Zero Anaphora Resolution in Chinese with Shallow Parsing

Ching-Long Yeh[1], Yi-Chun Chen[2]

[1] Department of Computer Science and Engineering, Tatung University

[2] 40 Chungshan N. Rd. 3rd. Section Taipei 104 Taiwan

chingyeh@cse.ttu.edu.tw, yjchen7@ms7.hinet.net

Abstract

Most traditional approaches to anaphora resolution are based on the integration of complex linguistic information and domain knowledge. However, the construction of a domain knowledge base is very labor-intensive and time-consuming. In this paper, we work on the output of a part-of-speech tagger and use shallow parsing instead of complex parsing to resolve zero anaphors in written Chinese. We employ centering theory and constraint rules to identify the antecedents of zero anaphors as they appear in the preceding utterances. We focus on the cases of zero anaphors that occur in the topic or subject, and object positions of utterances. The experimental result shows that the precision rates of zero anaphora detection and the recall rate of zero anaphora resolution with the method are 81% and 70% respectively.

Keywords

Anaphora Resolution; Zero Anaphora Detection; Antecedent Identification; Shallow Parsing; Centering Theory.

1   Introduction

In natural languages, expressions that can be deduced contextually by the reader are frequently omitted in texts. This is especially the case in Chinese, where a kind of anaphoric expression is frequently eliminated. This will be termed zero anaphor (ZA) hereafter, due to its prominence in discourse (Li and Thompson 1981). The omission may cause considerable problems in natural language processing systems. For example in a machine translation system, a Chinese text can not be translated properly into text in a target language without identifying the meaning of the omitted expressions first. In information extraction, the events related to some subjects omitted in texts can not be extracted effectively. In this paper, we aim at the resolution of zero anaphora in Chinese text.

An approach of anaphora resolution employs knowledge sources or factors, for example, gender and number agreement, c-command constraints, semantic information to discount unlikely candidates until a minimal set of plausible candidates is obtained (Grosz et al. 1995; Lappin and Leass 1994; Okumura and Tamura 1996; Walker et al. 1998; Yeh and Chen 2001). Anaphoric relations between anaphors and their antecedents are identified based on the integration of linguistic and domain knowledge. However, it is very labor-intensive and time-consuming to construct grammatical and domain knowledge base. Another approach

employs statistical models or AI techniques, such as machine learning, to compute the most likely candidate (Aone and Bennett 1995; Connoly et al. 1994; Ge et al. 1998; Seki et al. 2002). This approach can sort out the above problems. However, it heavily relies upon the availability of sufficiently large text corpora that are tagged, in particular, with referential information (Stuckardt 2002).

A recent approach is the search for inexpensive, fast and reliable procedures of anaphora resolution (Baldwin 1997; Ferrández et al. 1998; Kennedy and Boguraev 1996; Mitkov 1998). The approach relies on reliable and cheaper NLP tools such as part-of-speech (POS) tagger and shallow parsers. In this paper, we adopt this approach. The task of ZA resolution can be divided into two phases: first detecting the occurrences of zero anaphors in text, and then finding their antecedents in the discourse. A POS tagger and the following shallow parser are used to accomplish the task of the first phase. We then employ the centering theory (Grosz et al. 1995) to develop a rule-base as the basis to determine the antecedents of zero anaphors found in the first phase. We have carried out an experiment using a number of news articles as the test data. The result shows that the precision rate of zero anaphora detection is 80% and within the detected zero anaphors, 70% can be resolved correctly.

In the following sections we first briefly describe the nature of zero anaphora in Chinese. In Section 3 we describe in details the shallow parsing method. In Section 4 we describe the ZA resolution method. In Section 5, we show the experiments and result. Finally our conclusions are summarized, and future works are suggested.

2   Zero Anaphora in Chinese

As mentioned in Section 1, zero anaphors are generally noun phrases that are understood from the context and do not need to be specified. For example in (1), the topic of the utterance (1a) is張三 'Zhangsan' which is eliminated in the second utterance.

