# 國 立 交 通 大 學

## 資訊科學與工程研究所

## 博 士 論 文

# 以權重學習與知識擷取為基礎之中文指代消解研究

# Chinese Anaphora Resolution Based on Weight Learning and Knowledge Acquisition

研 究 生：吳典松

指導教授：梁　婷　博士

中 華 民 國 九 十 九 年 十 二 月

# 以權重學習與知識擷取為基礎之中文指代消解研究

# Chinese Anaphora Resolution Based on Weight Learning and Knowledge Acquisition

研 究 生：吳典松　　Student : Dian-Song Wu

指導教授：梁　婷 博士　　Advisor : Dr. Tyne Liang

國 立 交 通 大 學

資 訊 科 學 與 工 程 研 究 所

博 士 論 文

A Dissertation Submitted to

Institute of Computer Science and Engineering

College of Computer Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

in

Computer Science

December 2010

Hsinchu, Taiwan, Republic of China

中 華 民 國 九 十 九 年 十 二 月

# 以權重學習與知識擷取為基礎之中文指代消解研究

學生：吳典松　　指導教授：梁 婷 博士

國立交通大學 資訊科學與工程研究所

## 摘　　要

指代是一種常見的語言現象，用於避免篇章中相同敘述的重複。指代消解是指在篇章中辨識指代詞所對應的先行詞的程序。指代消解在許多自然語言處理的應用中扮演著不可或缺的角色，例如機器翻譯、文件摘要及資訊萃取。

在相關研究中，指代消解的方法多依靠語法規則、語意或語用的線索來辨識指代詞，而近年來多以統計或分類方法為研究方向。然而，在以規則為基礎的方法中，特徵分數的選取多依靠人工的方式來指定權重值，錯誤會因為主觀性的偏見而產生。另一方面，在以分類為基礎的方法中，每個候選詞在做選擇時彼此間是視為獨立的關係，因而無法獲得相對的偏好程度。為了克服這些問題，我們提出以權重學習與知識擷取為基礎之中文指代消解方法。

在本論文中，我們針對中文文件中的代名詞指代、零指代以及限定性名詞指代進行處理，並且根據個別性質提出不同的方法。我們使用詞彙知識擷取和特徵值測量來消解代名詞指代，詞彙知識擷取以抽取相關語意特徵為目的，例如，性別、數量及搭配相容性。特徵值測量則是以亂度為基礎的權重分配來選取先行詞。在 1343 個指代實例中進行實驗顯示，我們所提出的方法相對於以規則為基礎的方法獲得 7%的改善，消解成功率為 82.5%。

在零指代消解問題中，我們應用案例式推理及樣式概念化來克服建構推論機制及詞彙特徵不足的問題。在 1051 個指代實例中進行實驗顯示獲得的 F-score 為 79%，相對於以重心理論為基礎的方法獲得 13%的改善。

在限定性名詞指代消解問題中，我們使用特徵值測量的方式將所有候選詞同時進行評估，另外也利用以網頁搜尋為基礎的方法加上外部詞典的輔助，來進行語意相容性的判別。在 426 個指代實例中進行實驗顯示，我們所提出的方法相對於以分類器為基礎的方法獲得 4.7%的改善，消解成功率為 72.5%。

***關鍵字***：指代消解、特徵權重學習、知識擷取、網路探勘

# Chinese Anaphora Resolution Based on Weight Learning and Knowledge Acquisition

Student: Dian-Song Wu Advisor: Dr. Tyne Liang

Institute of Computer Science and Engineering

National Chiao Tung University

## ABSTRACT

Anaphora is a commonly observed linguistic phenomenon and used to avoid repetition of expressions in discourses. Anaphora resolution denotes the process of identifying the antecedent of an anaphor in a context. Effective anaphora resolution plays an essential role in many applications of natural language processing such as machine translation, summarization, and information extraction.

In previous research, anaphora resolution methods have relied on syntactic rules, semantic or pragmatic clues to identify the antecedent. More recently, statistical-based or classification-based approaches are focused. However, in a rule-based approach, a salience score by manual weight assignment is usually adopted to select the antecedent. Errors may occur due to intuitive observations and subjective biases in selecting feature weight. On the other hand, the drawback of a classification-based approach is that it considers different candidates for the same anaphor independently. Thus it cannot effectively capture the preference relationships between competing candidates during resolution. To overcome these problems, we propose Chinese anaphora resolution methods based on weight learning and knowledge acquisition.

In this thesis, pronominal, zero, and definite anaphora in Chinese texts are addressed and different approaches are presented. We use lexical knowledge acquisition and salience measurement to resolve Chinese pronominal anaphora. The lexical knowledge acquisition is aimed to extract more semantic features, such as gender, number, and collocate compatibility. The presented salience measurement is based on entropy-based weighting on selecting antecedent candidates. The experimental results show that our proposed approach yields 82.5% success rate on 1343 anaphoric instances, enhancing 7% improvement while compared with the general rule-based approach presented.

As to Chinese zero anaphora, we apply case-based reasoning and pattern conceptualization to overcome the difficulties of constructing proper reasoning mechanisms and insufficiency of lexical features. The experimental results show that our proposed approach achieved competitive resolution by yielding 79% F-score on 1051 anaphoric instances and yielded 13% improvement while compared with the general rule-based approach.

We use two strategies to resolve Chinese definite anaphors. One is an adaptive weight salience measurement in such a way that the entire set of candidates can be estimated simultaneously. Another scheme is a Web-based knowledge acquisition model so that semantic compatibility extraction and multiple resources can be employed. The experimental results show that our proposed approach yields 72.5% success rate on 426 anaphoric instances, enhancing 4.7% improvement while compared with the result conducted by a conventional classifier.

***Keywords***: Anaphora Resolution, Feature Weight Learning, Knowledge Acquisition, Web Mining

# ACKNOWLEDGEMENT

（誌　謝）

博士論文能夠完成，首先要感謝的是我的指導教授梁婷老師，感謝她在我的研究生涯中孜孜不倦的教誨與指導，並且在研究方法與論文寫作技巧上提供許多寶貴的意見，使我在學術研究的道路上獲益良多。

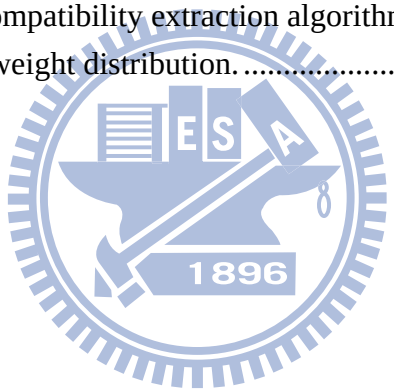
此外，也感謝系上楊武教授、彭文志教授與胡毓志教授在各階段口試時所提供的寶貴建議。同時感謝台灣大學資工系陳信希教授、清華大學資工系張俊盛教授、中研院資科所陳克健教授以及大同大學資工系葉慶隆教授在口試過程中提供許多寶貴的建議，使本篇論文趨於完善。感謝資訊擷取實驗室裡的許多研究夥伴們，多謝你們在這段期間對我的協助與鼓勵。

最後要感謝的是我的家人的支持與鼓勵，尤其是父母親與妻子素吟的無悔付出與關心，還有女兒怡萱和兒子彥德都是我最大的精神支柱。感謝愛犬小黑一生的忠心守護與陪伴，僅以此論文，獻給我所摯愛的家人們。

# TABLE OF CONTENTS

# LIST OF FIGURES





# LIST OF TABLES

# Chapter 1

# Introduction

## 1.1 Background and Motivation

In natural language communication, anaphora plays an essential role in the cohesion of discourses. Anaphora denotes the phenomenon of referring back to previously mentioned entities in a text. The referring expression is called an anaphor and the entity to which it refers is its antecedent [33]. Anaphors are used to avoid repetition of expressions in discourses. Different kinds of anaphoric expressions can be utilized in the context, such as pronominal anaphors, zero anaphors, and definite anaphors [33][46][52]. Followings are the definitions for each addressed anaphora.

*Definition* 1.1: Pronominal anaphora denotes that the preceding antecedents are referred by succeeding third personal pronouns including singular and plural forms.

*Definition* 1.2: Zero anaphora denotes that the preceding antecedents are referred by succeeding ellipses which function as subjects or objects.

*Definition* 1.3: Definite anaphora denotes that the preceding antecedents are referred by succeeding definite noun phrases.

For example, we have the text like "示威群眾$_i$與警察對峙，他們$_i$抗議違建的拆除行動，$\phi_i$並與警察發生衝突。這些人$_i$在衝突中毆打警察$_j$，$\phi_i$更接著搶走了他們$_j$的配槍。"

(The demonstrating people confronted the policemen. They$_i$ protested the dismantling action and had conflicts with the policemen. These people beat up the policemen in the conflict and then took away their$_j$ guns.)

Here “他們$_i$” (they$_i$) and “他們$_j$” (they$_j$) are pronominal anaphors referring to “示威群眾$_i$” (demonstrating people) and “警察$_j$” (the policemen) ,respectively. $\phi_i$ is a zero anaphor referring to “示威群眾$_i$” (demonstrating people) and the definite anaphor “這些人$_i$” (these people) also refers to “示威群眾$_i$” (demonstrating people).

