Topic Identification in Chinese Discourse Based on Centering Model

Yi-Chun Chen and Ching-Long Yeh
Department of Computer Science and Engineering
Tatung University
40 Chungshan N. Rd. 3rd. Section
Taipei 104 Taiwan
yjchen7@ms7.hinet.net chingyeh@cse.ttu.edu.tw

Submitted on 9 January, 2007 Accepted on 2 November, 2007

Abstract

In this article we are concerned with identifying topics of utterances in texts, which are discourse elements reflecting the links between a sentence and its context. The information carried by the topics can be used to contribute to a number of natural language processing applications, such as information retrieval, text categorization and discourse segmentation etc. However, the phenomenon of zero anaphora frequently occurs in Chinese texts, topics in utterances are frequently omitted from expressions, due to their prominence in discourse. For solving the problem of zero anaphora, we employ the key elements of the centering model of local discourse coherence to extract structures of discourse segments and then identify the topics of utterances in discourse.

Keywords

Topic Identification, Zero Anaphora Resolution, Centering Model

1. Introduction

One of the most striking characteristics in a topic-prominent language like Chinese is the important element, “topic,” in a sentence which can represent what the sentence is about (Li and Thompson 1981). That is, if we can identify the topics of utterances, we can obtain the most salient information embedded in text. The most salient element in computational linguistics can be referred to as the *focus* (Sider 1979) or *cente*r (Grosz and Sidner 1986; Grosz *et al*. 1995; Walker 1998). The concept behind theories of focus or center relies on the observation that a discourse is structured around a central topic. The topic usually remains prominent for a few utterances until the topic shifts to a new one. Another key concept is that the center of an utterance is typically pronominalized. This hypothesis affects the interpretation of pronouns which often refer to the center established in the preceding utterances within a discourse segment (Grosz *et al*. 1995).

In this article, we tend to identify the topic of each utterance within a discourse based on the centering model (Grosz *et al*. 1995). However, in many natural languages, elements that can be easily deduced by the reader are frequently omitted from expressions in texts. The elimination of anaphoric expressions is termed zero anaphor (ZA) which often occurs in Chinese texts, due to their prominence in discourse (Li and Thompson 1981). Accordingly, to identify the topic of each utterance in a discourse, we have to solve the problem of zero anaphora resolution. For example in (1) the subject of the utterance (1a) is 小柯 'Xiaoke,' which is eliminated in the following utterances (1b), (1c) and (1d).[1]

(1) a. 小柯^{i} 看到 助理 的 留言，
Xiaoke^{i} kandao zhuli de liuyan.
Xiaoke see assistant GEN message
Xiaoke noticed the message from the assistant.

b. φ^{i}_{1} 發現 他 的 事 已經 曝光，
φ^{i}_{1} faxian ta de shi yijing puguang.
(Xiaoke) find he GEN thing already expose
(Xiaoke) found that his incident has been exposed.

c. φ^{i}_{2} 急忙 走出 辦公室，
φ^{i}_{2} jimang zouchu bangongshi.
(Xiaoke) hurry step-out office
(Xiaoke) hurriedly stepped outside of the office.

d. φ^{i}_{3} 看到 了 一 群 記者 向 他 衝 過來 。
φ^{i}_{3} kandao le yi qun jizhe xiang ta chong guolai.
(Xiaoke) see ASPECT a group reporter face he rush come-over
(Xiaoke) saw that a group of reporters rushed to him.

In the following sections we first describe the role of topic in Chinese grammar and that Chinese is a topic-prominent language. In Section 3 we introduce the centering model. In Section 4 we briefly describe the nature of zero anaphora in Chinese and the method of zero anaphora resolution. In Section 5 we describe the method topic identification and then show the experiments and result. Finally our conclusions are summarized, and future works are suggested.

[1] We use a φ^{a}_{b} to denote a zero anaphor, where the subscript *a* is the index of the zero anaphor itself and the superscript *b* is the index of the referent. A single φ without any script represents an intrasentential zero anaphor. Also note that a superscript attached to an NP is used to represent the index of the referent.