(1) a. 張三 ^{i} 驚慌的往外跑，

　　　Zhangsan jinghuang de wang wai pao
　　　Zhangsan frightened CSC towards outside run
　　　Zhangsan frightened and ran outside.

　　b. ϕ_1^i 撞到 一個人 ^{j}，

　　　zhuangdao yi ge ren
　　　(he) bump-to a person
　　　(He) bumped into a person.

　　c. 他$_2^i$ 看清 了 那人 ^{j} 的 長相，

　　　ta   kanqing le na ren de zhangxiang
　　　he see-clear ASPECT that person GEN appearance
　　　He saw clearly that person's appearance.

　　d. ϕ_3^i 認出 那人 ^{j} 是 誰。

　　　renchu na ren shi shei
　　　(he) recognise that person is who
　　　(He) recognized who that man is.

In addition to zero anaphors, anaphors can be pronominal and nominal forms, as exemplified by 他 'He' and 那個人 'that person' in (1c) and (1d), respectively (Chen 1987)[1].

According to Li and Thompson (Li and Thompson 1981), zero anaphors can be classified as intrasentential or intersentential. In the intrasentential case, the antecedent exists in the same sentence, or the zero anaphor can be understood and does not need to be expressed, such as the φ in (2) while antecedent and anaphors are located in different sentences in the intersentential case, such as the φ^i and in (1b) and (1d).

(2) 張三 參加 比賽 φ 贏得 一 台 電腦。
　Zhangsan canjia bisai　yingde yi tai diannao
　Zhangsan enter competition (he) win a CL computer
　Zhangsan entered a competition and (he) win a computer.

In the intersentential case, antecedent and anaphors are located in different sentences. Depending upon the distance between the sentences containing antecedent and anaphor, it can further be divided into two types: immediate and long distance. The former is where the sentence containing the antecedent is immediately followed by the one containing the anaphor, such as ϕ_1^j in (3b) and ϕ_1^k in (3d). On the other hand, for the long distance type, the sentence containing the antecedent and anaphors, , are not in immediately succeeding order, such as ϕ_1^i in (3e).

(3) a. 螃蟹^i 有 四 對 步足^j，
　　　pangxie you si dui buzu
　　　crab have four-pair walking-foot
　　　A crab has four pairs of feet.

　b. ϕ_1^j 俗稱 「腿兒」，
　　　sucheng tuier
　　　(they) common-called "tuier"
　　　(They) are commonly called "tuier."

　c. 由於 每 條 「腿兒」 的 關節^k 只能 向 下 彎曲，
　　　youyu mei tiao tuier　de guanjie zhineng xiang xia wanqu
　　　since every "tuier" ASSOC joint only can towards down bend
　　　Since every "tuier"'s joint can only bend downwards,

　d. ϕ_1^k 不能 向 前後 彎曲，
　　　buneng xiang qianhou wanqu
　　　(it) not can towards forward-backward bend
　　　(it) can't bend backward or forwards.

　e. ϕ_1^i 爬行 時，
　　　paxing shi
　　　(it) crawl ASPECT

---

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.

When (it) crawls,

f. ϕ_2^i 必須 先 用 一 邊 步足 的 指尖 抓 地，

    bixu xian yong yi bian buzu de zhijian zhua di

    (it) must first use one-side walking-foot ASSOC fingertip grasp-on ground

    (it) must use the tips of feet on one side to grasp the ground.

g. ϕ_3^i 再 用 另一 邊 的 步足 直伸 起來，

    zai yong ling yi bian de buzu zhishen qilai

    (it) then use another one-side ASSOC walking-foot straight-rise upwards

    (It) then uses the feet on the other side to move upwards.

h. ϕ_4^i 把 身體 推 過去。

    ba shenti tui guoqu

    (it) BA body push get-through

    (It) pushes the body towards one side.

3   Sentence Parsing

Full parsing is used to provide an as detailed as possible analysis of the sentence structure and to build a complete parse tree for the sentence, while shallow parsing is limited to parsing smaller constituents such as noun phrases or verb phrases (Abney 1996; Li and Roth 2001). In this section, we show you some examples of full parsing and describe our method of shallow parsing in Chinese.

3.1   Full Parsing

Many traditional approaches to parsing natural language sentences aim to recover complete, exact parses based on the integration of complex syntactic and semantic information. They search through the entire space of parses defined by the grammar and then seek the globally best parse referring to some heuristic rules or manual correction. For example, the sentence (4) taken from Sinica Treebank (Sinica Treebank 2002) is annotated as below.

