

The training data comes from two sources: the sampling data of NTCIR-6; and the documents relevant to the NTCIR-6 sampling data, which are collected from the Web (see also Section 3.2). Firstly, potential opinionated sentences are identified from the documents. This is achieved by selecting sentences containing both known named entities, which would be the opinion holders, and matched opinion words, which would be the opinion words. Dependency paring is applied to these sentences to identify the verbs, which are dependent on the opinion holders. The identified verbs are collected. Suppose a verb V occurs in the opinionated sentences m times and in the whole training text n times, the probability of V serving as an opinion operator, labeled as $p\_op(V)$, is estimated by,

$$p\_op(V) = m / n \tag{3}$$

A verb which is associated with a large $p\_op(V)$ value and occurs more than a threshold number of times are included in the opinion operator lexicon. Some opinion operators are given below:

警告 (*warning*) 強調 (*emphasize*) 駁斥 (*refute*)

指出 (*point out*) 稱讚(*praise*) 批评(*criticize*)

说(*said*) 表示(*said*) 答 (*answer*) 发表(*announce*)

Note that, some opinion operators also play the role of opinion indicators. For example, sentences with opinion operator 稱讚(*praise*) always bring

positive information while sentences with 批评 (*criticize*) are always negative.

Opinion indicators are mainly conjunctions, adverbs and adverbial phrases. These types of words and phrases in an opinion sentence are collected and manually selected. Typical opinion indicators include:

- Negational conjunctions, such as 但是(*but, however*), 儘管(*Though*) etc., indicate that the orientation of the following clause is different with the preceding.
- Continual conjunctions, such as 幷且, 而且 (and), 特别(*especially*) etc., indicate that the orientation of the following clause is the same as the preceding.
- Adverbs and adverbial phrases directly indicate the polarity of the opinionated sentence, e.g. 遺憾的是 (*It is regrettable*), 過于(*excessive*), etc.
- Opinion operators directly indicate the polarity of the opinionated sentence as discussed above.

The knowledge of opinion operator and opinion indicator is useful for identification of opinionated sentences from the input text and to determine their polarities. During opinion analysis, the opinion indicators are firstly located in the sentences. The knowledge of opinion indicator is then used to determine the possible polarity and the possible sentimental orientation change in the sentences. Meanwhile, the opinion operators are also identified. This is used to determine an opinionated sentence. Also, the opinion operators are used for opinion holder recognition.

3.2 Knowledge of Opinion Words

Opinion words play a key role in opinionated sentence identification and opinion polarity determination. Different from the usage of opinion operators and opinion indicators, which is relatively regular, the use of opinion words are more versatile. The lack of a comprehensive opinion word lexicon renders opinion word acquisition difficult. Opinion words are generally classified into three types.

1. A context-independent opinion word (labeled as *CFOW*) whose polarity is constant irrespective of context, e.g. 完美(*perfect*) is always positive and 惡劣(*bad*) is always negative.
2. A context-dependent opinion word (labeled as *CDOW*) whose polarity is determined by the context and topic, e.g. 可笑的 is positive when it is used in the context of talk shows meaning *burlesque*; but it has a negative sense when it is used in the context of politics meaning (*absurd)*.
3. A neutral word carries an opinion polarity, when it is associated with some opinion objects. For example 大 expresses positive orientation (*great*) when collocated with 成績 (*achievement*). On the contrary, 大 brings negative orientation when collocated with 困難(*difficulty*). For practical reason, this is also classified as CDOW.

