International Journal of Innovative Computing, Information and Control
Volume **11**, Number **4**, August **2015**

 ISSN 1349-4198
pp. **1239–1255**

RESEARCH ON OPINION HOLDER EXTRACTION OF UYGHUR

Long Yu[1], Xiangchao Duan[2,*], Shengwei Tian[3], Turgun·Ibrahim[2] and Askar·Hamdulla[3]

[1]Network Center
[2]School of Information Science and Engineering
Xinjiang University
No. 14, Shengli Road, Tianshan District, Urumqi 830046, P. R. China
{ yulong_xju; Turgun_xju }@126.com
*Corresponding author: ywLuo_pleasant@126.com

[3]School of Software
Xinjiang University
No. 999, Northwest Road, Saybagh District, Urumqi 830008, P. R. China
{ swTian_xju; Askar_xju }@126.com

Received October 2014; revised February 2015

Abstract. *Opinion holder extraction aims to extract the entities that express opinions in opinion sentences. On the basis of analyzing the Uyghur grammatical characteristics and rules, the Uyghur comments are viewed as research objects, and a fine-grained three-layer model of opinion holder extraction is proposed. CRFs (Conditional Random Fields) model is used to identify all the opinion holder candidates of each comment, combined with the manual heuristic rules and Uyghur name composition rules. Then the opinion sentences are classified into four different types according to the classification algorithm, and different extraction methods are put forward for the corresponding opinion holder type, respectively. The expansion rules are introduced to mend holder extraction results. The experimental results show that the average precision rate is 80.14%, and the average recall rate is 84.39%, which indicate the efficiency and feasibility of the proposed method of opinion holder extraction.*

Keywords: Opinion holder extraction, Opinion mining, Opinion sentences, Heuristics rules

1. **Introduction.** Opinion is the subjective evaluation for certain things or certain facts made by entities, such as people, organizations, and institutions. With the rapid development of Web 2.0, more and more people start to express their views online; they comment the commodities bought via the Internet, and publish their views and opinions on the Blog or news platform. These comments play an important role in business, government and customers. How to extract useful information from a great number of comments has become the focus of researches. Therefore, opinion mining gradually becomes a research hotspot of natural language processing, computational linguistics and text processing. Opinion mining is to extract the opinion holder, topic, opinion word and determine the polarity and strength of opinion word by analyzing the numerous comments information. From the current researches of many scholars, opinion mining contains the following four main subtasks:

(1) Topic extraction; (2) Opinion holder extraction; (3) Claim selection; (4) Sentiment analysis.

Opinion holder extraction is an important subtask of opinion mining, and a significant component of opinion question answering system. Opinion holder extraction aims

to extract entities that express opinions in opinion sentences, and it is very important for sentiment analysis. However, research work of opinion holder extraction is relatively small compared with the other three subtasks. How to accurately extract the opinion holders of commentary statements has become a problem. Opinion holder extraction is an important element of opinion analysis; opinion analysis is incomplete without opinion holder extraction. Opinions holder extraction has an important reference value for merchants and online shoppers, and also plays an important role in government's public opinion supervision at the same time. For major languages such as English and Chinese, opinion mining has achieved a rich harvest, but the research for minority languages like Uyghur is relatively small. The lack of research on opinion holder extraction has a deep influence on the further mining of minority language information, and also hinders the development of minority language information technology. We researched the basic theory of opinion holder extraction of opinion mining; this will lay the foundation of the future application like opinion mining systems.

In recent years, many scholars studied on opinion holder extraction deeply, and obtained fruitful results. Some scholars adopted syntactic analysis tree based method [1-6], and combined different models (Maximum Entropy Ranking, CRFs) to identify opinion holders. These experiments have achieved very good results, but these methods rely on the results of syntax analysis heavily, while the current parsing technology of minority language such as Uyghur is not as good as English, so the methods based on syntactic analysis have certain language limitations. Many scholars considered that machine learning methods are also the better ways to extract opinion holder [7-12], and they combined machine learning models (Maximum Entropy, Support Vector Machine, CRFs) with heuristic rules, used opinion indicators, opinion operators, adverbs, and opinion words as features to extract opinion holders. The extraction results of these methods depend largely on the comprehensiveness of heuristic rules, as well as the size of experimental corpus. Xu et al. [13] used some heuristics to recognize the core of opinion holders, and adopted some heuristics rules and patterns to expand the opinion holder from its core. The recognition of opinion holders is highly dependent on the comprehensiveness of heuristics rules. Kim and Hovy [14] established a mapping table with annotated corpus, and mapped the topic word to the opinion holder according to semantic role. This method can get a high accuracy, but it will miss the opinion holders outside of the mapping table, and the coverage is relatively small. Wiegand and Klakow [15] proposed a method of opinion holder extraction based on convolution kernels, defined the opinion holder boundary which is simple and clear. They used different kernel forms in the extraction process and made a comparison. Kim et al. [16] put forward an anaphor resolution based opinion holder identification method exploiting lexical and syntactic information. The experiment achieved good results, but this method did not consider the relation between antecedent and personal pronoun, and ignored the semantic features. Zhang et al. [17] proposed a method to automatically identify opinion sentences and opinion holders in Internet public opinion, they established a series of related resources to opinion analysis, used opinion operator as a key indicator to extract opinion holder of a given sentence, and utilized pattern matching to expand opinion holders. Dipankar and Sivaji [18] defined the annotation agreements of opinion holder, and proposed two different holder identification strategies for Bengali blog sentences: baseline model and syntax model. This method used part of speech and indicative verbs as features, but some other features are neglected such as semantic and position. Elarnaoty et al. [19] utilized a method which is independent from any lexical parsers. They used semi-supervised pattern recognition technology to analyze features and extract opinion holders. The precision rate of this method is high, but the recall rate is low, and the opinion holders of some opinion sentences are empty. Most of

the previous works are based on the following two kinds of methods: heuristic rule based and machine learning based methods. All of the methods have achieved good results, but they also have their own drawbacks. In order to solve these disadvantages, we need to explore new methods to extract opinion holders.

In this paper, we used Uyghur commentary statements of open fields as the research object, and proposed a fine-grained three-layer model of opinion holder extraction on the basis of analyzing the Uyghur grammatical characteristics and rules. The model refines the task of opinion holder extraction, and takes co-reference holder extraction work into account, which increase the precision rate, the recall rate, and the F1 values of opinion holder extraction. There is no very mature parser for Uyghur so far, which brings great difficulties to opinion holder extraction of Uyghur. The proposed model is independent from syntactic parser; we refined the extraction tasks, combined CRFs and heuristic rules to extract opinion holder candidates furthest, and solved the disadvantages of heuristic rule based methods.