(4) 他 終於 找到 一 份 工作 了。

    ta zhongyu zhaodao yi fen gongzuo le

    he final find a CL job ASPECT

    He finally found a job.

→ S(agent:NP(Head:Nhaa:他)|time:Dd:終於|Head:VC2:找到|goal:NP(quantifier: DM:一份|Head:Nac:工作)|particle:Ta:了)

S(agent:NP(Head:Nhaa:*he*)|time:Dd:*finally*|Head:VC2:*find*|goal:NP(quantifier: DM:*a*|Head:Nac:*job*)|particle:Ta:*le*)

The sentence structure in Sinica Treebank is represented by employing head-driven principle, that is, each sentence or phrase has a head leading it. A phrase consists of a head, arguments and adjuncts. One can use the concept of head to figure out the relationship among the phrases in a sentence. In the example (4), the head of the NP (noun phrase) , 他 'he,' is the *agent* of the verb, 找到 'find'. Although the head-driven principle may prevent the ambiguity of syntactical analysis (Chen *et al*. 1999), to choose the head of a phrase

automatically may cause errors. Another example (5) is extracted from the Penn Chinese TreeBank (The Penn Chinese Treebank Project 2000).

(5) 張三 告訴 李四 王五 來 了。

    Zhangsan gaosu Lisi Wangwu lai le
    Zhangsan tell Lisi Wangwu come ASPECT
    Zhangsan told Lisi that Wangwu has come.

→ (IP (NP-PN-SBJ (NR 張三))
     (VP (VV 告訴)
     (NP-PN-OBJ (NR 李四))
     (IP (NP-PN-SBJ (NR 王五))
     (VP (VV 來)
     (AS 了)))))
   (IP (NP-PN-SBJ (NR *Zhangsan*))
      (VP (VV *tell*)
        (NP-PN-OBJ (NR *Lisi*))
        (IP (NP-PN-SBJ (NR *Wangwu*))
         (VP (VV *come*)
       (AS *le*))))))

The Penn Chinese TreeBank provides solid linguistic analysis for the selected text, based on the current research in Chinese syntax and the linguistic expertise of those involved in the Penn Chinese Treebank project to annotate the text manually.

3.2 Shallow Parsing

Shallow (or partial) parsing which is an inexpensive, fast and reliable method does not deliver full syntactic analysis but is limited to parsing smaller syntactical related constituents (Abney 1991; Abney 1996; Li and Roth 2001; Mitkov 1999). For example, the sentence (6a) and can be divided as (6b):

(6) a. 花蓮 成爲 熱門 的 旅遊 地點。

     Hualian chengwei remen de luyou didian
     Hualian become popular NOM tour place
     Hualien became the popular tourist attraction.
   b. [NP 花蓮] [VP 成爲] [NP 熱門 的 旅遊 地點]
    [NP Hualien] [VP became] [NP the popular tourist attraction]

Given a Chinese sentence, our method of shallow parsing is divided into the following steps: First the sentence is divided into a sequence of POS-tagged words by employing a segmentation program, AUTOTAG, which is a POS tagger developed by CKIP, Academia Sinica (CKIP 1999). Second the sequence of words is parsed into smaller constituents such as noun phrases and verb phrases with phrase-level parsing. Each phrase is represented as a word list. Then the sequence of word lists is transformed into *triples*, [*S,P,O*]. For example in (7), (7b) is the output of sentence (7a) produced by AUTOTAG and (7c) is the *triple* representation.

(7) a. [花蓮(Nc) 成爲(VG) 熱門(VH) 的(DE) 旅遊(VA) 地點(Na)]

  b. [[花蓮], np], [[成為], vp], [[熱門,的,旅遊,地點], np]
  c. [[花蓮], [成為], [熱門,的,旅遊,地點]]

The definition of *triple* representation is illustrated in Definition 1.The *triple* here is a simple representation which consists of three elements: *S*, *P* and *O* which correspond to the *Subject* (noun phrase), *Predicate* (verb phrase) and *Object* (noun phrase) respectively in a clause.

Definition 1:
A Triple T is characterized by a 3-tuple:
T = [*S*, *P*, *O*] where
- *S* is a list of nouns whose grammatical role is the subject of a clause.
- *P* is a list of verbs or a preposition whose grammatical role is the predicate of a clause.
- *O* is a list of nouns whose grammatical role is the object of a clause.