We built our initial opinion word lexicon (also referred to as sentimental lexicon) from two linguistic resources: (a) The *Lexicon of Chinese Positive Words* [Shi et al. 2005], which consists of 5,054 positive words and the *Lexicon of Chinese Negative Words* [Ling and Zhu 2005], which consist of 3,493 negative words; (b) The opinion word lexicon provided by National Taiwan University (NTU) which consists of 2,812 positive words and 8,276 negative words [Ku et al. 2005]. The segmentation of some "opinion words" listed in the NTU lexicon is different from our system, e.g. "不充分的" (*insufficient*) is regarded as one negative opinion word in the NTU lexicon; but the same is segmented as 不(*not*) followed by a positive word 充分的(*sufficient*) in Opinmine. For this reason, we used the first resource as the base and enriched it using selected words from the NTU lexicon instead of blindly merged the two. The words in the enriched lexicon were then manually classified into CFOWs and CDOWs.

Knowledge of a CDOW opinion word includes the word itself and its contextual behavior. The annotated corpus provided by NTCIR is rather small and is insufficient for effective training. Thus, we collected additional documents relevant to the 32 topics specified by NTCIR-06 as the unsupervised training data. In the description file for each NTCIR topic, there is a <TITLE> element for title words and a <CONC> element for topic-relevant words. For example, in Topic 7, the title words are 鈴木一朗 新人王 美國職棒大聯盟 and the topic-relevant words are 鈴木一朗 打擊王 新人王 MVP 金手套 國民榮譽賞 銀棒獎 盜壘王 西雅 圖水手隊 歐力士隊. We used the Web crawler developed by The Hong Kong Polytechnic University to retrieve relevant documents from the Internet. Documents, which contain the title words and more than 50% topic-relevant words, are deemed relevant. As a result, 5,800 documents were collected, which is about 8.3 times of the number of documents in the NTCIR corpus (i.e. 700).

The learning of opinion words is based on several hypotheses.

Hypothesis 1: The polarity of a CFOW remains the same in all related documents and the polarity of a CDOW is dependent on the domain and its context [Xu et al. 2007].

Hypothesis 2: Opinion words in *and* conjunctions usually have similar polarity; and *but* conjunctions opposite [Hatzivassiloglou et al. 1997].

Hypothesis 3: In news text, the polarity of continual or parallel opinionated sentences or clauses remains the same unless a negation is detected.

Based on these hypotheses, we designed the opinion word knowledge learning algorithm, which involves 11 steps:

Step 1. Extract opinion indicators and the listed opinion words from the NTCIR training corpus;
Step 2. Tag the unlisted (i.e. new) opinion words as CDOWs and append them to the opinion word lexicon;
Step 3. Learn the global and topic-related local polarity behaviors of the CDOWs. Suppose a CDOW, *w,* occurs *t* times in the training text; and in which *t_pos* times are positive, *t_neu times* neutral and *t_neg* times negative (i.e. *t_pos*+*t_neu*+*t_neg*=*t*), the global positive polarity ratio of *w*, *p_pos(w)*, is calculated by:,

$$p\_pos(w) = t\_pos / t \quad (4)$$

Similarly, the global neutral polarity ratio, *p_neu(w)*, and global negative polarity ratio, *p_neg(w),* are calculated. Further, suppose *w* occurs *t_i* times in the text of topic *i*, and in which *t_i_pos* times are positives, *t_i_neu times* neutral and *t_i_neg* times negative (i.e. *t_i_pos*+*t_i_neu*+*t_i_neg*=*t_i)*, the local positive polarity ratio of *w* to topic *i*, *p_i_pos(w)*, is calculated by:

$$p\_i\_pos(w) = t\_i\_pos / t\_i \quad (5)$$

The local neutral polarity ratio, *p_i_neu(w)*, and local negative polarity ratio, *p_i_neu(w),* are calculated in a similar way. These values describe the global and topic-related local polarity behaviors of *CDOW*s.
Step 4. Identify the opinion indicators. Determine their polarities and the associated negation in the non-annotated text;
Step 5. Match the listed CFOWs and CDOWs in the non-annotated text, and mark the polarity of CFOWs in the sentences. As for the polarity of CDOWs, the corresponding global and local polarity ratios are marked.
Step 6. If a CDOW occurs in a sentence and its polarity is determined by the adverbial or verb opinion indicators, the polarity of the CDOW in this sentence is assigned as the value suggested by the opinion indicators. If a CDOW co-occurs with a CFOW in the same sentence or the neighboring continual/parallel sentence, which is determined by conjunction opinion operators, the polarity of the CDOW follows the polarity of the CFOW, or the opposite polarity is assigned if a negation is detected.
Step 7. Update both the global and local polarity ratio of CDOWs by using the results in Step 6.
Step 8. Based on Hypothesis 3, determine the polarities of the CDOWs beyond the ones considered in Step 6. As such, if the context of the CFOW is positive, the CDOW is also positive unless a negation is detected. Suppose S_j is the *j-th* sentence in topic *i*, the polarity of S_j is estimated by the polarities of the CFOWs and CDOWs in this sentence, viz:

$$pts(S_j) = \sum_{S_j} p\_pos(w_{CFOW}) + \sum_{S_j} \alpha_1 \cdot (p\_pos(w_{CDOW}) - p\_neg(w_{CDOW})) + \alpha_2 \cdot (p\_i\_pos(w_{CDOW}) - p\_i\_neg(w_{CDOW})) \quad (6)$$

where, $p\_pos(w_{CFOW})$ is the value of positive ratio of a CFOW. Its value equals to 1 if w_{CFOW} is a positive word and -1 if w_{CFOW} is negative; α_1 and α_2 ($\alpha_1 + \alpha_2 = 1$) are parameters for weighting global and local polarities of a CDOW. A large value (>0) of $pts(S_j)$ implies S_j tends to be positive. Note that negation has been taken into account when determining the values for each element in Equation 6.

Suppose w_{cd} is a CDOW and its polarity is unknown. The polarity of S_j and the polarity of the neighboring sentences S_{j-1} and S_{j+1} (labeled as $pts(S_{j-1})$ and $pts(S_{j+1})$, respectively) are used to determine the polarity of w_{cd}, viz:

$$p(w_{cd}) = 0.5 \cdot pts(S_{j-1}) + pts(S_j)^* + 0.5 \cdot pts(S_{j+1}) \quad (7)$$

where $pts(S_j)^*$ is the polarity trend of S_j by following Equation 6 but ignoring the contribution of w_{cd}. If the value of $p(w_{cd})$ is greater than a empirical threshold *z (z>0)*, w_{cd} is considered positive in sentence *j*; if $p(w_{cd}) < -z$, w_{cd} , it is determined as negative; otherwise, w_{cd} is neutral.
Step 9. After the polarities of all CDOWs in the non-annotated text are determined, update both the global and local CDOW polarity ratios.
Step 10. Recognize the adjectives detected in the opinionated sentence but not listed in the CFOW and CDOW lexicons. The adjectives are collected as new CDOWs. Repeat Step 6 to Step 9 to estimate both their global and local polarity ratios.
Step 11. Repeat Step 10 until no new opinion word is found.

In this way, the opinion word lexicon is expanded and the polarity ratios of CDOWs are learned by using both supervised learning and unsupervised learning. This opinion word knowledge is essential to opinion sentence analysis (see Section 5).

4 SAM: Sentence Analysis Module

SAM is comprised of two sub-modules: the opinion sentence analysis sub-module, which extracts opinionated sentences from a given input text and determines the polarity of each opinionated sentence; and the sentence-topic relevance estimation sub-module, which determine which topic each opinionated sentence is most relevant to.