2. Topic Prominence in Chinese

In Chinese, in addition to the grammatical relations of "subject" and "direct object," the description of Chinese must also include the important element, "topic" (Li and Thompson 1981). The topic of a sentence often comes first in the sentence, and it always refers to something about which the writer assumes that the person reading the sentence has some knowledge. The subject of a sentence is the noun phrase that has a "doing" or "being" relationship with the verb in the sentence. For example in (2), 那台電腦 'that computer' is the topic, while 張三 'Zhangsan' having a "doing" relationship with the verb 修 'fix' is the subject of the sentence.

(2) 那 台 電腦 張三 修 過 了。
na tai diannao Zhangsan xiu guo le.
that CL computer Zhangsan fix ASPECT CRS
That computer Zhangsan has fixed.

By distinguishing topics and subjects in sentences, we have the four following types of sentences: sentences with both subject and topic, sentences in which the subject and the topic are identical, sentences with no subject, and sentences with no topic, which are exemplified in (3) to (6), respectively (Li and Thompson 1981).

(3) 那 本 書 我 已經 讀 過 了。
na ben shu wo yijing du guo le.
that CL book I already read ASPECT CRS
That book I have already read.
(4) 張三 打 我 。
Zhangsan da wo.
Zhangsan hit I
Zhangsan hit me.
(5) 衣服 燙 完 了。
yifu tang wan le.
cloth iron finish CRS
The clothing (someone) has finished ironing it.
(6) 進來 了 一 個 人。
jin-lai le yige ren.
enter-come ASPECT one CL person
A person came in.

Sentence (6) is an example of a "presentative sentence." In such sentence, the subject is usually an indefinite noun phrase, which cannot occur in sentence-initial position and cannot be a topic. Instead, the indefinite subject noun phrase must be placed after the verb. Besides, topics in Chinese sentences must be either definite or generic (Li and Thompson 1981). Consequently, the only noun phrase in (6), 一個人 'one person' is clearly the subject of the verb 進來 'come in', but it is not the topic because it is neither definite nor generic. It introduces a previously unknown entity, i.e., new information, into the discourse.

Another type of sentence is without a topic because the topic can be understood by the reader and is omitted from expressions in the context (Li and Thompson 1981; Huang 1994). For example in (1), the utterances (1b) to (1d) do not contain the subjects/topics but refer to 小柯 'Xiaoke' in the preceding utterance (1a). The situation in which noun phrases are unspecified is the *topic chain*, where the topic established in the first utterances serves as the referent for the unrealized topics in the chain of utterances following it.

Topic, as a discourse element, can simply relate to some part in the preceding utterance, introduce a subtopic which is related to what has been discussed, or reintroduce a topic that has been mentioned earlier. All of the above cases except the example (6)[2] involve a noun phrase that refers to an object mentioned earlier in the sentence or in a previous sentence. This noun phrase is called an anaphor. In addition to topic, anaphors in general can occur in other positions in a Chinese sentence and can be expressed in one of zero, pronominal and nominal forms. In our work in this article, we focus on the topic identification, while only the occurrences of ZAs are resolved.

3. Centering Model

Centering has its computational foundations established by Grosz and Sidner (Grosz 1977; Sidner 1979) and was further developed by Grosz *et al*. (Grosz *et al*. 1983; Grosz and Sidner 1986). Within the framework of the centering model, each utterance U in a discourse segment has two structures associated with it, called forward-looking centers, $C_f(U)$, and backward-looking center, $C_b(U)$. The forward-looking centers of U_n, $C_f(U_n)$, depend only on the expressions that constitute that utterance. They are not constrained by features of any previous utterance in the discourse segment (DS), and the elements of $C_f(U_n)$ are partially ordered to reflect relative prominence in U_n. The more highly ranked an element of $C_f(U_n)$, the more likely it is to be $C_b(U_{n+1})$. The highest ranked element of $C_f(U_n)$ that is realized[3] in U_{n+1} is the $C_b(U_{n+1})$.