In the step of *triple* transformation, the sequence of word lists as shown in (7b) is transformed into triples by employing the Triple Rules. The Triple Rules is built by referring to the Chinese syntax. There are four kinds of Triples in the Triple Rules, which corresponds to five basic clauses: subject + transitive verb + object, subject + intransitive verb, subject + preposition + object, and a noun phrase only. The rules listed below are employed in order:

Triple Rules:
Triple1(S,P,O) ➔ np(S), vtp(P), np(O).
Triple2(S,P,*none*) ➔ np(S), vip(P).
Triple3(S,P,O) ➔ np(S), prep(P), np(O).
Triple4(S,*none,none*) ➔ np(S).

The vtp(P) denotes that the predicate is a transitive verb phrase, which contains a transitive verb in the rightmost position in the phrase; likewise the vip(P) denotes that the predicate is an intransitive verb phrase, which contains an intransitive verb in the rightmost position in the phrase. In the rule Triple3, the prep(P) denotes that the predicate is a preposition. The Triple4 is employed only if a sentence contains only one noun phrase and no other constituent. If all the rules in the Triple Rules failed, the ZA Triple Rules are employed to detect zero anaphor (ZA) candidates.

ZA Triple Rules:
Triple1z1(*zero*,P,O)➔ vtp(P), np(O).
Triple1z2(S,P,*zero*)➔ np(S), vtp(P).
Triple1z3(*zero*,P,*zero*)➔ vtp(P).
Triple2z1 (*zero*,P,*none*)➔ vip(P).
Triple3z1(*zero*,P,O) ➔ prep(P), np(O).
Triple4z1(*zero*,P,O) ➔ co-conj(P), np(O).

The zero anaphora in Chinese generally occurs in the topic, subject or object position. The rules Triple1z1, Triple2z1, and Triple3z1 detect the zero anaphora occurring in the topic or subject position. The rule Triple1z2 detects the zero anaphora in the object position and Triple1z3 detect the zero anaphora occurring in both subject and object positions. In the

Triple4, the co-conj(P) denotes a coordinating conjunction appearing in the initial position of a clause. For example in (8), there are two *triples* generated. In the second *triple*, *zero* denotes a zero anaphor according to Triple1z1.

(8) 張三 參加 比賽 贏得 冠軍。
    Zhangsan canjia bisai yingde guanjun
    Zhangsan enter competition win champion
    Zhangsan entered a competition and won the champion.
 → [[[張三], [參加], [比賽]], [[*zero*], [贏得], [冠軍]]]
    [[[*Zhangsan*], [*enter*], [*competition*]], [[*zero*], [*win*], [*champion*]]]

The Figure 1 illustrates the detailed procedure of Triple transformation. The input is a sequence of word lists after phrase-level parsing. The input sequence is scanned from the leftmost word list in the sequence and the Triple Rules are employed to generate a new Triple. If a new Triple is generated, the remaining sub-sequence is taken as a new input, or the ZA Triple Rules is employed to generate a new Triple. If no other word list is left to be processed, the procedure stops, or otherwise, the procedure continues to process the remaining sub-sequence.
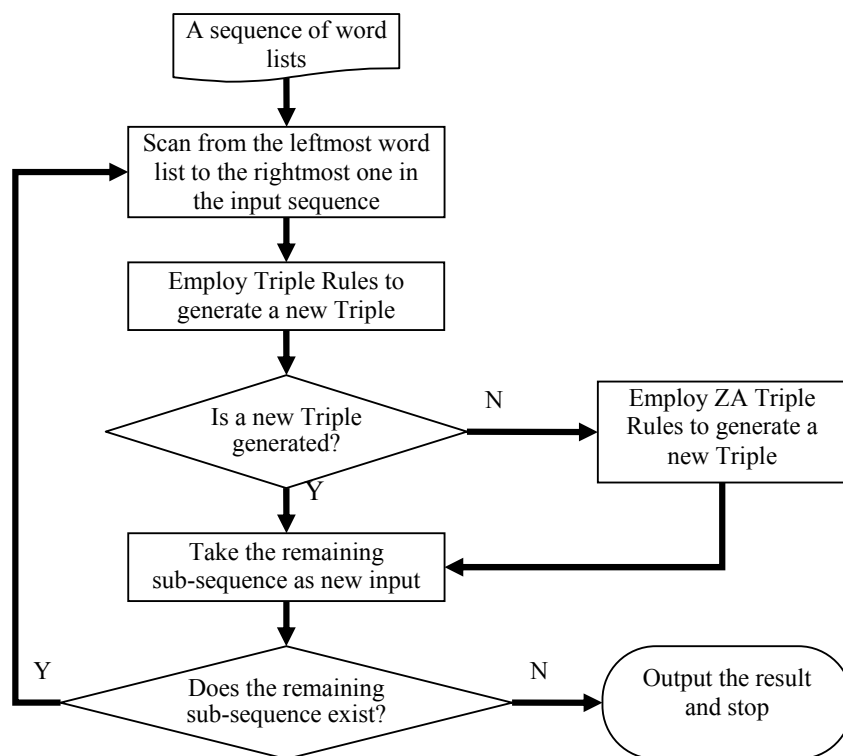
Figure 1. The procedure of Triple transformation