The resolution to the addressed anaphoric expressions is to identify the antecedent of an anaphor in a context [33]. It relies on the employment of the lexical knowledge, context information, and real-world knowledge extracted from both the tackled contexts as well as outer resources. In addition, the presentation and comprehension of anaphora are determined by the connection of shared knowledge or background knowledge between the readers and writers. Thus, effective resolution should be able to infer the relationship between antecedents and anaphors [46]. In fact, effective anaphora resolution facilitates many applications of natural language processing (NLP). It helps the message understanding of a generated summary by a summarizer as well as translated message by a machine translator.

English anaphora resolution has been a research focus in natural language processing for decades [9][11][12][16]. The Anaphora Resolution Exercise[1] (ARE) was organized to develop discourse anaphora resolution systems and evaluate them in a common and consistent environment. The first edition of the Anaphora Resolution Exercise was held in conjunction with the 6th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2007) and focused only on English pronominal anaphora and noun phrase (NP) coreference. In addition, systems such as GuiTAR [41], JavaRAP [42], and MARS [34] are implemented for English anaphora resolution. In contrast to profound studies of English texts, efficient Chinese anaphora resolution has not been widely addressed [8]. Difficulties involved are mainly attributed to the

[1] http://clg.wlv.ac.uk/events/ARE/

following factors. First, morphological clues are rare for determining gender or number of Chinese nouns [48]. Second, no capitalization feature to identify proper nouns. Third, no sufficient ontology, such as WordNet, is available for identifying hypernymy or hyponymy relation between concepts.

Essentially, anaphora resolution can be resolved by using either rules or statistical models [2][3][18][20][22][28]. A rule-based approach is based on a manual-weight salience score which evaluates each candidate. Errors may occur due to intuitive observations and subjective biases in selecting feature weight. Recently, statistical-based or classification-based approaches have been addressed [38][39][45]. Nevertheless, classification-based approaches force different candidates for the same anaphor to be considered independently [13]. Only a single candidate is evaluated at a time and the resolution proceeds in the reverse order of sentences until an antecedent is found. This may cause a real antecedent to be neglected once the classifier labels a candidate to be positive. In addition, the lack of adequate lexical or commonsense knowledge is the other obstacle to achieve accurate resolution results [40][43]. Hand-crafted lexicons are the most common resources for acquiring lexical knowledge, yet it suffers form the coverage problem. So our resolution considers the employment of the web corpus except the existing ontologies.

## 1.2 Research Objectives

In this thesis, pronominal, zero, and definite anaphora in Chinese texts are addressed and different approaches are presented. We use lexical knowledge acquisition and salience measurement to resolve Chinese pronominal anaphora. The lexical knowledge acquisition is aimed to extract more semantic features, such as gender, number, and collocate compatibility by employing multiple resources. The

presented salience measurement is based on entropy-based weighting on selecting antecedent candidates.

As to Chinese zero anaphora, we propose case-based reasoning (CBR) and pattern conceptualization to overcome the difficulties of constructing proper reasoning mechanisms and insufficiency of lexical features. As all cases are represented with the patterns containing semantic tags for their nouns and grammatical tags for the verbs, such pattern conceptualization will be able to efficiently reduce data sparseness in the case base.

We use two strategies to resolve Chinese definite anaphors in written texts. One is an adaptive weight salience measurement for antecedent identification in such a way that the entire set of candidates can be estimated simultaneously. Another scheme is a Web-based knowledge acquisition model to extract useful lexical knowledge so that semantic compatibility extraction and multiple resources can be employed to enhance the resolution performance.

## 1.3 Organization of the Thesis

The remainder of the thesis is organized as follows. In Chapter 2, we investigate the related resolution strategies in anaphora resolution literature. Chapter 3 presents our Chinese pronominal anaphora resolution using lexical knowledge and entropy-based weight in details. Chapter 4 illustrates the method of Chinese zero anaphors resolution using case-based reasoning and pattern conceptualization. In Chapter 5, we describe the Chinese definite anaphora resolution method and the web-based knowledge acquisition model. We conclude and propose future research directions in Chapter 6.

# Chapter 2

# Related Work

In this chapter, we describe a generic anaphora resolution process and investigate different computational strategies used for anaphora resolution. These computational strategies are grouped into two classes: rule-based approaches and statistical or machine learning approaches.

## 2.1 A Generic Anaphora Resolution Process

Anaphora resolution has been considered one of the most challenging problems in NLP. The difficulty of the problem lies in its dependence on sophisticated semantic and world knowledge. Anaphora resolution systems usually aim to resolve anaphors which have noun phrases as their antecedents [33]. A generic anaphora resolution process is described as follows [15]:

Step 1: Select noun phrases to be resolved:

The Selection of noun phrases can rely on linguistic or semantic information, such as NPs that are related to a specific semantic class.

Step 2: Extract features for the selected noun phrases:

Features may be lexical, syntactic, semantic, or other heuristics. Systems can select sophisticated features that require complex NLP tools, or more superficial features acquired through shallow processing.

Step 3: (optional) Determine if the noun phrase is new in the discourse:

A system can include a module for determining whether a NP is anaphoric, before trying to find an antecedent for it. Such modules can be useful when the anaphora resolution model adopted by the system returns an antecedent in

all cases.

Step 4: Create the set of antecedent candidates:

Systems consider as possible antecedents only the NPs that occur before the anaphor in the text. Some systems consider them all, while others impose a maximum number of previous sentences to be considered.

Step 5: (optional) Filter unreasonable candidates:

Some systems exclude candidates that do not conform to some basic constraints, for example number or gender agreement.

Step 6: Score or search antecedent candidates:

This is the core part of an anaphora resolution system. It is the module that interprets the features extracted in Step 2 and determines whether two NPs are anaphorically related based on them. This module can be built by a set of hand-made heuristics or a machine-learning algorithm. Most resolution models rank all antecedents according to a computed score or a set of rules, while other systems search in a particular order for a candidate that conforms to a set of constraints.

## 2.2 Rule-based Approaches

Most traditional approaches are based on hand-crafted rules concerning constraints like syntactic parallelism, semantic parallelism, proximity, or parsing results [2][3][18][20][22][28][32][47].

Hobbs' algorithm is the first syntax-oriented method presented in this research domain [18]. From the result of syntactic tree, they check the number and gender agreement between antecedent candidates and a specified pronoun. The proposed algorithm is based on various kinds of syntactic constraints on pronominal entities

which are used to search the tree. The search is done in an optimal order that performs a breadth-first search of the syntactic tree for an antecedent, accepting the first candidate which meets selectional constraints.

In RAP (Resolution of Anaphora Procedure) proposed by Lappin and Leass [22], the algorithm applies to the syntactic representations generated by McCord's Slot Grammar parser, and relies on salience measures derived from the syntactic structure. It does not make use of semantic information or real world knowledge in choosing among the candidates.

A modified version of RAP system is proposed by Kennedy and Boguraev [20]. The algorithm does not require full syntactic parsing process but has comparable result to the algorithm of Lappin and Leass. It depends only on part-of-speech tagging with a shallow syntactic parse indicating grammatical role of NPs and containment in an adjunct or noun phrase. The method was applied to personal pronouns, reflexives and possessives. The major idea is to construct coreference equivalence classes that have an associated value based on a set of ten factors. An attempt is then employed to resolve every pronoun to one of the previously introduced discourse referents by taking into account the salience value. In addition, CogNIAC (COGnition eNIAC) [2] is a system developed at the University of Pennsylvania to resolve pronouns with limited knowledge and linguistic resources. It presents a high precision pronoun resolution system that is capable of greater than 90% precision with 60% recall for some pronouns.

A knowledge-poor approach is proposed by Mitkov [32] and the approach can also be applied to different languages (English, Polish, and Arabic). The main components of this method are so-called “antecedent indicators” which are used for assigning scores (2, 1, 0, -1) against each candidate noun phrases. They play a

decisive role in tracking down the antecedent from a set of possible candidates. In addition, a set of filtration and evaluation rules are used to resolve anaphora in Chinese financial texts [49]. Another rule-based approach was described in [48] to resolve pronominal anaphora in Chinese texts by using number, gender, grammatical roles, and distance features. To obtain further structured relationship between anaphors and antecedents, Converse [8] used full parsing results from the Penn Chinese Treebank and obtained 77.6% accuracy. Similarly, Yang et al. [53] proposed pronominal resolution using the syntactic information extracted from the parse trees.

Recently, Yeh and Chen [55] presented ZA resolution with partial parsing based on centering theory and obtained 66% F-score in 150 news articles. On the other hand, Converse [8] applied full parsing results but obtained unsatisfactory ZA resolution since only few features were used by the Hobbs algorithm which is originally designed for resolving English anaphora. The main drawbacks of rule-based approaches are attributed to intuitive observations and subjective biases in selecting feature weight. The accuracy is not always guaranteed by heuristics. Moreover, it takes laborious effort to designate new rules whenever the test data vary from original ones.

## 2.3 Statistical or Machine Learning Approaches

Recently, statistical or machine learning techniques have been employed in anaphora resolution [24][38][39][45]. To deal with insufficient knowledge acquired from a given corpus, the World Wide Web has been also widely used as a corpus [4][5][31][36].