The set of forward-looking centers, C_f, is ranked according to discourse salience. The highest ranked member of the set of forward-looking centers is referred to as the preferred center[4], C_p (Brennan *et al*. 1987). The preferred center of the utterance U_n represents a prediction about the C_b of the following utterance U_{n+1} and is the most preferred antecedent of an anaphoric or elliptical expression in U_{n+1}. Hence, the most important single construct of the centering model is the ordering of the list of forward-looking centers (Walker *et al*. 1994; Strube and Hahn 1996). In addition to the structures for centers, C_b, and C_f, the theory of centering specifies a set of constraints and rules (Grosz *et al*. 1995, Walker *et al*. 1994).

Constraints

For each utterance U_i in a discourse segment $U_1, \ldots, U_m$:

[2] The example (6) is also a case of inverted sentence (Hu 1995) and a cataphor occurs in the subject position where the referent is made to 一個人 ‘one person’ mentioned subsequently in the text (Mikov 2002).

[3] An utterance U, realizes c if c is an element of the situation described by U, or c is the semantics interpretation of come subpart of U.

[4] The notion of preferred center corresponds to Sider’s notion of expected focus (Sidner 1983)

1. U_i has exactly one C_b.
2. Every element of $C_f(U_i)$ must be realized in U_i.
3. Ranking of elements in $C_f(U_i)$ guides determination of $C_b(U_{i+1})$.
4. The choice of $C_b(U_i)$ is from $C_f(U_{i-1})$, and can not be from $C_f(U_{i-2})$ or other prior sets of C_f.

Backward-looking centers, C_bs, are often omitted or pronominalized and discourses that continue centering the same entity are more coherent than those that shift from one center to another. This means that some transitions are preferred over others. These observations are encapsulated in two rules (Grosz *et al*. 1995; Walker *et al*. 1990; Walker *et al*. 1994):

Rules

For each utterance U_i in a discourse segment $U_1, \ldots, U_m$:

I. If any element of $C_f(U_i)$ is realized by a pronoun in U_{i+1} then the $C_b(U_{i+1})$ must be realized by a pronoun also.

II. Sequences of continuation are preferred over sequence of retaining; and sequences of retaining are to be preferred over sequences of shifting.

Rule I represents one function of pronominal reference: the use of a pronoun to realize the C_b signals the hearer that the speaker is continuing to talk about the same thing. Psychological research and cross-linguistic research have validated that the C_b is preferentially realized by a pronoun in English and by equivalent forms (i.e. zero anaphora) in other languages. Rule II reflect the intuition that continuation of the center and the use of retentions when possible to produce smooth transitions to a new center provide a basis for local coherence. The transition states are further described in the next section (Grosz *et al*. 1995).

The typology of transitions from U_{i-1} to U_i is based on two factors: whether the $C_b(U_i)$ is the same as $C_b(U_{i-1})$, and whether this discourse entity, $C_b(U_i)$, is the same as the $C_p(U_i)$:

1. $C_b(U_i) = C_b(U_{i-1})$, or $C_b(U_{i-1})$ is undefined.
2. $C_b(U_i) = C_p(U_i)$

If both factors, (1) and (2), hold, a pair continuations across U_n and across U_{n+1}. If (1) holds but (2) does not, the utterances are in a retaining transition, which corresponds to a situation where the speaker is intending to shift onto a new entity in the next utterance. If (1) does not hold, the utterances are in one of the shifting transition states depending on whether (2) holds (Brennan *et al*. 1987, Walker *et al*. 1994).