4   ZA Resolution Method

The ZA resolution method we develop is divided into three parts. First each sentence of an input document is translated into triples as described in Section 3. Second, ZA identification verifies that each ZA candidate is annotated in triples by employing ZA identification constraints. Third antecedent identification identifies the antecedent of each detected ZA by using rules based on the centering theory.

4.1   Centering Theory

In the centering theory (Grosz *et al*. 1995; Walker *et al*. 1994; Strube and Hahn 1996), each utterance U in a discourse segment has two structures associated with it, they are called forward-looking centers, $C_f(U)$ and backward-looking centers, $C_b(U)$. 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. 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*". The superlative element of $C_f(U_n)$ may become the C_b of the following utterance, $C_b(U_{n+1})$..

In addition to the structures for centers, C_b, and C_f, the centering theory 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 $U_1, \ldots, U_m$:
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. 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:

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 (Grosz *et al*. 1995).

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.

For example in (9), the subject of the utterance (9b) is eliminated, and its antecedent is identified as the subject of the preceding utterance (9a) according to the centering theory.

(9) a. 電子股 ^{i} 受 美國 高科技股 重挫 影響，
      dianzigu shou meiguo gaokejigu zhongcuo yingxiang
      Electronics stock receive USA high-tech stock heavy-fall affect
      Electronics stocks were affected by high-tech stocks in USA.
  b. φ^i 持續 下跌。
      chixu xiadie
  (Electronics stocks) continue fall
  (Electronics stocks) continued falling down.

4.2 Zero Anaphora Resolution

The process of analyzing Chinese zero anaphora is different from general pronoun resolution in English because zero anaphors are not expressed in discourse. The task of ZA resolutions is divided into two phases: first ZA detection and then antecedent identification. In this paper, we focus on the cases of ZA occurring in the topic or subject, and object positions.

In the ZA detection phase, we use the ZA Triple Rules described in 3.2 to detect omitted cases as ZA candidates denoted by *zero* in *triples*. Table 1 shows some examples corresponding to the ZA Triple Rules.

ZA Triple Rule	Example
Triple1^{z1}(*zero*,P,O)	φ 撞到 一 個 人 (1b) zhuangdao yi ge ren (he) bump-to a person (He) bumped into a person.
Triple1^{z2}(S,P,*zero*)	張三 喜歡 φ 嗎 Zhangsan xihuan ma Zhangsan like (somebody or something) Q Does Zhangsan like (somebody or something)?
Triple1^{z3}(*zero*,P,*zero*)	φ 喜歡 φ xihuan (he) like (somebody or something) (He) likes (somebody or something).
Triple2^{z1} (*zero*,P,*none*)	φ 去 購物 了 qu gouwu le (he) go shopping ASPECT (He) has gone shopping.
Triple3^{z1}(*zero*,P,O)	φ 在 那邊 zai nabian (he) in there (He) is there.
Triple4^{z1}(*zero*,P,O)	φ 跟 小朋友 玩 gen xiaopengyou wan (he) with child play (He) is playing with little children.

