中文零代詞解析與應用

Chinese Zero Anaphora Resolution and Its Applications

研究生：陳貽浚（Yi-Chun Chen）

指導教授：葉慶隆（Prof. Ching-Long Yeh）

大同大學

資訊工程研究所

博士論文

Ph.D. Dissertation
Department of Computer Science and Engineering
Tatung University

中華民國九十四年七月

July 2005

中文零代詞解析與應用

博士生： 陳貽浚

指導教授： 葉慶隆

大同大學

資訊工程研究所

博士論文

中華民國九十四年七月

Chinese Zero Anaphora Resolution and Its Applications

Student's name: Yi-Chun Chen
Advisor's name: Ching-Long Yeh

Ph.D. Dissertation
Department of Computer Science and Engineering
Tatung University

July 2005

大同大學

資訊工程研究所

博士學位論文

中文零代詞解析與應用

陳貽浚

經考試合格特此證明

博士學位論文考試委員

張俊盛

陳煇煌

葉慶隆

胡毓志

指導教授

葉慶隆

所長

包蒼龍

中華民國 94 年 7 月 31 日

誌 謝

回想起這六年博士班的日子，一路走來，無論在學業、工作或生活上的點點滴滴，都必定是我此生最難忘，也是收穫最豐富的時光，感謝所有關愛我的人，謝謝你們給我的提攜與幫助。從進入大同資訊工程所碩士班開始，轉眼已經過了八年，首先要感謝的是我的指導教授葉慶隆老師，老師在研究上的教導與協助，引領學生進入了自然語言處理的領域，而老師對學生關心和鼓勵，更讓學生在作研究時可以勇往直前，謹此再向敬愛的恩師致上最深的感謝。此外，感謝所上的師長們在學問及智慧上的啟蒙。還要感謝校內外口試委員：陳輝煌教授、謝忠健教授、虞台文教授、林金城教授、張俊盛教授、胡毓志教授、梁婷教授等，對於論文不足之處加以指正與建議，使本論文更加完善，在此致上深忱的謝意。我也要感謝研究室裏的學弟妹們平日的幫忙與照顧，讓我感受到許多的人情溫暖。

在大同攻讀博士學位期間，特別要感謝我的父母，在這些年來持續地支持與鼓勵，讓我在學習的路上無後顧之憂。尤其要感謝的是我新婚的老婆，在撰寫論文的期間，經常適時的予以體諒與安慰，讓我可以順利的的完成論文。最後致上我最深的感激給我的家人，期盼將來能有小小的成就，以回報家人對我的付出。

陳貽浚

民國九十四年七月

大同大學

ABSTRACT

Anaphora resolution is the task of determining the antecedent of an anaphor which can be zero, pronominal and nominal forms. It plays an increasingly important role in a number of natural language processing applications including machine translation, information retrieval, text summarization, etc. In this thesis, we aim to investigate computational resolution of zero anaphora in Chinese text and apply the resolution method on NLP applications for examining its performance. The work of zero anaphora resolution is divided into two steps: First, we investigate linguistic behavior of Chinese zero anaphora and computational approaches to anaphora resolution for developing the method of Chinese zero anaphora resolution. Second, the zero anaphora resolution system is implemented according to results of the first step. On completing the implementation, an evaluation of the system is performed on real news articles. Because zero anaphors are not expressed on the surface text, our resolution method is first to detect zero anaphors in each utterance, and then identify their antecedents in the preceding utterance.

After the method of zero anaphora resolution is carried out, we adopt the resolution method as a basis for improving the accuracy of NLP applications. A text categorization system integrates the zero anaphora resolution process to recover the omissions of anaphors in query text. An information retrieval system employs a topic identification method to resolve the omissions of topics of documents in the text collection for creating better indices. The topic identification method is developed by employing the notion of the centering model and the zero anaphora resolution method and is further used to create the metadata of XML Topic Maps. The experiments of these applications demonstrate on text collection taken from several newspapers, such as China Times Express and Central Daily News.

摘 要

在一般口語表達或文章書寫時，我們常利用代詞(anaphor)來取代之前已經提過的詞語，而代詞解析(Anaphora resolution)就是在文句中找出代詞所參考的先行詞的方法。代詞解析的技術也在一些自然語言處理的應用上扮演著重要的角色。在本篇論文中，我們提出了一種基於重心理論(centering theory)中文零代詞解析方法，並將此法使用在文件分類(text classification)、資訊擷取(information retrieval)與主題地圖(topic map)資訊建立等自然語言處理的應用上。中文零代詞解析的研究分為兩個階段：首先，我們參考了中文語言學中描述零代詞現象的文獻，以及代詞解析的計算機理論後，發展出中文零代詞的解析方法；接下來，我們採用此解析方法實作出中文零代詞解析系統，並以真實的新聞文件作實驗，來驗證系統的解析能力。

為了測試中文零代詞解析在自然語言處理應用上的效果，我們分別製作了一個文件分類系統與資訊擷取系統，並以中時晚報與中央日報等新聞為測試文件作實驗。文件分類的實驗中，我們先將輸入的查詢文件作過零代詞解析，再觀察分類正確率的提升程度。資訊擷取的實驗中，我們提出了一種基於零代詞解析技術的文句主題辨識(topic identification)方法，並進一步利用此法辨識出測試文件中每個文句的主題，再觀察資訊擷取召回率(recall rate)與準確率(precision rate)的提升程度。除了這二個自然語言處理應用之外，我們也利用了上述的主題辨識方法，提出一種主題地圖資訊建立方法，試圖以自動取得文件主題的方式，建立主題地圖中的主題資訊。

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Motivation and Goal

In natural languages, elements that can be deduced contextually by the reader are frequently omitted from expressions in texts. The elimination of anaphoric expressions referring to the elements mentioned previously in context is termed zero anaphor (ZA) which often occurs in Chinese, due to their prominence in discourse [Li and Thompson 1981]. ZAs are generally noun phrases (NPs) that are understood from the context and do not need to be specified. As shown in Example (1.1) the subject of Utterance (1.1a) is 小柯 ‘Xiaoke,’ which is eliminated in the following Utterance (1.1b), (1.1c) and (1.1d). The omission may cause considerable problems in natural language processing (NLP) applications. For example in a machine translation system, a Chinese text can not be translated properly into text in a target language without identifying the occurrences of the omissions first. In information extraction, the events having subjects omitted in texts can not be identified correctly. Zero anaphora resolution is the process of detecting the occurrences of ZAs and identifying their antecedents. It plays an important role in a number of Chinese NLP applications, including machine translation, information retrieval, question answering and text summarization etc.

(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 this thesis, our goal is to develop the method of Chinese zero anaphora resolution and then apply the method on NLP applications for evaluating its performance. First of all, we investigate related linguistic studies and computational approaches of anaphora resolution to develop the method. Second, we concentrate our attention on the implementation of a zero anaphora resolution system employing the resolution method in the first part. On completing the implementation, we then carry out an evaluation of the results of the system. Third, for verifying the efficiency of the zero anaphora resolution method in real NLP applications, we develop a text categorization system integrating the zero anaphora resolution process to show the accuracy improvement. An information retrieval system which uses topic identification to obtain more valuable information embedded in text is also implemented. The topic identification method employs the notion of the centering model and the zero anaphora resolution method to identify the topic of each utterance in text, and can be further used to create the metadata of XML Topic Maps (XTM) [Pepper and Moore 2001]. The experiments of these applications demonstrate on real news articles, which are collected from several newspapers, such as China Times Express and Central Daily News.

1.2 Methodology

There are several methods of anaphora resolution. One method is to integrate different knowledge sources or factors (e.g. gender and number agreement, c-command constraints, semantic constraints) that reduce unlikely candidates until a minimal set of plausible candidates is obtained [Carbonell and Brown 1988, Lappin and Leass 1994, Okumura and Tamura 1996, Walker *et al*. 1998, Yeh and Chen 2001]. The relationships 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 from both linguistic and domain sources. Another method employs statistical models or AI techniques, such as machine learning, to compute the most likely candidates [Aone and Bennett 1995, Connoly *et al*. 1994, Ge *et al*. 1998, Seki *et al*. 2002]. The problems mentioned previously can be sorted out by using this method. However, it heavily relies upon the availability of tagged text corpora that are sufficiently large, 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, Yeh and Chen 2003]. The approach relies on reliable and cheaper NLP tools such as part-of-speech (POS) tagger and shallow parsers. In this thesis, we adopt this approach, which works on the output of a part-of-speech tagger and uses shallow parsing instead of complex syntactic and semantic analysis to resolve ZAs in Chinese text. Since

ZAs are not expressed in text, the task of zero anaphora resolution is divided into two phases: first detecting the occurrences of ZAs in text, and then finding their antecedents in the discourse. For developing our shallow parser, we establish the *Triple Rules* to transform utterances in texts into triples[1] and detect ZAs. A POS tagger and the following shallow parser are used to accomplish the task of the first phase. We then employ the centering model [Grosz *et al*. 1995] to develop a rule-base as the basis to determine the antecedents of ZAs found in the first phase.

In the implementation of this zero anaphora resolution method, we employ a Chinese word segmentation and POS annotation system [CKIP 2003], termed AUTOTAG[2] hereafter, developed by CKIP, Academia Sinica, to segment Chinese into lexical items and annotate their POS information of the input text stream. A shallow parser employing the *Triple Rules* is developed to parse smaller syntactical constituents such as noun phrases and verb phrases [Abney 1996, Li and Roth 2001, Mitkov 1999], which are transformed into *triple* representations. Finally, the key elements of the centering model of local discourse coherence are employed to identify the antecedents of ZAs.

In the thesis, the applications employing the zero anaphora resolution method includes text categorization, information retrieval and the creation of topic maps [Yeh and Chen 2003a, Yeh and Chen 2004b, Yeh and Chen 2004c]. The text categorization system integrates the zero anaphora resolution process to recover the occurrences of ZAs in the input query text for the accuracy improvement of categorization. In the categorization system, the term frequency / inverse document frequency term weighting scheme is

[1] The *triple* here is a 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. See Chapter 5 for details of triple representations and the *Triple Rules*.

[2] The Chinese word segmentation and POS annotation system [CKIP 2003] is a revised online system. An earlier system, AUTOTAG, is also developed by CKIP and released as a downloadable program in 1999 [CKIP 1999].

utilized to calculate the weight of each term extract from training data and the *k*-nearest neighbor classifier is used to classify the input query documents. The information retrieval system we developed uses topic identification to extract the topic of each utterance for creating better indices of the articles in the test collection. The topic identification method is further adopted to identify topic chains in text for creating the metadata of topic maps. The metadata includes two child elements of the occurrence, *resourceRef* and *resourceData* [Pepper and Moore 2001]. Once the topic chains of a document are identified, the metadata of a topic node in topic maps or the information relevant to the topic of the document can also be created.

1.3 System Overview and Application Architecture

The zero anaphora resolution system we develop in this thesis is mainly divided into two phases: ZA detection that is to parse each utterance and recognize the omitted elements as ZAs, and antecedent identification that is to determine the antecedent of each ZA, as illustrated in Figure. 1.1.

As shown in the figure, the ZA detection phase starts by word segmentation and POS tagging. By taking Chinese text as the input, the word segmentation and POS tagging program consults the lexicon and some heuristic rules to segment Chinese words and annotate their POS information. Then the shallow parser employs chunking rules to parse smaller syntactical constituents of each utterance, and then transforms the utterances into triple representations according to *Triple Rules*. In this step, each ZA candidate is detected by the shallow parser. The antecedent identification program takes the output of the shallow parser and utilizes ZA identification constraints to eliminate non-ZA cases. Finally, the antecedent identification rules are employed by the antecedent identification program to determine the antecedents of ZAs.

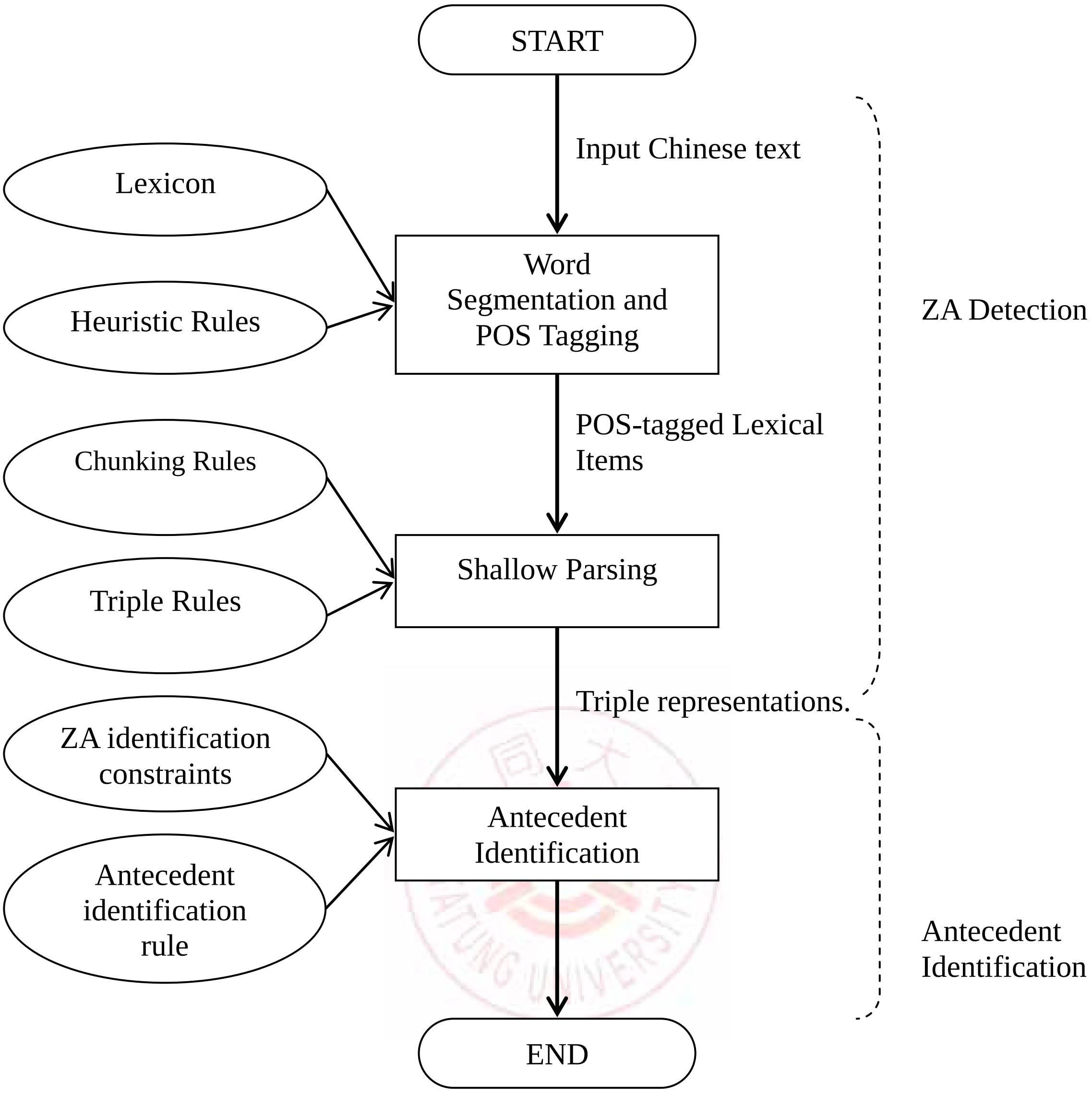


Figure 1.1: System overview of the Chinese zero anaphora resolution system

The applications including a text categorization system and an information retrieval system employ the zero anaphora resolution method recover the omissions of anaphors for improving the accuracy. As shown in Figure 1.2, the occurrences of ZAs in the query text are resolved first, and the ZA-resolved text is then taken as the input of the query program. Figure 1.3 illustrates the information retrieval system. We employ the topic identification method[3] which is based on the zero anaphora resolution method and the centering model [Yeh and Chen 2004b] to identify topics of utterances in text for creating better indices of the text collection.

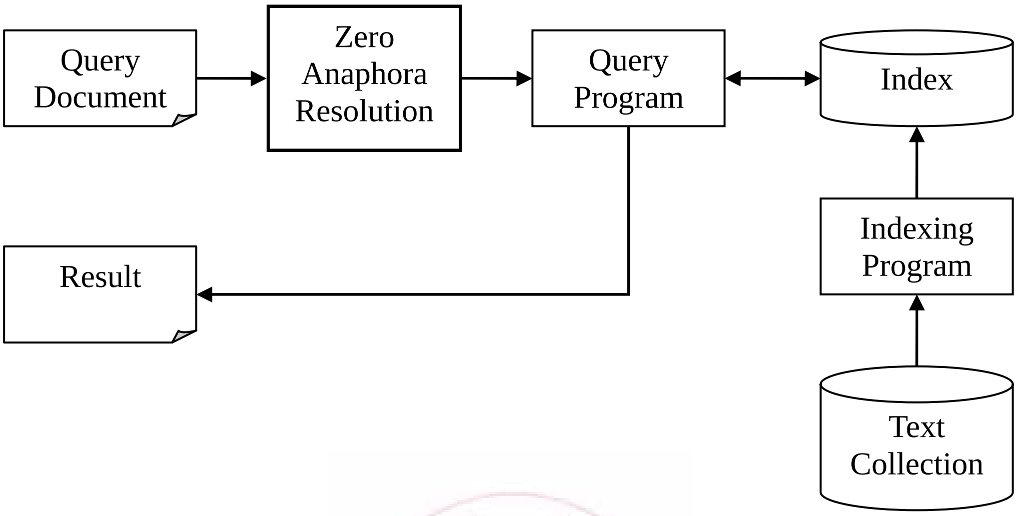


Figure 1.2: System Architecture of the Text Categorization System

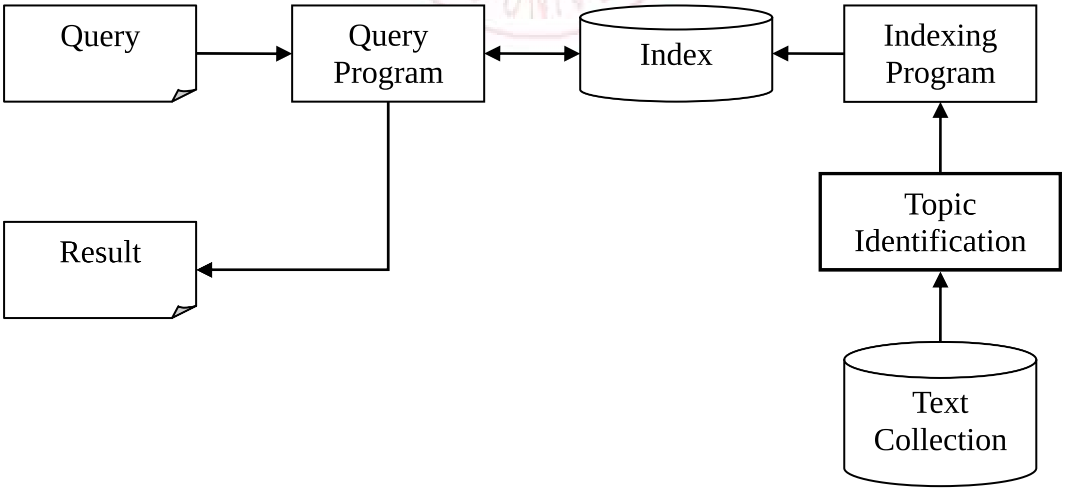


Figure 1.3: System Architecture of the Information Retrieval System

[3] See Chapter 6 for details.