4. Zero Anaphora Resolution

The process of analyzing Chinese zero anaphora is different from pronominal and nominal anaphora resolution. There are two obvious facts for explaining the difference: (i) ZAs are not expressed in text. To perform the task of resolution, ZAs have to be detected first. (ii) The surface information of a ZA itself is null. Because the anaphor is zeroed, the morphological and lexical information such as number and gender agreement used in pronominal and nominal anaphora resolution cannot be utilized in zero anaphora resolution.

Therefore, we divide the task of zero anaphora resolution in Chinese into two phases: first ZA detection by using the POS and syntactic information, and then antecedent identification by employing the centering model (Grosz *et al*. 1995; Brennan *et al*. 1987; Yeh and Chen 2003; Chen 2005).

4.1 Zero Anaphora in Chinese

Zero anaphors generally refer to noun phrases that can be understood in the preceding utterances and do not need to be specified in a discourse. For resolving ZAs, we have to detect them first. Referring to the linguistic studies, especially in (Liao 1992), Liao indicated that the omission of noun phrases is heavily related to the verbs in sentences and only the elements governed by the verbs can be omitted. We adopt this notion, and use the lexical and syntactical knowledge to perform the task of ZA detection.

Since the omission of noun phrases rely on the features of verbs (Liao 1992), the first step is to get the POS information of the constituents[5], especially the verbs, of an utterance. According to the classification of verb phrases in (Li and Thompson 1981), the types of verbs in Chinese are intransitive (no object), transitive (one object) and ditransitive (two objects). This classification can be employed to detect whether the ZA is embedded in the object position of an utterance, such as the example (7), in which the verb 掉進去 'fall-in-to' in (7d) is a transitive verb whose object is omitted.

(7) a. 張三^{i} 騎 著 他 的 新 腳踏車，
Zhangsan^{i} qi zhe ta de xin jiaotache.
Zhangsan ride DUR he GEN new bicycle
Zhangsan was riding his new bicycle.

b. 因爲 φ^{i}_{1} 太 開心，
yinwei φ^{i}_{1} tai kaixin.
because (he) too happy
Because (he) was too happy.

c. φ^{i}_{2} 沒 看到 前面 的 大 水溝^{j}，
φ^{i}_{2} mei kandao qianmian da shuigou.
(he) not see front NOM big gutter
(He) did not see the front big gutter.

d. φ^{i}_{3} 就 掉進去 φ^{j}_{1} 了 。
φ^{i}_{3} jiudiaojinquφ^{j}_{1} le.
(he) fall-in-to (it) CRS
(He) fell in to it.

(8) 北京 鴨 (被) 烤熟 了。
Beijing ya (bei) kao shou le.
Beijing duck bake well CRS
The Beijing duck was baked well.

[5] Here, because the anaphor is zeroed and the lexical information of itself cannot be obtained, we take the POS information of the remaining constitutes of an utterance.

The subject of a sentence is the noun phrase that has a *doing* or *being* relationship with the verb in that sentence. Each verb requires a specific type of noun phrase to be its subject in a simple sentence (Li and Thompson 1981). Therefore, we could simply detect the ZA occurring in the subject position of an utterance. Consider the example (7), the subjects of the verbs in the utterances (7b), (7c) and (7d) are not specified and are obviously omitted.

In the aspect of considering topic and subject omission in Chinese by reviewing linguistic background about topics and subjects in Section 2, there are four types of sentences: (i) sentences with both subject and topic, (ii) sentences in which the subject and the topic are identical, (iii) sentences with no subject, and (iv) sentences with no topic. In the types of (i) and (ii), no ZA occurs in these types of sentences. The sentences with no subject are regarded as *passive* sentences, e.g. the example (8), which is a passive sentence with 被 'bei' omitted (Hoede *et al*. 2002). From this perspective, the type of (iii) can be treated are the sentences in which the subject and the topic are identical and no ZA needs to be processed. The sentences with no topic include presentative sentences and sentences with ZAs embedded. The presentative sentences discussed in (Li and Thompson 1981) are taken as the cases of exophora (Halliday and Hasan 1976) or inverted sentences (Hu 1995). In our work, we do not deal with the problem of exophora and inverted sentences that are other issues in linguistics and NLP, but focus on zero anaphora resolution. Therefore, the detection of ZAs occurring in the topic or subject position is treated as the detection of subject omission.