A statistical approach is introduced Dagan and Itai [11], they employ the information on a corpus for disambiguating pronouns which is an alternative solution

to the syntactical dependent constraints knowledge. Their experiment makes an attempt to resolve references of the pronoun “it” in sentences randomly selected from the corpus. The model uses a statistical feature of the co-occurence patterns obtained in the corpus to find out the antecedent. The antecedent candidate with the highest frequency in the co-occurence patterns are selected to match the anaphor.

Ge et al. [16] proposed a probabilistic model for resolving third-person pronouns. The model consists of a probability equation, which is initially conditioned on a number of features and is then simplified to handle the sparseness of the training data. This approach consists of decomposing the probability equation for the model by discarding dependencies between features. The decomposition is done by making use of Bayes' rule, the chain rule and certain independence assumptions.

Bunescu [5] present an approach to solving definite descriptions in unrestricted text based on searching the web for a particular type of lexico-syntactic patterns. Using statistics on these patterns, they intend to recover the antecedents for a predefined subset of definite descriptions occurring in two types of anaphoric relations: identity anaphora and associative anaphora. Moreover, Modjeska et al. [36] utilized web search and lexico-syntactic patterns to solve the out-of-vocabulary problem in hand-crafted lexicon. They presented a machine learning approach to other-anaphora, which uses a Naive Bayes classifier and two sets of features. The first set consists of standard morpho-syntactic, recency, and semantic features based on WordNet. The second set also incorporates semantic knowledge obtained from the Web via lexico-syntactic specific to other-anaphora. Adding this knowledge resulted in an improvement of 11.4% points in the classifier’s F-measure, yielding a final F-measure of 56.9%.

In supervised learning, Guan et al. [17] adopt maximum entropy method and sort

the Chinese personal pronoun into two classes referring to personal entity and referring to non-personal entity. The two classes would be treated in different anaphora resolution process called PARS and IARS. This approach can solve the problem of the Chinese personal pronoun referring to inhuman entity effectively.

The traditional learning model for anaphora resolution considers the antecedent candidates of an anaphor in isolation, and thus cannot effectively capture the preference relationships between competing candidates for its learning and resolution. To deal with this problem, a twin-candidate model for anaphora resolution is proposed [54]. The main idea behind the model is to recast anaphora resolution as a preference classification problem. Specifically, the model learns a classifier that determines the preference between competing candidates, and chooses the antecedent of a given anaphor based on the ranking of the candidates.

Ng and Cardie [39] utilized C4.5 decision tree classifier for the task of coreference resolution. Bergsma and Lin [4] presented a SVM-based approach by using general features and path-coreference data which were extracted from a large parsed corpus to compensate for a paucity of data. Such approach successfully resolves 75% of 1078 anaphoric instances in English texts. Zhao and Ng [57] presented a decision tree classification approach to Chinese anaphoric zero pronouns resolution and obtained 43% F-score in 205 texts.

The major drawback of a classification-based approach is that it forces different candidates for the same anaphor to be considered independently since only a single candidate is evaluated at a time. In contrast with a classifier, a ranking approach can directly concern with the entire set of candidates at once and compare different candidate antecedents by assigning salience scores [13].

# Chapter 3

# Pronominal Anaphora Resolution

Pronominal anaphora is commonly used in texts. Pronominal anaphora resolution requires not only morphological and syntactic analysis but also semantic features related to candidate NPs and verbs. In general, pronouns do not carry enough semantic information. This fact forces the use of the semantic information provided by the verbs accompanied by the anaphor and the antecedent. Traditional approaches based on limited knowledge have used morphological agreement and syntactic restrictions in order to reject incompatible candidates [37]. We include the semantic information defining compatibility relations between nouns (subjects and objects) and verbs through collocation patterns in order to be applied in the resolution process.

In this chapter, a hybrid approach using two strategies is presented to resolve pronominal anaphors in Chinese written texts [50]. One is a web-based acquisition model to extract useful lexical knowledge, such as gender, number, and collocate compatibility. Another is an adaptive weight salience measurement for antecedent identification. The experimental results show that our proposed approach yields 82.5% success rate on 1343 anaphoric instances, enhancing 7% improvement while compared with the general rule-based approach presented by Wang and Mei [48].

The subsequent sections of the chapter are organized as follows. Section 3.1 introduces pronominal anaphora in Chinese texts and some of the problems encountered. Section 3.2 describes the proposed method by using lexical knowledge and entropy-based weight in detail. Section 3.3 presents the resolution results and comparisons. Section 3.4 gives a summary of our study.

## 3.1 Chinese Pronominal Anaphora

Pronominal anaphora resolution relies on the constraints between pronouns and antecedents, such as gender, number, grammatical role and animacy. As mentioned above, a general Chinese person's name does not always carry gender information and a Chinese noun does not have morphological differences for indicating its singularity or plurality.

In addition, identifying the referent of a pronoun in Chinese texts is not always trivial if insufficient real-world knowledge is incorporated. Table 3-1 lists two subsequent sentences where each word is followed by its part-of-speech, the first pronoun "他們 $_1$" (they$_1$) refers to "示威群眾" (demonstrating people) while the second pronoun "他們 $_2$" (their$_2$) refers to "警察" (policemen). So it is necessary for an anaphora resolver to check collocate compatibility between anaphors and their candidates.

Table 3-1. A collocate compatibility example.

| |
|---|
| 示威(VA)[2]群眾(Na)與(Caa)警察(Na)對峙(VH)，他們 $_1$(Nh)抗議(VE)違建(Na)的(DE)拆除(VC)行動(Na)，並(Cbb)與(P)警察(Na)發生(VJ)衝突(Na)。這些(Neqa)人(Na)在(P)衝突(Na)中(Ng)毆打(VC)警察(Na)，更(D)接著(D)搶走(VC)了(Di)他們 $_2$(Nh)的(DE)配槍(Na)。 |
| (The demonstrating people confronted the policemen. They$_1$ protested the dismantling action and had conflicts with the policemen. These people beat up the policemen in the conflict and then took away their$_2$ guns.) |

[2] A detailed description of part-of-speech tag set used in this thesis is available at http://ckipsvr.iis.sinica.edu.tw/category_list.doc. For example, "Na" denotes a common noun and "VA" means an intransitive verb.

## 3.2 PA Resolution Framework

Figure 3-1 illustrates the presented pronominal anaphora resolution which is incorporated with three external resources, namely web search results, CKIP lexicon[3], and Wikipedia name profile. The resolution is implemented in the training phase and the testing phase. The training phase involves lexical knowledge acquisition and feature weight learning. Three kinds of lexical knowledge are addressed, namely, gender, number, and collocate compatibility. In feature weight learning, an entropy-based approach is employed. The testing phase concerns text preprocessing, antecedent candidate identification, feature extraction, and antecedent identification. The following subsections describe each component and the resolution procedure.

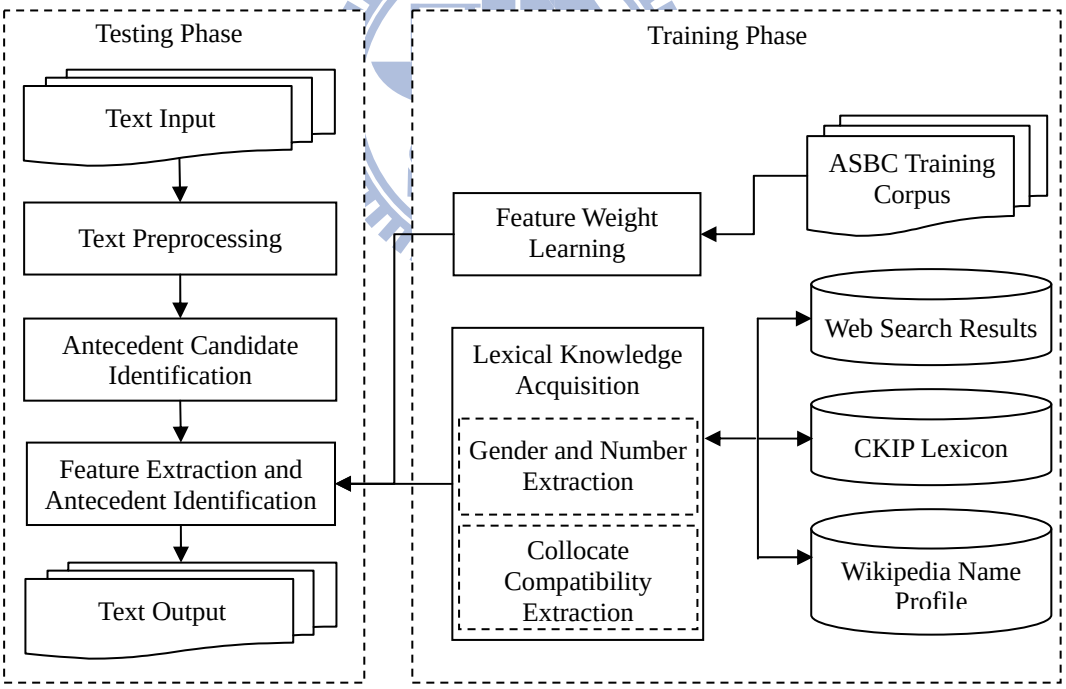



Figure 3-1. The presented Chinese pronominal anaphora resolution procedure.

