ZERO ANAPHORA RESOLUTION IN CHINESE WITH PARTIAL PARSING BASED ON CENTERING THEORY *

Ching-Long Yeh and Yi-Chun Chen
Department of Computer Science and Engineering
Tatung University
40 Chungshan N. Rd. 3rd. Section
Taipei 104
Taiwan
chingyeh@cse.ttu.edu.tw d8806005@mail.ttu.edu.tw

ABSTRACT

Most traditional approaches to anaphora resolution are based on the integration of complex syntactic information and domain knowledge. However, to construct 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 a partial parsing instead of a complex parsing to resolve zero anaphors in Chinese text. We employ centering theory and constraint rules to identify the antecedents of zero anaphors appeared in the preceding utterances. In this paper, we focus on the cases of zero anaphors that occur in the topic or subject, and object positions of utterances. The 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.

1. INTRODUCTION

In Chinese text, anaphoric expressions are frequently eliminated, termed zero anaphor (ZA) hereafter, due to their prominence in discourse [1]. Zero anaphors are generally noun phrases that are understood from the context and do not need to be specified. Since zero anaphors are not expressed in discourse, the task of ZA resolution is divided into two phases, first detecting zero anaphors and then identifying their antecedents. In this paper, we focus on the most important cases of ZA that occur in the topic or subject, and object positions of utterances.

The methods of anaphora resolution can be classified into traditional and alternative approaches. The former integrates different knowledge sources or factors (e.g. gender and number agreement, c-command constraints, semantic information) that discount unlikely candidates until a minimal set of plausible candidates is obtained [2,3,4,5,6]. 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 a domain knowledge base. The latter employs statistical models or AI techniques, such as machine learning, to compute the most likely candidate [7,8,9,10]. 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 [11].

A recent trend is in search of inexpensive, fast and reliable procedures for anaphora resolution [12,13,14,15]. The approach relies on cheaper and more reliable NLP tools such as part-of-speech (POS) tagger and shallow parsers. Our approach can be included in these approaches, which rely on limited knowledge and only need partial syntactic parsing of text. The resolution process works from the output of a POS tagger enriched with annotations of grammatical function of lexical items in the input text stream. The partial parsing technique is used to detect zero anaphors and identifies the noun phrases preceding the anaphors as antecedents. We have carried out an experiment using a number of news articles. The result shows that the precision rate of zero anaphora detection is 80% and within the detected zero anaphors 70% can be resolved correctly.

* This research was supported by the Taiwan National Science Council under Contract No. NSC92-2422-H-036-322.

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 frightened and ran outside.
b. ϕ_1^i 撞到 一個 人^{j}
(He) bumped into a person.
c. 他$_2^i$ 看清 了 那 人^{j} 的 長相
He saw clearly that person's appearance.
d. ϕ_3^i 認出 那 人^{j} 是 誰
(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 [16][1].

According to Li and Thompson [1], 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)[2] while antecedent and anaphors are located in different sentences in the intersentential case[3], such as the ϕ_1^i and ϕ_3^i in (1).

(2) 房子 ϕ 蓋好 了
The house, (someone) has finished building it.

3. ZA RESOLUTION METHOD

The ZA resolution method we develop is divided into three parts. First we use a POS tagger to produce the tagged result of an input document. Second is ZA detection that identifies occurrences of ZA within utterances by employing detection rules based on the result of partial parsing. Third is antecedent identification that identifies the antecedent of each detected ZA using rules based on the centering theory.

3.1. Partial Parsing

Partial (or shallow) parsing does not deliver full syntactic analysis but is limited to parsing smaller constituents such as noun phrases or prepositional phrases [17,18]. For example, the sentence (3) can be divided as follows:

(3) 花蓮 成爲 熱門 的 旅遊 地點。
Hualien became the popular tourist attraction.
→ [NP 花蓮] [VP 成爲] [NP 熱門 的 旅遊 地點]
[NP Hualien] [VP became] [NP the popular tourist attraction]

In our work, we use a number of simple noun phrase rules to identify the noun phrases in the output produced by AUTOTAG which is a POS tagger developed by CKIP, Academia Sinica [19]. For example, the result of (3) produced by AUTOTAG is as below.

(4) 花蓮 成爲 熱門 的 旅遊 地點。
→ [花蓮(Nc) 成爲(VG) 熱門(VH) 的(DE) 旅遊(VA) 地點(Na) 。]

There are about 47 POS tags used in AUTOTAG, including 13 noun POS tags, 17 POS verb tags and 1 preposition POS tag. We create simple noun phrase, verb phrase prepositional phrase rules in DCG [20].

3.2. Centering Theory

In the centering theory [2,21,22], 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. Grosz *et al*., in their paper [2], 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 [2,21].

Constraints

---

[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] The intrasentential case, sentence (2), is taken from Li and Thompson (Li and Thompson, 1981).

[3] We do not tend to explain the detail of ZA in Chinese in this paper. See Yeh's dissertation [24] for a detailed account.

For each utterance U_i in a discourse segment U_1, …, 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 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:

Rules

For each utterance U_i in a discourse segment U_1, …, 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 [2].