4.2 Rules of Zero Anaphora Resolution

In the ZA detection phase, we employ POS information and simple syntactic relations to establish the ZA detection rules for detecting omitted cases as ZA candidates. The ZA detection rule 1 is adopted to detect the ZAs occurring in the topic or subject position, while the ZA detection rule 2 is adopted to detect the ZAs occurring in the object position in an utterance. In the ZA detection rules 3, we further consider the case of prepositions and coordinating conjunctions for detecting the ZAs occurring in the topic or subject position.

ZA detection rules

For each utterance U_i in a discourse segment $U_1, \ldots , U_m$:

1. If no noun phrase appears before a verb phrase in U_i, then an omission of topic or subject is detected as a ZA candidate.
2. For each utterance U_i in a discourse segment $U_1, \ldots , U_m$: If a transitive verb phrase appears in the leftmost position of U_i, then an omission of object is detected as a ZA candidate.
3. If U_i consists of a preposition or a coordinating conjunction in the initial position of a clause, and followed by a noun phrase, then an omission of topic or subject is detected as a ZA candidate.

In the phase of antecedent identification, we concentrate on the resolution of ZA, and we first design the ZA identification constraints for filtering out the non-anaphoric cases[6] from

[6] The non-anaphoric cases such as exophora or cataphora are the different research issues from the zero anaphora resolution. In our work, we do not intend to eliminate all non-anaphoric cases but to filter out some less complicated ones.

the ZA candidates which are detected in the phase of ZA detection. In the case of cataphora, because the first utterance has neither preceding utterances nor previous elements to be referred to as antecedents, the candidates detected in this utterance cannot be anaphors. By the observation of the test data, a news article sometimes has 據說 'it is said' as its first utterance, which is a case of expohora. Therefore, the ZA identification constraint 1 is employed to eliminate the exophora or cataphora. In addition, the constraint 2 includes some cases might be incorrectly detected as ZAs, such as passive sentences or inverted sentences (Hu 1995).

ZA identification constraints

For each ZA candidate *c* in a discourse:

1. *c* can not be in the first utterance in a discourse segment (exophora or cataphora)
2. ZA does not occur in the following case:
   NP + *bei* + NP + VP + *c* (passive)
   NP (topic) + NP (subject) + VP + *c* (inverted)

Most lexical knowledge such as person, number and gender employed in pronoun resolution in English cannot be utilized in zero anaphora resolution because the ZA itself is not expressed in text. In the antecedent identification, we employ the concept of *centers*[7] which are of the key elements of the centering theory (Grosz *et al*. 1995; Brennan *et al*. 1987) to establish the antecedent identification rule for identifying the antecedent of each ZA.

Antecedent identification rule:

```
For each ZA $z$ in a discourse segment $U_1, \dots, U_m$:
If $z$ occurs in $U_i$, and no ZA occurs in $U_{i-1}$
  then choose the preferred center of $U_{i-1}$ as the antecedent
Else if only one ZA occurs in $U_{i-1}$
  then choose the antecedent of the ZA in $U_{i-1}$ as the antecedent of $z$
Else if more than one ZA occurs in $U_{i-1}$
  then choose the antecedent of the ZA in $U_{i-1}$ as the antecedent of $z$ according to the
  forward-looking center ranking criterion
End if
```

Forward-looking center ranking criterion:

Topic > *Subject* > *Object* > *Others*

In the rules of the center model[8], they stipulate that if there is only one pronoun in an utterance, this pronoun should be the backward-looking center. In addition, if the next sentence also contains a pronoun, the pronoun refers to the one in the preceding utterance. The preferred center is the most preferred discourse entity referred by a pronoun for local coherence of a discourse. Psycholinguistic research (Gordon *et al*. 1993) and cross-linguistic research (Kameyama 1986, Walker *et al*. 1994) have validated that the backward-looking center is preferentially realized by a pronoun in English and by equivalent forms (*i.e.* zero pronouns) in other languages (Grosz et al 1995).