This article is structured as follows. We summarize the relevant Uyghur grammatical characteristics and rules in Section 2. We describe the proposed fine-grained three-layer model of opinion holder extraction in Section 3. We conduct a further experiment to investigate its effectiveness and make a comparison with other methods in Section 4. Finally, we conclude this paper in Section 5.

2. **Uyghur Grammatical Characteristics and Rules.**

Uyghur is a kind of adhesive language; it belongs to the Turkic languages of Altai language family. Uyghur has complex morphological changes and grammatical forms; the different grammatical functions of Uyghur are realized by connecting different affixes before or after a stem. Uyghur has its own unique characteristics in stem affix, word order and personal pronoun, etc. In order to make the results of opinion holder extraction more accurate, and reduce the information omissions of opinion holder extraction, we summarized the following Uyghur related grammatical characteristics and rules and apply them to the process of opinion holder extraction.

2.1. **Uyghur name composition rules.**

(1) Uyghur names have no specific surname, Uyghur people adopt the patronymic linkage naming system, and he (or she) uses his (or her) father's given name as their surname. The full name of Uyghur people is composed by his (or her) given name and his (or her) father's given name. The forms of Chinese names and Uyghur names are different. For Chinese names, surname is in the first place, given name is behind the surname, that is surname + given name, for example: "李红 (Li Hong)". The order of Uyghur name is opposite, given name is in the first place, surname is behind the surname, and they are connected by a separation dot, for example: someone's given name is "سەنەم (sanam)", his (or her) father's given name is "مامات (mamat)", so his Uyghur name is "سەنەم . مامات (sanam·mamat)". In the process of opinion holder extraction, if two Uyghur names are connected by a separation dot, we will combine the two names into one name.

(2) Uyghur names essentially consist of 2-4 syllables, there are no names of one syllable, and names of five syllables are rare.

(3) Uyghur names are essentially constituted by the Arabic and Persian words; although there are some changes in the structure, the stems are still Arabic or Persian.

(4) Uyghur female names are generally constituted by the combinations of some common words such as "گۈل (gul)", "قىز (kiz)", "خان (han)" with other words; the common words also can be used as names alone.

(5) Uyghur male names are generally constituted by the combinations of some common words such as "مۇھەممەد (Mohammed)", "ئاخون (ahon)" with other words; the common words also can be used as names alone.

2.2. **Word order rule of Uyghur sentence.** The normal word order of Uyghur sentence is "subject + object + predicate", namely the predicate is behind the subject and object; in addition, attributive and adverbial are located before each central word, for example (Uyghur writing order is from right to left):

ئۇ قىشنى ناھايىتى ياماق كۆرىدۇ.

(He hates winter very much.)

However, there is a special case for Uyghur; the predicate can be before the object, also can be behind the object. For example:

ئالىمنىڭ ئېيتىشچە ئۈرۈمچىنىڭ قاتنىشى بەك ناچار.

ئالىم ئۈرۈمچىنىڭ قاتنىشى بەك ناچار ئېيتتى.

(Alim said the traffic of Urumqi is too bad.)

The above two examples express the same meaning, "ئېيتىشچە (said)" can be in the second position (the second word of the first sentence); "ئېيتتى (said)" can also be in the final position (the last word of the second sentence). This sentence contains an opinion holder, "ئېيتىشچە, ئېيتتى (said)" is the opinion indicator, its position has a great influence on determining the opinion holder, so in the experiment, we will give full consideration to the unique characteristics of predicate location and apply them to the opinion holder extraction process.

2.3. **Characteristics of stem and affix.** Stem is the linguistic unit that can express the basic meaning of a word, and cannot be decomposed any more. Affix is the morpheme connected to the root or stem to derive new words. Removing the configuration affixes from one word, the remaining part is the stem. Uyghur words are constituted by connecting affixes and suffixes to stem according to requirements. Uyghur affixes have certain grammatical functions, Uyghur is a kind of adhesive language, and so one word can have more than one configuration morphemes to express complex semantics. For Uyghur, affixes can express the meanings that words express, and the meanings of some Uyghur verbs are often expressed by the affixes appended after the subject, for example:

تۇرسۇنلىشىچە

(Tursun considers)

From the above example we can see that affix "لىشىچە" appends after the name "تۇرسۇن (Tursun)", it also expressed the meaning "Tursun considers", that is, the affix takes the place of a word; it also expresses the meaning of the word. In Uyghur expressions, this form accounts for a high proportion, but the stem "تۇرسۇن (Tursun)" is the real opinion holder in "تۇرسۇنلىشىچە (Tursun considers)", so we need to extract the stems of opinion holder candidates to improve the precision rate of the experiment.

2.4. **Case forms of personal pronouns.** The features of Uyghur personal pronouns are mainly reflected by the case forms. The case forms of personal pronouns contain 10 forms, such as nominative. Different case form has different "case affix"; Uyghur personal pronouns consist of different singular and plural forms and different rhetorical significance types (normal, respectful, honorific, and scornful) according to different person. The distribution of Uyghur personal pronouns, singular and plural, and rhetorical significance is shown in Table 1.

As we can see from the table, the biggest difference between Uyghur and English personal pronouns is that there is no distinction of Uyghur third person singular. In English,

TABLE 1. Singular, plural and rhetorical meaning of Uyghur personal pronouns

First person	singular		مەن/I, ئۆزۈم/myself
	plural		بىز/we, ئۆزىمىز/ourselves
Second person	singular	normal	سەن/you, ئۆزۈڭ/yourself
		respectful	سىز/you, ئۆزىڭىز/yourself
		honorific	سىز/you
	plural	normal	سىلەر/you, ئۆزۈڭلار/yourselves
		respectful	سىلى(ئۆزلىرى)/yourselves
		scornful	سەنلەر/you
Third person	singular		ئۇ/he, she, it ئۆزى/himself, herself, itself
	plural		ئۇلار/they

the third person contains "he", "she" and "it", but for Uyghur, the third person can mean both men and women, and it can also refer to things, which brings a big difficulty to opinion holder extraction. However, the singular and plural forms of Uyghur personal pronouns have obvious features, the plural forms of Uyghur personal pronouns are usually represented by the affixes, if a word is connected by the affixes like "لار", "لەر", we will affirm the word is plural.

3. Opinion Holder Extraction.

3.1. **Introduction of opinion holder.** Opinion holder extraction is to identify people organizations or institutions that express opinions in evaluation sentences. Opinion holders are generally named entities, including:

(1) Person name, for example: "ساۋۇت (sawut)";

(2) Organization name, for example: "دۇنيا سەھىيە تەشكىلاتى (World Health Organization)";

(3) Title, for example: "ئىقتىسادشۇناس (economist)".