1.4 Scope of Thesis

In this thesis, we concentrate on resolving ZAs whose antecedents are noun phrases in Chinese written text. We are not concerned with non-anaphoric cases (e.g. exophora[4] or cataphora[5]) [Halliday and Hasan 1976] that may have anaphoric forms in text but their antecedents are not expressed in the preceding utterances.

In the scope of zero anaphora investigated in this thesis, of the two types of ZAs, intra- and inter-sentential, we are only concerned with the latter. In our zero anaphora resolution procedure, we will ignore the occurrence of intra cases in the test texts. Furthermore, we do not intend to resolve long distance ZAs. By our observation, we found that they occurred far less frequently than their non-zero counterparts in the real texts. The decision not to resolve them is based on this experimental result. Another reason for not resolving them is that, from the computational point of view, it is impractical to spend a lot of effort on a few cases.

We do not aim at developing a general account for anaphoric phenomena occurring in various kinds of text. Instead we focus on investigating anaphors occurring in narrative texts and the test corpus selected from for this work is Chinese news articles.

This work, though it includes the implementation of a Chinese shallow parser, does not intend to invent any new idea on shallow parsing apart from its treatment of zero anaphora resolution. The Chinese shallow parser is mainly taken as the framework to parse syntactical constituents and employ the *Triple Rules* we establish for ZA detection. Thus we adopt concepts from existing shallow parsing techniques to develop our shallow parser so that it can detect ZA occurring in utterances. To accomplish the task of zero anaphora resolution, in addition to the shallow parser, the system involves a POS tagger

[4] Exophora is reference of an expression directly to an extralinguistic referent in which the referent does not require another expression for its interpretation.

[5] Cataphora arises when a reference is made to an entity mentioned subsequently.

and an antecedent identification component. We do not intend to develop a POS tagger with Chinese word segmentation process, which is another issue in NLP, but use an available POS-tagger in the system.

In the applications of NLP, we build an information retrieval system and an text categorization system by employing well-known methods like term frequency / inverse document frequency (TFIDF) word weighting scheme [Salton and Buckley 1988] and *k*-nearest neighbor (*k*-NN) classifier [Yang *et al*. 2002, Ko and Seo 2002]. We do not focus on comparing different traditional approaches on these applications, but integrate the zero anaphora resolution procedure into them to present the accuracy improvement.

1.5 Contributions

The main contribution of this thesis is a Chinese zero anaphora resolution method and system employing a set of computational rules based on the centering model. In contrast to other anaphor resolution work that integrates complex linguistic information or relies on the labor-intensive and time-consuming construction of a domain knowledge base [Lappin and Leass 1994, Okumura and Tamura 1996, Walker *et al*. 1998], our rules were established by integrating simple Chinese syntax to detect ZAs. In antecedent identification, according to the observations on real data, we adopt the centering model of local discourse coherence The experiments we carried out for the zero anaphora resolution not only show the effectiveness of the method, but also can be used as an essential procedure for a number of NLP applications.

This work focuses on investigating the resolution of ZA in Chinese, which contrasts with previous work on other languages, like anaphora resolution in English and Spanish [Lappin and Leass 1994, Mitkov 1998, Palomar *et al*. 2001], or zero anaphora resolution in Japanese [Okumura and Tamura 1996, Walker et al. 1998]. This work can provide a starting basis towards the study of anaphor resolution in a multilingual environment.

In practical terms, we develop different NLP applications integrating zero anaphora resolution. The work is to experiment on real Chinese news articles and the results show the method of zero anaphora resolution would make contribution to the performance of Chinese text categorization and information retrieval.

1.6 Thesis Organization

The rest of this thesis is organized as followed. First, surveys on linguistic aspects and the methods of anaphora resolution are given in Chapters 2 and 3. In Chapter 4, we describe the zero anaphora resolution method based on the centering model. The implementation and evaluation of the Chinese zero anaphora resolution system including the shallow parser are described in Chapter 5. After the zero anaphora resolution system is built, the NLP applications integrating the zero anaphora resolution procedure are illustrated in Chapter 6. In Chapter 7 the experiments and results of the applications are presented. Finally, Chapter 8 summarizes the results and suggests areas for future research.

CHAPTER 2

RELATED LINGUISTIC BACKGROUND

In this chapter, we briefly introduce the linguistic background related to the research in this thesis. We start by describing the role of topic in Chinese grammar and that Chinese is a topic-prominent language. We then introduce various kinds of anaphors in Chinese, including zero, pronominal and nominal anaphors. After introducing the above concepts, we present a survey of linguistic studies on the Chinese zero anaphora. We also illustrate the difference between anaphora and non-anaphora cases (e.g. cataphora and exophora) in these studies.

2.1 Topic Prominence in Chinese

One of the most striking characteristics in a topic-prominent language like Chinese is that in addition to the grammatical relations of “subject” and “direct object,” the description of Chinese must also include the important element, “topic,” which can represent what the sentence is about [Li and Thompson 1981]. The topic of a sentence always comes first in the sentence, and it always refers to something about which the writer assumes 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. In Example (2.1), 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.

(2.1) 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 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 Example (2.2) to (2.5), respectively [Li and Thompson 1981].

(2.2) 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.

(2.3) 張三 打 我 了。

Zhangsan da wo le.

Zhangsan hit I CRS

Zhangsan hit me.

(2.4) 衣服 燙 完 了。

yifu tang wan le.

cloth iron finish CRS

The clothing (someone) has finished ironing it.

(2.5) 進來 了 一 個 人。

jin-lai le yige ren.

enter-come ASPECT one CL person

A person came in.

Example (2.5) 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 [Li and Thompson 1981]. Instead, the indefinite subject noun phrase must be placed after the verb. Besides, topics in Chinese sentences must be either definite or generic. Consequently, the only noun phrase in this sentence, 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]. As shown in Example (2.6), Utterance (2.6b) to (2.6e) do not contain the subjects/topics but refer to 張三 ‘Zhangsan’ in the preceding Utterance (2.6a). 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.

(2.6) a. 張三^i 好不容易 從 台南 調回 台北 ，

Zhangsan^i haoburongyi cong Tainan diaohui Taipei.

Zhangsan difficult from Tainan transfer-back Taipei

Zhangsan managed to get transferred from Tainan back to Taipei.

b. ϕ^i_1 結 了 婚 ，

ϕ^i_1 jie le hun.

(he) get ASPECT marry

(He) got married.

c. ϕ^i_2 成 了 家 ，

ϕ^i_2 cheng le jia.

(he) start ASPECT family

(He) started a family.

d. ϕ^{i}_{3} 有 一 個 小 兒子 ，

ϕ^{i}_{3} you yi ge xiao erzi.

(he) have one CL little son

(He) has a baby son.

e. ϕ^{i}_{4} 還 蠻 幸福 美滿 的。

ϕ^{i}_{4} hai man xingfu meiman de.

(he) quite happy content SFP

(He) is actually quite happy and contented.

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 Example (2.5)[6] 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 sentence. In Chinese, anaphors can be in one of zero, pronominal and nominal forms. In the next section, we give an overview of various kinds of anaphors.

2.2 Zero Anaphora in Chinese

As mentioned previously, ZAs are generally noun phrases that are understood from the context and do not need to be specified. In Example (2.7), the topic of Utterance (2.7a) is 張三 ‘Zhangsan’ which is eliminated in the second utterance. In addition to ZAs,

[6] Example (2.5) 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].

anaphors can be pronominal and nominal forms, as exemplified by 他 ‘He’ and 那個人 ‘that person’ in Utterance (2.7c) and (2.7d), respectively [Chen 1987][7].

(2.7) a. 張三^{i} 驚慌 的 往 外 跑 ，

Zhangsan^{i} jinghuang de wang wai pao.

Zhangsan frightened CSC towards outside run

Zhangsan frightened and ran outside.

b. ϕ^{i}_{1} 撞到 一個 人^{j} ，

ϕ^{i}_{1} zhuangdao yi ge ren.

(he) bump-to a person

(He) bumped into a person.

c. 他$^{i}_{2}$ 看清 了 那人^{j} 的 長相 ，

ta$^{i}_{2}$ kanqing le na ren de zhangxiang.

he see-clear ASPECT that person GEN appearance

He saw clearly that person’s appearance.

d. ϕ^{i}_{3} 認出 那 人^{j} 是 誰 。

ϕ^{i}_{3} renchu na ren shi shei.

(he) recognise that person is who

(He) recognized who that man is.

According to Li and Thompson [Li and Thompson 1981], ZAs can be classified as intrasentential or intersentential. In the former case, the ZA and its antecedent exist in the

[7] We use a ϕ^{b}_{a} to denote a ZA, where the subscript *a* is the index of the ZA itself and the superscript *b* is the index of the antecedent. A single ϕ without any script represents an intrasentential ZA. Also note that a superscript attached to a noun phrase is used to represent the index of the antecedent.

same sentence. In the later case, and the ZA can be understood and does not need to be expressed. Consider Example (2.8), the ϕ refers to 張三 ‘Zhangsan’ of the first clause in the same sentence. In Example (2.7), the ZAs in Utterance (2.7b) and (2.7d) refer to the same antecedent 張三 ‘Zhangsan’ in the different sentences.

(2.8) 張三 參加 比賽 ϕ 贏得 一 台 電腦 。

Zhangsan canjia bisai ϕ 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^{j}_{1} in Utterance (2.9b) and ϕ^{k}_{1} in Utterance (2.9d)[8]. 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^{i}_{1} in Utterance (2.9e) whose antecedent occurs four sentences away in Utterance (2.9a).

(2.9) a. 螃蟹^{i} 有 四 對 步足^{j} ，

pangxie^{i} you si dui buzu.

crab have four-pair walking-foot

A crab has four pairs of feet.

b. ϕ^{j}_{1} 俗稱 「腿兒」 ，

[8] Example (2.9) is taken form Yeh’s thesis [Yeh 1995].

ϕ^{j}_{1} 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^{k} 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^{k}_{1} 不能 向 前後 彎曲 ，

ϕ^{k}_{1} buneng xiang qianhou wanqu.

(it) not can towards forward-backward bend

(it) can’t bend backward or forwards.

e. ϕ^{i}_{1} 爬行 時 ，

ϕ^{i}_{1} paxing shi.

(it) crawl ASPECT

When (it) crawls.

f. ϕ^{i}_{2} 必須 先 用 一 邊 步足 的 指尖 抓 地 ，

ϕ^{i}_{2} 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^{i}_{3} 再 用 另 一 邊 的 步足 直伸 起來 ，

ϕ^{i}_{3} 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^{i}_{4} 把 身體 推 過去 ，

ϕ^{i}_{4} ba shenti tui guoqu.

(it) BA body push get-through

(It) pushes the body towards one side.

Since Chinese has no inflection, conjugation, or case markers, the pronominal system is relatively simple, as shown in Table 2.1 [Li and Thompson 1981]. A third-person pronoun can be used to replace an intersentential ZA, except for first- and second-person pronouns, without changing the meaning of the sentence. As shown in Example (2.9), all of the ZAs can be replaced by third person pronouns. Though the resulting meaning of each sentence is unchanged, the whole discourse becomes less coherent.

Table 2.1: Pronominal system in Chinese

Number	**Person**	**Pronoun**
singular	first	我 (I)
singular	second	你, 妳 (you)
singular	third	他, 她, 它 (he/she/it)
plural	first	我們 (we)
plural	second	你們, 妳們 (you)
plural	third	他們, 她們, 它們 (they)

2.3 Linguistic Studies on Chinese Zero Anaphora

In Chinese text, an element that can be deduced contextually by the readers is often expressed as a zero form. The phenomenon is termed *ellipsis* in linguistics [Lü 1946, Liao 1992]. Chen [Chen 1987] further analyzed this phenomenon and divided it into *zero anaphora* and *zero cataphora*. The anaphor refers to an element mentioned previously, while the cataphor refers to an element in the following sentences.

Lü [Lü 1946] observed that there are a lot of ellipses embedded in Chinese texts. He started to discuss ellipsis by considering the subject and object of a sentence. As shown in Example (2.10), the subject in Utterance (2.10b) referring to 我 'I' in Utterance (2.10a) is omitted. In his paper [Lü 1986], he further discussed that an ellipsis can refer to only one referent in context, or the omission is treated as an implication. However, in his opinion, it is sometimes difficult to distinguish these two phenomena.

(2.10) a. 我^{i} 學生 也 教 過 多 了 ，

wo^{i} xuesheng ye jiao guo duo le.

I student also teach ASPECT many CRS

I have also taught many students.

b. ϕ^{i}_{1} 沒有 教 過 你 這樣 的，

ϕ^{i}_{1} meiyou jiao guo ni zheyang de.

(I) not teach ASPECT you such SFP

(I) had never taught one like you.

Lü [Lü 1996] indicated that, in general, a sentence has subject with it, but it exists some cases without expressing the subject. The cases include *question-answering dialogue*, *imperative*, *generic subject*, and *natural phenomenon*, which are shown in

Example (2.11) to (2.14) respectively.

(2.11) a. 他^i 收下 了 嗎 ？

ta^i shouxia le ma.

he accept CRS Q

Did he accept (it) ?

b. ϕ^i_1 收下 了，

ϕ^i_1 shouxia le.

(he) accept CRS

(He) accepted (it).

(2.12) φ 走 吧 ！

φ zou ba.

(we) go SA

Let’s go.

(2.13) 活 到 老 學 到 老。

huo dao lao xue dao lao.

live to old learn to old

It is never too old to learn.

(2.14) 下 雪 了 。

xia xue le.

descend snow CRS

It is snowing.

Example (2.12), (2.13), and (2.14) are cases of exophora [Halliday and Hasan 1976], in which the referents are referred as a specific person in a given situation or do not refer to anything specific [Lappin and Leass 1994]. In this thesis, our work focuses on the resolution of zero anaphora as clarified in Chapter 1, and therefore, we do not deal with the cases of exophora.

Li and Thompson [Li and Thompson] indicated a salient feature of Chinese grammar is the fact that noun phrases understood from context do not need to be specified, termed zero pronouns. The utterances containing zero pronouns are perfectly grammatical in the appropriate contexts. Because the noun phrases are mentioned previously in a discourse, there is no need to specify them subsequently. For example, the ϕ^{i}_{1} in Utterance (2.7b) refers to 張三 'Zhangsan' in Utterance (2.7a). In imperative sentences, they explained that the conditions governing the occurrence of the second person pronouns are precisely the same as those governing the occurrences of pronouns in a question-answering dialogue. For example, suppose the host *A* has just poured a cup of hot tea, and the guest *B*, not knowing it is hot, seems to want to touch it; *A* might say a perfectly normal utterance like Example (2.15), in which the ϕ^{i} refers to 妳 'you' and ϕ^{j} refers to the cup of hot tea before *B*'s eyes. They further analyze the zero anaphora by taking discourse into account and provide the notion of topic chain, as exemplified in Example (2.6).

(2.15) ϕ^{i} 別 碰 ϕ^{j} ！

ϕ^{i} bie peng ϕ^{j}.

don't touch

Don't touch (it).

Liao [Liao 1992] discussed the zero anaphora in Chinese by focusing on the discourse analysis including context, intentions and assumptions of communication, foreground and background knowledge, social behavior, *etc*. He indicated that some other researches might mention the sentences without subjects or objects, but most of them are only on the basis of the sentence structures. He clarified 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. The elements termed *valent*s which are governed and bounded by the verbs can be subjects, direct objects, indirect objects or prepositional objects. The number of valents depends on the types of verbs. In Example (2.16) to (2.18), the three sentences respectively present the cases of one-valent, two-valent and three-valent verbs whose valents are underlined. However, the number of valents of the same verb could be different because of the ambiguity of the verb. This study does not provide clear rules for the purpose of resolution, but it is helpful to the establishment of ZA detection rules.

(2.16) 他 跑 了 三 千 公尺。

ta pao le san qian gong chi.

he run CRS three thousand meter

He has run for three thousands meters.

(2.17) 他 已經 離開 台北 了。

ta yijing likai Taipei le.

he already leave Taipei CRS

He had already left Taipei.

(2.18) 他 送 她 一 份 禮物。

ta song ta yi fen liwu .

he give she a present

He gave her a present.

Liao clarified that the ellipsis of a valent is made by the contexts including background knowledge, the given situation and textual (not contextual) information, and the readers can recover the ellipsis form the contexts. He also indicated that ellipsis of a valent is a means often used for surface coherence of a discourse. By his observation, the topic chain is a main condition for the ellipsis, which most frequently occurs in narrative genre. [Liao 1992].

Chen [Chen 1987] proposed a notion of *continuity* of referents in discourse to give a specific account for zero anaphora. Continuity of referent has *micro continuity* and *macro continuity*. The micro continuity has to do with the position of the antecedent and anaphor in their respective sentences. The notion of micro continuity is similar to the centering model of local discourse coherence [Grosz 1995][9]. The macro continuity considers the linear and hierarchical relationship between sentences containing antecedent and anaphor in the discourse structure. Chen also mentioned the phenomenon of cataphora, in contrast to anaphora, where the referent is referred to an element specified in the following sentences. Considering Example (2.19), the ϕ^{i}_{1} and ϕ^{i}_{2} both refer to the noun phrase 羅伯特 'Robert' specified in the subsequent Utterance (2.19c).

(2.19) a. ϕ^{i}_{1} 回到 家 ，

ϕ^{i}_{1} hui dao jia.

(he) back-to home

When (he) came back home.

b. ϕ^{i}_{2} 放下 書包，

ϕ^{i}_{2} fangxia shubao.

(he) put-down schoolbag

(He) put down the schoolbag.

c. 羅伯特^{i} 便 著手 設計 心目 中 新 的 美國 國旗 。

[9] The details of the centering modeling the local coherence of discourse are discussed in Section 3.2.

Luobote^{i} bian zhuoshou sheji xinmu zhong xin de meiguo guoqi.

Robert thereupon start design mind in new NOM U.S.A. national flag

Robert thereupon started to design the new national flag of the U.S.A. in mind.