[3] CKIP (Chinese Knowledge Information Processing Group) lexicon is available at http://www.aclclp.org.tw/use_ckip_c.php

### 3.2.1 Text Preprocessing

Text preprocessing includes sentence segmentation, POS tagging, named entity identification, and noun phrase chunking. The sentence segmentation and POS tagging are processed by CKIP Chinese word segmentation system[4]. The named entity identification is done by applying the hybrid approach presented in [48]. In an experiment of 150 news documents selected from Academia Sinica Balanced Corpus (ASBC)[5], this approach yields 94% precision and 93% recall on person name identification, and 89% precision and 84% recall on organization name identification.

A finite state machine chunker is constructed to recognize noun phrases by their head nouns which may be common nouns, proper nouns, location nouns, temporal nouns, or pronouns [56]. In Chinese, a head noun (as indicated in italics in Table 3-2) occurs at the end of a noun phrase. Except for noun phrase chunks, the chunker is also able to recognize verbal nominalization and transformation by utilizing heuristics described in [14]. As shown in Table 3-2, all the chunks, including the one containing the verb "放鬆" (relax), will be treated as antecedent candidates and will be assigned with semantic feature values like gender, animate and number useful at antecedent candidate identification .

[4] CKIP Chinese word segmentation system is available at http://ckipsvr.iis.sinica.edu.tw/
[5] Academia Sinica Balanced Corpus is available at http://www.sinica.edu.tw/SinicaCorpus/

Table 3-2. Chinese noun phrase examples.

| Types | Noun phrase examples |
|---|---|
| Common noun | 每(Nes)位(Nf)用戶(Na)的(DE)個人(Na)***資料****(Na)*<br>(every subscriber's individual ***information***) |
| Proper noun | 委員會(Nc)主席(Na)***劉生明****(Nb)*<br>(committee chairman ***Liu Shengming***) |
| Location noun | 相當(Dfa)有名(VH)的(DE)***公園****(Nc)*<br>(a very famous ***park***) |
| Temporal noun | 十月(Nd)六日(Nd)***早上****(Nd)*<br>(in the ***morning*** of October 6) |
| Verbal nominalization | 心情(Na)的(DE)***放鬆***(VHC)<br>(the ***release*** of mood) |
| Transformation case | 放鬆(VHC)的(DE)***狀態***(Na)<br>(the relaxed ***condition***) |

### 3.2.2 Antecedent Candidate Identification

Table 3-3 lists the target pronominal anaphors to be resolved in this thesis. Unlike English pronouns, Chinese pronouns remain the same in expressing nominative and accusative cases. Table 3-4 lists the positional distribution of 692 anaphor-antecedent pairs in our training data and it shows that 94% of antecedents are in two sentences ahead of anaphors. Some antecedent candidates can be explicitly filtered out by applying the following heuristics. Here, CAN denotes an item in the

candidate set preceding the corresponding pronominal anaphor (PA). A CAN will be filtered if it satisfies any of the following patterns.

1. Conjunction pattern: PA[c]CAN or CAN[c]PA

   c∈{跟，和，與，同，及，向，對，面對，或，或是，或者，亦或，以及，還是，還有}

2. Verb pattern: PA[Vt]CAN or CAN[Vt] PA

   Vt denotes a transitive verb in a sentence.

3. Preposition pattern: PA[p]CAN or CAN[p]PA

   p∈{在，對，到，朝，給，向，比}

Table 3-3. The target pronominal anaphors.

| | Singular | Plural | Possessive(Singular) | Possessive(Plural) |
|---|---|---|---|---|
| Male | 他(he, him) | 他們(they, them) | 他的(his) | 他們的(their, theirs) |
| Female | 她(she, her) | 她們(they, them) | 她的(her, hers) | 她們的(their, theirs) |
| Neutral | 它(it) | 它們(they, them) | 它的(its) | 它們的(their, theirs) |

Table 3-4. The positional distribution of anaphor-antecedent pairs.

| Relative Position* | (a) | (b) | (c) | (d) |
|---|---|---|---|---|
| Number of pairs | 48 | 367 | 651 | 672 |
| Ratio | 6.9% | 53.0% | 94.0% | 97.1% |

* Relative Position:

(a) Pairs are in the same clause.

(b) Pairs are in the same sentence.

(c) Antecedents are in the previous sentence.

(d) Pairs are in the same paragraph.

### 3.2.3 Lexical Resources

We use two lexical resources to acquire number and gender features for anaphora resolution. One is the CKIP lexicon [1] and out of which we annotated 5,697 nouns with gender and number. For example, “硬漢” (tough guy) and “姑丈” (uncle) are marked as male nouns; “反對黨” (opposition party) and “考察團” (investigation group) are marked as plural. Table 3-5 shows the statistics of the annotated data in the tagged lexicon. The other resource, denoted as “Wikipedia Name Profile”, was constructed by extracting 780 common Chinese person names from Wikipedia[6] and, for each name, the gender and role are tagged by hands. For instance, (“羅大佑” (Luo Da You), “男” (male), “歌手” (singer)) and (“劉墉” (Liu Yong), “男” (male), “作家” (writer)) are two entries stored in the Name Profile.

Table 3-5. Gender and number statistics in the CKIP lexicon.

| | Gender | | | Number | |
|---|---|---|---|---|---|
| Type | Male | Female | Neutral | Singular | Plural |
| Number of entries | 502 | 515 | 4860 | 5345 | 352 |

### 3.2.4 Lexical Knowledge Acquisition

Lexical knowledge acquisition involves the extraction of gender, number, and collocate compatibility from reliable patterns constructed at training phase. In this subsection, we describe detail extraction procedures as follows.

The gender extraction aims to classify each noun phrase to be male, female or

[6] Common Chinese person names are available at http://zh.wikipedia.org/w/index.php?title=%E4%BA%BA%E5%90%8D%E8%A1%A8&variant=zh-tw

unknown with the help of so-called gender-indicating pattern (GP) and Web mining results. All the gender modifiers are mined from the Web in advance by implementing the procedure as shown in Figure 3-2. Moreover, there are six kinds of GPs (denoted as "$GP_i$" and $1 \le i \le 6$) and each GP is utilize to identify the occurrence of masculine pattern or feminine pattern as shown in Figure 3-3.

**Algorithm 3.1. The gender modifier mining algorithm**

**Input:** Randomly select 100 male name $m_i$ and 100 female names $f_i$, respectively

**Output:** Top 5 clue words for male and female, respectively

**Procedure Gmod():**

Step 1: Submit each name to the search engine Google and acquire at most 50 snippets

Step 2: Retain nouns, verbs, adjectives, and adverbs in snippets as set $W=\{w_1,w_2,\ldots, w_n\}$

Step 3: For each $w_i \in W$ do

Calculate $cnt_m$: the frequency that $w_i$ appears with male names

Calculate $cnt_f$: the frequency that $w_i$ appears with female names

Step 4: Select the set $W_m=\{ w_1,w_2,\ldots, w_i \}$, where $\frac{cnt_m}{cnt_f + cnt_m} > 0.8$

Select the set $W_f=\{ w_1,w_2,\ldots, w_j \}$, where $\frac{cnt_f}{cnt_f + cnt_m} > 0.8$

Step 5: Use Bayesian Parameter Learning (Equation (1)) [35] and rank words in the ascending order of $\sigma^2$. The frequencies of words collocating with male names and female names are $\alpha - 1$ and $\beta - 1$, respectively

$$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \quad (1)$$

Step 6: Output top n clue words from $W_m$ and $W_f$, respectively

Figure 3-2. The gender modifier mining algorithm.

**Algorithm 3.2. The gender-indicating pattern identification algorithm**

**Input:** 1. A candidate noun phrase $N_i$

2. The count of masculine patterns $C_m$ and feminine patterns $C_f$

**Output:** The number feature $f_{gnd}$, where $f_{gnd} \in$ {male, female, unknown}

**Procedure Gender():**

Step 1: Search *Attachment titles pattern* ($GP_1$): $N_i$ is followed by a gender-marked title

(a). If $GP_1$ is $N_i$+[先生], then $C_m$++

(b). Else if $GP_1$ is $N_i$+ [女士|小姐|夫人], then $C_f$++

Step 2: Search *Opposite roles pattern* ($GP_2$): $N_i$ acts as a possessive of some specific nouns

(a). If $GP_2$ is $N_i$+[的]+ [太太|妻子|夫人|老婆|女友|未婚妻], then $C_m$ ++

(b). Else if $GP_2$ is $N_i$+[的]+ [先生|丈夫|老公|男友|未婚夫], then $C_f$ ++

Step 3: Search *Reflexives pattern* ($GP_3$): $N_i$ is followed by a reflexive

(a). If $GP_3$ is $N_i$+[他自己], then $C_m$++

(b). Else if $GP_3$ is $N_i$+[她自己], then $C_f$++

Step 4: Search *Possessives pattern* ($GP_4$): $N_i$ is followed by a possessive

(a). If $GP_4$ is $N_i$+[他的], then $C_m$++

(b). Else if $GP_4$ is $N_i$ +[她的], then $C_f$++

Step 5: Search *Complement derivation pattern* ($GP_5$): Person nouns are subjects and gender-marked nouns are in the predicate position. Gender-marked nouns are identified by using the tagged CKIP lexicon

(a). If $GP_5$ is $N_i$+[是]+Modifier+Male-noun, then $C_m$++

(b). Else if $GP_5$ is $N_i$+[是]+Modifier+Female-noun, then $C_f$++

Step 6: Search *Gender-modifier pattern* ($GP_6$): $N_i$ is modified by a gender-modifier which is mined by **Gmod()** as shown in Figure 3-2

(a). If $GP_6$ is gender-modifier +$N_i$ and gender-modifier like "英俊" (handsome), then $C_m$++

(b). Else if $GP_6$ is gender-modifier +$N_i$ and gender-modifier like "溫柔" (tender), then $C_f$++

Step 7: Calculate the feature value $f_{gnd}$ =*Gender*($N_i$)

$$Gender(N_i) = \begin{cases} male, \ if \ \rho_{male} > \rho_{female} \\ female, if \ \rho_{female} < \rho_{male} \\ unknown, otherwise \end{cases} \quad (2)$$

$$\rho_{male} = \frac{C_m}{C_m + C_f}$$

$$\rho_{female} = \frac{C_f}{C_m + C_f}$$

Step 8: Output $f_{gnd}$

Figure 3-3. The gender-indicating pattern identification algorithm.