However, opinion holders are not limited to them; they can also be some common noun phrases, for example: "زاۋۇت-ھاگىزىن (manufacturers)", "ئوقۇغۇچىلار (students)", "ئامېرىكا رەھبىرى (Ame-rican leaders)". Personal pronouns can also be opinion holders, for example: "ئۇ (he)", "ئۇلار (they)", "مەن (I)", so in the sentences that personal pronouns are opinion holders, personal pronouns are direct opinion holders, and we need to find the real opinion holders by contextual information, the objects that pronouns refer to. Opinion holders can be divided into two categories: explicit opinion holder and implicit opinion holder. We can find the true opinion holder in the sentence of explicit opinion holder, while there are no entities that express opinions in the sentence of implicit opinion holder; we also call this type default opinion holder.

Most researchers classify opinion holder into explicit type and implicit type, this kind of classification is simple to realize and easy to process corpus, but it is too general to get good experimental result. In order to refine the opinion holder extraction tasks, we classify opinion holder into the following four types more particularly: (1) Indicative verb type; (2) Co-reference type; (3) Punctuation mark type and (4) Implicit type. After analyzing a lot of Uyghur opinion sentences, we find that these four types of opinion holder can cover all of opinion holder types. We classify opinion holder into these four types, and put forward different extraction methods according to different types. This kind of classification will make the extraction process easy to realize and get better results.

(1) Indicative verb type

Indicative verb is a very critical feature to determine opinion holder, because it clearly

points to the opinion expresser, the opinion holder usually co-occurred with indicative verbs certain pattern, for example:

ئەركەن • توختنىڭچە بۇ ئاپپرات بەك ياخشى ئىكەن.

(Arken·toheti thinks this camera is very good.)

We also call this type **Type1**.

(2) Co-reference type

Opinion holder of co-reference type is a common type of sentence opinion. In these sentences, personal pronouns are the direct opinion holders for instance:

ئادىل بىر ئاپپرات ئالدى،ئۇنىڭچە باھاسى ناھايىتى مۇۋاپىق.

(Adili bought a camera, and he thought that the price is very appropriate.)

From the above example we can see that "ئۇ (he)" is the direct opinion holder, but we do not know who is the real opinion expresser only through pronoun "ئۇ (he)", so we need to use anaphora resolution technology to find the real opinion holder, namely the antecedent of pronoun. We also call co-reference type **Type2**.

(3) Punctuation mark type

Punctuation is also an important feature to determine opinion holder; there are two crucial punctuation marks: colon and quotation mark. Generally, there are no indicative verbs in these opinion sentences, because the punctuation mark takes the place of indicative verb, and plays the role of indicative verb. For example:

داغۇت:ئىقتىساد تەرەققىي قىلىشنىڭ سۈرئىتى ئاستىلىۋاتىدۇ.

(Dawut: the speed of economic development is slower.)

Obviously, "داغۇت (Dawut)" is the opinion holder of the sentence. We also call punctuation mark type **Type3**.

(4) Implicit type

If there is no real opinion expresser in one opinion sentence or there is a personal pronoun as the opinion holder, but there is no antecedent of the personal pronoun in the opinion sentence, we call the opinion sentence implicit type. For example:

بۇ يانفون بەك ياخشى ئىكەن.

(This telephone is very nice.)

مەن بۇ يەرنىڭ ھاۋاسىنى ناھايىتى ياخشى كۆرىمەن.

(I really like the weather here.)

In the first example, the sentence has no opinion expresser, in the second example, "مەن (I)" is the direct opinion holder, but there is no antecedent, we do not know who the real opinion holder is.

After analyzing the position and other characteristics of opinion holder in the corpus, we developed the following heuristic rules to identify the opinion holder candidates.

Rule1 Opinion holder must be a named entity.

Rule2 Opinion holder always occurs in the beginning or near the end of the sentence.

Rule3 Opinion holder has a strong association with certain indicative verb.

Rule4 Opinion holder always co-occurred with indicative verbs with certain pattern.

Rule5 Opinion holder usually appears before the colon, or in the beginning or end of the quotation mark.

Table 2 shows some opinion sentences of types and the applicative rules of each type.

In the process of the experiment, we fully considered the position feature, part of speech feature of opinion holder and indicative verb, and we also thought about contextual information, semantic feature, distance feature, opinion word and opinion holder modifier. Combining Uyghur grammatical characteristics and rules, we used CRFs model, heuristic

TABLE 2. Examples and applicative rules of each type

Type	Example	Applicative rules
Indicative verb type	ئەركەن • توختنىڭچە بۇ ئاپپارات بەك ياخشى كىكەن. (Arken·toheti thinks this camera is very good.)	Rule 1, Rule 2, Rule 4
Co-reference type	سەردار بىر تال لەپتوپ كومپيۇتېر سېتىۋالدى،ئۇ كومپيۇتېرنىڭ ئىقتىدارى يامان كىكەن (Seldan bought a laptop; he thought the performance is good.)	Rule 1, Rule 3, Rule 4
Punctuation mark type	داۋۇت:ئىقتىساد تەرەققىي قىلىشنىڭ سۈرئىتى ئاستىلاۋاتىدۇ. (Dawut: the speed of economic development is slower.)	Rule 1, Rule 5
Implicit type	مەن بۇ تېپتىكى يانفوننى بەك ياقتۇرىمەن. (I like this telephone very much.)	