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 (5), the subject of the utterance (5b) is eliminated, and its antecedent is identified as the subject of the preceding utterance (5a) according to the centering theory.

(5) a. 電子股^{i} 受 美國高科技股 重挫 影響，
Electronics stocks were affected by high-tech stocks fallen heavily in America.
b. ϕ^{i} 持續 下跌。
(Electronics stocks) continued falling down.

3.3. 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 simple syntactic relations to detect omitted cases as ZA candidates::

ZA detection rules:

1. For each utterance U_i in a discourse segment U_1, … , U_m: 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, … , 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.

After we employ the ZA detection rules to detect omitted cases as ZA candidates and then we further use the ZA detection constraints to filter out non-anaphoric cases:

ZA detection constraints

For each ZA candidate *c* in a discourse:

1. *c* can not be in the first utterance in a discourse segment (exophora[4] or cataphora[5])
2. ZA does not occur in the following case (inverted sentence):
NP + (coordinating conjunction / preposition) + NP + VP + *c*

In the antecedent identification phase, we employ the concept, '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 [6]:

Antecedent identification rule:

For each zero anaphor *z* in a discourse segment U_1, … , 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

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

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 [1], 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. Since the antecedent identification rule is corresponding to the concept of centering theory.

4. 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 first only employ ZA detection rules as the baseline, and then plus ZA detection constraint to see the result. 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.

4.1. ZA Detection

By employing the ZA detection rules and constraints mentioned in Section 3.3, zero anaphors occur in topic or subject, and object positions can be detected. In the experiment, we first only employ ZA detection rules, and then include the ZA detection constraints to see the improvement. The result shows 100% of topic or subject, and object omission cases are detected, and the precision rates calculated using equation 1 show in the Table 1.

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

The main errors of ZA detection occur in the experiment when parsing inverted sentences and non-anaphoric cases (e.g. exophora or cataphora) [18,23]. 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 ZA is in the following utterances, but we can detect about 60% cataphora in the test corpus by employing ZA detection constraint 1.

4.2. Antecedent Identification

In this phase, we take the output of employing ZA detection rules and 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 section 3.3 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 [2]. By comparing with the simple antecedent identification rule, the antecedent identification rule based on the centering theory (see section 3.3) determines the antecedents according to grammatical role criteria. For example, in the discourse segment (6), the zero anaphors are detected in the utterances (6b) and (6c). According to the antecedent identification rule, the noun phrase, 基隆醫院 'Kee-lung General Hospital,' whose grammatical role is corresponding to the zero anaphor ϕ_1^i in (6b) is identified as the antecedent. Subsequently, the antecedent of the zero anaphor ϕ_2^i in (6c) is identified as the antecedent of ϕ_1^i in (6b), 基隆醫院

(6) a. 基隆醫院 ^{i} 爲 擴大 服務 範圍，
Kee-lung General Hospital aims to increase service coverage.

b. ϕ_1^i 積極 提升 醫療 服務 品質 及 標準化，
(It) actively improves the service quality of medical treatment and standardization.

c. ϕ_2^i 獲 衛生署 認可 爲 辦理 外勞體檢醫院。
(It) is certified by Department of Health as a hospital which can handle physical examinations of foreign laborers.

Table 2 shows the recall rates and precision rates of ZA resolution. 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.

Table 1: Results of ZA detection

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

Table 2: Results of ZA resolution

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

5. 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 partial parsing instead of a complex parsing to resolve zero anaphors in Chinese text. 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 detection 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 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 [3].

Another future works is to enhance the partial parsing technique we used in this paper, for example, enhance the output of text chunking, without analyzing each phrase structure in an utterance but divide each clause within an utterance into syntactically correlated parts of words. We will further extend our approach to dealing with other omission cases, such as indirect object omission and take more experiments on texts from other domains.

References

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

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

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

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

[5] Walker, M. A. 1998. Centering, anaphora resolution, and discourse structure. In Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince, editors, Centering in Discourse. Oxford University Press.

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

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

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

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

[10] Kazuhiro Seki, Atsushi Fujii, and Tetsuya Ishikawa. 2002. A Probabilistic Method for Analyzing Japanese Anaphora Integrating Zero Pronoun Detection and Resolu-

tion. *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, pp.911-917.

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

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

[13] A. Ferrández, Manuel Palomar, Lidia Moreno. 1998. Anaphor Resolution in Unrestricted Texts with Partial Parsing. *Proceedings of the 18.th International Conference on Computational Linguistics (COLING'98)/ACL'98 Conference*, pages 385-391. Montreal, Canada.

[14] Kennedy, Christopher and Branimir Boguraev. 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.

[15] 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.

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

[17] Steven Abney. 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.

[18] Mitkov, Ruslan. 2002. *Anaphora Resolution*, Longman.

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

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

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

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

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

[24] Ching-Long Yeh. 1995. *Generation of Anaphors in Chinese*, Ph.D. thesis, University of Edinburgh.

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

[26] C. L. Sider. 1983. Focusing in the comprehension of definite anaphora. *Computational Models of Discourse*, MIT Press.

[27] Walker, M. A. 1989. Evaluating Discourse Processing Algorithms. *Proc. Of ACL '89*, Vancouver, Canada.