Figure 3-4 illustrates the overall three-layer gender feature extraction for each $N_i$ of an input document $D_i$ and it is described as follows:

Step 1: If $N_i$ is matched with the tagged CKIP lexicon or Common Name Profile[7], then return the corresponding gender.

Step 2: Else Search $D_i$ with the help of gender-indicating patterns and gender information statistics *Gender($N_i$)* defined in Equation (2). If the gender feature can be decided as male or female, then return the corresponding gender.

Step 3: Else transform $N_i$ to queries according to each kind of *GPs*. For example, "$N_i$ +[先生]", "$N_i$+[他自己]". Search the Web corpus for each gender-indicating pattern and calculate *Gender($N_i$)*. If the gender feature can be decided as male or female, then return the corresponding gender.

Step 4: For other cases, the gender feature is marked unknown.

[7] Common Chinese person names are available at http://zh.wikipedia.org/w/index.php?title=%E4%BA%BA%E5%90%8D%E8%A1%A8&variant=zh-tw

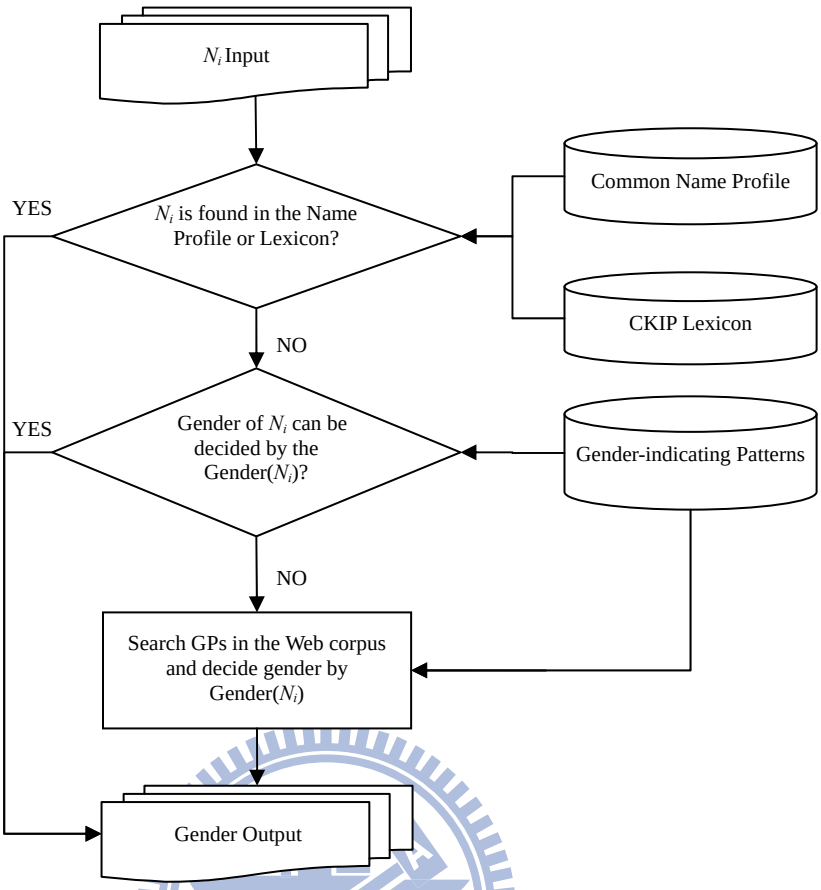



Figure 3-4. The gender extraction procedure.

The number extraction is presented to facilitate resolving plural anaphors. With the collection of numerical and quantitative clue words, the extraction is implemented as shown in Figure 3-5.

**Algorithm 3.3. The number extraction algorithm for assigning the number feature to each candidate noun phrase of input sentences**

**Input:** 1. A candidate noun phrase *NP*
2. The set of quantifiers Q
3. The set of collective quantifiers P={群，夥，堆，對，隊，些，組，伙，雙，疊，批}
4. The set of plural numerals R={全部，所有，數個，許多，有些，少數，少許，多數，諸多，一些，這些，那些，若干}

**Output:** The number feature $f_{num}$, where $f_{num} \in$ {singular, plural, unknown}

**Procedure Number():**

Step 1: Identify head noun *HNP* of the candidate noun phrase *NP*

Step 2: **If** *NP* satisfies any of the following conditions, **then** return $f_{num}$ = singular
(i) *HNP* is a person name
(ii) *NP* contains a title
(iii) *NP*∈{[這|那|該|某|一]+{*Q*-*P*}+noun }

Step 3: **Else if** *NP* satisfies any of the following conditions, **then** return $f_{num}$ = plural
(i) *HNP* is an organization name
(ii) The last character of *NP*∈{們，倆}
(iii) *NP* contains plural numbers+*Q*
(iv) *NP* follows *r,* where *r*∈*R*

Step 4: For other cases, return $f_{num}$ = unknown

Figure 3-5. The number extraction algorithm.

The presented collocate compatibility extraction measures binding strength between candidates and anaphors. We consider three types of collocate patterns, namely subject-verb, verb-object, and possessive-noun, and use collocate statistics to evaluate the preference of candidates. The collocate compatibility extraction is implemented as shown in Figure 3-6.

1. For each pronominal anaphor, replace it with its antecedent candidates accordingly.
2. According to the role (subject or object) of the anaphor in its context, one

collocate pattern is extracted for each candidate.

For instance, consider Table 3-1 mentioned above, anaphors “他們$_1$” and “他們$_2$” are the subject-verb and possessive-noun patterns, respectively. Therefore the collocate patterns for “他們$_1$” are “群眾遊行” and “警察遊行”. Accordingly, “群眾的配槍” and “警察的配槍” are patterns for “他們$_2$”. For each candidate, its collocate compatibility with the anaphor is calculated by Equation (3). In the case of the anaphor “他們$_1$”, three queries are formed for each candidate and they are submitted to Google search engine. For candidate “群眾”, the *pattern query* is “群眾遊行”. Accordingly, the *candidate query* and the *attach query* are “群眾” and “遊行” , respectively.

**Algorithm 3.4. The collocate compatibility extraction algorithm**

**Input:** A candidate noun phrase *can,* a pronominal anaphora *ana*

**Output:** The value of Sem_Com for pair *can* and *ana*

**Procedure Col_Com():**

Step 1: Consider *ana* in the following cases:

Case 1: subject -*verb* //ana is an subject of a verb

Case 2: *verb*-object //ana is an object of a verb

Case 3: possessive-*noun* //ana is a possessive of a noun

Step 2: If Case 1, then

*pattern=can+verb*, *candidate=can*, *attach=verb*

Step 3: Else if Case 2, then

*pattern=verb+ can*, *candidate=can*, *attach=verb*

Step 4: Else if Case 3, then

*pattern=can+ noun*, *candidate=can*, *attach= noun*

Step 5: Acquire the numer of pages $cnt_{pat}$ by submitting *pattern* as query

Acquire the numer of pages $cnt_m$ by submitting *candidate* as query

Acquire the numer of pages $cnt_n$ by submitting *attach* as query

Step 6: Calculate $Col\_Com(can, ana) = \log_2 \frac{p(cnt_{pat})}{p(cnt_m) \times p(cnt_n)}$ (3)

$$p(cnt_{pat}) = \frac{cnt_{pat}}{cnt_{total}}$$

$$p(cnt_m) = \frac{cnt_m}{cnt_{total}}$$

$$p(cnt_n) = \frac{cnt_n}{cnt_{total}}$$

where $cnt_{\text{total}}$ is the number of Google pages

Step 7: Output the value of $Col\_Com(can, ana)$

Figure 3-6. The collocate compatibility extraction algorithm.

### 3.2.5 Feature Set

There are seventeen features concerned at our antecedent identification as follows. *C* denotes an antecedent candidate and *P* denotes the pronominal anaphor. For each feature, we set its value to be 1 if an antecedent candidate satisfies the

feature constraint; otherwise we set its value to be 0.