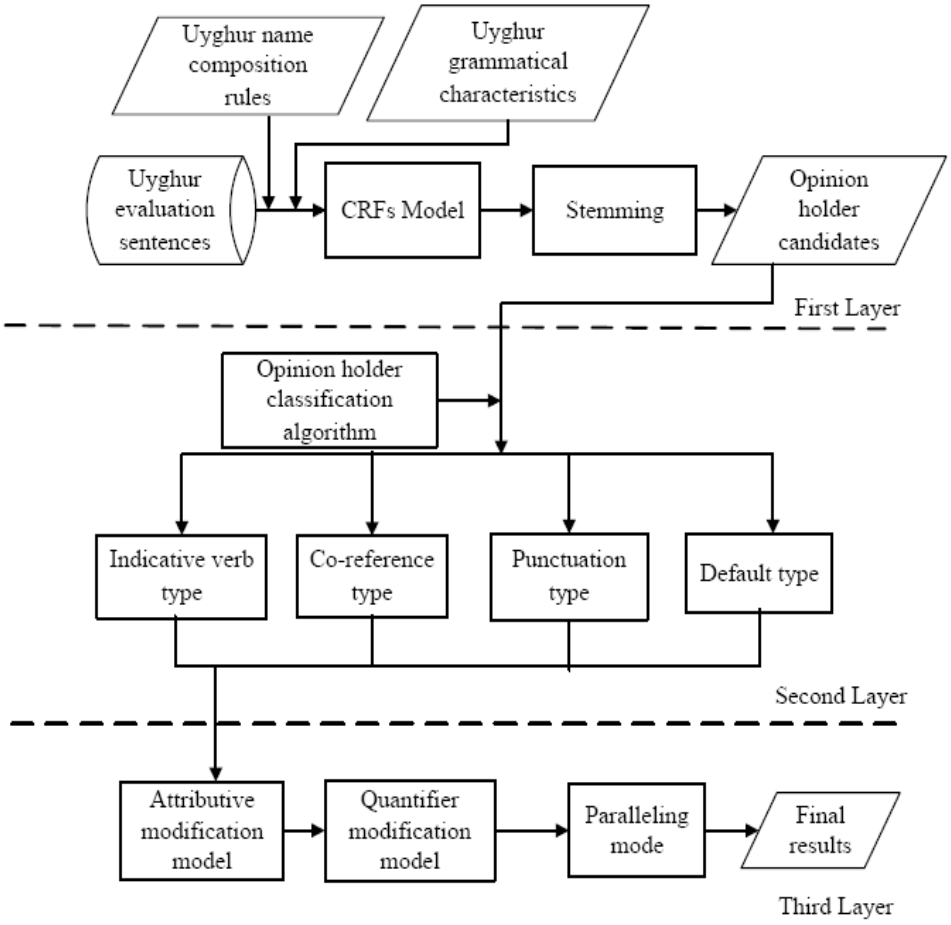

FIGURE 1. Three-layer model of opinion holder extraction

rules to extract opinion holder candidates, then the classification algorithm was introduced to classify the opinion sentences, and different extraction methods are put forward for the corresponding opinion holder type. At last, the expansion rules are proposed to mend opinion holder extraction results. The three-layer model of opinion holder extraction is shown in Figure 1.

3.2. **Candidate extraction.** CRFs (Conditional Random Fields) [20] is a kind of undirected graph model that uses the given input sequence to predict the output sequence, and it is widely used in the sequence labeling in recent years. Given a set of observing sequences that needs to be labeled, we can use CRFs to predict the joint probability distribution of the sequence to be labeled, for the observation sequence x and the state

sequence y, we can define a linear CRFs model as follows:

$$P(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{i,n} \lambda_n f_n(y_{i-1}, y_i, x) + \sum_{i,n} \mu_n g_n(y_i, x)\right) \tag{1}$$

Z is a normalization factor:

$$Z(x) = \sum_{y} \exp\left(\sum_{i,n} \lambda_n f_n(y_{i-1}, y_i, x) + \sum_{i,n} \mu_n g_n(y_i, x)\right) \tag{2}$$

$f_n(y_{i-1}, y_i, x)$ is the feature of output node i and $i-1$ in the observation sequence x, each $g_n(y_i, x)$ is the feature of input node and output node i, λ and μ are the weights of feature functions.

In this paper, we combined Uyghur grammatical characteristics, rules and CRFs model to extract all the opinion holder candidates for each comment. For feature selection, we considered the part of speech, position, opinion word and indicative verb, and also took the semantics, contextual information and other features into account. Table 3 shows some of the features we selected in the process of opinion holder extraction.

TABLE 3. Part of the features of opinion holder extraction

Feature Type	Feature Name	Feature Description
Named entity	FIsPer	Is the current word a person name?
	FIsLoc	Is the current word a location name?
	FIsOrn	Is the current word an organization name?
	FIsBegin	Is the current word in the beginning of the sentence?
	FIsEnd	Is the current word in the end of the sentence?
Part of speech	FPOS	Part of speech of the current word
	FIsNoun	Is the current word a noun?
	FIsProN	Is the current word a personal pronoun?
Punctuation mark	FBeColon	Does the current word appear before a colon?
	FAfQuo	Does the current word appear after quotation mark?
Indicative verb	SHasInVerb	Is there an indicative verb?
	FDisINVerb	The distance between the current word and indicative verb
	PosInVerb	The position of indicative verb
Opinion word	POSOfOp	Part of speech of the opinion word
	FDisOp	The distance between the current word and opinion word
	PosOfOp	The position of opinion word

In order to reduce the occasionality of experimental results and solve the sparsity of experimental data, we used 5-fold cross-validation to extract opinion holder candidates, the corpus were divided into 5 parts averagely, we used 4 parts of which as the training data in turn, and the rest is viewed as the test data.

3.3. **Opinion holder extraction algorithm.** After extracting all the opinion holder candidates of each opinion sentence, we classified all the opinion sentences according to opinion holder classification algorithm, and put forward different extraction methods for the corresponding opinion holder type, respectively, and the specific classification algorithm is shown as follows:

```
Input: allSentenceSet, Output: Type1Set, Type2Set, Type3Set, DefaultSet
1 Type1Set = Null, Type2Set = Null, Type3Set = Null, DefaultSet = Null
2 for each sentence in allSentence
3     if (HolderCanSet != Null && ProNounSet != Null &&IndiVerbSet != Null)
4         Type2Set.Add(sentence)
5     else if (HolderCanSet != Null&& ProNounSet == Null && IndiVerbSet != Null)
6         Type1Set.Add(sentence)
7     else if (HolderCanSet != Null && PuncSet != Null && IndiVerbSet == Null)
8         Type3Set.Add(sentence)
9     else
10        DefaultSet.Add(sentence)
11 end
12 return Type1Set, Type2Set, Type3Set
```

allSentenceSet is the set of all the opinion sentences, HolderCanSet is the set of opinion holder candidates of each comment, ProNounSet is the set of personal pronouns of each comment, PuncSet is the set of punctuation marks, IndiVerbSet is the set of the indicative verbs of each comment (in this paper, we summarized the indicative verb set, whether a sentence contains indicative verbs is judged by the indicative set, the specific indicative verb set is shown in Section 3.3.1), Type1Set is the set of opinion sentences of indicative verb type, Type2Set is the set of opinion sentences of co-reference type, Type3Set is the set of opinion sentences of punctuation mark type, and DefaultSet is the set of opinion sentences of default type.

In the process of experiment, the sentences of corpus are all opinion sentences, namely there are no non-opinion sentences in the corpus, so we will not extract the opinion holders of default type. We can get all the sentences of explicit opinion holder according to the classification algorithm, and the rest are the sentences of implicit opinion holder. There are no specific person names or organization names in the sentences of implicit opinion holder, so if the opinion holder candidate set of one sentence is empty, we treat the sentence as default type, and label the opinion holder as default opinion holder.