4.1 Opinion Sentence Analysis

Our opinion sentence analysis algorithm makes use of multiple features:

Entity-related feature, F_{entity}: A typical opinion sentence entails a holder, an object, an operator and an opinion expression based on opinion words or indicators. The holder and object of an opinion are persons, organizations, countries or their corresponding pronouns. Effectively, these are entities. Thus, we incorporate entity-related feature in opinionated sentence identification. Given a sentence S_j in topic *i*. Suppose potential entities $en_1 \ldots en_t$ are identified in S_j with their corresponding entity hood, $EH(en_1)$ …

$EH(en_t)$. Entity hood is a value ranges from 0 to 1; and the larger $EH(en)$ is, the higher is the probability that en being a true entity. Suppose one potential entity en_m ($1 \le m \le t$) occurs $f_{en\text{-}m}$ times in the training text, in which $f_{en\text{-}m\text{-}op}$ times in opinionated sentences, the confidence of en_m occurring in an opinionated sentence, labeled as $con(en_m)$, is estimated as follows,

$$con(en_m) = EH(en_m) \cdot \frac{f_{en-m-op}}{f_{en-m}} \cdot (1 - \frac{1}{f_{en-m}}) \quad (8)$$

$con(en_m)$ ranges from 0 to 1; and the larger this value is, the higher is the probability that en_m occurs in an opinionated sentence. The entity-related feature for opinionated sentence identification, labeled as F_{entity}, is then calculated by,

$$F_{entity}(S_j) = \sum_{m=1}^{t} con(en_m) \quad (9)$$

Topic feature. F_{topic}. An opinionated sentence usually follows the same topic as the news reports, where it appears. This implies that a sentence consisting of more topic words has a higher probability to be an opinionated sentence. Suppose S_j entails n_j common words describing a topic i, which is given a description file, the topic-related feature for opinionated sentence identification, labeled as F_{topic}, is estimated by,

$$F_{topic}(S_j) = \frac{n_j}{|S_j|} \cdot \frac{n_j}{|i|} \quad (10)$$

where, $|S_j|$ is the word numbers of S_i and $|i|$ is the number of topic words of i. A larger value of F_{topic} implies that S_j has a higher probability to be opinionated.

Opinion word feature, $F_{opinion\text{-}word}$, This feature accounts for opinion words, which is labeled as $F_{opinion\text{-}word}$. It considers the number of opinion words in S_j and the corresponding polarity probabilities as follows:

$$F_{opinion\_word} = \sum_{S_j} w_{CFOW} + \sum_{S_j} \begin{matrix} \alpha_1 \cdot (p\_pos(w_{CDOW}) + p\_neg(w_{CDOW})) \\ + \alpha_2 \cdot (p\_i\_pos(w_{CDOW}) + p\_i\_neg(w_{CDOW})) \end{matrix} \quad (11)$$

$F_{opinion\_word}$ estimates the probability that a sentence being opinionated. It incorporates both the positive and negative sentence probabilities. A larger value of $F_{opinion\_word}$ means that S_j has a higher probability to be an opinionated sentence.

Indicator feature, $F_{indicator}$ considers the number of opinion indicators in a sentence, S_j. It is based on the assumption that a sentence containing more opinion indicators would more likely be an opinionated sentence.

Opinion operator feature, $F_{operator}$, covers opinion operators. The more identified opinion operators a sentence has, the more likely it carries an opinion. Suppose there are k opinion operators, $o_1 \ldots o_k$ in a sentence, S_j. $F_{operator}$ is determined as follows,

$$F_{operator}(S_j) = \sum_{m=1}^{k} p\_op(o_m) \quad (12)$$

where $p\_op(o_m)$ is the probability of o_m serves as a opinion operator (see Equation 3).

Support Vector Regression (in short SVR) is used to integrate the five features into a linear expression [Scholkopf et al. 1998]. The five features for each sentence in the training text are calculated. The regression function in SVR is trained by values of these features. Similarly, this SVR function is then applied to the testing data. As such, it determines whether a testing sentence is opinionated and if so, its polarity.

Moreover, once an opinionated sentence is identified, we locate the opinion operators, which in turn are used to determine the opinion holders. Based on the characteristics of the NTCIR annotated data, we define several heuristic rules for locating opinion operators. Following are a few examples:

- The distance between an opinion operator and a neighboring entity/pronoun is always less than five words.
- Opinion operators typically appears adjacent to punctuations, such as "*quote*(" ")" and "*colon* (:)".
- If no conjunction/negation opinion indicator or specified punctuations is detected, there will be only one opinion operator in one clause or one sentence.