[7] The centers include forward-looking centers, the backward-looking center (Grosz *et al*. 1995) and the preferred center (Brennan *et al*. 1987).

[8] The centering model includes two rules as described in Section 3.

Referring to the notions of the center model, we create the antecedent identification rule according to three perceptions described as follows: (i) If there is only one ZA occurring in an utterance, to choose the preferred center in the preceding utterance as the antecedent of the ZA. (ii) If there are two ZAs respectively occurring in two successive utterances, the co-reference is made. (iii) If the preceding utterance contains more than one ZA, the ZAs are ranked with the same ranking criterion for forward-looking centers.

Grosz *et al*., in their paper (Grosz *et al*. 1995), assume that grammatical roles are the major determinant for ranking the forward-looking centers, with the order "*Subject* > *Object(s)* > *Others*". In Chinese, the concept of subject seems to be less significant while the topic in a sentence appears to be crucial in explaining the structure of ordinary sentences in the language (Li and Thompson 1981). By adopting the concept of grammatical roles and topic-prominence in Chinese, we order the grammatical roles in Chinese with topic having the highest priority and the order is referred to as the forward-looking center ranking criterion. This criterion is not only used to rank forward-looking centers but also employed to choose the antecedent of the ZA in the antecedent identification rule.

5. Topic Identification

5.1 Topic Identification Method

As mentioned previously, topics are significant and valuable information embedded in text, and the information might further be taken as the useful data or knowledge in some NLP applications. But in Chinese, the topics in utterances are frequently omitted from expressions, due to their prominence in discourse. Therefore, to obtain the topic of each utterance in a discourse, we have to resolve the problem of zero anaphora.

Grosz *et al*., in their paper (Grosz *et al*. 1995), reported on that psychological research and cross-linguistic research have validated that the backward-looking center is preferentially realized by a pronoun in English and by equivalent forms (i.e. zero anaphora) in other languages. By adopting this notion, the key elements of the centering model of local discourse coherence and the vital characteristic, topic-prominence, in Chinese, we establish the topic identification rule for identifying the topics in text.

When a ZA occurs in the utterance U_i, the antecedent of the ZA in the preceding utterance is identified as the topic of U_i. Otherwise, if the transition relation, center shifting, occurs, topic will not be identified as any of the element in the preceding utterance but the element in the current utterance according to forward-looking center ranking criterion described in Section 4.2.

Topic identification rule:

For identifying each topic *t* in a discourse segment consisting of utterances U_1, … , U_m:

If at least one ZA occurs in U_i

then refer to forward-looking center ranking criterion to choose the ZA to be resolved and then get its antecedent as the *t* according to Antecedent identification rule.

```
Else if no ZA occurs in $U_i$
  then refer to forward-looking center ranking criterion to choose one element of $U_i$
  as the $t$
End if
```

We now take the example (9) to identify each topic of the utterances (9a) to (9d) by employing the topic identification rule. As shown in Table 1, the topic of (9a) is 電子股 'Electronics stocks,' and the topic of (9b) is omitted identified as the antecedent of φ^i_1, 電子股 'Electronics stocks.' Similarly, the topic of (9d) is 證券股 'Securities stocks,' which is referred to as the antecedent of the zeroed topic of (9c). Therefore, we can obtain the topics of this example in the second column of Table 1.