3.3.1. *Opinion holder extraction of indicative verb type.* Indicative verb is a very important feature to identify opinion holder, it is an indicator of opinion holder, and has a strong association with opinion holder. The indicative verb usually co-occurs with opinion holder certain pattern in opinion sentences. To improve the precision, we summarized the indicative verb set, and indicative verb can be a single word, also can be an affix. Indicative verb set contains 106 words or affixes; among them, there are 15 positive indicative verbs, 19 negative indicative verbs, and 72 neutral indicative verbs. Table 4 lists some of Uyghur indicative verbs.

For opinion sentence of indicative verb type, we proposed a weighted decision function model to select the best opinion holder candidate as the final opinion holder. On the

TABLE 4. Part of Uyghur indicative verbs

Type	Indicative verbs
Positive	“ماختىماق (commend)”, “ئامراق (like)”, “ئالقىشلىماق (praise)”, “مۇئەييەنلەشتۈرمەك (confirm)”, “قارشى ئالماق (welcome)”, “رازى بولماق (satisfied)” ⋯
Negative	“قارشى تۇرماق (oppose)”, “بىزار قىلماق (hate)”, “ئەيىبلىمەك (censure)”, “ئىنكار قىلماق (deny)”, “رەت قىلماق (refuse)”, “تەنقىد قىلماق (criticize)” ⋯
Neutral	“كۆرسىتىشچە (point out)”, “ھېس قىلىشچە (feel)”, “قارىشچە (think)”, “ئىپادە بىلدۈرۈشچە (consider)”, “ئېيتىشچە (say)”, “بايان قىلماق (expound)” ⋯

basis of analyzing the position, part of speech, and co-occurrence with indicative verb of opinion holder, we selected three features as the elements of weighted decision function. If opinion sentence S_i has opinion holder candidates $h_1, h_2, \cdots, h_m$, we can select the best opinion holder h_i by weighted decision function model. Formulas (3), (5), (6) are the feature functions, h_i is the ith opinion holder candidate, $n(h_i, v_j)$ is the co-occurrence times of opinion holder candidate h_i and indicative verb v_j, N is the occurrence times of all the indicative verbs in the corpus, $[-3, +3]$ is the size of sliding window that regards h_i as the center, the radius of window is determined by experiment, if the window is too small, the contextual information will be omitted, if the window is too big, the relation strength between opinion holder and indicative verb will be weakened, the meaning of slide window will be lost. The determination method of the window is shown in Section 4.2.

(a) Position feature function of opinion holder

$$Position(h_i) = \begin{cases} 1, & \text{if } h_i \text{ appears at the begining or the end of the sentence} \\ 0, & \text{others} \end{cases} \tag{3}$$

(b) Feature function of indicative verb

$$IndiVerb(h_i, v_j) = \begin{cases} 1 + \frac{n(h_i,v_j)}{N} * \log(1 + \frac{n(h_i,v_j)}{N}), \\ [-3, +3] \text{ or } v_j \text{ appears at the end of the sentence} \\ 0, \text{ others} \end{cases} \tag{4}$$

(c) POS feature function of opinion holder

$$POS(h_i) = \begin{cases} 1, & \text{if } h_i \text{ is a noun or a noun phrase} \\ 0, & \text{others} \end{cases} \tag{5}$$

(d) Weighted decision function

$$DeFun(h_i) = \alpha Position(h_i) + \beta IndiVerb(h_i, v_j) + \gamma POS(h_i) \tag{6}$$

$Position(h_i)$, $IndiVerb(h_i, v_j)$, $POS(h_i)$ are calculated through formulas (3), (4), (5), α, β, γ are the weights of feature functions, and $\alpha + \beta + \gamma = 1$, α, β, γ represent the importance degree in the weighted decision function, the values of α, β, γ are determined in Section 4.2, we calculate the $DeFun(h_i)$ value of each opinion holder candidate by the weighted decision function, and select the maximal one as the final opinion holder.

3.3.2. *Opinion holder extraction of co-reference type.* Co-reference is a common phenomenon of natural language; in the opinion sentences of co-reference type, personal pronouns (major third person) are the direct opinion holder, person names or organization names are the indirect opinion holder, namely the antecedents of pronouns. We do not know who the real opinion holder is only through pronouns, so we need to use anaphora resolution technology to find the real opinion holder.

Strube and Muller [21] adopted a decision tree based approach to pronoun resolution in spoken dialogue. They presented a set of features designed for pronoun resolution in spoken dialogue and determine the most promising features. Li et al. [22] proposed a supervised pronoun anaphora resolution system based on factorial hidden Markov models (FHMMs). Cheery and Bergsma [23] adopted an unsupervised Expectation Maximization approach to pronoun resolution. Compared with their methods, the proposed co-reference method in this paper is more pertinent and meticulous; we proposed these rules on the basis of Uyghur grammatical characteristics.

Personal pronouns and antecedents have a very intimate relation in semantic, distance and other features. We make the quantification dispose to the semantic, position, singular and plural, distance features:

(1) Consistency of semantic. Semantic value is represented with SenWeight, the antecedent of personal pronoun is a noun or noun phrase, so the personal pronoun must stay the same with antecedent; "ئۇ (he)" cannot indicate location name or organization name, for example: "ئۇ (he)" cannot indicate "ئۈرۈمچى (Urumqi)", because the former points to a person, the later points to a location name. If opinion holder candidate h_i has consistency with personal pronoun, semantic value is 2; otherwise the value is 0.

(2) Consistency of singular and plural. The value of singular and plural is represented with NumWeight, opinion holder can be a person, also can be a group, so the personal pronoun must stay the same with antecedent in singular and plural form. For example: "ئۇ (she)" cannot point to "ئوقۇغۇچىلار (students)", because the former points to a single person, the later points to a group. The attribute of singular and plural can be identified by a lot of modifiers, such as "نۇرغۇنلىغان (many)", "ھەر بىر (each)". If opinion holder candidate h_i has consistency with personal pronoun in singular and plural form, the value of singular and plural is 3; otherwise the value is 0. We set the weight of singular and plural to be 3, slightly larger than the position weight of opinion holder and the weight of semantic feature, the reason is that singular and plural information is the most important one in the four features, and it is relatively easy to determine. If the value of singular and plural is too large, the effects of other features will be reduced.

(3) Position feature of opinion holder. The value of position feature is represented with PosWeight, position feature of opinion holder is a two-value function, if opinion holder candidate h_i is at the beginning or the end of a sentence, the value of PosWeight is 1; otherwise the value is 0.

(4) Distance feature. Distance feature includes sentence distance (DisSenWeight) and reference distance (DisWordWeight). If opinion holder candidate h_i is in sub-sentence ss_i, the personal pronoun is in sub-sentence ss_j, the DisSenWeight value is $|i-j|$; if opinion holder candidate h_i and personal pronoun are in the same sub-sentence, the DisSenWeight value is 0. Reference distance indicates the relative distance between opinion holder candidate h_i and personal pronoun, not the word numbers between them. DisWordWeight value is calculated with Formula (7):

$$DisWordWeight = \frac{n}{N} * SenNum \tag{7}$$

n is the word numbers between opinion holder candidate h_i and personal pronoun, N is the total word numbers of the sentence, $SenNum$ is the sub-sentence numbers of the opinion sentence.