Table 1. Examples of zero anaphora

After ZA candidates are detected by employing the ZA Triple Rules, the ZA identification constraints are utilized to filter out non-anaphoric cases. In the ZA identification constraints, the constraint 1 is employed to exclude the exophora [2] or cataphora[3] which is different from anaphora in texts. The constraint 2 includes some cases might be incorrectly detected as zero anaphors, 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
2. ZA does not occur in the following case:
    NP + *bei* + NP + VP + c
    NP (topic) + NP (subject) + VP + c

In the antecedent identification phase, we employ the 'backward-looking center' of centering theory to identify the antecedent of each ZA. First we use noun phrase rules to obtain noun phrases in each utterance, and then the antecedent is identified as the most prominent noun phrase of the preceding utterance (Yeh and Chen 2001):

Antecedent identification rule:
 For each zero anaphor z in a discourse segment $U_1, \dots , U_m$:
 If z occurs in U_i, and no zero anaphor occurs in U_{i-1}
    then choose the noun phrase with the corresponding grammatical role in U_{i-1} as the antecedent
 Else if only one zero anaphor occurs in U_{i-1}
    then choose the antecedent of the zero anaphor in U_{i-1} as the antecedent of z
 Else if more than one zero anaphor occurs in U_{i-1}
    then choose the antecedent of the zero anaphor in U_{i-1} as the antecedent of z according to grammatical role criteria: *Topic > Subject > Object > Others*
 End if

Due to topic-prominence in Chinese (Li and Thompson 1981), topic is the most salient grammatical role. In general, if the topic is omitted, the subject will be in the initial position of an utterance. If the topic and subject are omitted concurrently, the ZA occurs. The antecedent identification rule corresponds to the concept of centering theory.

5   Experiment and Result

In this section we describe the experiment and result of the two-phase zero anaphora resolution described in the preceding section. In the ZA detection phase, we only take the result of employing the ZA Triple Rules as the baseline at first, and then include ZA identification constraints to see the difference. In the antecedent identification phase, we also use a rule without involving the centering theory to pit our method against to show improvement. The test corpus is a collection of 150 news articles contained 998 paragraphs, 4631 utterances, and 40884 Chinese words.

---

[2] Exophora is reference of an expression directly to an extralinguistic referent in which the referent does not require another expression for its interpretation.
3 Cataphora arises when a reference is made to an entity mentioned subsequently.

5.1 ZA Detection

By employing the ZA Triple Rules and ZA identification constraints mentioned previously, zero anaphors occur in topic or subject, and object positions can be detected. In the experiment, we first only employ the ZA Triple Rules, and then include the ZA identification constraints to see the improvement. Because the ZA Triple Rules cover each possible topic or subject, and object omission cases, the result shows that the zero anaphors are over detected. The Table 1 shows the precision rates calculated using equation 2.

$$\text{Precision rate of ZA detection} = \frac{\text{No. of ZA correctly detected}}{\text{No. of ZA candidates}} \quad (1)$$

The main errors of ZA detection occur in the experiment when parsing inverted sentences and non-anaphoric cases (e.g. exophora or cataphora) (Mitkov 2002; Hu 1995). Cataphora is similar to anaphora, the difference being the direction of the reference. In this paper, we do not deal with the case that the referent of a zero anaphor is in the following utterances, but we can detect about 60% cataphora in the test corpus by employing ZA identification constraint 1.

5.2 Antecedent Identification

In this phase, we take the output of employing the ZA Triple Rules and ZA identification constraints, and further to identify the antecedents of zero anaphors. We first use a simple antecedent identification rule without involving the centering theory and then employ the antecedent identification rule mentioned in 4.2 to show the improvement:

> **Simple Antecedent identification rule:**
> For each zero anaphor z in a discourse segment $U_1, \ldots , U_m$: If z occurs in U_i then choose the noun phrase in U_{i-1} having the longest distance from z as the antecedent.

The simple antecedent identification rule does not consider the ranking of centers in the centering theory (Grosz *et al*. 1995). By comparing with the simple antecedent identification rule, the antecedent identification rule based on the centering theory (see 4.2) determines the antecedents according to grammatical role criteria. For example, in the discourse segment (10), the zero anaphors are detected in the utterances (10b) and (10c). According to the antecedent identification rule, the noun phrase, 基隆醫院 'Kee-lung General Hospital,' whose grammatical role corresponds to the zero anaphor ϕ_1^i in (10b) is identified as the antecedent. Subsequently, the antecedent of the zero anaphor ϕ_2^i in (10c) is identified as the antecedent of ϕ_1^i in (10b), 基隆醫院.