The following heuristics are used to recognize pinion holders:

- It must be a recognized entity or pronoun.
- It must collocate and strongly associated with certain identified opinion operators.
- It always occurs in the beginning of a sentence or near the beginning or end of a quotation.

In addition to recognizing the above heuristics, their contexts are analyzed. Based on some manually compiled heuristic rules, the modifier of the entity and its neighboring entities are analyzed to determine whether they will be combined together as the opinion holder. For example, if a person entity is detected and its neighboring words may play as its affix, these words will be combined as a single person entity. This rule is illustrated in the following sentence, where the affix of the core of opinion holder (bolded) is merged together to be an opinion holder, which is bracketed.

[美國/nt (*USA*) 國防部長/n (*defense minister*) **柯恩**/nr(*Kern*)] 今天(*today*) 在(*at*) 澳洲/ns (*Australia*) 坎培拉(*Canberra*) 表示(*said*)

For another example, if two neighboring entities with the same type (name, country and so on) are interrupted by a pause mark (、) or a conjunction indicating continuation (such as 和 与 *and*) and an opinion operator is identified around each entity, these two entities are parallel opinion holders; and they are associated with the nearest opinion operator. For example, [中央軍委副主席**胡錦濤**] (*Vice president* ***Hu***)、[中共國防部部長**遲浩田**] (*Defense minister*

Chi在與柯恩會談時 (*in the meeting with Kern*)*強調 (emphasize)*,

Finally, the polarities of the sentences are determined by using the knowledge of opinion indicators/ negations and the positive polarity trend estimated by Equation 6. Sentences with positive polarity greater than a predefined threshold (>0) are regarded as positive and conversely, the ones with values lower than that (<0) are regarded negative.

4.2 Sentence-Topic Relevance Estimation

The Support Vector Regression (SVR) Model is adopted in the sentence-topic relevance estimation sub-module; and is based on the following observations: In general, a sentence, which contains more content words and more named entities, especially with more common topic words, has a higher probability of being relevant to the topic. Also, the position information of a sentence in the document is useful in estimating the relevance between sentence and topic. Ten features in total are used by our current SVR model to estimate the relevance between a sentence and a given topic.

Given a sentence S_j from document D in topic I. The following features are designed or selected as the support vectors:

(1) $V_{cov\text{-}s\text{-}title}$ is the coverage of the content words in S_j appearing in the title words of topic *i,* which is given in the element of <TITLE> in the description document. Large $V_{cov\text{-}s\text{-}title}$ favors relevance.

(2) $V_{cov\text{-}s\text{-}toipic}$ is the coverage of the content words in S_j appearing in the topic words of topic *i,* which is given in the element of <CONC>). Large $V_{cov\text{-}s\text{-}toipic}$ favors relevance.

(3) $V_{cov\text{-}s\text{-}rel\text{-}support}$ is the coverage of the content words in S_j appearing in the relevant definition of topic *i*, which is given in the first half of the element <REL>. Large $V_{cov\text{-}s\text{-}rel\text{-}support}$ favors relevance.

(4) $V_{cov\text{-}s\text{-}rel\text{-}reject}$ is the coverage of the content words in S_j appearing in the non-relevant definition of topic *i*, which is given in the second half of the element <REL>. Higher coverage leads to larger penalty, hence more irrelevant.

(5) We also consider the coverage of the entities in S_j appearing in the entities of the TITLE, CONC, REL elements of topic *i*. These features are similar to features (1)-(4). The difference is that they are the coverage of entities in S_j to the entities in the corresponding element rather than to the words.

(6) $V_{sen\text{-}num\text{-}entity}$ is the percentage of entities in S_j. A sentence with more entities always brings more information and thus it has higher probability to be relevant.