The total weight value of opinion holder candidate is calculated with Formula (8):

$$\begin{aligned} TotalWeight(h_i) = {} & SenWeight + NumWeight + PosWeight \\ & - DisSenWeight - DisWordWeight \end{aligned} \tag{8}$$

After quantifying the above features, we compute the total weight value of opinion holder candidates with the following algorithm:

Step1: Confirm the sentence unit of personal pronoun.

Step2: Calculate SenWeight + NumWeight + PosWeight of each opinion holder candidate of the sentence unit. If there are no opinion holder candidates in the sentence unit, turn to Step6.

Step3: If opinion holder candidate h_i is adjoined with h_{i+1}, and h_i is before h_{i+1}, and there are indicative verbs in the sentence unit, we will filter out opinion holder candidate h_{i+1}.

Step4: Calculate DisSenWeight value and DisWordWeight value of each opinion holder candidate of the sentence unit.
Step5: Select opinion holder candidate that has the maximal TotalWeight value as the final opinion holder.
Step6: Expand one sentence unit towards left of personal pronoun, calculate TotalWeight value of each opinion holder candidate in the sentence unit, and select the maximal one as the final opinion holder. If there are no opinion holder candidates in the sentence unit, turn Step7.
Step7: Expand one sentence unit towards right of personal pronoun, calculate TotalWeight value of each opinion holder candidate in the sentence unit, and select the maximal one as the final opinion holder. If there are no opinion holder candidates in the sentence unit, the extraction process will be finished.

3.3.3. *Opinion holder extraction of punctuation mark type.* In the opinion sentences of punctuation mark type, punctuation marks take the place of indicative verbs. So in the opinion sentences of punctuation mark type, if opinion holder candidate h_i is before and adjoined a colon, or after and adjoined quotation marks, we take h_i as the opinion holder of the sentence.

3.4. **Expansion rules.** After extracting opinion holders by the algorithms of Section 3.3, we find that some opinion holders are single word. In fact, a single word cannot express the meaning of opinion holder completely, so it is necessary to find the integrated opinion holder through extended rules.

(1) Expansion rule of attribute modification
In one opinion sentence, a named entity may have the attribute modification, for example: "دىپلوماتىيە مىنىستىرلىقنىڭ بايانچىسى خوك لېي (diplomacy ministry spokesman Hong Lei)", "دىپلوماتىيە مىنىستىرلىقنىڭ بايانچىسى (diplomacy ministry spokesman)" is the modification of "خوك لېي (Hong Lei)", "بايانچىسى خوك لېي دىپلوماتىيە مىنىستىرلىقنىڭ (diplomacy ministry spokesman Hong Lei)" is the real opinion holder.
(2) Expansion rule of quantifier
The modification of opinion holder may be a quantifier, while the quantifier changes the original meaning of opinion holder, for example: "بەزى ئوقۇغۇچىلار (some students)", "بىر قىسىم سودىگەرلەر (part of manufacturers)", the quantifier limits the boundary of opinion holder. So we use the expansion rule of quantifier to determine the accurate boundary of opinion holder, and improve the precision of experiment.
(3) Expansion rule of paralleling model
In the process of opinion holder extraction, we find that some opinions are expressed by two persons, the two opinion holders are connected with conjunction, for example:

قۇرنان بىلەن ئەركىنىڭچە شىنجاڭنىڭ مەنزىرىسى بەك چىرايلىق.

(Kurban and Arken think that the landscape of Xinjiang is beautiful.)

Obviously, both "قۇربان (Kurban)" and "ئەركىن (Arken)" are opinion holders, they are connected with "بىلەن (and)", but we can only extract one of them, to improve the precision rate and reduce the information omission, we introduce expansion rule of paralleling model. If opinion holder candidates h_i and h_{i+1} are determined to be opinion holders, and they are connected with conjunctions such as "بىلەن (and)", we take both h_i and h_{i+1} as opinion holders.

4. Experimental Results and Discussion.

4.1. **Experimental corpus and part of speech labeling.** The experimental corpus comes from some big Uyghur business websites, and we collected 2 129 opinion sentences that cover product reviews, news reviews and person reviews. Table 5 shows the proportions of four opinion holder types in the corpus. We adopted the Uyghur word part of speech labeling system which is developed by our laboratory, and amended the labeling results manually. Table 6 shows the number of tokens, number of opinion holders, number of names and number of indicative verbs in the corpus. The precision of word part of speech labeling reaches over 95%, which satisfies experimental demand. Compared with English and Chinese, the corpus collection of Uyghur is more difficult, so we are still working on the collection work.

Table 5. Proportions of four opinion holder types

Type	Indicative verb type	Co-reference type	Punctuation mark type	Implicit type
Number	715	528	300	586
Proportion (%)	33.58	24.8	14.09	27.52

Table 6. Some detailed data of corpus

Type	Token	Opinion holder	Name	Indicative verb
Number	23419	1543	1672	1307

4.2. **Experimental results and analysis.** We use the common evaluation standards (Precision, Recall, and F1-measure) of natural language processing to evaluate the extraction result. The opinion holder extraction results of three types are shown in Table 7.

Table 7. Results of opinion holder extraction

Type	Precision (%)	Recall (%)	F1-measure (%)
Type1	80.74	83.33	82.01
Type2	73.31	79.36	76.21
Type3	86.36	90.48	88.37
Average	80.14	84.39	82.20

We can see from the above table that the average precision rate, the average recall rate, and the average F1 values of three types are over 80%, which indicate the efficiency and feasibility of the proposed method of opinion holder extraction. The precision rate of punctuation mark type is 86.36%, the reason is that the sentence structure of punctuation mark type is simple; it is easy to determine the opinion holder of punctuation mark type. The precision rate of co-reference type is 73.31%, the lowest one in three types; that is because the sentence structure of co-reference type is complicated, and we need to analyze the semantic, position, singular and plural, distance features in the process of opinion holder extraction.