In brief, these linguistic studies investigated phenomena of Chinese anaphora and principles or constraints proposed for the interpretation and use of anaphors. The results of these studies are not sufficient for natural language processing purposes because they are not represented in computational forms, and furthermore, the works did not demonstrate the effectiveness of the results. However, we could observe our test texts by referring to these studies to establish our zero anaphora resolution rules.

2.4 Summary

In this chapter, we first illustrated that the importance of the element, topic, in Chinese and the characteristics of ZAs. A survey on relevant linguistic studies show how complicated the factors are for the use of anaphors in Chinese. The linguistic studies also reveal a fact that most of the factors affecting the use of anaphors are discourse-oriented. Although the interpretation of zero anaphora in these studies is from the viewpoint of linguistics and is not sufficient for natural language processing purposes, these notions might help us to establish the resolution rules. Since our zero anaphora resolution system is built based on the observation on test texts, we therefore make some simplifications on the zero anaphora resolution.

CHAPTER 3

COMPUTATIONAL ANAPHORA RESOLUTION IN TEXT

After introducing what linguistic studies have been done on Chinese anaphora, in this chapter, we focus on the computational treatment of anaphora resolution. We first describe the different phases of anaphora resolution processing. Then we introduce theories used in anaphora resolution and approaches employing these theories. Approaches based on statistical or artificial intelligence (AI) techniques such as machine learning are also discussed in this chapter.

3.1 The Process of Anaphora Resolution

Most of the anaphora resolution systems deal with resolution of anaphors whose antecedents are noun phrases because it is a much more complicated task to identify anaphors whose antecedents are verb phrases, clauses, sentences or even paragraphs/ discourse segments. Typically, all noun phrases preceding an anaphor are initially regarded as potential candidates for antecedents [Mitkov 1999]. In this section, we illustrate the knowledge required for anaphora resolution and then introduce the task of automatic anaphora resolution.

3.1.1 Knowledge Needed for Anaphora Resolution

The task of anaphora resolution requires considerable knowledge sources to support it – morphological, lexical, syntactic, semantic, discourse and even real-word knowledge [Mitkov 2002]. Morphological and lexical information is needed for identifying anaphoric pronoun and as input to further syntactic processing. Sometimes anaphors can be resolved simply based on lexical information such as gender and number agreement As shown in Example (3.1), the noun phrase *Greene* is selected as an antecedent of pronoun he because the other candidates *no letters*, *Catherine* and *Switzerland* are eliminated based on a gender or number mismatch.

(3.1) *Greene*^{i} had no letters from Catherine while in Switzerland and *he*$^{i}_{1}$ feared the silence.

Gender agreement is a useful criterion in English when the candidates for the anaphor are proper female or male names, nouns referring to humans, nouns representing professions which cannot be referred to by *it*, gendered animals, or word such as ship which can be referred to by *she* or *it*. Similarly, number agreement helps to eliminate candidates that do not carry the same number as the anaphor. The number and gender agreement can be used in English and even more discriminative in languages such as German or Russian, where nouns denoting inanimate objects are routinely marked for neuter, feminine, or masculine gender [Mitkov 2002]. However, the agreement cannot be employed in zero anaphora resolution because anaphors are zeroed and cannot be discriminate their gender and number.

In Example (3.1), *Greene*, *no letters*, *Catherine* and *Switzerland* should be parsed as noun phrases for resolution processing. The example shows not only the importance of morphological and lexical information but also demonstrate the significance of syntactic knowledge. Syntactic knowledge is essential for anaphora resolution. In addition to providing information about clauses and constitutes (e.g. NPs, VPs), syntactic analysis which parses the grammatical roles (e.g. subject, object) in a sentence play an important role in the different rules used in anaphora resolution processing. Consider the simplified ZA resolution rule: a ZA only refers to the subject NP in the preceding clause within the same sentence. The rule relied on syntactic information about the clause boundaries, constitutes and the grammatical role of each constitute. As shown in Example (3.2), 張三 ‘Zhangsan’ would be identified as the antecedent of the ZA in the second clause by employing the rule.

(3.2) 張三 喜歡 李四 ϕ 不 喜歡 王五。

Zhangsan xi huan Lisi ϕ bu xi huan Wangwu.

Zhangsan like Lisi (he) not like Wangwu

Zhangsan likes Lisi but (he) does not like Wangwu.

(3.3) a. *The petrified kitten*^{i} refused to come down from the tree.

b. *It*$^{i}_{1}$ gazed beseechingly at the onlookers below.

Morphological, lexical and syntactic knowledge is important in anaphora resolution; however, there are other anaphora cases cannot be resolved. In Example (3.3), gender or number agreement rules can filter out neither *the petrified kitten* nor *the tree* as an antecedent candidate of *It* in Utterance (3.3b), because both candidates are gender neutral. Semantic information is required for selecting *the petrified kitten* in as the proper antecedent: the agent of the verb *gaze* is animate and the noun *kitten* is animacy. In a computational system, such information would rely on a knowledge base such as a dictionary or ontology. The majority of anaphora resolution systems, however, have no means of performing complex semantic processing. Such systems work with surface constituents and are based on their resolution strategies on the output of partial or full syntactic parsing [Baldwin 1997, Ferrandez *et al*. 1998, Kennedy and Boguraev 1996, Lappin and Leass 1994, Mitkov 1998, Yeh and Chen 2003].

Although the morphological, lexical, syntactic and semantic criteria for antecedent identification are strong, they are still not always sufficient to distinguish a set of possible candidates [Mitkov 2002]. In the case of antecedent ambiguity, the most salient element is usually the suitable antecedent among the candidates. The most salient element in computational linguistics can be referred to as the *focus* [Sider 1979] or *center*[10] [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].

(3.4) a. Tuesday morning had been like any other.

b. *Lisa*^{i} had packed *her*$^{i}_{1}$ schoolbag, teased *her*$^{i}_{2}$ 12-year-old brother James and bossed *her*$^{i}_{3}$ seven-year-old sister Christine.

c. After breakfast at 8:25, *she*$^{i}_{4}$ walked down the stairs of the family's first floor flat and shouted: "I'm off to school now – bye Mum, bye Dad, I will see you later".

In Example (3.4), the pronoun *her* and *she* refer to the center *Lisa* established previously in the first clause within Sentence (3.4a). It is unlikely that the pronoun *she* in Sentence (3.4c) would refer to her sister, though *Christine* is the nearest antecedent candidate.

Anaphora resolution systems required the knowledge mentioned in the above paragraphs offers an illustration of the complexity of natural language understanding, but there is yet another difficulty to consider. In Example (3.5) and (3.6), the resolution of the pronominal anaphors would be done if further world knowledge were available.

[10] Centers are the key elements of the centering model which is further discussed in detail in the following sections.

(3.5) *The soldiers*^{i} shot at the women and *they*$^{i}_{1}$ missed.

(3.6) The soldiers shot at *the women*^{i} and *they*$^{i}_{1}$ fell.

Integrating general real-world knowledge into a practical anaphora resolution system is a very labor-intensive and time-consuming task. Consequently, most systems do not utilize the extra-linguistic knowledge except very narrow domains [Mitkov 2002].

3.1.2 Realization of Anaphora Resolution

The automatic anaphora resolution consists of two main stages: anaphora detection and antecedent identification. The task of anaphora detection is to recognize the anaphors expressed in zero, pronominal and nominal forms whose antecedents have to be identified. Antecedent identification including two steps, candidate locating and antecedent selection, is to identify the antecedents of the anaphors detected in the first stage [Seki *et al*. 2002, Yeh and Chen 2004].

In the stage of anaphora detection, pronominal anaphors or pronouns such as *she*, *he* and *it* are easy to be detected because they are obviously expressed on the surface of text. However, when a pronoun *it* does not refer to anything specific, termed *pleonastic* [Lappin and Leass 1994], this anaphor cannot be further resolved its antecedent and should be filtered out in anaphora detection. For example, “It’s raining.” The word it is obviously a pronoun, but it has no antecedent. Nominal anaphors are more problematic to be detected. Definite noun phrases are potentially anaphoric, which often refer to preceding noun phrases, as exemplified in Example (3.7). The definite noun phrases *The Queen* in Utterance (3.7b) refers to the noun phrase *Queen Elizabeth* in Utterance (3.7a). Another example in Chinese as shown in Example (3.8), 這間機械廠 ‘this machinery factory’ in Utterance (3.8b) refers to 烏拉重機械廠 ‘Urals machinery factory’ in Utterance (3.8a). Although definite noun phrases are possibly anaphoric and can be detected by definite articles, it is not every definite noun phrase surely anaphoric. In Example (3.9), the noun phrase *The Duchess of York* is not anaphoric and does not refer to *Queen Elizabeth*. Therefore, similar to recognition of pleonastic pronouns, an anaphora resolution system should have the ability to recognize the definite noun phrases that are not anaphoric [Mikov 2002]. Zero anaphors cannot be detected with the surface information of themselves because the anaphoric expression is omitted, as exemplified in Example (3.2). ZA detection relies on the analysis of the constituents and their relationships in utterances. Besides, the non-anaphoric cases like cataphors should be recognized and eliminated from ZA detection. The method of ZA resolution is further discussed in the next chapter.

(3.7) a. *Queen Elizabeth*^{i} attended the ceremony.

b. *The Queen*$^{i}_{1}$ delivered a speech

(3.8) a. 尼古拉^{i} 在 烏拉重機械廠^{j} 擔任 副廠長，

Nigula^{i} zai wulazhongjixiechang danren fuchangzhang.

Nicola in Urals-machinery-factory occupy vice-factory-director

Nicola occupied a vice-factory-director in Urals machinery factory.

b. 他$^{i}_{1}$ 在 這 間 機械廠$^{j}_{1}$ 認識 了 芬娜。

ta$^{i}_{1}$ zai zhe jia jixiecjang$^{j}_{1}$ renshi ASPECT Fenna.

He in this machinery-factory know Faina

He knew Faina in this machinery factory.

(3.9) a. Queen Elizabeth attended the ceremony.

b. The Duchess of York was there too.

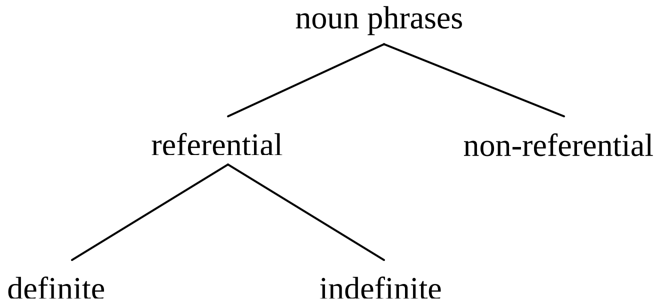


Figure 3.1: Referential and non-referential noun phrases

In Chinese, only noun phrases that are referential can be definite such as 那部汽車 ‘that car’ or indefinite such as 一個人 ‘one person’. The situation is shown in Figure 3.1. The difference is that a definite noun phrase refers to an entity that the reader has understood, while an indefinite noun phrase refers to an entity that the reader does not already know. Non-referential noun phrases never take classifier phrases [Li and Thompson 1981], and therefore, if a noun phrase has a classifier phrase, it must be a referential noun phrase. Because an anaphor refers to an entity expressed in the preceding utterances as mentioned in Chapter 2, we may ignore indefinite cases in nominal anaphora detection.

When the anaphors are detected, the anaphora resolution system has to identify their antecedent candidates. In this thesis, as clarified in Chapter 1, we concerned with processing anaphors whose antecedents are nominal phrases. Antecedent identification including two steps, candidate locating and antecedent selection. The first step is typically to identify all noun phrases in the preceding text as antecedent candidates of an anaphor within a certain search scope. Since anaphoric relations often operate within a discourse segment, the search scope is often limited to a discourse segment containing the anaphor

(Kenny and Boguraev 1996).[11] In the step of antecedent selection, the resolution rules are based on different knowledge sources, for example, gender and number agreement, c-command constraints, and semantic information, which are referred to as anaphora resolution factors [Mitkov 2002]. Anaphora resolution systems could employ these factors to filter out certain noun phrases from the proper antecedent 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, Wu 2003].

3.2 Theories Used in Anaphora Resolution

After illustrating the process of anaphora resolution, some theories and models including centering model, binding theory, rhetorical structure theory, and discourse representation theory (DRT) that have been used in anaphora resolution are introduced in this section.

3.2.1 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

[11] The related discourse theories or models employed in anaphora resolution are illustrated in detail in the following sections.

$C_f(U_n)$ that is realized[12] 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[13], 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 $U_1, \ldots, U_m$:

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

[12] 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.

[13] The notion of preferred center corresponds to Sider's notion of expected focus [Sidner 1983]

et al. 1994]:

Rules

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

I. 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 reflects 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 [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 $C_p(U_i)$:

Factors of Transitions

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 Factor (1) and (2) hold, a pair continuations across U_n and across U_{n+1}. If Factor (1) holds but Factor (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 Factor (1) does not hold, the utterances are in one of the shifting transition states depending on whether Factor (2) holds [Brennan *et al*. 1987, Walker *et al*. 1994]. The definition of transition states is summarized in Table 3.1.

Consider Example (3.10), the centering structures contain C_b, C_f and C_p where the set of C_f are partially ordered to reflect relative prominence in each utterance are presented in Table 3.2. The first two transition states of Utterance (3.10a) and (3.10b) are CONTINUE corresponding to the two factors. In Utterance (3.10c), the transition state is RETAIN because of "$C_b(U_{1c}) \neq C_p(U_{1c})$.". SMOOTH-SHIFT is the last transition state of Example (3.10) while "$C_b(U_{1d}) = C_p(U_{1d})$" and "$C_b(U_{1d}) \neq C_b(U_{1c})$" hold.

(3.10) 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.

Table 3.1: Transition states

	$C_b(U_i) = C_b(U_{i-1})$ or $C_b(U_{i-1})$ is undefined	$C_b(U_i) \neq C_b(U_{i-1})$
$C_b(U_i) = C_p(U_i)$	CONTINUE	SMOOTH-SHIFT
$C_b(U_i) \neq C_p(U_i)$	RETAIN	ROUGH-SHIFT

Table 3.2: Centering structures and transition states of Example (3.10)

(1a)	C_b: undefined	CONTINUE
	C_f: [電子股，美國高科技股]	
	C_p: 電子股	
(1b)	C_b: 電子股	CONTINUE
	C_f: [ZA (電子股)]	
	C_p ZA (電子股)	
(1c)	C_b: 電子股	RETAIN
	C_f: [證券股]	
	C_p: 證券股	
(1d)	C_b: 證券股	SMOOTH-SHIFT
	C_f: [ZA (證券股)]	
	C_p: ZA (證券股)	

3.2.2 Binding Theory

The binding theory which defines the syntactic constraints on coreference that exist between the noun phrases in a sentence is part of the principle and parameters theory [Chomsky 1981]. It accounts for the interpretation of anaphors including reflexive, pronominal and nominal anaphors. The binding theory regards reflexive anaphors in English as short-distance anaphors which refer to antecedents in a *local domain*. Since reflexive anaphors are bound by their antecedents in the local domain, they are often called *bound anaphors*. In contrast, pronominal anaphors are long-distance anaphors which permit antecedents to come only outside their local domain. Nominal anaphors here are lexical or referential expressions including all nominals headed by a common or proper noun such as *the man*. Their antecedents are independently and must be free in every domain. The most difficult area of the binding theory is the formulation of the notion local domain. Arriving at a useful definition of the local domain in structural terms has been an active area of research [Correa 1988, Mitkov 2002], and we do not intend to further discuss this problem in the thesis.

The following Example (3.11), (3.12) and (3.13) show the three classes, reflexive, pronominal and nominal anaphors, respectively.

(3.11) *Betty* looked at *herself* in the mirror.

(3.12) Betty looked at her in the mirror.

(3.13) Betty said *the puppy* was *the most wonderful gift*.

In the Chomsky's axiomatic statement of the binding theory [Chomsky 1981], two noun phrases are said to be coreferential if they bear the same referential index. The indexing rule massively over-generates logical forms of utterances from their surface representation, and indiscriminately assigns unwarranted coreference relations. The annotated structures produced by the rule are subject to a number of well-formed conditions as shown below, which are constraints on the assigned coreference relations.

The most elementary condition is the *agreement condition*, and the main component of the theory is given by the *binding axioms*. The notions of *binding* and the *c-command* used in it are also illustrated below.

Agreement condition

If NP_1 and NP_2 are co-indexed, then they must agree in person, gender, and number agreements.

Binding axioms

(i) Reflexive and reciprocal anaphors must be bound within its local domain.

(ii) A pronominal anaphor must be free within its local domain.

(iii) A lexical referential expression must be free in every domain.

Notion of binding

x binds *y* if and only if:

(i) *x* c-commands *y*

(ii) *x* is co-indexed with *y*

Notion of c-command

A node *A* c-commands a node *B* if and only if [Haegeman 1994]:

(i) *A* does not dominate *B*

(ii) *B* does not dominate *A*

(iii) The first branching node dominating *A* also dominates *B*.

Figure 3.2 shows the notion of c-command, it can be seen as follows [Mitkov 2002]:

- *B* c-commands *C* and every node that *C* dominates.
- *C* c-commands *B* and every node that *B* dominates.
- *D* c-commands *E* and *J*, but neither *C*, nor any of the nodes that *C* dominates.
- *H* c-commands *I* and no other node.