(10) a. 基隆 醫院 ^{i} 爲 擴大 服務 範圍，
       Jilong yiyuan wei kuoda fuwu fanwei
       Kee-lung hospital for expand service coverage
       Kee-lung General Hospital aims to expand service coverage.
  b. ϕ_1^i 積極 提升 醫療 服務 品質 及 標準化，
      jiji tisheng yiliao fuwu pinzhi ji biaozhunhua
     (it) active improve medical-treatment service quality and standardization

(It) actively improves the service quality of medical treatment and standardization.

c. ϕ_2^i 獲 衛生署 認可 為 辦理 外勞 體檢 醫院。

    huo weishengshu renke wei banli wailao tijian yiyuan

(it) obtain Department-of-Health certify to-be handle foreign-laborer physical-examination hospital

(It) is certified by Department of Health as a hospital which can handle physical examinations of foreign laborers.

Table 3 shows the recall rates and precision rates of ZA resolution calculated using equation 2 and equation 3. Errors occur in the phase when a zero anaphor refers to an entity other than the corresponding grammatical role or the antecedent of the zero anaphor in the preceding utterance.

$$\text{Precision rate of ZA resolution} = \frac{\text{No. of antecedent correctly identified}}{\text{No. of ZA candidates}} \tag{2}$$

$$\text{Recall rate of ZA detection} = \frac{\text{No. of antecedent correctly identified}}{\text{No. of ZA occurred in text}} \tag{3}$$

Cases ZAs	ZA Triple rules	ZA Triple rules + constraints
No. of ZAs	2216	2216
ZA Candidates	3400	2754
Precision Rate	65.2%	80.5%

Table 2. Results of ZA detection

Cases Accuracy	simple antecedent identification rule	employ centering theory
Recall Rate	65.8%	70%
Precision Rate	55.3%	60.3%

Table 3. Results of ZA resolution

6   Conclusions

In this paper, we develop an inexpensive method of Chinese ZA resolution that works on the output of a part-of-speech tagger and uses a shallow parsing instead of a complex parsing to resolve zero anaphors in Chinese texts. In our preliminary experiment, we deal with the cases of topic or subject, and object omission. The precision rate of ZA detection is 81% and the recall rate of ZA resolution is 70%. The errors of ZA resolution are in the following cases:

1. Out of the grammatical role criteria (ranking of forward-looking centers): When a ZA refers to an entity other than the corresponding grammatical role or the antecedent of the zero anaphor in the preceding utterance.
2. Out of local coherence: The antecedent of a ZA is mentioned in more previous utterances.
3. Cataphora: When a ZA refers to an antecedent mentioned in the succeeding utterances.
4. Other non-anaphoric cases: Depending on the background knowledge of readers, the referent of a ZA does not require expression in the text.

In case 3 and 4, we do not tend to treat non-anaphoric cases in this paper, but we can detect about 60% cataphora and exophora and 50% inverted sentences in the test corpus by employing ZA identification constraints.

We have performed the method and experiment on ZA resolution in the previous sections. The result is promising to some extent; however, there are still some problems that need further investigation, such as pronoun resolution and the applications of ZA resolution.

In the task of pronoun resolution, because the pronominal anaphors are expressed in discourse, the detection rules are unnecessary to the task of pronoun resolution. We may modify the antecedent identification rule mentioned in 3.3 to identify the antecedents of pronominal anaphors occurring in utterances and some anaphora resolution factors can be used, such as gender and number agreement (Lappin and Leass 1994).

Another line of research to be undertaken in the future is the enhancement of the shallow parsing technique we used in this paper. For example, one might enhance the output of text chunking, without analyzing each phrase structure in an utterance but by dividing each clause within an utterance into syntactically correlated parts of words. We would also further extend our approach to dealing with other omission cases, such as verb omission and conduct more experiments on texts from other domains.

7   Acknowledgement

8   References

Abney, Steven, 1991, Parsing by chunks, In Robert Berwick, Steven Abney, and Carol Tenny, editors, *Principle-Based Parsing*, Kluwer Academic Publishers.

Abney, Steven, 1996, Tagging and Partial Parsing, In: Ken Church, Steve Young, and Gerrit Bloothooft (eds.), *Corpus-Based Methods in Language and Speech*, An ELSNET volume, Kluwer Academic Publishers, Dordrecht.