The experimental results of weights selection of weighted decision function in the opinion holder extraction process of indicative verb type are shown in Figure 2; α, β, γ represent the importance degree in the weighted decision function, in the 10 combinations of α, β, γ, we can know that when $\alpha = 0.3$, $\beta = 0.4$, $\gamma = 0.3$, the precision rate, the

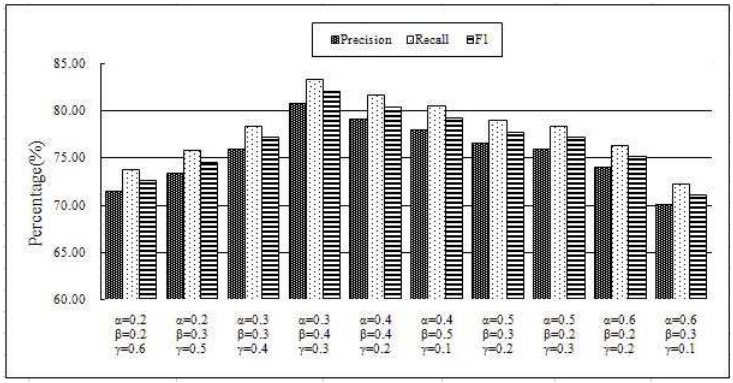

FIGURE 2. Weights selection result of weighted decision function

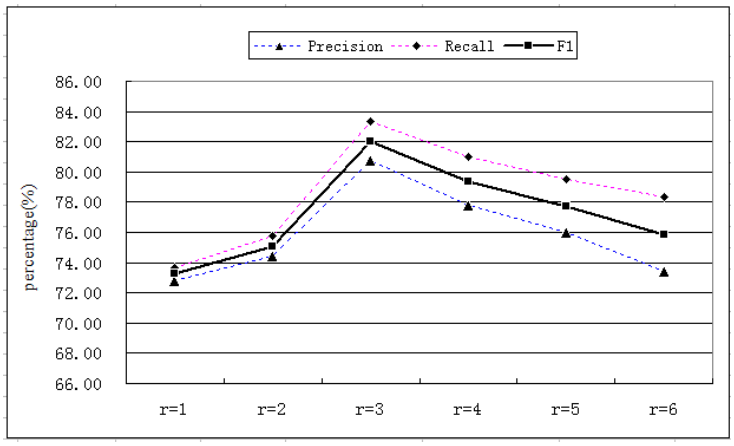

FIGURE 3. Experimental result of window radius selection

recall rate and F1 value are the best. $\beta = 0.4$ proves that the importance of indicative verb position is bigger than other two features.

Figure 3 is the experimental result of window radius selection in the opinion holder extraction process of indicative verb type; from the figure we can see that the precision rate, the recall rate and F1 value reach the highest when $r = 3$. If the window is too small, the contextual information will be omitted, if the window is too big, the relation strength between opinion holder and indicative verb will be reduced, and the meaning of slide window will be lost.

In order to verify the efficiency and feasibility of the proposed method in opinion holder extraction, we also use the methods of Ku et al. [12] and Elarnaoty et al. [19] to do the experiments with the same corpus, and compare the extraction results with ours. The results comparison is shown in Figure 4.

From the comparison of the experimental results, we can see that the precision rate, recall rate and F1-measure value are increased compared with Lun's method, the average precision rate is increased by 6.26%, the average recall rate is increased by 7.41%, the average F1-measure value is increased by 6.80%. The reason is that the proposed method refines the task of opinion holder extraction, the classification of opinion sentences is more meticulous, and different extraction methods are put forward for the corresponding opinion holder type, the extraction task is clearer. The average precision rate is reduced by 5.25% compared with Mohamed's method, while the average recall rate is increased by 13.05%, the recall rate of Mohamed's method is very low, and the opinion holders of some sentences of the explicit opinion holder type are empty, in fact, we can find the real

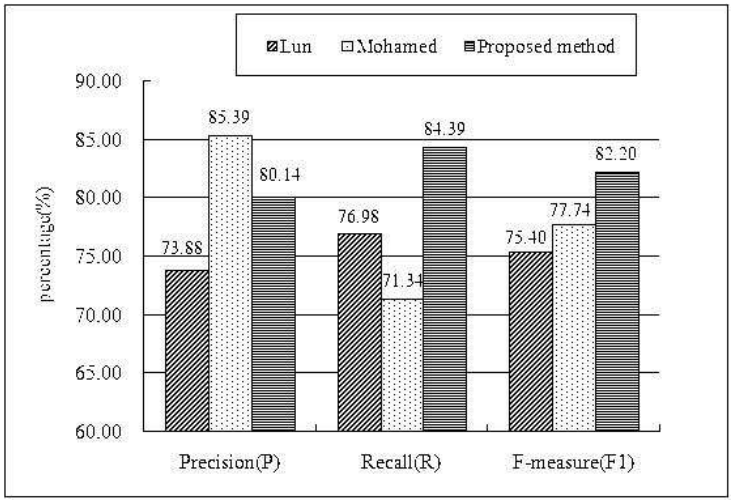

FIGURE 4. Experimental results comparison of opinion holder extraction

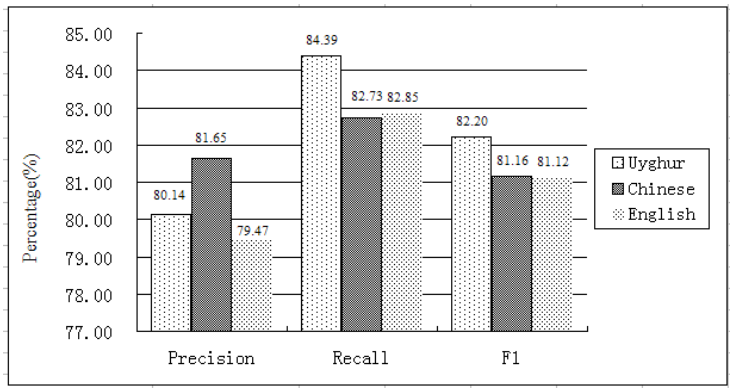

FIGURE 5. Experimental results comparison among Uyghur, Chinese and English

opinion holders in these sentences. The whole performance of proposed method is better than Mohamed's method.

In order to verify the efficiency and feasibility of the proposed method on other languages, we do the same experiment with Chinese corpus and English corpus, and make a comparison with Uyghur corpus. We collect 6 516 Chinese opinion sentences and 7 837 English opinion sentences, the comparison results of experiments are shown in Figure 5. In the experiments of Chinese and English, we still used the three-layer model to extract opinion holder, while we summarized different grammatical characteristics and rules which are suitable for Chinese and English. We consult Chinese linguistic experts and English linguistic experts, and summarized grammatical characteristics and rules for Chinese and English; these grammatical characteristics and rules have big differences with Uyghur grammatical characteristics and rules. For example: the word order of Chinese sentence is "subject + predicate + object", we can judge that opinion holder is always located before the indicative verb in Chinese opinion sentences according to this rule. For English, there is gender distinction of third person singular "he, she, it", so in the extraction process of co-reference type, gender is an important feature to extract for English opinion sentences. For Chinese corpus, the average precision rate is 78.35%, the average recall rate is 83.63%, and the average F1 is 80.90%. For English corpus, the average precision rate is 79.47%, the average recall rate is 82.85%, and the average F1 is 81.12%, which indicate the proposed method is applicative for different languages.