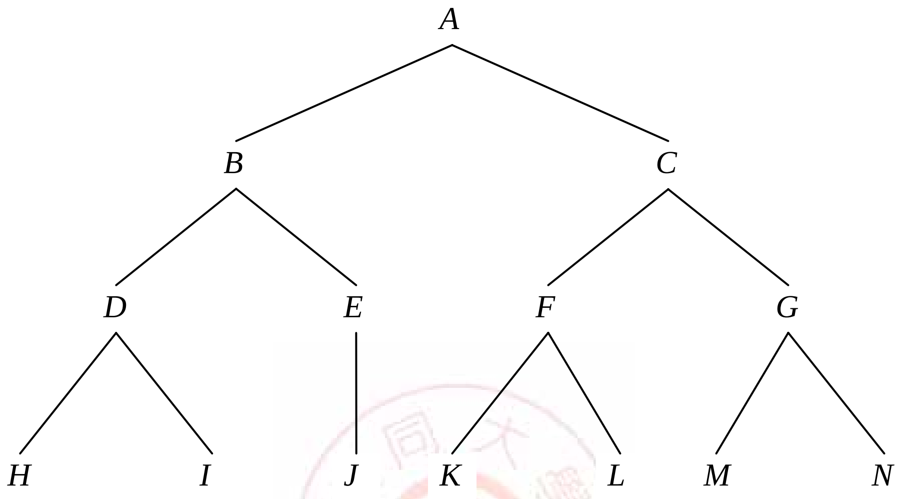


Figure 3.2: The notion of c-command

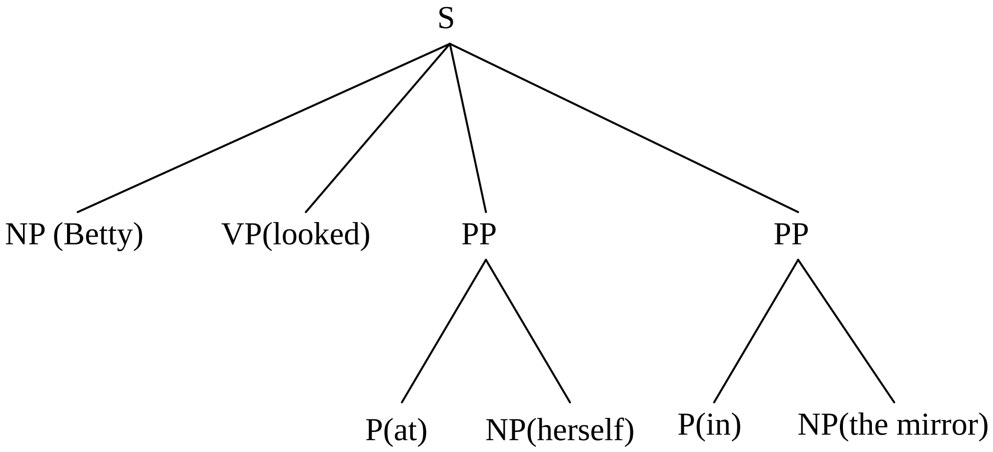


Figure 3.3: A tree representation of the example (3.11)

In Figure 3.3, which presents an examination of c-command relations in Example (3.11), the NP *herself* is c-commanded by the NP *Betty* in the tree. According to the binding axioms of the binding theory, the reflexive anaphor *herself* refers to *Betty* in this sentence.

3.2.3 Rhetorical Structure Theory

A discourse consists of segments, which are defined by communicative purpose. Mann and Thompson developed rhetorical structure theory (RST) to reflect the discourse structure [Mann and Thompson 1988]. It defines a number of relations to identify particular functional relationships between two non-overlapping spans of text: the nucleus (N) and the satellite (S). A relation definition contains constraint fields on N, S and on the combination of N and S and a field specifies the effect the writer intends to achieve in the relation. A relation can exist alone or together with other relations to form a schema. By applying schemas iteratively, a text can be analyzed as a functionally dependent structure. Marcu further employed the distinction between the nuclei and the satellites that pertain to discourse relations of RST to build rhetorical structure trees of discourse segments [Marcu 1996]. For example, consider the following Example (3.14):

(3.14) a. [Although discourse markers are ambiguous,^{1}]

b. [one can use them to build discourse trees for unrestricted texts:^{2}]

c. [this will lead to many new applications in natural language processing.^{3}]

Assume now that we can infer that *Although* marks a CONCESSIVE relation between satellite 1 and nucleus either 2 or 3, and the colon, an ELABORATION relation between satellite 3 and nucleus either 1 or 2. By ruling out the CONCESSION relation between 1 and 3, and the ELABORATION relation between 3 and 1, we will obtain the discourse tree of the sample text, as shown in Figure 3.4.

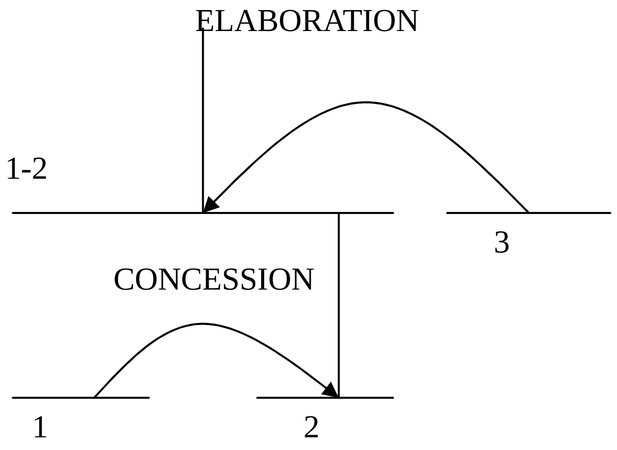


Figure 3.4: The rhetorical structure tree of the sample text (3.14)

3.2.4 Discourse Representation Theory

Discourse Representation Theory (DRT) [Kamp 1981] was originally designed as a principled method to cope with two related problems: The fact that intersentential and intrasentential pronouns seem to call for two entirely different types of explanation, and a well-known problem in connection with the interpretation of full noun phrases in so-called donkey-sentences as exemplified in the examples below [Hess 1991]. It introduces a new level of representation, discourse representation structures (DRSs), which derived systematically from the syntactic structure of the sentences of a discourse. On this level, all the textual references are resolved, whether they be intersentential or intra-sentential, even if they span the entire discourse. Consider Example (3.15), the Utterance (3.15a) would correspond to the DRS-diagram on Figure 3.5, which is semantically equivalent to the first-order logic formula (3.15b).

Discourse referents are accessible beyond the clause or sentence where there are expressed because the semantic interpretation procedure in DRT begins by adding the syntactic interpretation of a new sentence to the existing DRSs. Example (3.16) presents a case of intersentential anaphors, and its DRS-diagram is shown on Figure 3.6. Sentences with conditionals are more complex, and require a new construct in DRSs. Two DRSs, one each for antecedent and consequent, are linked with an appropriate connector, written as a “⇒”, to form a complex DRS as shown on Figure 3.7, which expresses the DRS-diagram of Utterance (3.17a).

(3.15) a. John owns a donkey.

b. $\exists\, x, y$: John (x) $\wedge$ donkey (y) $\wedge$ owns (x, y)

(3.16) a. John owns a donkey. He beats it.

(3.17) a. If John owns a donkey, he beats it.

b. $\exists\, x, y$: (John (x) $\wedge$ donkey (y) $\wedge$ owns (x, y)) $\Rightarrow$ beats (x, y)

x y

John (*x*)
donkey (*y*)
owns (*x*, *y*)

Figure 3.5: DRS-diagram of the example (3.15a)

x y u v

John (*x*)
donkey (*y*)
owns (*x*, *y*)
$u = x$
$v = y$
beats (*u*, *v*)

Figure 3.6: DRS-diagram of the example (3.16)

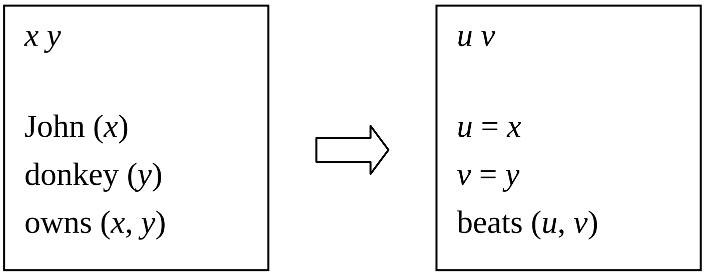


Figure 3.7: DRS-diagram of the example (3.17a)

3.3 Approaches of Anaphora Resolution

Approaches of anaphora resolution can be distinguished by the computational strategy they used. In this section, we introduce these approaches by grouping them into the approaches based on integration of linguistic and knowledge sources, the approaches based on statistical models, and know-poor approaches that are the latest trends.

3.3.1 Approaches Based on Linguistic and Knowledge Sources

This section introduces that the methods of integration of different knowledge sources (e.g. syntactic, semantic, and discourse knowledge) or factors (e.g. gender and number agreement, and c-command constraints) are to identify anaphors and their antecedent candidates, and then discount unlikely candidates until a minimal set of plausible candidates is obtained.

Carter reports on a shallow processing approach, Shallow Processing Anaphor Resolver (SPAR), relying heavily on linguistic knowledge such as syntax, semantics and local focusing [Sidner 1979] and limiting the extent and use of world knowledge can achieve accurate results in ambiguity resolution [Carter 1987, Carter 1990]. SPAR was implemented to resolve the problem of anaphor resolution in the task of paraphrasing simple English stories. It processed stories sentence by sentence, resolving the ambiguities in each sentence integrating the information in them into context, and outputting a paraphrase of the sentence before moving on to the next one. The paraphrase is intended to show not only what referents have been assigned to the anaphors, but also, by using near!synonyms and varying the syntactic structures, what lexical and structural choices have been made.

Carbonell and Brown propose a multi-strategy approach to anaphor resolution [Carbonell and Brown 1988]. They hypothesize that anaphor resolution may be accomplished through the combination of a set of strategies rather than by a single monolithic method. They concentrate on resolving intersentential anaphora by considering this type of anaphora to be more frequent and more crucial in designing interactive natural language interfaces. A general framework for anaphor resolution is proposed based on the integration of multiple knowledge sources: sentential syntax, case-frame semantics, dialogue structure and general world knowledge. The approach is based on a set of constraints and preferences. The constraints are local anaphor constraints and precondition/postcondition constraints. The former corresponds to the agreement constraints such as gender and number, while the latter uses real-world knowledge and pragmatics. In Example (3.15) which is taken form their paper [Carbonell and Brown 1988], *he* refers to *Tom*, as John no longer has the apple. The postcondition on *give* is that the actor no longer have the object being given, which conflicts with the precondition on *eat* that the actor have the item being eaten, if the actor is assumed to be John. The preferences in this paper are case-role persistence preference, semantic alignment preference, syntactic parallelism preference, syntactic topicalization preference and intersentential recency preference which are applied to each of the candidates deduced by the constraints.

(3.14) a. John gave *Tom* an apple.

b. *He* ate the apple.

Lappin and Leass in their paper [Lappin and Leass 1994] present an algorithm for identifying the noun phrase antecedents of third person pronouns and lexical anaphors (reflexives and reciprocals). The algorithm refered to as RAP (Resolution of Anaphora Procedure) identifies both intrasentential and intersentential antecedents of pronouns in text. RAP applies to the syntactic structures of McCord's Slot Grammar parser [McCord 1990, McCord 1993], and relies on measures of salience derived from the syntactic structure and a simple dynamic model of attentional state to select the antecedent NP of a pronoun from a list of candidates. It does not utilize semantic conditions or real-world knowledge to identify the antecedents among the candidates.

Okumura and Tamura [Okumura and Tamura 1996], present a method to handle complex sentences with the centering theory [Grosz *et al*. 1995] and descried the framework that identifies the antecedents of zero pronoun in Japanese discourses. They adopt the *partition* approach which divides a complex sentence into several simple sentences manually, and then the intrasentential ellipsis is regards as the intersentential case. Because a complex sentence is transform into several simple sentences, they also extend the search scope of the antecedents for resolving the intersentential ellipsis occurring in the complex (pre-partitioned) sentences.

Tetreault and Allen present an automated corpus-based analysis using RST [Mann and Thomson 1988] to aid in pronoun resolution [Tetreault and Allen 2003]. The work builds on preliminary research discussed in [Tetreault 2002] in which the RST-tagged Treebank [Carlson *et al*. 2001 corpus of Wall Street Journal articles merged with coreference information is constructed to provide a testing ground. With this test-bed system, they evaluate two algorithms based on leading theories of decomposing discourse: centering [Grosz and Sidner 1986] and Veins theory [Cristea *et al*. 1998][14]. They

[14] Veins theory [Cristea et al., 1998] is an extension of the centering Theory from local to global

concluded that the clausal segmentation has the promise of improving pronoun resolution based on these two theories.

Guenthner and Lehmann, in their paper [Guenthner and Lehmann 1983], propose rules for pronoun resolution which operate in the restricted context of relational database query dialogues or might be extendable to other types of dialogues. Their system constructs a DRS for a dialogue and applies morphological (gender and number agreement), syntactic, semantic and pragmatic factors, in that order, to accessible antecedent candidates until only one candidate remains.

3.3.2 Approaches Based on Statistical Models or Machine Learning

Traditional approaches to anaphora resolution require a great deal of knowledge in natural language understanding such as morphology, syntax, semantics, discourse and general world knowledge but to integrate all this knowledge is vary difficult for the development of anaphora resolution systems. An alternative approach is to utilize statistic models or some AI techniques such as machine learning.

Ge *et al*. propose a statistical framework for resolution of pronominal anaphors [Ge *et al*. 1998]. They incorporate multiple four anaphora resolution factors including the distance between the pronoun and the proposed antecedent, gender/number/animaticity of the proposed antecedent, governing head information and noun phrase repetition into the framework. These factors are combined into a single probability to identify the antecedent. This program does not use hand-crafting rules but rely instead on the corpus of Penn Wall Street Journal Treebank that has been marked with co-reference information.

discourse. It identifies domains of referential accessibility for each discourse unit over discourse structure trees as defined in RST. The theory assumes that only a subset of the clauses preceding the anaphor are actually relevant to successfully interpreting the anaphor. This subset (domain of referential accessibility) is determined by the interaction of the tree hierarchy and whether a clause is a nucleus or a satellite.

Seki *et al*. propose a probabilistic method of zero pronouns resolution in Japanese [Seki *et al*. 2002]. They focus on intersentential zero pronouns by consider this type to be major referential expressions in Japanese. The method integrates two probability parameters to perform zero pronoun detection and resolution in a single framework. The first parameter quantifies the degree to which a given case is a zero pronoun, while the second parameter quantifies the degree to which a given entity is the antecedent for a detected zero pronoun. Their system is tested by using the Kyotodaigaku Text Corpus version 2.0 [Kurohashi and Nagao 1998], in which 20,000 articles in Mainichi Shimbun newspaper articles in 1995 were analyzed by JUMAN and KNP [Kurohashi and Nagao 1998b, Kurohashi 1998] (the morph/syntax analyzers used in the system) and revised manually.

Aone and Bennett [Aone and Bennett 1995] describe an approach to build an trainable anaphora resolution system. In their approach, a corpus of Japanese newspaper articles manually tagged with discourse information is taken as training examples for the system. The machine learning algorithm called MLR (the Machine Learning-based Resolver), which employs the C4.5 decision tree algorithm by Quinlan [Quinlan, 1993]. They evaluate and compare the results of the MLR with those produced by the MDR (the Manually-Designed Resolver) which is reported previously as a robust, extensible, and manually trainable system in [Aone and Mckee 1993].

Müller, Rapp and Strube [Müller *et al*. 2002] investigate the practical applicability of a weakly supervised machine learning algorithms, Co-Training [Blum and Mitchell 1998], for the task of building a classifier for reference resolution. They apply Co-Training to the problem of reference resolution in German text from the tourism domain. In their system two levels of features which are the features assigned to noun phrases (e.g. grammatical function, semantic class, gender and number agreement, *etc*.) and features assigned to the potential coreference relation (e.g. distance between anaphor and antecedent) are considered.Because the deficiency of supervised machine learning approaches is the need for an unknown amount of annotated training data for optimal performance, they are concerned with the question if Co-Training can significantly reduce the amount of manual labeling work and still produce a classifier with an acceptable performance. However, they report that the results of the experiment are mostly negative.

Preiss [Preiss 2002] compares the performance of the Kennedy and Boguraev anaphora resolution algorithm [Kennedy and Boguraev 1996][15] to the performance of a memory-based machine learning algorithm TiMBL (version 3.0) [Daelmans *et al*. 2000]. For each pronoun, the machine learning algorithm is given the Kennedy and Boguraev feature set for a list of candidate antecedents, but it does not have access to the salience weights for the features. The experiment is performed on a manually annotated test corpus taken from the first corpus of the written section of the BNC (British National Corpus).In this paper, they report on the results showing that there is no significant difference between their performance.

3.3.3 Knowledge-poor Approaches

In Section 3.4.1 and 3.4.2, we have introduced the approaches relying on heavily on linguistic information and domain (or general) knowledge, and the approaches based on sufficiently tagged corpora. In this section, we discuss the latest trend, knowledge-poor approaches, which are inexpensive, fast and reliable procedures of anaphora resolution.

Kennedy and Boguraev [Kennedy and Boguraev 1996] report on an algorithm for anaphora resolution which is a modified and extended version of that developed by [Lappin and Leass 1994]. The motivation for developing the modified version is to make

[15] Kennedy and Boguraev's approach is a knowledge-poor approach which is further discussed in the following section.

it available to a wide range of text processing frameworks. Therefore, their algorithm does not require in-depth, full, syntactic parsing of text, but works instead from the output a part of speech tagger, enriched only with annotations of grammatical function of lexical items in the input text stream. The system uses a phrase-level grammer to identify NP constituents and, following Lappin and Leass [Lappin and Leass 1994], refers to the salience factors for ranking antecedent candidates. The work does employ full syntactic parsing while retains a degree of quality and accuracy in pronominal anaphora resolution comparable to that reported in [Lappin and Leass 1994].

Baldwin, in his paper [Baldwin 1997], presents a pronoun resolution system, which is designed around the assumption that there is a sub-set of anaphors that do not require general world knowledge or sophisticated linguistic processing for successful resolution. What distinguishes CogNIAC from other algorithms is that it does not resolve a pronoun in circumstances of ambiguity, that is, only pronouns with sufficiently high confidence can be resolved. This results in the system that produces high precision, but unsatisfying recall.

Mitkov [Mitkov 1998] propose a robust, knowledge-poor approach to resolving pronouns in technical manuals, which works as follows: it takes as an input the output of a text processed by a POS tagger, identifies the noun phrases which precede the anaphor within a distance of 2 sentences, checks them for gender and number agreement with the anaphor and then applies the so-called antecedent indicators (preferences) to the remaining candidates by assigning a positive or negative score. The antecedent indicators have been identified empirically and are related to salience (definiteness, givenness, indicating verbs, lexical reiteration, section heading preference, non-prepositional noun phrases), to structural matches (collocation, immediate reference), to referential distance or to preference of terms. Their evaluation shows that the results are better than the approaches such as [Baldwin 1997] selected for comparison and tested on the same data.

Ferrández *et al*. [Ferrández *et al*. 1998] report on an approach similar to Kennedy and Boguraev's approach [Kennedy and Boguraev 1996]. They works on the output of a POS tagger and apply a partial parsing from the formalism: Slot Unification Grammar [Ferrandez *et al*. 1997], which has been implemented in Prolog. In addition to pronoun resolution, they also resolve other types of anaphora such as one-anaphora (e.g. a blue one) and surface-count anaphora (e.g. your daughter and she). By comparing to the pronoun resolution in [Kennedy and Boguraev 1996], the accuracy is improved from 75% to 83%. For one-anaphora and surface-count anaphora, because their system runs on a small test corpus (9600 words), there are not sufficient anaphors to be taken for evaluation (only 5 anaphors with 80% accuracy).