Aone, Chinatsu and Bennett, Scott William, 1995, Evaluating automated and manual acquisition of anaphora resolution strategies, *Proceedings of the 33rd Annual Meeting of the ACL*, Santa Cruz, New Mexico, pages 122–129.

Baldwin, Breck, 1997, CogNIAC: high precision coreference with limited knowledge and linguistic resources, ACL/EACL workshop on Operational factors in practical, robust anaphor resolution.

Chen, F.-Y., Tsai, P.-F., Chen, K.-J. and Huang, C.-R., 1999, Sinica Treebank, *Computational Linguistics and Chinese Language Processing (CLCLP)*, 4(2): 87-104.

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.

CKIP, 1999, 中文自動斷詞系統 Version 1.0 (Autotag), *http://godel.iis .sinica.edu.tw /CKIP/*, Academia Sinica.

Connoly, Dennis, Burger, John D. and Day, David S., 1994, A Machine learning approach to anaphoric reference, *Proceedings of the International Conference on New Methods in Language Processing*, 255-261, Manchester, United Kingdom.

Ferrández, A., Palomar, Manuel and Moreno, Lidia, 1998, Anaphor Resolution in Unrestricted Texts with Partial Parsing, *Proceedings of the 18th International*

Conference on Computational Linguistics (COLING'98)/ACL'98 Conference, pages 385-391. Montreal, Canada.

Gazdar, G. and Mellish, C., 1989, *Natural Language Processing in PROLOG – An Introduction to Computational Linguistics*, Addison- Wesley.

Ge, Niyu, Hale, John and Charniak, Eugene, 1998, A statistical approach to anaphora resolution, *Proceedings of the Sixth Workshop on Very Large Corpora*, pages 161 –170

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., 1995, Centering: A Framework for Modeling the Local Coherence of Discourse, *Computational Linguistics,* 21(2), pp. 203-225.

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

Kennedy, Christopher and Boguraev, Branimir, 1996, Anaphora for everyone: pronominal anaphora resolution without a parser, *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*, 113-118. Copenhagen, Denmark.

Lappin, S. and Leass, H., 1994, An algorithm for pronominal anaphor resolution, *Computational Linguistics*, 20(4).

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

Li, X. and Roth, D., 2001, Exploring Evidence for Shallow Parsing, *Proceedings of Workshop on Computational Natural Language Learning*, Toulouse, France.

Mitkov, Ruslan, 1998, Robust pronoun resolution with limited knowledge, *Proceedings of the 18th International Conference on Computational Linguistics (COLING'98)/ACL'98 Conference*. Montreal, Canada.

Mitkov, Ruslan, 1999, Anaphora resolution: the state of the art, Working paper (Based on the COLING'98/ACL'98 tutorial on anaphora resolution), University of Wolverhampton, Wolverhampton.

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.

Seki, Kazuhiro, Fujii, Atsushi, and Ishikawa, Tetsuya, 2002, A Probabilistic Method for Analyzing Japanese Anaphora Integrating Zero Pronoun Detection and Resolution, *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, pp.911-917.

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.

Sinica Treebank, 2002, URL *http://turing.iis.sinica.edu.tw/treesearch/*, Academia Sinica.

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

Stuckardt, Roland, 2002, Machine-Learning-Based vs. Manually Designed Approaches to Anaphor Resolution: the Best of Two Worlds, *Proceedings of the 4th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC2002)*, University of Lisbon, Portugal, pages 211-216.

The Penn Chinese Treebank Project, 2000, URL *http://www.cis.upenn.edu/~chinese/*. Linguistic Data Consortium, University of Pennsylvania.

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., 1994, Japan Discourse and the Process of Centering, *Computational Linguistics,* 20(2): 193-233.

Yeh, Ching-Long and Chen, Yi-Chun, 2001, An empirical study of zero anaphora resolution in Chinese based on centering theory, *Proceedings of ROCLING* XIV, Tainan, Taiwan.

Yeh, Ching-Long and Chen, Yi-Chun, 2003, Using Zero Anaphora Resolution to Improve Text Categorization, *Proceedings of PACLIC 17*, Sentosa, Singapore.

9   Appendix: Abbreviations

In the word-by-word translation, some markers are abbreviated as below. We follow the abbreviations used in [1].

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)