(9) a. 電子股 ^{i} 受 美國高科技股 重挫 影響，
dianzigu ^{i} shou meiguo gaokejigu zhongcuo yingxiang.
Electronics stock receive USA high-tech stock heavy-fall affect
Electronics stocks were affected by high-tech stocks fallen heavily in America.

b. φ^i_1 持續 下跌，
φ^i_1 chixu xiadie.
(Electronics stocks) continue fall
(Electronics stocks) continued falling down.

c. 證券股 ^{j} 也 有 相對 回應，
zhengquanqu ^{j} ye you xiangdui huiying.
Securities stocks also have relative response
Securities stocks also had response.

d. φ^j_1 陸續 下殺 至 跌停。
φ^j_1 luxu xiasha zhi dieting.
(Securities stocks) continue fall by close.
(Securities stocks) fell by close one after another.

Utterance	Topic
(9a) 電子股 ^{i} 受 美國高科技股 重挫 影響，	電子股
(9b) φ^i_1 持續 下跌，	電子股
(9c) 證券股 ^{j} 也 有 相對 回應，	證券股
(9d) φ^j_1 陸續 下殺 至 跌停。	證券股

Table 1: Examples of zero anaphora

5.2 Experiment and Result

We employ the rules of zero anaphora resolution mentioned in Section 4.2 for resolving the occurrences of ZAs. The ZA detection rules are used to detect omitted cases as ZA candidates. The ZA identification constraints and the antecedent identification rule are used to eliminate the non-anaphoric cases of these candidates and to identify the antecedents respectively. After the occurrences of ZAs are resolved, the topic identification rule is employed to identify each topic of each utterance in text.

For evaluating the topic identification method, we take part articles of China Times Express and Central Daily News form CIRB030 (Chen and Chen 2004) as the test corpus. The test corpus contains more than more than 30,000 utterances in 592 news articles of China Times Express and 30 news articles of Central Daily News.

The recall rates and precision rates of zero topic resolution are 0.67 and 0.64 respectively calculated using equation 1 and equation 2. Most errors occur when a zero topic does not refer to the preferred center in the preceding utterance, or refers to other entity in the more previous utterance.

$$\text{Precision rate of zero topic resolution} = \frac{\text{No. of antecedent correctly identified}}{\text{No. of zero topic identified}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

$$\text{Recall rate of zero topic resolution} = \frac{\text{No. of antecedent correctly identified}}{\text{No. of zero topic occurred in text}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

6. Conclusions

In this article, we propose a method of topic identification based on the centering model. According to observations on real texts, we found that to identify the topics in Chinese context is much related to the issue of zero anaphora resolution. We use a zero anaphora resolution method to resolve the problem of ellipsis in Chinese text and then employ the notions of the centering model of local discourse coherence to identify the topics of utterances in discourse. The result is promising to some extent; however, there are still some works that need further investigation, such as establishment of NLP applications. The information carried by topics can be used to contribute to some information retrieval, text categorization and discourse segmentation *etc*. One of our future work is to establish some NLP applications and then apply the method on these applications for evaluating its performance. Also, we will further work on the improvement of accuracy of zero topic resolution.

7. Acknowledgement

We give our special thanks to CKIP, Academia Sinica for making great efforts in computational linguistics and sharing the Chinese word segmentation system to academic research. We would like to thank to LIPS Lab. of LIS department and NLP Lab. of CSIE department in NTU, National Science Council, and ROCLING for providing CIRB030 for NLP research.

References

Brennan, S., Friedman, M. and Pollard, C., 1987, A centering approach to pronouns, in *Proceedings of the 25th annual meeting of the ACL*, pages 155-162.

Chen, Kuang-hua and Chen, Hsin-His, 2004, Overview of CIRB030 Information Retrieval Test Collection, *http://lips.lis.ntu.edu.tw/cirb/index.htm*.

Chen, P, 1987, *Hanyu lingxin huizhi de huayu fenxi* (a discourse approach to zero anaphora in chinese) (in chinese), Zhongguo Yuwen (Chinese Linguistics), pages 363-378.

Gordon, Peter C., Grosz, B. J., and Gilliom, L. A., 1993, Pronouns, names and the centering of attention in discourse. *Cognitive Science*, 17(3):311-348.