3.4 Summary

In this chapter, the process of anaphora resolution including considerable knowledge sources and different phases of the anaphora resolution procedure is illustrated. We also introduce the state-of-the-art theories and formalisms used in anaphora resolution in Section 3.2. Approaches of anaphora resolution are introduced by grouping them into by different the computational strategies they employed.

Most traditional approaches are based on complex linguistic information and domain knowledge. However, it is difficult to integrate such a great deal of information including morphology, syntax, semantics, discourse, and world knowledge for developing a robust anaphora resolution system. An alternative approach employing statistical models or AI techniques can sort out the above problems, but it heavily relies on the availability of sufficiently text corpora that are tagged, in particular, with referential information. A recent approach is the search for inexpensive, fast and reliable procedures of anaphora resolution as described in Section 3.3.3. This approach does require full syntactic parsing or comprehensive semantic information, but instead employs a POS tagger and shallow parsing to detect and resolve anaphora occurring in text.

CHAPTER 4

ZERO ANAPHORA RESOLUTION BASED ON THE CENTERING MODEL

4.1 Introduction

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].

4.2 Zero Anaphora in Chinese Utterances

Zero anaphors, as mentioned previously, 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[16], 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).[17] This classification can be employed to detect whether the ZA is embedded in the object position of an utterance, such as Example (4.1), in which the verb 掉進去 'fall-in-to' in Utterance (4.1d) is a transitive verb whose object is omitted.

(4.1) 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 <u>大 水溝</u>^{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 jiudiaojinqu ϕ^j_1 le.

[16] 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.

[17] These three types of verbs are corresponding to one-valent, two-valent and three-valent verbs as mentioned previously in Chapter 2 [Liao 1992].

(he) fall-in-to (it) CRS

(He) fell in to it.

(4.2) 北京 鴨 (被) 烤熟 了。

Beijing ya (bei) kao shou le.

Beijing duck bake well CRS

The Beijing duck was baked well.

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 Example (4.1), the subjects of the verbs in Utterance (4.1b), (4.1c) and (4.1d) 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 Chapter 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. Example (4.2), 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.

In addition to the verbs, we also consider the *coverb*s, prepositions, and coordinating conjunctions. The coverb introduces a noun phrase and the phrase formed by the coverb plus the noun phrase generally precedes the main verb and follows the subject or topic [Li and Thompson 1981]. The syntactic structure is: subject/topic + converb + noun phrase + verb + (noun phrase), as exemplified in Example (4.3) and (4.4). The ZA occurs in subject or topic position in this syntactic structure is also shown in Example (4.5). In the cases of prepositions, and coordinating conjunctions, subject or topic followed by a preposition phrase or a coordinating conjunction might be omitted, as shown in Example (4.6) and Utterance (4.7b) respectively [Huang 1994].

(4.3) 他 跟 我 說話。

ta gen wo shuohua.

he with I talk

He talked with me.

(4.4) 你 替 我 買 票 吧 。

ni ti wo mai piao ba.

you instead-of I buy ticket SA

You buy the ticket instead of me, OK?

(4.5) a. 喬治^{i} 拿 著 玫瑰花 跟 戒指，

Qiaozhi na zhe meiguihua gen jiezhi

George hold DUR roses and ring

George took the roses and the ring.

b. ϕ^{i}_{1} 向 瑪莉 求婚 ，

ϕ^{i}_{1} xiang Mali qiuhun

(he) face Mary propose-to

(he) proposed to Mary.

(4.6) φ 跟 他 愛人 在 一起 。

φ gen ta airen zai yiqi.

with he spouse exist together

(He) is together with his wife.

(4.7) a. 張三^{i} 已經 七十 歲 了，

Zhangsan yijing qishi sui le.

Zhangsan already seventy year CRS

Zhangsan is already seventy years old.

b. ϕ^{i}_{1} 爲 了 生活費 ，

ϕ^{i}_{1} wei le sheng huo fei

(he) for ASPECT living-expenses

For the sake of living expenses

c. 他$^{i}_{2}$ 還是 每 天 辛苦 工作 。

ta$^{i}_{1}$ haishi meitian xinku de gongzuo

he still everyday hard work

He still works hard everyday.

4.3 Rules of Zero Anaphora Detection

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 and 4, we further consider the case of the coverb, prepositions, and coordinating conjunctions for detecting the ZAs occurring in the topic or subject position.

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.
3. For each utterance U_i in a discourse segment U_1, … , U_m: If the syntactic structure , coverb + noun phrase + verb, precedes U_i and no noun phrase appears before the coverb, then an omission of topic or subject is detected as a ZA candidate.
4. For each utterance U_i in a discourse segment U_1, … , U_m: 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.

4.4 Rules of Antecedent Identification

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[18] from the ZA candidates which are detected in the phase of ZA detection. In the case of

[18] 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.

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 cases:

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*[19] 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 consisting of utterances U_1, … , 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

[19] The centers include forward-looking centers, the backward-looking center [[Grosz *et al*. 1995] and the preferred center [Brennan *et al*. 1987]. (See Section 3.3 for more detailed discussion)

```
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[20], 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].

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* >

[20] The centering model includes two rules as described in Section 3.3.

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.

4.5 Summary

The rules for zero anaphora resolution including ZA detection and the antecedent identification are presented in this Chapter. Since the ZA is an expression without being specified in surface text, the information of itself is null. In ZA detection, we utilized the POS information of the constituents of the utterance and their grammatical relation to establish the ZA detection rules. We also further designed the ZA identification constraints for filtering out the non-anaphoric cases which might be incorrectly detected as ZAs. In antecedent identification, the centering model is employed as a basis for establishing the antecedent identification rule.

CHAPTER 5

IMPLEMENTATION AND EVALUATION OF THE ZERO ANAPHORA RESOLUTION METHOD

5.1 Introduction

In previous chapters, we surveyed the related linguistic studies and computational strategies for anaphora resolution and established some rules. To show how these rules work in a real system, in this chapter, we illustrate the implementation of our Chinese zero anaphora resolution system.

The resolution system works on the output of a POS tagger and employs ZA detection rules for detecting 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. As shown in Figure 5.1, it includes two components of zero anaphora resolution, ZA detection and antecedent identification, and works as follows: First, we take the AUTOTAG to segment Chinese lexical items and annotate their POS information in the input text stream. Second the sequence of POS-tagged words is parsed into smaller constituents with phrase-level parsing. Each constituent is represented as a word list. Then the sequence of word lists is transformed into *triple*s, [*S*, *P*, *O*][21], and each ZA candidate is detected by employing the ZA detection rules and marked as *zero*. Third, the system takes the output of the shallow parser and uses the ZA identification constraints to eliminate non-ZA cases. Finally, the antecedent identification rules are employed to determine the antecedents.

[21] The definition of *triple* and its transformation is further illustrated in the next section.

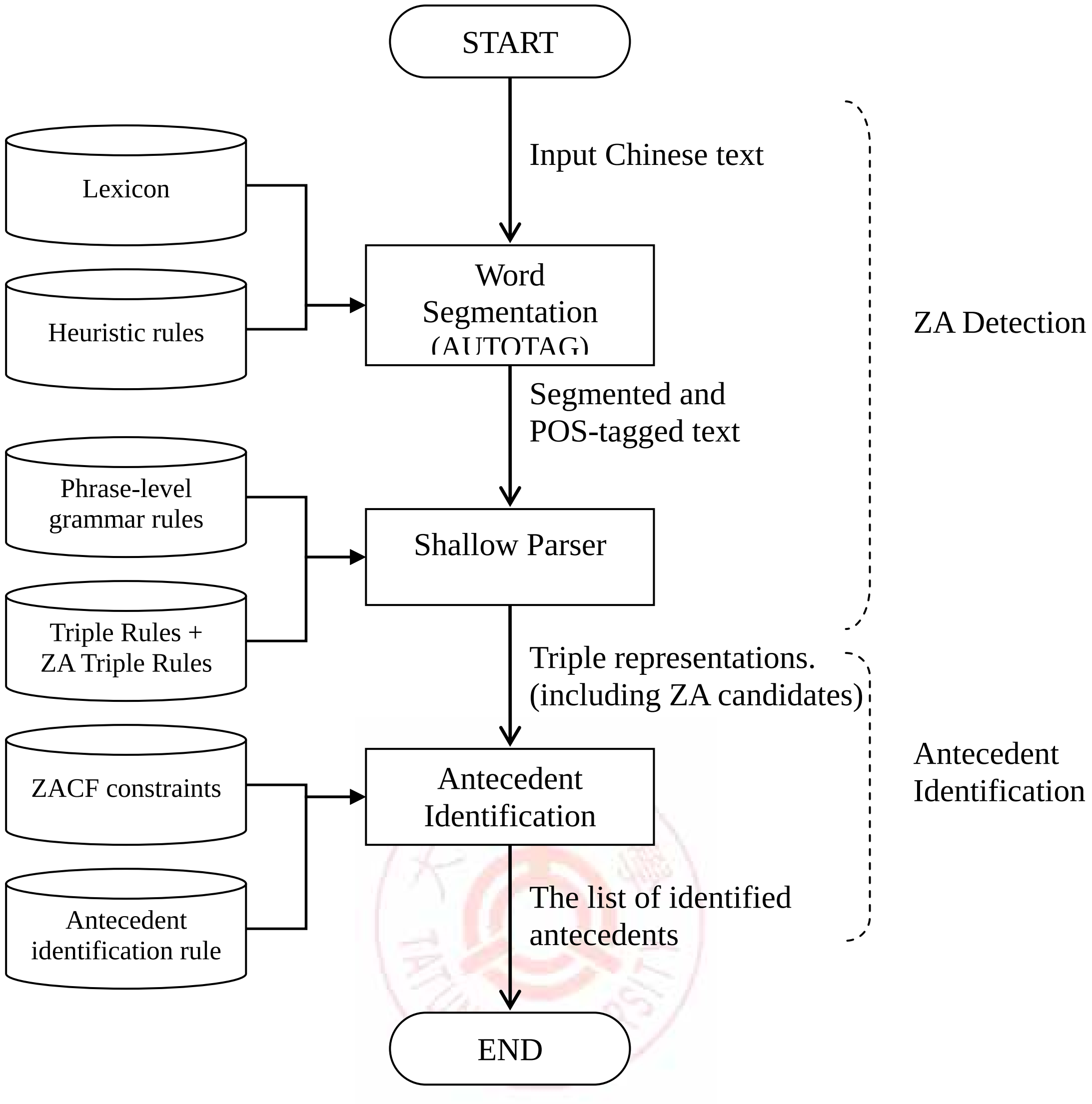

START
Input Chinese text
Lexicon
Heuristic rules
Word Segmentation (AUTOTAG)
ZA Detection
Segmented and POS-tagged text
Phrase-level grammar rules
Triple Rules + ZA Triple Rules
Shallow Parser
Triple representations. (including ZA candidates)
ZACF constraints
Antecedent identification rule
Antecedent Identification
Antecedent Identification
The list of identified antecedents
END

Figure 5.1: Diagram of our Chinese zero anaphora resolution system

5.2 A Shallow Parser with Zero Anaphora Detection

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 present some examples of full parsing and then describe our method of shallow parsing in Chinese.

5.2.1 Full Parsing

Many traditional approaches to parsing natural language sentences aim to recover complete and 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, Utterance (5.1a) taken from Sinica Treebank [Sinica Treebank 2002] is annotated as the following expression (5.1b).

(5.1) a. 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.

b. 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 Example (5.1), 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, to choose the head of a phrase automatically may cause errors [Chen *et al.* 1999]. Another example, Example (5.2), is extracted from the Penn Chinese TreeBank [The Penn Chinese

Treebank Project 2000].

(5.2) a. 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.

```
b. (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

5.2.2 Shallow Parser

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, Utterance (5.3a) and can be divided as the expression (5.3b).

(5.3) 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. Second a shallow parser parses the sequence of words is into smaller constituents such as noun phrases and verb phrases with phrase-level parsing[22] and then transforms them into triples, [*S*,*P*,*O*]. As shown in Example (5.4), (5.4b) is the output of Utterance (5.4a) produced by AUTOTAG and (5.4c) is the triple representation.

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

b. [[花蓮], np], [[成爲], vtp], [[熱門,的,旅遊,地點], np]

c. [[花蓮], [vtp(成爲)], [熱門,的,旅遊,地點]]

The definition of triple representation is illustrated below. The triple here is a simple

[22] There are about 47 POS tags used in AUTOTAG, in which 13 POS tags belong to the noun tag set and 17 POS tags belong to the verb tag set. We create simple rules of noun phrases and verb phrases in DCG (Definite Clause Grammar) [Gazdar and Mellish 1989]. (See Appendix A)

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 of Triple

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 (5.1b) is transformed into triples by employing the Triple Rules. The Triple Rules are built by referring to the Chinese syntax. There are four kinds of triples in the Triple Rules, which corresponds to four basic clauses: subject + transitive verb + object, subject + intransitive verb, subject + coverb + object[23], 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,*none*) → np(S), coverb(P).

[23] Note that the clause ‘subject + coverb + object’ is taken from the former part of the syntactic structure containing a coverb: subject/topic + converb + noun phrase + verb + (noun phrase) (See Section 4.2). We leave the later part ‘noun phrase + verb + (noun phrase)’ from this structure for the rules Triple1 and Trple2 and the rules Triple3 is further employed to establish its ZA Triple rule for detecting the subject/topic omission.

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 coverb(P) denotes that the predicate is a coverb. The Triple4 is employed only if an utterance contains only one noun phrase and no other constituent. If all the Triple Rules failed, the ZA Triple Rules are employed to detect ZA candidates.

ZA Triple Rules
Triple1^{z1}(*zero*,P,O)→ vtp(P), np(O).
Triple1^{z2}(S,P,*zero*)→ np(S), vtp(P).
Triple1^{z3}(*zero*,P,*zero*)→ vtp(P).
Triple2^{z1}(*zero*,P,*none*)→ vip(P).
Triple3^{z1}(*zero*,P,*none*) → coverb(P).
Triple4^{z1}(*zero*,P,O) → co-conj(P), np(O).
Triple4^{z2}(*zero*,P,O) → prep(P), np(O).

The zero anaphora in Chinese generally occurs in the topic, subject or object position. The rules Triple1^{z1}, Triple2^{z1} and Triple3^{z1} detect the ZAs occurring in the topic or subject position. The rule Triple1^{z2} detects the ZAs in the object position and the rule Triple1^{z3} detect the ZAs occurring in both subject and object positions. In the rules Triple4^{z1} and Triple4^{z2}, the co-conj(P) and prep(P) denote a coordinating conjunction or a preposition appearing in the initial position of a clause. As shown in Example (5.5), there are two *triples* generated. In the second *triple*, *zero* denotes a ZA according to Triple1^{z1}. Some examples corresponding to the ZA Triple Rules are also presented in Table 5.1.

(5.5) a. 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.

b. [[[張三], [參加], [比賽]], [[*zero*], [贏得], [冠軍]]]

[[[Zhangsan], [enter], [competition]], [[*zero*], [win], [champion]]]

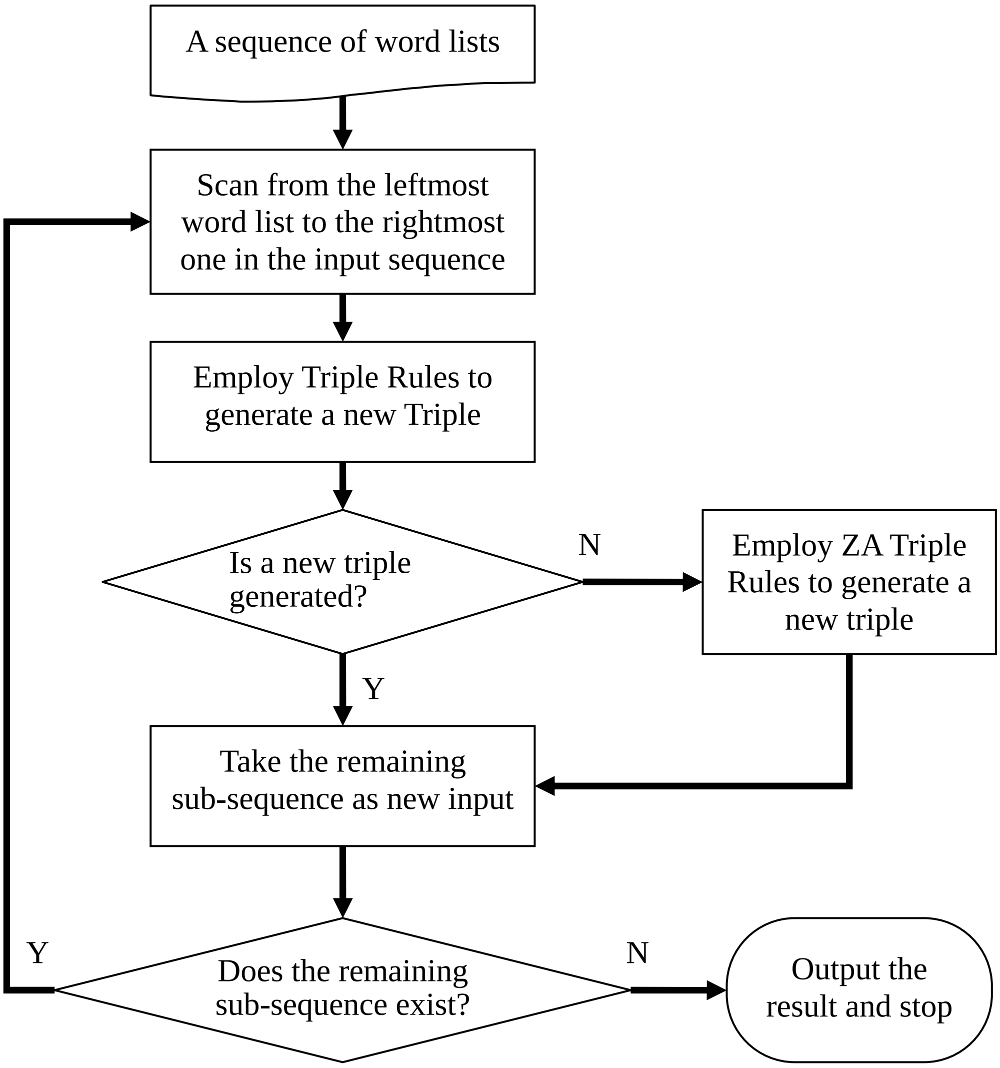

A sequence of word lists
Scan from the leftmost word list to the rightmost one in the input sequence
Employ Triple Rules to generate a new Triple
Is a new triple generated?
N
Employ ZA Triple Rules to generate a new triple
Y
Take the remaining sub-sequence as new input
Y
Does the remaining sub-sequence exist?
N
Output the result and stop

Figure 5.2: The procedure of Triple transformation

Figure 5.2 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.