(7) $V_{position}$ caters for positional information. Normally, the sentences in the beginning paragraphs especially the first paragraph is more relevant to the topic. Similarly, the beginning sentences in a given paragraph have relative higher relevant probability than the others. Suppose a document has p paragraphs and the *k-th* paragraph, p_k, has n sentences, the support vector of position information for the *i-th* sentence in p_k, $V_{position}$, is estimated by,

$$V_{position} = (1 - \frac{k-p}{p}) + 0.5 \cdot (1 - \frac{i-n}{n}) \qquad (13)$$

(8) The feature based on the centroid of the document. A sentence, which is more similar to the centroid of a document, is naturally more relevant to the topic of the document. Estimation of this feature in Opinmine is built on top of the algorithm proposed by [Radev et al. 2003]. Suppose there are N documents related to topic *i*; and a word t appears in one of the document d $tf(t,d)$ times and t appears in n_t documents of topic *i*. Thus, we define the weight of t in the document d, labeled as *TF-IDF*(t), as follows:

$$TF - IDF(t) = tf(t,d) \cdot \frac{N}{n_t} \qquad (14)$$

The value of TF-IDF weights the centroid of a word in a document. The centroid of a sentence S_j is then estimated by summing the centroid of each content word in S_j. A larger sentence centroid implies the more relevant the sentence S_j is to the topic.

(9) The relevance between a document and a topic. A sentence has higher probability to be relevant to a given topic if its document is relevant to the topic. This hypothesis indicates that the estimation of sentence-topic relevance must consider both global and local relevance. Similar to the feature set (1)-(6), the coverage of content words/entities in a document d appearing in the TITLE, CONC, REL elements of topic *i* are used as support vectors.

(10) The percentage of entities in the document.

Using the features of annotated sentences as input and the corresponding annotated relevance as output, the SVR model is trained. This model is then applied to the training text.

5 Evaluation

Table 1 and Table 2 present the evaluation results of opinionated sentence determination, topic relevance judgment and polarity determination by using the lenient gold standard and strict gold standard, respectively. Results of precision, recall and F_1 of each category are shown. Also, the relative result ranking among the seven submissions is given.

The CUHK system, i.e. Opinimine, achieves good performance in polarity judgment and relevance determination. This result has shown the effectiveness

of the incorporation of multi features. However, the F_1 performance of opinionated sentence determination is slightly disappointing as it is affected by the low recall.

	Opinionated			Relevance			Polarity		
	P	*R*	F_1	*P*	*R*	F_1	*P*	*R*	F_1
CUHK	0.82	0.52	0.64	0.80	0.83	0.81	0.52	0.33	0.41
Rank	1	6	6	1	3	1	1	4	1

Table 1. Lenient Evaluation of Opinion Analysis

	Opinionated			Relevance			Polarity		
	P	*R*	F_1	*P*	*R*	F_1	*P*	*R*	F_1
CUHK	0.34	0.58	0.43	0.47	0.90	0.62	0.20	0.60	0.30
Rank	1	7	3	1	3	1	1	4	1

Table 2. Strict Evaluation of Opinion Analysis

The evaluation results of opinion holder identification using the lenient and strict gold standards are listed in Table 3.

	Lenient			Strict		
	P	*R*	F_1	*P*	*R*	F_1
Sentence-based	0.65	0.75	0.70	0.71	0.79	0.74
Holder-based	0.74	0.93	0.83	0.79	0.81	0.80

Table 3. Evaluation of Identifying Opinion Holder

Opinmine achieves the best performance on opinion holder identification in both sentence-based and holder-based evaluation. This result benefits from: 1. the effectiveness of the name entity recognizer; 2. our consideration of the association between opinion holders and opinion operators; 3. the heuristic rules for opinion holder determination.