5. **Conclusions.** In this paper, we summarized Uyghur name composition rules, Uyghur word order rules, and other characteristics that can be applied to opinion holder extraction of Uyghur. A fine-grained three-layer model of opinion holder extraction is proposed, the model refines the tasks of opinion holder extraction, the classification of opinion sentences is more meticulous, and different extraction methods are put forward for the corresponding opinion holder type, the extraction task is clearer. We proposed a weighted decision function model to extract opinion holder of indicative verb type. On the basis of analyzing the relation of semantic, singular and plural, distance between antecedent and personal pronoun, we put forward a new algorithm to extract opinion holder of co-reference type. The experimental results show that the average precision rate is 80.14%, and the average recall rate is 84.39%, which indicate the efficiency and feasibility of the proposed method of opinion holder extraction.

In the future we will continue the further study in the following two aspects: (1) optimize the algorithm to the proposed method, and further improve the precision rate and recall rate of opinion holder extraction; (2) make a deep research on the sentences of multi-opinion holders.

Acknowledgments. This paper has been supported by the National Natural Science Foundation of China (Grant No. 61262064, 60963017, 61063026, 61063043, 61331011), National Social Science Foundation of China (Grant No. 10BTQ045, 11XTQ007).

REFERENCES

[1] B. Lu, Identifying opinion holders and targets with dependency parser in Chinese news text, *Proc. of the NAACL HLT 2010 Student Research Workshop, Association for Computational Linguistics*, Los Angeles, CA, USA, pp.46-51, 2010.

[2] S. M. Kim and E. Hovy, Identifying opinion holders for question answering in opinion texts, *Proc. of AAAI-05 Workshop on Question Answering in Restricted Domains, American Association for Artificial Intelligence*, Pittsburgh, US, pp.1367-1373, 2005.

[3] B. Lu, K. T. Benjamin and T. Jiang, Supervised approaches and dependency parsing for Chinese opinion analysis at NTCIR-8, *Proc. of NTCIR-8 Workshop Meeting*, Tokyo, Japan, pp.234-240, 2010.

[4] D. Dipankar and B. Sivaji, Emotion holder for emotional verbs-the role of subject and syntax, *Proc. of Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science*, Iasi, Romania, pp.385-393, 2010.

[5] L. Gui, R. F. Xu, J. Xu and C. X. Liu, A cross-lingual approach for opinion holder extraction, *Journal of Computational Information Systems*, vol.9, no.6, pp.2193-2200, 2013.

[6] A. Ancuta-Lenuta, L. Camelia, D. Mihaela and P. Rodica, Extracting opinion holders and targets in romanian texts, *Proc. of the 2014 IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*, Cluj Napoca, Romania, pp.37-42, 2014.

[7] M. Wiegand and D. Klakow, Generalization methods for in-domain and cross-domain opinion holder extraction, *Proc. of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics*, Avignon, France, pp.325-335, 2012.

[8] C. F. Wang, T. F. Ma, L. Q. Guo, X. J. Wang and J. W. Yang, PKUTM experiments in NTCIR-8 MOAT task, *Proc. of NTCIR-8 Workshop Meeting*, Tokyo, Japan, pp.228-233, 2010.

[9] K. Liu and J. Zhao, NLPR at multilingual opinion analysis task in NTCIR7, *Proc. of NTCIR-7 Workshop Meeting*, Tokyo, Japan, pp.226-231, 2008.

[10] W. Guo, R. Song and H. F. Lin, Opinion recognition and holder extraction based on SVM and distance-weighted computing, *Computer Engineering & Science*, vol.30, no.10, pp.125-128, 2008.

[11] R. Song, L. Hong and H. F. Lin, ChunkCRF-based opinion holder identification and application to opinion summarization, *Journal of Chinese Computer Systems*, vol.30, no.7, pp.1462-1466, 2009.

[12] L. W. Ku, C. Y. Lee and H. H. Chen, Identification of opinion holders, *Computational Linguistics and Chinese Language Processing*, vol.14, no.4, pp.383-402, 2009.

[13] R. F. Xu, K. F. Wong and Y. Q. Xia, Coarse-fine opinion mining – WIA in NTCIR-7 MOAT task, *Proc. of NTCIR-7 Workshop Meeting*, Tokyo, Japan, pp.307-313, 2008.

[14] S. M. Kim and E. Hovy, Extracting opinion, opinion holder and topics expressed in online media text, *Proc. of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, Association for Computer Linguistics*, Sydney, Australia, pp.1-8, 2006.

[15] M. Wiegand and D. Klakow, Convolution kernels for opinion holder extraction, *Proc. of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, Association for Computational Linguistics*, Los Angeles, CA, USA, pp.795-803, 2010.

[16] Y. Kim, Y. Jung and S. H. Myaeng, Identifying opinion holders in opinion text from online newspapers, *Proc. of 2007 IEEE International Conference on Granular Computing*, San Jose, CA, USA, pp.699-702, 2007.

[17] Y. F. Zhang, F. Long and B. Lv, Identifying opinion sentences and opinion holders in Internet public opinion, *Proc. of 2012 International Conference on Industrial Control and Electronics Engineering*, Saint Joseph, USA & Xi'an, China, pp.1668-1671, 2012.

[18] D. Dipankar and B. Sivaji, Finding emotion holder from Bengali blog texts-an unsupervised syntactic approach, *Proc. of the 24th Pacific Asia Conference on Language, Information and Computation*, Sendai, Japan, pp.621-628, 2010.

[19] M. Elarnaoty, S. AbdelRahman and A. Fahmy, A machine learning approach for opinion holder extraction in Arabic language, *International Journal of Artificial Intelligence & Applications*, vol.3, no.2, pp.45-63, 2013.

[20] C. Sutton, A. McCallum and K. Rohanimanesh, Dynamic conditional random fields: Factorized probabilistic models for segmenting and labeling sequence data, *Journal of Machine Learning Research*, vol.8, no.3, pp.692-723, 2007.

[21] M. Strube and C. Muller, A machine learning approach to pronoun resolution in spoken dialogue, *Proc. of the 41st Annual Meeting on Association for Computational Linguistics*, Sapporo, Japan, pp.168-175, 2003.

[22] D. C. Li, T. Miller and W. Schuler, A pronoun anaphora resolution system based on factorial hidden Markov models, *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics*, Portland, USA, pp.1169-1178, 2011.

[23] C. Cheery and S. Bergsma, An expectation maximization approach to pronoun resolution, *Proc. of the 9th Conference on Computational Natural Language Learning*, Ann Arbor, USA, pp.88-95, 2005.