Table 5.1: Examples of zero anaphora

ZA Triple Rule	Example
Triple1^{z1}(*zero*,P,O)	φ 撞到 一 個 人 φ 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 φ ma Zhangsan like (somebody or something) Q Does Zhangsan like (somebody or something)?
Triple1^{z3}(*zero*,P,*zero*)	φ 喜歡 φ φ xihuan φ (he) like (somebody or something) (He) likes (somebody or something).
Triple2^{z1} (*zero*,P,*none*)	φ 去 購物 了 φ qu gouwu le (he) go shopping ASPECT (He) has gone shopping.

Triple3^{z1}(*zero*,P,*none*)	ф 向 張三 借 書 ф xiang Zhangsan jie shu (he) face Zhangsan borrow book (He) borrowed a book from Zhangsan.
Triple4^{z1}(*zero*,P,O)	ф 和 小朋友 ф han xiaopengyou wan (he) with child play (He) is playing with little children.
Triple4^{z2}(*zero*,P,O)	ф 在 公園 ф zai gongyuan (he) in park (He) is in the park.

5.3 Antecedent Identification

After ZA candidates are detected by the shallow parser, the antecedent identification component works as follows: First, the ZA identification constraints are utilized to filter out non-anaphoric cases. Then the antecedent identification rule is used to identify the antecedents of ZAs.

The output of the shallow parser is a list of utterances, and each utterance is represented as triples that might contain *zero*s. The antecedent identification program adopts ZA candidate filtering constraints, termed ZACF constraints hereafter, which is established according to the ZA identification constraints for eliminate non-ZA cases from these *zero*s. The ZACF constraint 1 is used to filter out the cases of cataphora and exophora, while the ZACF constraint 2 is used to eliminate the cases of passive sentences or inverted sentences.

ZACF constraints

1. The *zero*s in the first utterance of whole input text stream are ignored.

2. In the following cases, the *zeros* are ignored:

   [[np], [coverb(*bei*)], [*none*]] + [[np], [vtp], [*zero*]]

   [[np], [*none*], [*none*]] + [[np], [vtp], [*zero*]]

In the step of identifying antecedents, the antecedent identification program parses each utterance of the input stream one after another until no remaining utterance needs to be processed. Each input utterance is taken to be verified with the ZACF constraints by the program for deciding that it might be ignored or be further processed. The utterance is then extracted its noun phrases as forward-looking centers, which are ordered as follows: the subject (S) of the leftmost triple, the object (O) of the rightmost triple, and then the elements deduced predicates (P) of the remaining triples from the left to right. The order is according to forward-looking center ranking criterion described previously in Chapter 4. The highest ranked center is taken as the default preferred center of the utterance and the default backward looking center of the subsequent utterance. Finally, each utterance has the ranked forward-looking center list and the default preferred center with it.

Because we focus on the resolution of intersentential ZAs, the intrasentential ZAs embedded in an utterance like Example (5.5) are ignored. Therefore, the antecedent identification program only processes: (i) the first triple and the last triple of an utterance if more than one triple exists in an utterance, or (ii) only one triple existing in an utterance. In the first case, the program checks whether the first tuple of the first triple is *zero* or not for resolving the ZA in the topic or subject position later, while the last tuple of the last triple is also processed similarly for resolving the ZA in the object position. In the case of only one triple existing in an utterance, the triple is checked its first and last tuples for resolving the ZA in the topic or subject, and object positions.

Before further performing process of determining the antecedents of *zero*s, we define the *Replace and Shift* (*RS*) operations, which are utilized to illustrating the procedure:

Replace and Shift operations

A. Replace the backward looking center of the subsequent utterance with the highest ranked center of the ranked forward looking center list of the current utterance.

B. Replace the preferred center with the second forward looking center (if exists) of the ranked forward looking center list.

C. Remove the first forward looking center of the ranked forward looking center list, and all the following forward looking centers are shifted up for one position.

D. Replace the backward looking center with a *null* marker.

E. Replace the preferred center with a *null* marker.

According to the antecedent identification rule, the procedure for determining the antecedents of *zero*s is described in the following three situations which work in order. (1) No *zero* exists in the preceding utterance: Take the default preferred center of the preceding utterance as the antecedent of the current *zero,* and then do RS operation A, B and C in order on the preceding utterance. (2) Only one *zero* exists in the preceding utterance: Take the backward looking center of the current utterance as the antecedent of the current *zero,* and then do RS operation D on the current utterance. If the backward looking center of the current utterance does not exist (has been *null*), take the preferred center of the preceding utterance as the antecedent of the current *zero*, and then do RS operation A, B and C in order on the preceding utterance. (3) More than one *zero* exists in the preceding utterance:[24] (a) If the preferred center of the preceding utterance is not null, do the procedure step by step. First, do RS operation D on the current utterance and RS

[24] This situation is more complicated, and we illustrate it with two sub- situations, (a) and (b).

operation E on the preceding utterance. Second, Rank the *zero*s in the preceding utterance into a ranked list with the forward-looking center ranking criterion, and then replace each *zero* with its antecedent. Third, concatenate this list as the former part with the forward-looking center list of the preceding utterance to be the new forward-looking center list of the preceding utterance. Fourth, do RS operation A and C in order on the preceding utterance. Fifth, take the backward looking center of the current utterance as the antecedent of the current *zero*, and then do RS operation A and C again in order on the preceding utterance. (b) If the preferred center of the preceding utterance is null, take the backward looking center of the current utterance as the antecedent of the current *zero*, and then do RS operation A and C in order on the preceding utterance.

5.4 Evaluation

After the zero anaphora resolution system is illustrated in the previous sections, we describe the experiment and result of the zero anaphora resolution in this section. In ZA detection, 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, we also use a rule without involving the centering model 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.

5.4.1 ZA Detection

By employing the ZA Triple Rules and ZA identification constraints, ZAs 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. Table 5.2

shows the precision rates calculated using Equation (5.1).

$$\text{Precision rate of ZA detection} = \frac{\text{No. of ZA correctly detected}}{\text{No. of ZA candidates}} \qquad (5.1)$$

The main errors of ZA detection occur in the experiment when parsing inverted sentences and non-anaphoric cases (e.g. exophora or cataphora). In this paper, we do not deal with the non-anaphoric cases, but we can employ ZA identification constraints, e.g. the ZACF constraint 1, to filter out about 60% cataphors in the test corpus.

5.4.2 Antecedent Identification

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 model and then employ the antecedent identification rule to show the improvement.

Simple Antecedent identification rule

For each ZA *z* in a discourse segment $U_1, \dots, U_m$: If *z* occurs in U_i then choose the noun phrase in U_{i-1} having the longest distance from *z* as the antecedent.

Table 5.2: Results of ZA detection

Cases ZAs	ZA Triple rules	ZA Triple rules + ZA identification constraints
No. of ZAs	2315	2315
ZA Candidates	3400	2754
Precision Rate	68%	84%

Table 5.3: Results of zero anaphora resolution

Cases Accuracy	Simple antecedent identification rule	Employing the centering model
Recall Rate	65.8%	70%
Precision Rate	55.3%	60.3%

The simple antecedent identification rule does not consider the ranking of centers in the centering model [Grosz *et al*. 1995]. By comparing with the simple antecedent identification rule, the antecedent identification rule is on the basis of the centering model. Considering Example (5.6), the ZAs are detected in Utterance (5.6b) and (5.6c) and both of their antecedents are 那顆蘋果 'that apple' in Utterance (5.6a). If we employ the simple antecedent identification, which does not concern with the centers of an utterance, ϕ^{i}_{2} would refer to 張三 'Zhangsan,' which is the noun phrase in the preceding utterance having the longest distance from the it.

(5.6) a. 那 顆 蘋果^{i} 已經 放 在 桌 上 幾 天 了，

na ke pingguo^{i} yijing fang zai zhuo shang ji tian le.

that CL apple already put at table on several day CRS

That apple has been put on the table for several days.

b. 張三 沒 吃 ϕ^{i}_{1} ，

Zhangsan mei chi ϕ^{i}_{1}.

Zhangsan not eat

Zhangsan didn't eat (it).

c. 李四 也 沒 動 ϕ^{i}_{2} 。

Lisi ye mei dong ϕ^{i}_{2}.

Lisi also not touch

Lisi also didn’t touch (it).

Table 5.3 shows the recall rates and precision rates of zero anaphora resolution calculated using Equation (5.2) and Equation (5.3). Main errors occur when a ZA does not refer to an entity that is the highest ranked forward-looking center in the preceding utterance nor the antecedent of the ZA is not in the preceding utterance.

$$\text{Precision rate of ZA resolution} = \frac{\text{No. of antecedent correctly identified}}{\text{No. of ZA identified}} \qquad (5.2)$$

$$\text{Recall rate of ZA resolution} = \frac{\text{No. of antecedent correctly identified}}{\text{No. of ZA occurred in text}} \qquad (5.3)$$

5.5 Summary

The implementation of our zero anaphora resolution system including ZA detection and antecedent identification is described in this chapter. We realized the rules in Chapter 4 established by investigating the related linguistic studies and computational strategies for anaphora resolution. The shallow parser is designed with Triple and ZA Triple rules, which are created according the ZA detection rules. We also illustrated the antecedent identification component implemented based on the ZA identification constraints and antecedent identification rules and evaluated their performances. The precision rate of ZA detection is 84% and the recall rate of zero anaphora resolution is 70%. The errors of ZA resolution are in the following cases:

1. Out of the forward-looking center ranking criterion (ranking of forward-looking centers): When a ZA refers to an entity in the preceding utterance but the entity is not the highest ranked forward-looking center.
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 thesis, but we can detect about 60% cataphora and exophora and 50% inverted sentences in the test corpus by employing ZA identification constraints.

CHAPTER 6

APPLICATIONS OF ZERO ANAPHORA RESOLUTION IN CHINESE

6.1 Introduction

Anaphora resolution plays an important role in a number of NLP applications, such as machine translation, information retrieval, question answering, text summarization and so on [Mitkov 2002]. Though in the previous similar works in English information retrieval [Bonzi and Liddy 1990], they report on that resolution of anaphors may not help in information retrieval, the recent investigations shows that the pronominal anaphora resolution can contribute the improvement of information retrieval or question-answering systems [Vicedo and Ferrández 2000, Edens *et al*. 2003, Watson *et al*. 2003]. In Chinese, topics or zero anaphors may carry more important information then pronouns do, due to their frequent occurrence in a Chinese discourse. For evaluating how the zero anaphora resolution performs on Chinese NLP applications, we integrate it into the NLP applications to show the performance.

In this chapter, we present a text categorization system that utilizes the zero anaphora resolution system to recover the omissions of anaphors in test text and then the resulting text is used as the input of the text categorization system [Yeh and Chen 2003a]. An information retrieval system employing the topic identification method [Yeh and Chen 2004b] to resolve the omissions of topics and then to extract the topics of documents in the text collection for creating better indices is also presented. At last, we report on a method of creating XML Topic Maps (XTM) [Pepper and Moore 2001] based on the topic identification method [Yeh and Chen 2004c].

6.2 Using Zero Anaphora Resolution to Improve Text Categorization

Text categorization is the task of classifying documents into a certain number of predefined categories. A list of keywords is used to represent a document or a class, so that a free document can be categorized by comparing its keyword list and those of document classes. A number of supervised learning algorithms such as the naive Bayes, k-nearest neighbors and Rocchio have been applied to this issue, which use pre-classified documents as training data [Joachims 1997, Schapire *et al*. 1998, Tsay and Wang 2000]. Without training data set, the unsupervised training methods use techniques of clustering which group similar documents into one cluster that no longer distinguishes between constituent documents [Ko and Seo, 2000]. The main difference between these two training methods is that supervised training method needs the pre-classified documents for the training data set. In general, the accuracy of text categorization based on supervised learning is better than based on unsupervised learning.

We employ the zero anaphora resolution method to recover the omissions of topic or subject, and object in utterances to improve the accuracy of text categorization. The text with ZAs resolved is taken as the input of a text categorization system. The new text categorization system works as below: First an input document with ZA resolved is taken as a ZA-resolved input document which each ZA in the text is replaced by its antecedent. Second the ZA-resolved input document is categorized by the *k*-nearest neighbor (*k*-NN) classifier.

6.2.1 Term Extraction

Due to the nature of Chinese language, there is no blank between words. In the term extraction, we employ Bi-gram model to extract terms from documents that belong to each class [Yang *et al*. 1993; Chen 2001]. The process starts from the first character of the sentence and combines two consecutive characters to form a bi-gram. Then it goes on to the second and repeats the grouping further on until the end. Consequently, all possible overlapping bi-grams are obtained. For example, the terms extracted from sentence “花蓮成爲熱門的旅遊地點” are: [花蓮, 蓮成, 成爲, 爲熱, 熱門, 門的, 的旅, 旅遊, 遊地, 地點].

6.2.1 *k*-NN Algorithm

As an instance-based classification method, *k*-NN has been known as an effective approach to a broad range of pattern recognition and text classification problems [Yang *et al*. 2002, Ko and Seo 2002]. In the *k*-NN algorithm, a new input instance should belong to the same class as their *k* nearest neighbors in the training data set. After all the training data is stored in memory, a new input instance is classified with the class of *k* nearest neighbors among all stored training instances.

$$w(t,d) = tf_{t,d} \times idf_t \tag{6.1}$$

$$idf_t = \log(\frac{N}{df_t}) \tag{6.2}$$

where

i) $tf_{t,d}$ is the within-document term frequency (TF).

ii) N is the number of all training documents.

iii) df_t is the number of training documents in which t occurs.

For the distance measure and the document representation, we uses the conventional vector space model, which represents each document as a vector of term weights, and the distance between two documents is measured using the cosine value of the angle between the corresponding vectors. We compute the weight vectors for each document using one of the conventional TFIDF (term frequency and inverse document frequency) term weighting schemes [Salton and Buckley 1988]. The weight of term *t* in document *d*, *w*(*t*,*d*), is calculated as a TFIDF value by the Equation (6.1) and (6.2).

Given a test document *d*, the *k*-NN classifier assigns a relevance score to each candidate category c_j using the following Equation (6.3):

$$s(c_j, d) = \sum_{d' \in R_k(d) \cap D_j} \cos(wd, wd') \tag{6.3}$$

where $R_k(d)$ denotes a set of the *k* nearest neighbors of document d and D_j is a set of training documents in class c_j.

6.3 Using Topic Identification to Improve Information Retrieval

Information retrieval is to identify documents, from text collections, which are relevant with respect to some query. In current information retrieval systems, users can query with an unordered set of keywords, a question or a sentence. A list of document links matching the query can be retrieved and ordered by relevancy between the query and the documents. In this section, we are concerned with a hypothesis that the discourse-level element, topic, could be used to contribute the calculations of information retrieval. We employ a topic identification method [Yeh and Chen 2004b] based on the centering model to recover the omissions of topics and extract the topics of documents in the text collection. Then the topic information is inserted into the text collection to create better indices for information retrieval.

As mentioned in Section 2.2, Chinese is a topic-prominent language, and the most important element, “topic,” of a sentence 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 valuable information embedded in text. The information can be further taken for improving the accuracy of some NLP applications like information retrieval. However, topics in utterances are frequently omitted from expressions in texts, due to their prominence in discourse. Accordingly, to identify the topic of each utterance in a discourse, we have to solve the problem of zero anaphora resolution. In this section, we illustrate a word-based information retrieval system which uses English and Chinese words as indexing terms.[25] The calculation of weighting of each indexing term is based on the TFIDF word weighting scheme [Salton and Buckley 1988]. Then we extract and insert the topic information into each document in the text collection to create indices and show the improvement of information retrieval.

6.3.1 Topic Identification

For identifying the topic of each utterance in text, we establish the topic identification rule on the basis of the centering model [Grosz *et al*. 1995]. 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 Chapter 4.

Topic identification rule:

```
For identifying each topic $t$ in a discourse segment consisting of utterances $U_1, \dots, U_m$:
If at least one ZA occurs in $U_i$
    then refer to forward-looking center ranking criterion to choose the antecedent of the ZA as the $t$
Else if no ZA occurs in $U_i$
```