6 Conclusion

Opinmine, the CUHK opinion analysis system, for NTCIR-6 pilot task is reported in this paper. Generally speaking, Opinmine achieves satisfactory performance, particularly in opinion holder identification. This shows the effectiveness of unsupervised learning on Web text and the incorporation of multiple features in the SVR model. Also, the favorable performance on sentence relevance determination justified our hypothesis on combining both document-topic and sentence-topic relevance estimation.

Although the evaluation results are in general favorable, they are only preliminary. More evaluations using larger datasets and analyses are required to reveal and understand the pros and cons of the different modules in Opinmine. Currently, the performance of opinion object identification is unsatisfactory. This means that our system is lack of the capability to analyze deep opinion information at the subjectivity level. This will be a core part of our further research will focus on this part.

Acknowledgements

This project is partially funded by the Strategic Grant (#4410001), The Chinese University of Hong Kong. We would like to thank Fu Guohong, City University of Hong Kong and Lu Qin, The Hong Kong Polytechnic University for providing the word segmentation and POS tagging tools. Also, thanks are due to the Chinese Computing Laboratory, The Hong Kong Polytechnic University, for providing the Web crawler.

References

[CSSA 2002] 中國社會科學院語言文字應用研究所《姓氏人名用字分析統計》, 2002. ("*Statistical Word Analysis of Chinese Surnames*", published by Chinese Academy of Social Science)

[Fu et. al. 2006] Guohong Fu, and Kang-Kwong Luke. Chinese POS disambiguation and unknown word guessing with lexicalized HMMs. *International Journal of Technology and Human Interaction*, Vol.2, No.1, pp.39-50, 2006

[Hatzivassiloglou et. al. 1997] Hatzivassiloglou, V. et al., K. Predicting the semantic orientation of adjectives. In *Proc. of ACL-EACL 97*, 1997

[Kim et. al. 2006] Soo-Min Kim and Eduard Hovy. Extracting opinions, opinion holders, and topics expressed in online news media text, In *Proc. of ACL Workshop on Sentiment and Subjectivity in Text*, pp.1-8, Sydney, Australia, 2006

[Ku et al. 2005] L. W. Ku, T. H. Wu, L. Y. Lee, and H. H. Chen. Construction of an evaluation corpus for opinion extraction. in *Proc. of NTCIR-5 Workshop*. pp. 513–520, Tokyo, Japan, 2005

[Lu et al. 2003] Q. Lu, S. T. Chan, R. F. Xu, et al. A Unicode based adaptive segmentor. In *Proc. of the 2nd SIGHAN Workshop on Chinese Language Processing at ACL 2003*, pp.164-167, Spain, 2003

[Lu et al. 2005] L.-W. Ku, L.-Y. Li, T.-H. Wu and H.-H. Chen. Major topic detection and its application to opinion summarization. In *Proc. of SIGIR 2005*, pp. 627-628, 2005

[Radev et al. 2003] Dragomir R. Radev, Jahna Otterbacher, Hong Qi, Daniel Tam. 2003. MEAD ReDUCs: Michigan at DUC 2003. In *Proc. of DUC 2003*, 2003

[Scholkopf et al. 1998] B.Scholkopf, et al. Support Vector Regression with Automatic Accuracy Control, *In Proc. of 8th International Conference on Artificial Neural Networks*, pp.111-116, 1998.

[Sekiy et al. 2007] Yohei Sekiy, David Kirk Evansz, Lun-Wei Ku et al. 2007. Overview of opinion analysis pilot task at NTCIR-6, in *Proc. of NTCIR-6 workshop*, Tokyo, Japan, May, 2007

[Shi et al. 2005] Jilin Shi and Yinggui Zhu, *Lexicon of Chinese Positive Words*, SiChuan Dictionary Press, 2005.

[Xu et al. 2007] Ruifeng Xu, Yunqing Xia, Kam-Fai-Wong and Wenjie Li, Annotating opinions in customer reviews, submitted to *ACL 2007 workshop on Language Annotation*, 2007.