1. *Same_Pro*: *C* and *P* are the same pronouns, for example, *C* is “她” (she) and *P* is “她” (she) as well.
2. *Reflexive*: *P* is a reflexive of *C*, such as “劉生明他自己” (Liu Shengming himself) in which “劉生明” (Liu Shengming) is an antecedent candidate.
3. *Role*: *C* is the agent of a verb, namely, *C* precedes a transitive verb or an intransitive verb.
4. *Parallel*: *C* and *P* are the same grammatical roles. For example, *C* and *P* are both subjects of sentences.
5. *Gender*: *C* and *P* are the same gender. The gender feature is identified by the way mentioned in the previous subsection *gender extraction*.
6. *Number*: *C* and *P* are the same number. The number feature is determined by the way mentioned in the previous subsection *number extraction*.
7. *Animate*: *C* is an animate entity and *P* is a male or female pronoun. We utilize the semantic class of CKIP lexicon to annotate animate entities. In addition, person names and organization names are regarded as animate entities, too.
8. *NE_Per*: *C* is a person name and *P* is a male or female pronoun. A person name is identified by using a classifier presented in (Liang, Yeh, & Wu, 2003).
9. *NE_Org*: *C* is an organization name and *P* is a plural pronoun. An organization name is identified by the way described above.
10. *Col_Com*: The value of *Col_Com*(*C,P*) is maximum. Equation (3) is used to calculate the value for each antecedent candidate.
11. *Same_Clause*: *C* and *P* are in the same clause. A clause is bounded by

punctuation marks like “，”, “。”, “；”, “！”, and “？”.

12. *Same_Sent*: *C* and *P* are in the same sentence. A sentence is bounded by punctuation marks like “。”, “；”, “！”, and “？”.
13. *Same_Para*: *C* and *P* are in the same paragraph.
14. *Clause_Lead*: *C* is the first noun phrase in the clause.
15. *Sent_Lead*: *C* is the first noun phrase in the sentence.
16. *Repeat*: *C* repeats more than once in the context.
17. *Definite*: *C* is a definite noun phrase. For example, “這本雜誌” (the magazine) is a definite noun phrase.

Table 3-6 shows the feature vectors associated with some antecedent candidates of the anaphor 他們$_2$ (they) in the example of Table 3-1. “警察” (policemen) is selected as the antecedent by applying the weighted salience measurement described in the following subsections.

Table 3-6. Feature vectors of antecedent candidates

| Antecedent candidate | Feature vector |
|---|---|
| “一些群眾” (some people ) | (0,0,1,0,0,1,1,0,0,0,0,1,1,1,1,1,0) |
| “衝突” (conflict) | (0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,1,0) |
| “警察” (policemen) | (0,0,0,1,0,1,1,0,0,1,0,1,1,0,0,1,0) |

### 3.2.6 Entropy-based Weight

The weight function in Equation (4) is motivated from the decision tree learning which utilizes the concept of entropy to select an attribute [35]. The entropy value denotes the uncertainty associated with a random variable. In our case, a feature with

lower entropy denotes that it can reduce uncertainty in selecting correct antecedents. Therefore, a feature with lower entropy is given a higher weight, and vice versa. During the training phase, positive instances were annotated manually. Other candidates between the positive pairs were used to form the negative instances.

$$
\begin{aligned}
&weight_i = 1 - entropy_i(S) \\
&entropy_i(S) = \sum_{j=1}^{v} \frac{|S_j|}{|S|} \times entropy(S_j) \\
&entropy(S) = -\frac{p}{p+n} log_2 \frac{p}{p+n} - \frac{n}{p+n} log_2 \frac{n}{p+n}
\end{aligned}
\tag{4}
$$

where

$S$: the set of training instances

$S_j$: the subset of training instances in which $fval_i$ has value $j$

$p$: the number of positive instances

$n$: the number of negative instances

### 3.2.7 Antecedent Identification

The task of antecedent identification is to select the most likely candidate from the candidate set by Equation (5). Each candidate is filtered by checking its gender, number, and animate agreement. "Agreement$_k$" is a binary function that has a value 0/1. It is noticed that the value of *Rank*(*can*,*ana*) will be set to be zero if one of the three agreements is zero. A candidate with the highest value is selected as the antecedent for the target definite anaphor. The antecedent identification is implemented as shown in Figure 3-7.

**Algorithm 3.4. The antecedent identification algorithm**

**Input:** A document $D$

**Output:** Anaphor-antecedent pairs $(ana, ant)_p$

**Procedure Resolve():**

Step 1: Build the internal representation data structure of input document $D$. For example, sentence offset, word offset, and word POS.

Step 2: **For each** sentence in $D$ **do**

Identify noun phrases in each sentence by the NP chunker described above

Step 3: Identify all target anaphors $ana_p$ throughout $D$

Step 4: **For each** $ana_p$ **do**

Collect candidate set $S$. All antecedent candidates $can_q$ in $S$ are in a distance of two sentences ahead of $ana_p$

**For each** candidate $can_q \in S$ **do**

(i) Assign feature values to $can_q$

(ii) Rank pairs by Equation (5)

$$Rank(can, ana) = \frac{\sum_{i=1}^{n}(fval_i \times weight_i)}{\sum_{j=1}^{n}(\max(fval_j) \times weight_j)} \times \prod_{k=1}^{3} agreement_k \quad (5)$$

where

*can*: a candidate for a specified anaphor

*ana*: an anaphor to be resolved

$fval_i$: the $i^{th}$ feature value

$max(fval_i)$: the maximum value of the $i^{th}$ feature value

$agreement_k$: number, gender, and animate agreement

$weight_i$: the $i^{th}$ feature weight is computed by Equation (4)

(iii) A candidate $can_q$ with the highest Rank value is selected as antecedent $ant_p$ for a definite anaphor

Step 5: Output $(ana, ant)_p$

Figure 3-7. The antecedent identification algorithm.

## 3.3 Experiments

We extract 307 news documents from ASBC as our resolution corpus and from this corpus 1343 anaphor-antecedent pairs are identified by experts. The resolution performance is evaluated in terms of success rate defined by Equation (6) and is implemented by five-fold cross-validation. Figure 3-8 illustrates the entropy-based weight for each feature. It is found that features with top five weights are *Reflexive*, *Animate*, *Col_Com*, *Gender* and *Sent_Lead*, respectively. This result indicates that *Reflexive*, *Animate* and *Gender* features are three dominant features for animate entities if anaphors are gender-marked. In addition, the *Col_Com* feature shows the significance of collocate compatibility in selecting antecedents. *Sent_Lead* justifies the fact that Chinese is a topic prominent language.

$$success\ \ rate = \frac{number\ \ of\ \ correct\ \ resolution\ \ cases}{total\ \ number\ \ of\ \ anaphora\ \ cases\ \ identified} \tag{6}$$

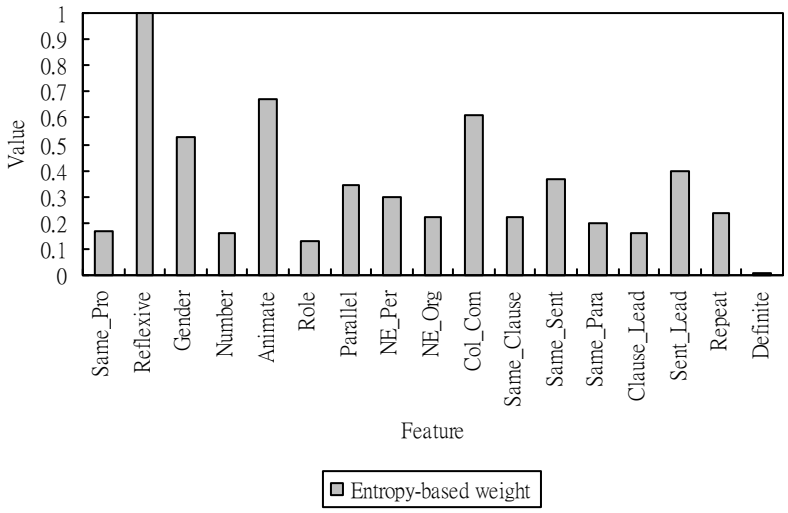



Figure 3-8. The entropy-based weight for each feature.

We implemented five resolution models for comparison. The baseline model was implemented by using number and gender agreement only, and the most recent subject was selected as the antecedent from a candidate set. The second model assigned equal-weight to all seventeen features and selected the top-weight candidate as the antecedent. The third and fourth models were implemented by considering four features only, namely number, gender, grammatical, and distance features. However, the third model assigned the features the same manual weight as described in [48] while the fourth model adopted our presented entropy-based weight. To evaluate how useful the collocate compatibility feature is in our method, we conduct a leave-one-out evaluation in the fifth model. The success rate decreases 4.7% when feature *Col_Com* is disabled.

Table 3-7 shows that our method yields 82.5% success rate on 1343 anaphoric instances by employing entropy-based weight scheme and lexical knowledge. It improves about 7% success rate while compared with a rule-based model like the one presented in [48].

Table 3-8 lists the distribution of each type of anaphors and individual success

rate. It is found that anaphors with gender-mark are more easily to be resolved than the neutral ones. Similar conclusion can be found for those singular anaphors.