[25] A Chinese word here is a meaningful word consisting of one or more Chinese characters, such 學校 ‘school’ and 加入 ‘join’.

```
    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 (3.10) to identify each topic of the utterances (3.10a) to (3.10d) by employing the topic identification rule. As shown in Table 6.1, the topic of (3.10a) is 電子股 ‘Electronics stocks,’ and the topic of (3.10b) is omitted identified as the antecedent of ϕ^i_1, 電子股 ‘Electronics stocks.’ Similarly, the topic of (3.10d) is 證券股 ‘Securities stocks,’ which is referred to as the antecedent of the zeroed topic of (3.10c). Therefore, we can obtain the topics of this example in the second column of Table 6.1.

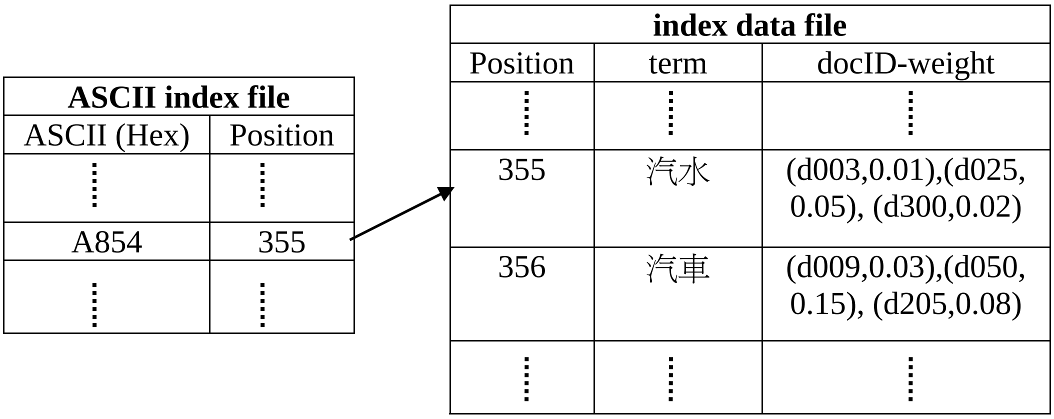
Table 6.1: Example of topic identification

	Utterance	Topic
(3.10a)	電子股^i 受 美國高科技股 重挫 影響， Electronics stocks^i were affected by high-tech stocks fallen heavily in America	電子股 Electronics stocks
(3.10b)	ϕ^i_1 持續 下跌。 (Electronics stocks)^i continued falling down.	電子股 Electronics stocks
(3.10c)	證券股^j 也 有 相對回應， Securities stocks^j also had respondence	證券股 Securities stocks
(3.10d)	ϕ^j_1 陸續 下殺 至 跌停。 (Securities stocks)^j fell by close one after another	證券股 Securities stocks

6.3.2 Information Retrieval System

The system is illustrated with two phases, document indexing and query processing including query matching and output ranking. In the document indexing phase, documents of a test collection are first segmented by the AUTOTAG [CKIP 2003], and each utterance of a document is transformed into a list of POS-tagged words separated by blanks. After the segmentation accomplished, each output document is taken as input to the system and is assigned a document number as identification (docID). Every word in an input document *d* is taken as an indexing term *t*, whose weight *w*(*t*,*d*) is calculated as a TFIDF [Salton and Buckley 1988] value by Equation (6.1) and (6.2).

The system creates an index data file which stores an indexing term list in the order of the ASCII (American Standard Code for Information Interchange) code of the each indexing term's first character. An indexing term was followed a sequence of docID-weight value pair. In addition to the index data file, the system also builds an ASCII index file, which stores the ASCII codes of all indexing terms' first characters and records their positions in the index data file.



ASCII index file

ASCII (Hex)	Position
⋮	⋮
A854	355
⋮	⋮

index data file

Position	term	docID-weight
⋮	⋮	⋮
355	汽水	(d003,0.01),(d025, 0.05), (d300,0.02)
356	汽車	(d009,0.03),(d050, 0.15), (d205,0.08)
⋮	⋮	⋮

Figure 6.1: An example of ASCII index file and index data file

Figure 6.1 shows an example that an ASCII code points to the indexing terms in the index data file. The ASCII code of the Chinese character *汽* is *A854*, whose position is *355* recorded in the ASCII index file. The position *355* is the starting position of indexing terms having the first character, *汽*, like *汽水* ‘soda-water’ and *汽車* ‘automobile’ in the index data file. By referring the ASCII index file, the system can obtain the starting position for efficiently searching the indexing terms.

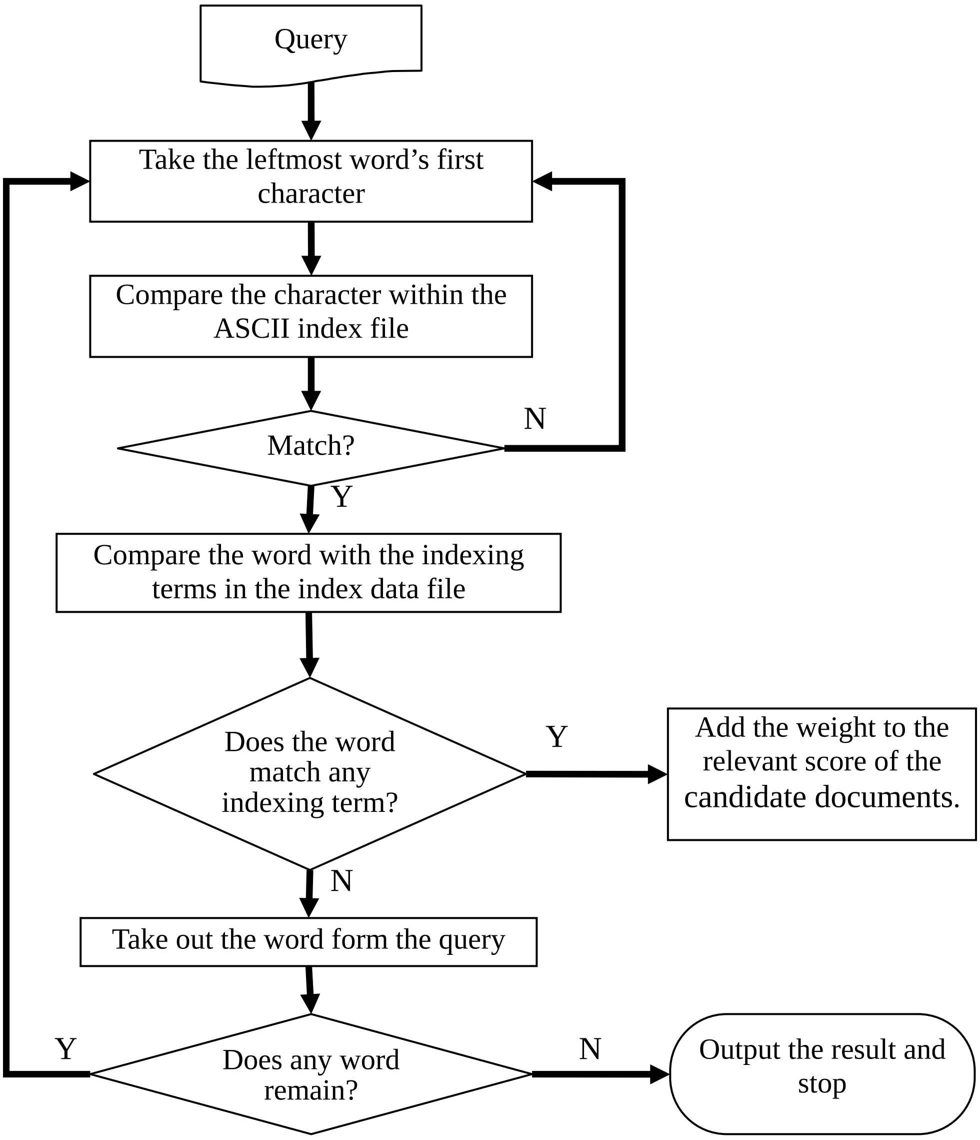

Query
Take the leftmost word’s first character
Compare the character within the ASCII index file
Match?
N
Y
Compare the word with the indexing terms in the index data file
Does the word match any indexing term?
Y
Add the weight to the relevant score of the candidate documents.
N
Take out the word form the query
Y
Does any word remain?
N
Output the result and stop

Figure 6.2: The procedure of query processing

In the information retrieval system, a query is a list of English or Chinese words separated by blanks or commas. As shown in Figure 6.2, the system first seeks each input word one after another. If the first character of the word of matches any ASCII code in the ASCII index file, the system gets the matched ASCII code and its position pointing to the index data file. Second, by referring the position of the matched ASCII code, the query word is taken to compare with the indexing terms in the index data file. If the query word matched an indexing term, the system gets the weights corresponding to candidate documents. After all the words of the query are processed, the system will then sum up the weights of the query to each document as the relevant score. Once all the candidate documents are selected that match the query, they are ranked by the relevant scores.

6.4 Creation of Topic Map by Identifying Topic Chain in Chinese

XML Topic Maps [Pepper and Moore 2001] enable multiple, concurrent views of sets of information objects and can be used to different applications. For example, thesaurus-like interfaces to corpora, navigational tools for cross-references or citation systems, information filtering or delivering depending on user profiles, etc. However, to enrich the information of a topic map or to connect with some document's URI is very labor-intensive and time-consuming. To solve this problem, we propose an approach based on zero anaphora resolution to identify and extract useful information in raw Chinese text. We first identify the topics of sentences in a document, and then assign this document into a topic node of the topic map and add the information of the document into the topic element simultaneously.

A topic map is composed of a number of topics, associations and occurrences [Biezunski *et al*. 1999]. A topic is a reification of subject in the real world. An association indicates the interrelationship between a pair of topics or even more parties. An occurrence connects the information relevant to a subject to the corresponding topic. The granularity of occurrence can range from a whole document to a span of text in document. In Chinese, a topic chain occurs in a span of text, where the topic of the beginning occurrence is referred by the succeeding utterances whose topics are omitted [Li and Thompson 1981]. Observing Chinese written texts, the span of text covered by a topic chain to a large extent represents the information about a certain subject. In other words, a topic chain can be used as the source of obtaining occurrences of topics. By employ our shallow parsing technique and zero anaphora resolution method, we may identify the topic chains in a text for further creating metadata in topic maps.

6.4.1 Topic Map

The purpose of a topic map is to convey knowledge about resources through a superimposed layer, or map, of the resources. A topic map captures the subjects of which resources speak, and the relationships between subjects, in a way that is implementation-independent. The key concepts in topic maps are *topics*, *associations*, and *occurrences* [Biezunski *et al*. 1999]. We now use the example in Figure 6.3 extracted from [Biezunski *et al*. 1999] to illustrate the relationship among topics, associations, and occurrences.

In Figure 6.3, an occurrence containing an addressable information resource, a URL, which is a reference resource of the topic "hamlet". In the example of associations, an association represents the relationship between *Shakespeare* and the play *Hamlet*. Because associations express relationships they are inherently multidirectional: If "Hamlet was written by Shakespeare", it automatically follows that "Shakespeare wrote Hamlet"; it is one and the same relationship expressed in slightly different ways. Instead of directionality, associations use roles to distinguish between the various forms of involvement members have in them. Thus the example above may be serialized using natural language as follows: “There exists a 'written by' relationship between Shakespeare (playing the role of 'author') and Hamlet (playing the role of 'work').” Relationships may involve one, two, or more roles [Biezunski *et al*. 1999].

```
<topic id="hamlet">
    <instanceOf><topicRef xlink:href="#play"/></instanceOf>
    <baseName> <baseNameString>Hamlet, Prince of Denmark </baseNameString>
    </baseName>
    <occurrence>
      <instanceOf> <topicRef xlink:href="#plain-text-format"/> </instanceOf>
      <resourceRef
        xlink:href="ftp://www.gutenberg.org/pub/gutenberg/etext97/1ws2610.txt"/>
    </occurrence>
</topic>
<association>
<instanceOf><topicRef xlink:href="#written-by"/> </instanceOf>
    <member>
      <roleSpec><topicRef xlink:href="#author"/></roleSpec>
      <topicRef xlink:href="#shakespeare"/>
    </member>
    <member>
      <roleSpec><topicRef xlink:href="#work"/></roleSpec>
      <topicRef xlink:href="#hamlet"/>
    </member>
</association>
```

Figure 6.3: An example of XML Topic Map

6.4.2 Creation of Metadata in Topic Maps

A topic chain is a frequently used grammatical structure in Chinese occurring in a span of text, where a referent is referred to in the first utterance and the following several utterances talking about the same referent but not overtly mentioning that referent [Li and Thompson 1981]. Topic identification here is similar to theme identification in [Rambow 1993]. The key elements of the centering theory, forward-looking centers and backward-looking center are employed to identify themes. The theme clearly corresponds to the backward-looking center: the theme, under a general definition, is what the current utterance is about; what utterances are about provides a link to previous discourse, since otherwise the text would be incoherent. The role of the backward-looking center is precisely to provide such a link.

In creation of topic maps, topic chains are used as the source of obtaining occurrences of topics. We employ the method of topic identification mentioned in Section 6.3 to identify the topic chains for developing the creation of metadata in a topic map. The metadata includes two child elements of the occurrence, *resourceRef* and *resourceData*. When the topic chains of a document are identified, we can add either the information of resourceRef to a topic node of a topic map or the information relevant to the topic of the document. Example (6.1) that is a short news article has a topic chain, 基隆醫院 ‘Kee-lung General Hospital,’ and we can add an occurrence containing the URL information of the news article to the topic 基隆醫院 ‘Kee-lung General Hospital’ as shown in Figure 6.4.

(6.1) a. 基隆醫院 ^{i} 爲 擴大 服務 範圍，

Jilong yiyuan^{i} wei kuoda fuwu fanwei.

Kee-lung hospital for expand service coverage

Kee-lung General Hospital aims to increase service coverage.

b. ϕ^{i}_{1} 積極 提升 醫療 服務 品質 及 標準化 ，

ϕ^{i}_{1} jiji tisheng yiliao fuwu pinzhi ji biaozhunhua.

(Kee-lung General Hospital) active improve medical-treatment service quality and standardization

(Kee-lung General Hospital) actively improves the service quality of medical treatment and standardization.

c. ϕ^{i}_{2} 獲 衛生署 認可 爲 辦理 外勞體檢醫院 。

ϕ^{i}_{2} huo weishengshu renke wei banli wailao tijian yiyuan.

(Kee-lung General Hospital) obtain Department-of-Health certify to-be handle foreign-laborer physical-examination hospital

(Kee-lung General Hospital) is certified by Department of Health as a hospital which can handle physical examinations of foreign laborers.