Grosz, B. J., 1977, The representation and use of focus in dialogue understanding *Technical Report 151*, SRI International.

Grosz, B. J. and Sidner, C. L., 1986, Attention, intentions, and the structure of discourse, *Computational Linguistics*, No 3 Vol 12, pp. 175-204.

Grosz, B. J., Joshi, A. K. and Weinstein, S., 1983, Providing a unified account of definite noun phrases in discourse, *Proceedings of 21st Annual Meeting of the ACL.*

Grosz, B. J., Joshi, A. K. and Weinstein, S., 1995, Centering: A Framework for Modeling the Local Coherence of Discourse, *Computational Linguistics*, 21(2), pp. 203-225.

Halliday, M. and Hasan, R., 1976, *Cohesion in English*, Longman.

Hoede, C., Li, X., Liu, X. and Zhang, L. Knowledge Graph Analysis of some Particular Problems in the Semantics of Chinese, *Memorandum No.1516*, Department of Applied Mathematics, University of Twente, Enschede, The Netherlands.

Hu, Wenze, 1995, *Functional Perspectives and Chinese Word Order*, Ph. D. dissertation, The Ohio State University.

Huang, Yan, 1994, *The Syntax and Pragmatics of Anaphora – A study with special reference to Chinese*, Cambridge University Press.

Kameyama, M., 1986, A property-sharing constraint in centering, in *Proceedings 24th Annual Meeting of the ACL*, pages 200-206, New York.

Li, Charles N. and Thompson, Sandra A., 1981, *Mandarin Chinese – A Functional Reference Grammar*, University of California Press.

Liao, Chiu-chung , 1992, *Liao, Chiu-chung wenji* (the collections of Liao, Chiu-chung's work) (in chinese), Beijing, Beijing Language Institute Press.

Mitkov, Ruslan, 2002, *Anaphora Resolution*, Longman.

Okumura, Manabu and Tamura, Kouji, 1996, Zero pronoun resolution in Japanese discourse based on centering theory, *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, 871-876.

Sidner, C. L., 1979, *Toward a Computational Theory of Definite Anaphora Comprehension in English Discourse*, Ph.D. thesis, MIT.

Sidner, C. L., 1983, Focusing in the comprehension of definite anaphora, *Computational Models of Discourse*, MIT Press.

Strube, M. and Hahn, U., 1996, *Functional Centering, Proceedings Of ACL '96*, Santa Cruz, Ca., pp.270-277.

Walker, M. A., 1989, Evaluating Discourse Processing Algorithms, *Proceedings Of ACL '89*, Vancouver, Canada.

Walker, M. A., 1998, Centering, anaphora resolution, and discourse structure. In Walker, M. A., Joshi, A. K. and Prince, E. F., editors, Centering in Discourse, Oxford University Press.

Walker, M. A., Iida, M. and Cote. S., 1990, Centering in Japanese discourse, *Proceedings of 13th International Conference on Computational Linguistics* (*COLING-90*), Helsinki.

Walker, M. A., Iida, M. and Cote. S., 1994, Japan Discourse and the Process of Centering, *Computational Linguistics*, 20(2): 193-233.

Yeh, Ching-Long and Chen, Yi-Chun, 2003, Zero Anaphora Resolution in Chinese with Partial Parsing Based on Centering Theory, *Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineering* (*NLP-KE'03*), Beijing, China.

Yi-Chun Chen, 2005, *Chinese Zero Anaphora Resolution and Its Applications*. *Ph.D. Dissertation*, Tatung University, Taipei, Taiwan.

Appendix: Abbreviations

In the word-by-word translation, some markers are abbreviated as below. We follow the abbreviations used in (Li and Thompson 1981).

Abbreviation	**Term**
ASSOC	associative (de)
ASPECT	aspect marker
BA	ba
BEI	bei
CL	classifier
CSC	complex stative construction (de)
GEN	genitive (de)
NOM	nominalizer (de)
Q	Question (ma)