Table 3-7. Performance evaluation.

| Method | Success rate |
|---|---|
| Baseline model | 51.6% |
| Rule-based (Equal-weighted) | 72.5% |
| Wang & Mei (2005) | 75.7% |
| Wang & Mei (2005) + entropy weight | 78.2% |
| Our method – *Col_Com* | 77.8% |
| Our method | 82.5% |

Table 3-8. Anaphoric types and their success rate.

| Type of anaphor | 他(的) | 他們(的) | 她(的) | 她們(的) | 它(的) | 它們(的) |
|---|---|---|---|---|---|---|
| # of identified instances | 825 | 207 | 162 | 30 | 88 | 31 |
| # of correctly resolved instances | 697 | 162 | 134 | 24 | 69 | 22 |
| Success rate | 84.4% | 78.2% | 82.7% | 80.0% | 78.4% | 73.3% |

## 3.4 Analysis and Summary

The resolution errors are summarized in Table 3-9. As we can see, most of errors are attributed to preprocessing and gender constraints. Preprocessing errors denote that the system is unable to extract all valid antecedents and incorporate them into the

set of competing candidates. In gender mismatch, one reason is that pronouns “他” (he) and “他們” (they) are often incorrectly used to identify female entities in Chinese texts. There is still room for improvement by carefully taking into account this kind of errors. Gender agreement should be applied more loosely and animate agreement can be enforced. Examples of gender mismatch cases are listed as follows:

(1) 依據傳說以古代那個楊貴妃，**他**皮膚很漂亮很漂亮。

(2) 我那個女兒**他**讀三年級，**他**每次看那個電視，

(3) 以有錢有閒的太太們為多，**他們**吃不多，

Table 3-9. Error analysis of PA.

| Error types | # of error instances | Ratio |
|---|---|---|
| POS tagging/chunking error | 70 | 30% |
| Gender mismatch | 49 | 21% |
| Inappropriate salience | 38 | 16% |
| Exceeding window size | 29 | 12% |
| Number mismatch | 28 | 12% |
| Multiple antecedents | 12 | 5% |
| Others | 9 | 4% |
| Total | 235 | 100% |

Our contributions are that we proposed three innovative methods for lexical knowledge acquisition and our study is the first one that utilizes entropy-based weight in anaphora resolution. Compared with the manual weight scheme, the presented entropy-based weight scheme is more capable to estimate the likelihood that a candidate turns out to be an antecedent. Moreover, the presented lexical knowledge

acquisition is indeed able to acquire more semantic information from contexts and Web resources. In comparison with a general rule-based approach, the presented resolution can achieve 7% improvement when lexical knowledge learning and entropy-based weight are implemented.

# Chapter 4

# Zero Anaphora Resolution

Zero anaphora is the major anaphora occurring in Chinese texts [52] [55]. It means that most of the anaphors appearing in Chinese texts can be unspecified if they are inferable from the contexts. The omitted grammatical constituent is called a zero anaphor (ZA). Zero anaphors may occur in a single sentence or in consecutive sentences. Essentially, the recovery of zero anaphors relies on contextual information, semantic inference, and world knowledge [23][46].

However, efficient Chinese ZA resolution has not been widely addressed. Hence, an effective ZA resolution is presented in this thesis with the aim to facilitate Chinese message understanding. A novel ZA resolution approach is proposed by applying case-based reasoning (CBR) and pattern conceptualization [51]. This is because CBR is able to exploit the previous experience that might be useful for the novel problem. In this thesis, we utilize the antecedent features of the retrieved cases to predict the antecedent of a novel case. As all cases are represented with the patterns containing semantic tags for their nouns and grammatical tags for the verbs, such pattern conceptualization will be able to efficiently reduce data sparseness in the case base. Moreover, the presented resolution is incorporated with a filtering mechanism to identify those non-anaphoric cases such as cataphora and non-antecedent instances in order to enhance the overall resolution performance. The experimental results show that our proposed approach achieved competitive resolution by yielding 79% F-score on 1051 ZA instances and yielded 13% improvement while compared with the general rule-based approach.

The subsequent sections of this chapter are organized as follows. Section 4.1 introduces the commonly-seen zero anaphora instances in Chinese texts. Section 4.2 describes the resolution approach by using CBR-based learning. Section 4.3 describes the procedure of zero anaphora resolution and the experimental results. Section 4.4 presents the final summary.

## 4.1 Chinese Zero Anaphora

According to [19][25], a Chinese sentence is generally integrated by complete syntactic components and expresses an intact meaning. It is composed of one or more clauses and is explicitly identified with punctuation marks like "。, ！, ？". A Chinese clause is an utterance which is identified with punctuation marks like "，, ； , ：, 。, ！, ？" and grammatically it may or may not be a complete syntactic component. As mentioned above, ZA is the most common anaphora displaying in Chinese texts and it can be intra-sentential when a ZA appears in a single-clause sentence or inter-sentential when it appears in multiple-clause sentences. In the following examples, we list some typical ZA and use "$\phi$" to denote zero anaphors which may play as subject or object roles in Chinese sentences and their referents are noun phrases.

(A) Inter-sentential ZA:

1. Subject-role case: The subject (like "Xiaoming" in the example) appears overtly once in the first clause, but later mentions of the same subject are left unspecified in a multiple-clause sentence.

(e1) 小明$_1$ 打開 在 地上 的 箱子， $\phi_1$ 拿出 兩 本 故事書 後， $\phi_1$ 回到 自己 的 房間。

(Xiaoming opened the box on the ground. (Xiaoming) took out two

storybooks. (Xiaoming) went back to his room.)

2. Object-role case: The object (like “new album” in the example) is unspecified in the second clause if it can be understood or inferred from the first clause in a multiple-clause sentence.

(e2) 張三 買 了 新 唱片$_2$，許多 朋友 都 向 他 借 $\phi_2$。

(Zhangsan bought a new album. Many of his friends borrowed (a new album) from him.)

(B) Intra-sentential ZA:

1. Subject-role case: The same subject (like “Lisi” in the example) is unspecified if it is shared from the previous verb in a single-clause sentence with one more verbs.

(e3) 李四$_3$ 參加 演講 比賽 $\phi_3$ 贏得 冠軍。

(Lisi participated in a lecture contest and (Lisi) won the first honor.)

2. Object-to-subject case: The subject (like “Wangwu” in the example) of the second verb is unspecified if it is the object of the first verb in a pivotal sentence.

(e4) 李四 允許 王五 $\phi_4$ 再 重做 一 份 報告。

(Lisi allowed Wangwu (and Wangwu) redo a report again.)

As mentioned previously, a Chinese sentence expresses one complete meaning. However, it is usually observed that a sentence might be incorrectly segmented into a sequence of clauses with punctuations like “，” and some of them are just a noun phrase or a prepositional phrase as shown in the following examples (e5 and ex6). So it is required for a ZA resolver to identify such kind of anaphoric relations in the adjacent clauses for a multiple-clause sentence.

(e5) 總理 斯洛德$_5$， $\phi_5$ 宣布 德國 將 舉行 議會 選舉。

(Premier Schroeder, (Premier Schroeder) declared that Germany will hold a council election.)

(e6) 人 的 生活 空間$_6$，$\phi_6$ 和 自然 環境 發生 了 對立。

(Human living space, (human living space) and environment brought about conflict.)

As mentioned above, the antecedent of a zero anaphor occurs in the previous expressions. However, there are also cases that antecedents are not specified in the previous context, called non-anaphoric zero anaphora (as shown in example (e7)). Therefore, effective zero anaphora resolution relies on not only the identification of antecedents but also the elimination of non-anaphoric cases.

Non-anaphoric zero anaphora case: In this example, $\phi_7$ refers to "time" but the antecedent "time" is not specified previously.

(e7) $\phi_7$ 過 了 兩 天 ，警察 找到 了 犯罪 的 證據。

(After two days, the police found the criminal evidence.)

Table 4-1 lists the positional distribution of 793 anaphor-antecedent pairs in our training data and it shows that 96.7% of antecedents are in a distance of two sentences.

Table 4-1. The positional distribution of anaphor-antecedent pairs.

| Relative Position* | (a) | (b) | (c) |
|---|---|---|---|
| Number of pairs | 710 | 767 | 789 |
| Ratio | 89.5% | 96.7% | 99.4% |

* Relative Position:

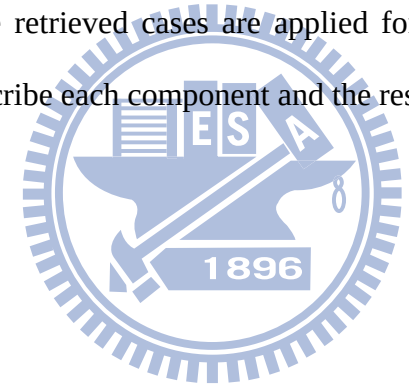
(a) Pairs are in the same complex sentence.

(b) Pairs are in two complex sentences.

(c) Pairs are in the same paragraph.

## 4.2 ZA Resolution Framework

Figure 4-1 illustrates the proposed ZA resolution at the training and testing phases. At training phase, the kernel case-based reasoning module is built in three major steps, namely, feature extraction, pattern conceptualization, and feature weight learning. As a result, a case base, which contains both anaphoric and non-anaphoric ZA cases, is constructed for case retrieval at testing phase. At the testing phase, an input text is processed by a pipeline of text preprocessing, zero anaphor detection, and antecedent (ANT) identification. Moreover, a weighted k-nearest-neighbor (WKNN) algorithm is presented to measure the similarities of cases at case retrieval. The antecedent features of the retrieved cases are applied for antecedent selection. The following subsections describe each component and the resolution procedure in detail.