```
<topic id="基隆醫院">
   <occurrence>
      <instanceOf>
         <topicRef xlink:href="# plain-text-format "/>
      </instanceOf>
      <resourceRef xlink:href="URL_Of_The_News_About_基隆醫院"/>
   </occurrence>
</topic>
```

Figure 6.4: An example of creation metadata in a topic map

6.5 Summary

Some Chinese NLP applications of zero anaphora resolution are presented in this chapter. We first described the text categorization system integrated with the process of zero anaphora resolution. In this system, each document of the test data set are recovered their omissions of topic or subject, and object in utterances by employing our zero anaphora resolution method. Second, we presented the information retrieval system, which utilizes the topic identification method to identify the topic of each utterance of a document for obtaining better indices. Since topics are often omitted in text, the topic identification method partially adopts the zero anaphora resolution method to resolve the zeroed topics. Third, we propose an approach to creating metadata of XML Topic Maps. The approach employs the similar topic identification method in the preceding application but focuses on identifying topic chains, which is a frequently used grammatical structure in Chinese occurring in a span of text talking about the same referent. The topic chains are used as the source of obtaining occurrences of topics in topic maps. In the next chapter, we will further demonstrate the evaluation of these applications by performing some experiments and showing the results.

CHAPTER 7

EXPERIMENTS AND RESULTS OF THE APPLICATIONS

7.1 Introduction

After illustrating the applications of zero anaphora resolution in the previous chapter, we further perform the experiments for evaluating these applications including text categorization, information retrieval and topic identification. Because different test data are required by these NLP applications, the experiments are demonstrated on different test collections. One is a collection of pre-classified news articles, and the other collections are taken from the test set of the Chinese Information Retrieval Benchmark, version 3.0 (CIRB030) [Chen and Chen 2004]. The former is used to evaluate the performance of the text categorization system, while the later is adopted as the test corpora of topic identification and information retrieval.

7.2 Text Categorization

We collect 300 news articles pre-classified into 8 categories as the test collection which contains about 132 thousands Chinese characters for the experiments of zero anaphora resolution and text categorization. Half of the test collection is taken to be resolved the ZAs by employing the ZA detection rules and antecedent identification rule, and these news articles are recovered the omitted elements in utterances. The other half is taken as the training data of the text categorization system.

In the ZA detection rules mentioned in Section 4.3, zero anaphors may occurs with verbs, coverbs, coordinating conjunctions, or prepositions in utterances. Because the process of zero anaphora resolution works on the output text of AUTOTAG [CKIP 2003] and the POS-tagging program can not recognize coverbs but identifies them as prepositions or coordinating conjunctions that are essential POS of coverbs in Chinese grammar, the results of ZA resolution are summarized in three cases, verbs, coordinating conjunctions and prepositions, shown in Table 7.1. In Table 7.1, the recall rates and the precision rates are calculated using Equation 5.2 and Equation 5.3 mentioned in Chapter 5.

The main errors of zero anaphora resolution occur when the antecedent of a ZA is not the highest ranked forward-looking center in the preceding utterance. For showing the improvement of integrating zero anaphora resolution into text categorization, we first run the text categorization system on the original data of the test collection as the baseline. Half of 300 news articles are taken as the training data set and the other half as the test data set. Without applying ZA resolution on test data set, the accuracy of categorization is 79% (118/150) which is calculated with Equation (7.1).

$$\text{Accuracy of text categorization} = \frac{\text{No. of the documents correctly classified}}{\text{No. of test documents}} \tag{7.1}$$

Table 7.1: Results of ZA resolution

Cases Accuracy	Verb	Coordinating conjunction	Preposition	Total
Recall Rate	67.4% (1076/1597)	80% (12/15)	67.3% (279/414)	67.4% (1367/2029)
Precision Rate	64.2% (1076/1676)	66.7% (12/18)	53.9% (279/518)	61.8% (1367/2212)

After the experiment performed on the original test collection, we then have the process of zero anaphora resolution on all the articles of the test data and perform the experiment of text categorization again. The result shows that the accuracy increases from 79% to 84% (126/150).

7.3 Topic Identification

As mentioned in Section 6.3.1, 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 established the topic identification rule in Section 6.3 for identifying the topics in text. For evaluating the topic identification method, we took a subset of the 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 average number of utterances of an article is 52, where 17 ZAs occurs in the topic position. The recall rates and precision rates of zero topic resolution are 0.67 and 0.64 respectively calculated using Equation (7.2) and Equation (7.3). Most errors occur when a zero topic does not refer to the topic in the preceding utterance, or refers to other entity in the more previous utterance.

$$\text{Recall rate of zero topic resolution} = \frac{\text{No. of antecedent correctly identified}}{\text{No. of zero topic occurred in text}} \tag{7.2}$$

$$\text{Precision rate of zero topic resolution} = \frac{\text{No. of antecedent correctly identified}}{\text{No. of zero topic identified}} \tag{7.3}$$

7.4 Information Retrieval

As described in Section 6.3, we use topic identification to improve information retrieval. For evaluating this approach, we selected 592 news articles of China Times Express as the test collection *A* according to the CIRB030 Answer Set.[26] The Answer Set is a list of document numbers (DOCNO) assigned the topic IDs and their relevance in four categories: “Highly Relevant”, “Relevant”, “Partially relevant”, and “Irrelevant.” Because many topic categories do not have enough relevant articles for observing the results, we further put 30 relevant news articles of Central Daily News into the test collection as the test collection *B*, which is the same as the test corpus used in the experiment of topic identification in the preceding section.

We performed an experiment to examine the effectiveness of using topic identification for information retrieval. In the experiment, we take the test collection *A* as input to the information retrieval system as the baseline, and then insert topic information to each news article to show the improvement. The keywords relevant to topics of CIRB030 Topic Set are taken as the queries for the test. The recall rates and R-precision rates of information retrieval are calculated using Equation (7.4) and Equation (7.5)

[26] CIRB030 developed by Kuang-hua Chen, National Taiwan University is a test collection designed to be used for performance evaluation of Chinese document retrieval. The test collection contains three parts: Document Set, Topic Set and Answer Set. It is a helpful and powerful tool for investigation of the developing systems and the developed systems.

respectively. Table 7.2 summarizes the number of relevant articles categorized by their topic ID and Table 7.3 shows the result of the experiment on the test collection *A*.

$$\text{Recall rate of information retrieval} = \frac{\text{No. of relevant articles retrieved}}{\text{No. of relevant articles for a query}} \tag{7.4}$$

$$\text{R - Precision rate of information retrieval} = \frac{\text{No. of relevant articles at top R articles retrieved}}{\text{No. of relevant articles for a query} (= \text{R})} \tag{7.5}$$

Table 7.2: Summary of the test collection A

TopicID	Title	Number of relevant articles
002	Joining WTO	3
006	Nobel Prizes in Physics	2
007	China Airlines Crash	1
009	Satellite ST1	3
013	Province-refining	5
014	Computer virus	10
018	Doomsday thought	2
019	Economic influence of the European monetary union	1
033	Clinton scandals	7
035	War crimes lawsuits	1
036	Nuclear power protests	3
039	College Admission Policy	6
046	Regulations and Damages from Drunken Driving	3
050	Teenager's Fashion	1
Total		48

Table 7.3: Results of the experiment on the test collection A

	Articles retrieved	Recall rate	R-precision rate
Baseline	5	40%	35%
	10	56%	
	20	72%	
After topic identification	5	42%	40%
	10	64%	
	20	82%	

Table 7.4: Results of the experiment on the test collection B

	Articles retrieved	Recall rate	Precision rate
Baseline	10	58%	58%
	20	80%	40%
After topic identification	10	65%	65%
	20	85%	43%

The experiment is performed repeatedly by replacing the test collection with the test collection *B*, and the keywords of six topic categories, which have ten relevant articles each, are taken as the queries. The recall rates and precision rates of information retrieval are calculated using Equation (7.4) and (7.6). Table 7.4 shows the result of the experiment on the test collection *B*.

$$\text{Precision rate of information retrieval} = \frac{\text{No. of relevant articles retrieved}}{\text{No. of articles retrieved}} \qquad (7.6)$$

7.5 Summary

We presented the experiments and results of the NLP applications including text categorization, information retrieval and topic identification for evaluating the performance of our zero anaphora resolution method.

In the text categorization system, each query text is recovered the occurrences of ZAs and then the resulting text is used as the new input query text. The result shows that ZA resolution method enhances the accuracy of text categorization from 79% to 84%. In the topic identification, we deal with the resolution of topic omission occurring in utterances. The recall rates and precision rates of zero topic resolution are 67% and 64% respectively. By adopting this topic identification method, the information retrieval system we developed can obtain the most important information embedded in text. The result of employing the information carried by zero topics in text to contribute the indexing of information retrieval is promising to some extent.

CHAPTER 8

SUMMARY AND FUTURE DIRECTIONS

8.1 Summary

In this thesis, we first developed the two-phase method of Chinese zero anaphora resolution: detecting the occurrences of ZAs in text, and then finding their antecedents in the discourse. In the phase of ZA detection, by referring to the Chinese syntax, relevant linguistics on zero anaphora and observation of real data, we considered the ZAs occurring with the predicates which are verbs, coverbs, prepositions, and coordinating conjunctions. The ZA diction rules were created corresponding to these cases. In the phase of antecedent identification, the notions of the centering model of local discourse coherence were adopted to choose the antecedents of ZAs. The notions were taken as a basis for establishing the antecedent identification rule. We also designed the ZA identification constraints for filtering out the non-anaphoric cases such as cataphora and exophora which might be incorrectly detected as ZAs.

Second we employed the above rules to implement the zero anaphora resolution system. The resolution system works on the output of a POS tagger and employs ZA detection rules for detecting omitted cases as ZA candidates. After each ZA candidate is detected, the ZA identification constraints are utilized to eliminate the non-anaphoric cases of these candidates and the antecedent identification rule is used to identify the antecedents. The system used the shallow parsing which is an inexpensive and reliable method does not deliver complex syntactic analysis but is limited to parsing smaller syntactical related constituents. The shallow parser was implemented on the Triple Rules and ZA Triple Rules, which could parse utterances into triples and also annotate the occurrences of ZAs. The evaluation of the resolution system was carried out on the test collection of news articles. The precision rate of ZA detection is 84% and the recall rate of zero anaphora resolution is 70%.

Third, we developed some NLP applications for examining the efficiency of the zero anaphora resolution method. A text categorization system was implemented with integrating the zero anaphora resolution process. The result showed that the accuracy increases from 79% to 84%. An information retrieval system which uses topic identification to obtain more valuable information embedded in text is designed and implemented. The experiments of the information retrieval system were performed on the test collections taken from CIRB030. On the different test collections, the precision rate and the recall rate are both improved. The topic identification method used in the information retrieval system employed the notion of the centering model and the zero anaphora resolution method to identify the topic of each utterance in text. We also proposed an approach based on this identification method to create the metadata of XML Topic Maps. For evaluating the topic identification method, we demonstrated the experiment on the test collection and the recall rates and precision rates of zero topic resolution are 0.67 and 0.64 respectively. In the following, the contributions of this thesis are summarized:

1. The computational rules based on the centering model for the Chinese zero anaphora resolution in contrast to the anaphora resolution work integrating complex linguistic information or relying on the labor-intensive and time-consuming construction of knowledge bases have been developed.
2. An implementation and evaluation of a Chinese zero anaphora resolution system by employing these rules has been demonstrated. The evaluation result shows that applying the centering model on Chinese zero anaphora is workable.

3. The experiments on Chinese NLP applications employing the zero anaphora resolution method have been demonstrated, and the results show that the method would make contribution to Chinese information retrieval and text categorization.

8.2 Future Directions

We suggest several issues related to this work, which require future investigation. These issues include the improvement on the ZA identification constraints for eliminate non-anaphoric cases, the resolution of other forms of anaphors like pronominal anaphors, the extended use of anaphora resolution in analysis of discourse coherence and other NLP applications involving anaphora resolution.

ZA identification. In the experimental results of zero anaphora resolution, we have found that some errors occur when the non-anaphoric cases like cataphora and exophora are detected as ZAs. Although the ZA identification constraints we have employed in the resolution system can filter out more than half of these cases in the test data, we still need to investigate more efficient constraints or approaches for solving this problem.

Coreference resolution. Coreference resolution is the task of identifying the expressions such as noun phrase, nominal and pronominal anaphors in text, which refer to the same entity. We may modify the antecedent identification rule mentioned in Section 4.4 to identify the antecedents of other kinds of anaphors occurring in utterances and some anaphora resolution factors can be used, such as gender and number agreement. However, the work of pronominal anaphora resolution also needs to consider the problem of elimination of non-anaphoric cases.

Analysis of discourse coherence. Centering is a computational model that relates focus of attention, choice of referring expression, and perceived coherence of utterances within a discourse segment. It measures coherence by the hearer’s inference load when interpreting a discourse segment. [Grosz *et al*. 1986, Grosz *et al*. 1995]. Our zero anaphora resolution method is on the basis of the centering model. The key elements of this model of local discourse coherence are employed to identify the antecedents of ZAs. After resolving the occurrences of ZAs in text, we may further measure the coherence of a discourse segment by examining the transition states, which are the relationship between attentional states of successive utterances. This issue has been discussed in other topic-orient language like Japanese and deserves more investigations in Chinese.

Other NLP applications. We have integrated the zero anaphora resolution process in some Chinese NLP applications in this thesis. There are still other applications involving the work of anaphora resolution such as question answering, machine translation and text summarization. Another future work is to build these systems by integrating our resolution method.

BIBLIOGRAPHY

[1] Abney, Steven, 1991, Parsing by chunks, in: Robert Berwick, Steven Abney, and Carol Tenny, editors, *Principle-Based Parsing*, Kluwer Academic Publishers.

[2] 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.

[3] 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.

[4] Aone, Chinatsu and McKee, Douglas, 1993, A language-independent anaphora resolution system for understanding multilingual texts". *Proceedings of the ACL'93*, 156-163.

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

[6] Biezunski, M., Bryan, M. and Newcomb, S., editors, 1999, *ISO/IEC 13250 Topic Maps: Information Technology -- Document Description and Markup Languages*.

[7] Blum, A. and Mitchell, T., 1998, Combining labeled and unlabeled data with Co-Training, in *Proceedings of the 11th Annual Conference on Learning Theory*, pages 92–100.

[8] Bonzi, S. and Liddy, E. D., 1990, The use of anaphoric resolution for document description in information retrieval. *Information Processing and Management*, 25(4): 429-441.

[9] 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.

[10] Byron, D. K., and Tetreault, J., 1999, A flexible architecture for reference resolution, in *Proceedings of 9th Conference on EACL*, pages 229–232.

[11] Carbonell, J. G. and Brown, R.D., 1988, Anaphora resolution: a multi-strategy approach, *Proceedings of the 12th International Conference on Computational Linguistics* (*COLING'88*), pages 96-101, Budapest, Hungary.

[12] Carlson, L., Marcu, D., and Okurowski, M. E., 2001, Building a discourse-tagged corpus in the framework of rhetorical structure theory, in *Proceedings of the 2nd SIGDIAL Workshop on Discourse and Dialogue*, *Eurospeech 2001*, Denmark.

[13] Carter, D. M., 1987, *Interpreting Anaphors in Natural Language Texts*, Chichester, Ellis Horwood.

[14] Carter, D. M., 1990, Control issues in anaphor resolution, *Journal of Semantics*, 7, pages 435-454.

[15] 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.

[16] Chen, Hongbiao, 2001, *Looking for Better Chinese Indexes: A Corpus-based Approach to Base NP Detection and Indexing*, Ph.D. thesis, Guangdong University of Foreign Studies.

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

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

[19] Chomsky, N., 1981, *Lectures on government and binding*. Foris, Dordrecht.

[20] CKIP, 1999, Zhong wen duan ci xi tong (中文自動斷詞系統) Version 1.0 (Autotag),

http://godel.iis .sinica.edu.tw /CKIP/, Academia Sinica.

[21] 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*, pages 255-261, Manchester, United Kingdom.

[22] Correa, Nelson, 1988, A Binding Rule for Government-binding Parsing, in *Proceedings of the 12th International Conference on Computational Linguistics* (*COLING'88*), pages 123-129.

[23] Cristea, D., Nancy Ide, and Laurent Romary, 1998, Veins theory: An approach to global cohesion and coherence, in *Proceedings of ACL/COLING*, pages 281–285, Montreal, Canada.

[24] Daelmans, W., J. Zavrel, K. van der Sloot, and A. van der Bosch, 2000, *TiMBL: Tilburg memory based learner bersion 3.0*, *Reference Guide*, *Technical Report ILK 00-01*, Tilburg University.

[25] Edens, Richard J., Helen L. Gaylard, Gareth J. F. Jones and Adenike M. Lam-Adesina, 2003, An Investigation of Broad Coverage Automatic Pronoun Resolution for Information Retrieval, *Proceedings of SIGIR*, 381-382.

[26] Ferrández, A., Palomar, M. and Moreno, L., 1997, Slot Unification Grammar, *Joint Conference on Declarative Programming, APPIA-GULP-PRODE*'97.

[27] Ferrández, A., Palomar, M. and Moreno, L., 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.

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

[29] 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.

[30] 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.

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

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

[33] 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*.

[34] 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.

[35] Guenthner, F. and Lehmann, H., 1983, Rules for pronominalization, in *Proceedingsof the First Conference on the European Chapter of the Association for Computational Linguistics*, pages 144–151, Pisa, Italy.

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

[37] Hess, Michael, 1991, Recent Developments in Discourse Representation Theory, in: M. King (ed.), *Communication with Men and Machines*, Geneva.

[38] 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.

[39] Hu, Wenze, 1995, *Functional Perspectives and Chinese Word Order*, Ph. D.

dissertation, The Ohio State University.

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

[41] Joachims, Thorsten, 1997, A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In Douglas H. Fisher, editor, *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pages 143-151, San Francisco, US.

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

[43] Kamp, H., 1981, A Theory of Truth and Semantic Representation, in: J.A.G. Groenendijk, T.M.V. Janssen, M.B.J. Stokhof (eds.), *Formal Methods in the Study of Natural Language*; Part 1, Mathematisch Centrum, Tract 135, Amsterdam, pp. 277-322.

[44] 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.

[45] Ko, Youngjoong and Seo, Jungyun, 2000, Automatic Text Categorization by Unsupervised Learning. *Proceedings of COLING-00, the 18th International Conference on Computational Linguistics*, pages 453-459.

[46] Kurohashi, S., 1998, *Japanese Dependency/Case Structure Analyzer KNP version 2.0b6* (in Japanese), Department of Informatics, Kyoto University.

[47] Kurohashi, S. and Nagao M., 1998a, Building a Japanese parsed corpus while improving the parsing system, in *Proceedings of The 1st International Conference on Language Resources & Evaluation*, pages 719–724.

[48] Kurohashi, S. and Nagao M., 1998b, *Japanese morphological analysis system JUMAN version 3.6 manual* (in Japanese), Department of Informatics, Kyoto University.

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

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

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

[52] Liao, Chiu-chung (廖秋忠), 1992, *廖秋忠文集*, 北京, 北京語言學院出版社.

[53] Lü, Shuxiang (呂叔湘), 1946, 從主語、賓語的分別談國語句子的分析, in 1989, *呂叔湘自選集*, 445-480, 上海, 上海教育出版社.

[54] Lü, Shuxiang (呂叔湘), 1986, 漢語句法的靈活性, *中國語文* 194:1-9.

[55] Lü, Shuxiang (呂叔湘), 1996, *現代漢語八百詞*, 北京,商務印書館.

[56] Markert, K., Nissim, M., and Modjeska, N., 2003, Using the web for nominal anaphora resolution, *EACL Workshop on the Computational Treatment of Anaphora*, pages 39-46.

[57] McCord, M., 1990, Slot grammar: a system for simpler construction of practical natural language grammars, *Natural Language and Logic: International Scientific Symposium*, edited by R. Studer, pages 118-145, Lecture Notes in Computer Science, Berlin, Springer Verlag.

[58] McCord, M., 1993, Heuristics for broad-coverage natural language parsing, *Proceedings of the APRA Human Language Technology Workshop*, University of Pennsylvania.

[59] 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.

[60] 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.

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

[62] Müller C., Rapp, S. and Strube, M., 2002, Applying Co-Training to Reference Resolution, in *Proceedings of the ACL’02*, pages 352-359, Philadelphia.

[63] 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)*, pages 871-876.

[64] Pepper, Steve and Moore, Graham, editors, 2001, XML Topic Maps (XTM) 1.0, *TopicMaps.Org Specification*.

[65] Preiss, Judita, 2002, Anaphora Resolution with Memory Based Learning, in *Proceedings of 5th Annual CLUK Research Colloquium* (*CLUK5*), pages 1-9.

[66] Quinlan, J. Ross, 1993, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers.

[67] Rambow, O., 1993, Pragmatic aspects of scrambling and topicalization in German: A Centering Approach, *IRCS Workshop on Centering in Discourse*. Univ. of Pennsylvania.

[68] Salton, G. and Buckley, C., 1988, Term Weighting Approaches in Automatic Text Retrieval, *Information Processing and Management*, 24(5):513-523.

[69] Schapire, R., Singer, Y. and Singhal, A., 1998, Boosting and rocchio applied to text filtering, *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval*, pages 215-223, Melbourne, Australia.

[70] 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.

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

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

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

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

[75] 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.

[76] Tetreault, J., 2002, Clausal structure and pronoun resolution, in *4th Discourse Anaphora and Anaphora Resolution Colloquium*, pages 217–220.

[77] Tetreault, J. and Allen, J., 2003, An Empirical Evaluation of Pronoun Resolution and Clausal Structure, *Proceedings of the 2003 International Symposium on Reference Resolution and its Applications to Question Answering and Summarization*, pages 1-8, Venice, Italy.

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

[79] Tsay, Jyh-Jong and Wang, Jing-Doo, 2000, Design and Evaluation of Approaches to

Automatic Chinese Text Categorization, *Computational Linguistics and Chinese Language Processing (CLCLP)*, 5(2): 43-58.

[80] Vicedo, J. L. and Ferrández, A., 2000, Importance of Pronominal Anaphora Resolution to Question Answering Systems, *Proceedings of 38th Annual Meeting of the Association for Computational Linguistics* (*ACL*), Hong-Kong, 555-562.

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

[82] 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.

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

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

[85] Watson, R., Preiss, J. and Briscoe, E.J., 2003, The Contribution of Domain-independent Robust Pronominal Anaphora Resolution to Open-Domain Question-Answering, *Proceedings of Int. Symposium on Reference Resolution and its Application to Question-Answering and Summarisation*, 75-82.

[86] Wu, D. S. 2003, *Automatic Pronominal Anaphora Resolution in English Texts,* Master thesis, National Chiao Tung University, Taiwan.

[87] Yang Y., Slattery S., and Ghani R. 2002. A study of approaches to hypertext categorization, *Journal of Intelligent Information Systems,* 18(2):219-241.

[88] Yang, Yun-Yen, Chen, Keh-Jiann, Hsieh, Ching-Chun and Chen, Shu-Mei, 1993, A Study of Document Auto-Classification in Mandarin Chinese, in *Proceedings of*

ROCLING VI, Hsinchu, Taiwan.

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

[90] 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.

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

[92] Yeh, Ching-Long and Chen, Yi-Chun, 2003b, 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.

[93] Yeh, Ching-Long and Chen, Yi-Chun, 2004a, Zero Anaphora Resolution in Chinese with Shallow Parsing, *Journal of Chinese Language and Computing* (Accepted).

[94] Yeh, Ching-Long and Chen, Yi-Chun, 2004b, Topic Identification in Chinese Based on Centering Model. *Proceedings of ACL Workshop on Reference Resolution and Its Applications*, Barcelona, Spain.

[95] Yeh, Ching-Long and Chen, Yi-Chun, 2004c. Creation of Topic Map by Identifying Topic Chain in Chinese. *Proceedings of ACM Symposium on Document Engineering* (*DocEng2004*), Milwaukee, Wisconsin, USA.

APPENDIX A

There are about 47 POS tags used in AUTOTAG, in which 13 POS tags belong to the noun tag set and 17 POS tags belong to the verb tag set, as shown in Table A.1. The noun phrase and verb phrase rules are created in DCG as shown below.

Noun phrase rules

n(N)→ na; nb; nc; ncd; nd; nep; neqa; neqb; nes; neu; nf; ng; nh.

vah(V)→ va; vh.

np([N])→ n(N).

np([N1,Ns])→ n(N1), (de, np(Ns) ; np(Ns) ; [ ]).

np([V1,Ns])→ vah(V1), (de;[]), np(Ns) .

Verb phrase rules

v(N)→ va; vac; vb; vc; vcl; vd; ve; vf; vg; vh; vhc; vi; vj; vk; vl; v_2; v_11.

vp([V])→ v(V).

vp([V1,Vs])→ n(V1), (de, vp(Vs) ; vp(Vs) ; [ ]).

Table A.1: Noun and verb tag set

Set	**POS tag**
Noun	Na, Nb, Nc, Ncd, Nd, Nep, Neqa, Neqb, Nes, Neu, Nf, Ng, Nh
Verb	VA, VAC, VB, VC, VCL, VD, VE, VF, VG, VH, VHC, VI, VJ, VK, VL, V_2, V_11

APPENDIX B

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

Table B.1: Abbreviations

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)