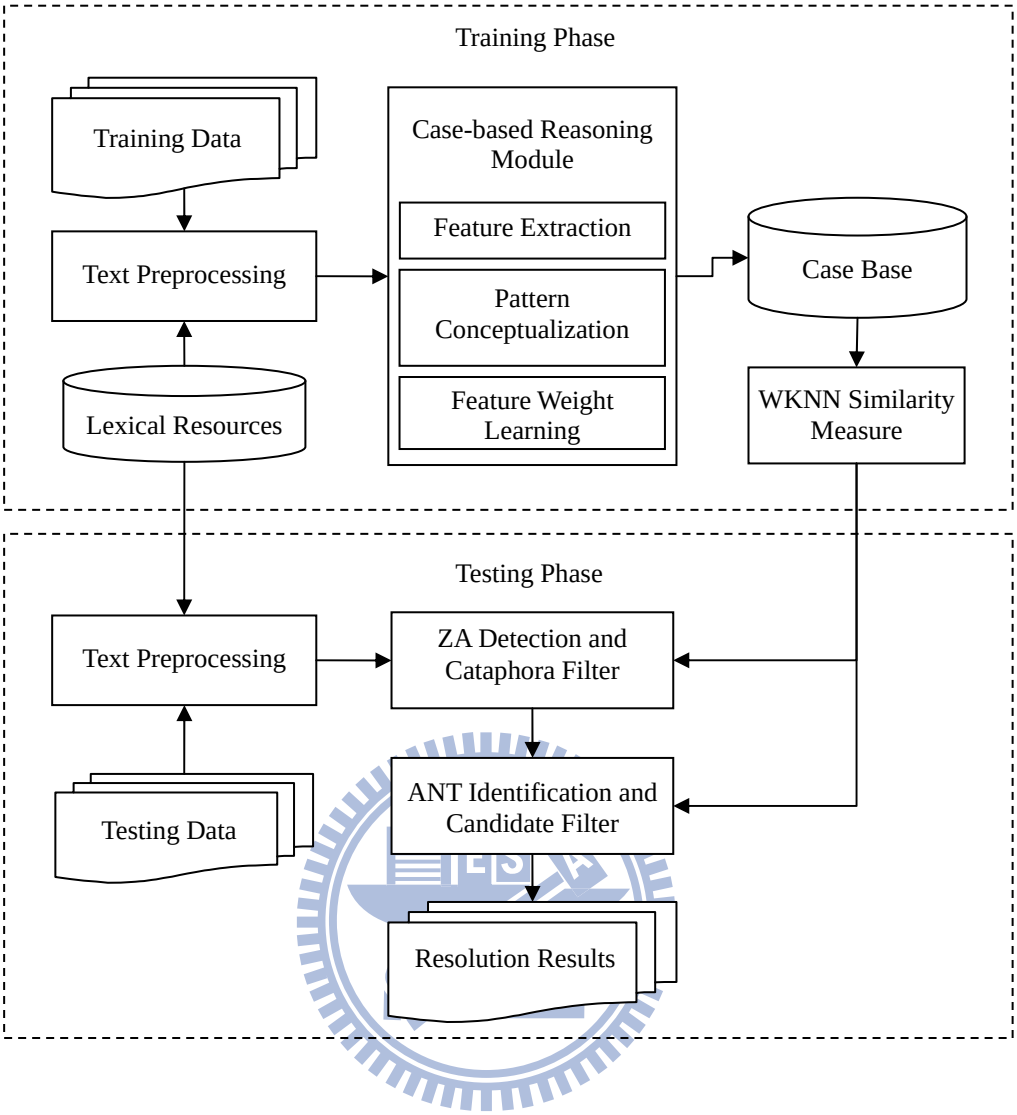



Figure 4-1. The presented Chinese zero anaphora resolution procedure.

### 4.2.1 CBR Approach

CBR is an incremental learning technique that has been successfully used for building knowledge systems and aiding knowledge acquisition [1][7][29]. The main concept of CBR is to exploit the previous experience that might be useful for the novel problem. In this thesis, we utilize the antecedent features of the retrieved cases to predict the antecedent of a novel case. In the case base, those anaphoric cases (treated as positive cases) will be encoded with more features than the non-anaphoric cases (treated as negative cases) and all the cases will be transformed into conceptual

patterns. By measuring the similarity between the novel case and the stored cases, we can check whether a give sentence contains a ZA or not. The most similar case will be reused for antecedent selection if it is a positive case.

For instance, an omission occurs before the verb "宣布" (announce) in the following example (e8). A positive case is extracted from case base, as shown in example (e9), to infer the corresponding antecedent.

(e8) 議員(Na)討論(VE)細節(Na)後(Ng)，ϕ 宣布(VE)明年(Nd)將(D)舉行(VC)大選(Na)。

(After discussing the details, (Councilor) announced that there will be an election next year.)

(e9) 主席(Na)整理(VC)意見(Na)後(Ng)，ϕ 決定(VE)明天(Nd)表決(VE)修正案(Na)。

(After collecting opinions, (Chairman) decided that the amendment will be decided by vote tomorrow.)

### 4.2.2 Outer Lexical Resources

Two outer resources are used to acquire informative features such as semantic classes of nouns and verbs during ZA resolution. The resources used are CKIP lexicon [6] and the Academia Sinica Bilingual WordNet (SinicaBOW)[8]. There are four kinds of verbs regarded as animate verbs; namely, {cognition}, {communication}, {emotion}, and {social}. CKIP lexicon contains 80,000 entries annotated with syntactic categories and corresponding semantic classes. There are 8 semantic classes selected from CKIP lexicon. During processing of noun phrases, head nouns of noun phrases are tagged with semantic classes. The classes can be divided into physical

[8] SinicaBOW is a Mandarin-English bilingual database based on the framework of English WordNet and language usage in Taiwan. A detailed description is available at http://bow.sinica.edu.tw/

entities and nonphysical entities as listed in Table 4-2.

Table 4-2. Semantic classes selected from CKIP lexicon.

| Entity | Semantic Classes |
|---|---|
| Physical | Mankind, Places, Artifacts, and Matter |
| Nonphysical | Events, Temporal, Principles, and Mental |

### 4.2.3 Feature Extraction

A case for example (e9) in the case base is represented in the form as shown in Table 4-3. It contains both ZA template and ANT template used as a ZA resolution method. During the training phase, the case base contains examples collected from the training corpus and annotated with ZA markers (denoted as "$\phi$") by human experts. Table 4-4 shows an input test case in which $\phi$ occurs before the verb "宣布" (announce). A detail description of features for ZA template and ANT template is shown in Table 4-5.

Table 4-3. Case representation in the case base.

| | |
|---|---|
| Content of a case in the case base | 主席(Na)整理(VC)意見(Na)後(Ng)，$\phi$ 決定(VE)明天(Nd)表決(VE)修正案(Na)。<br>(After collecting opinions, (Chairman) decided that the amendment will be decided by vote tomorrow.) |
| Implementation level<br>(ZA template) | PRE_NPS: N<br>ROLE: subject<br>POS: VE<br>FIRST_VERB: Y<br>CLASS_VERB: report<br>SEN_DIST: 2<br>PRE_VERB: Y<br>PRE_PREP: N<br>PRE_CONJ: N<br>PRE_ZA: N<br>CON_PAT: $\phi$ (VE)[temporal](VE)[events] |
| Implementation level<br>(ANT template) | TOPIC: Y<br>ROLE: subject<br>NUM: singular<br>GND: neutral<br>POS_HEAD: Na<br>NE: N<br>DEF: N<br>EMB: N |

| | |
|---|---|
| | CLASS: mankind<br>SEN_DIST: 1<br>OFFSET: 1<br>CUR_ZA: N<br>RPT: N<br>CON_PAT: [mankind](VC)[mental] |

Table 4-4. Input case representation.

| | |
|---|---|
| Content of an input case | 議員(Na)討論(VE)細節(Na)後(Ng)，ϕ 宣布(VE)明年(Nd)將(D)舉行(VC)大選(Na)。<br>(After discussing the details, (Councilor) announced that there will be an election next year.) |
| Implementation level<br>(ZA template) | PRE_NPS: N<br>ROLE: subject<br>POS: VE<br>FIRST_VERB: Y<br>CLASS_VERB: report<br>SEN_DIST: 2<br>PRE_VERB: Y<br>PRE_PREP: N<br>PRE_CONJ: N<br>PRE_ZA : N<br>CON_PAT: ϕ (VE)[temporal](VC)[events] |

Table 4-5. Description of template features.

| | Feature | Description |
|---|---|---|
| ZA template | PRE_NPS | If the preceding clause is a noun phrase then Y; else N. |
| | ROLE | Grammatical role of the ZA: subject, object, or other. |
| | POS | Part-of-Speech of the related verb. |
| | FIRST_VERB | If the related verb is the first one then Y; else N. |
| | CLASS_VERB | Semantic class of the related verb. |
| | SEN_DIST | The ZA occurs in the i-th clause of a complex sentence. |
| | PRE_VERB | If the ZA is followed by a verb then Y; else N. |
| | PRE_PREP | If the ZA is followed by a preposition then Y; else N. |
| | PRE_CONJ | If the ZA is followed by a conjunction then Y; else N. |
| | PRE_ZA | If a ZA occurs in the preceding clause then Y; else N. |
| | CON_PAT | The conceptual pattern of a sentence in which a ZA occurs. |
| ANT template | TOPIC | If the ANT is the first noun phrase of a complex sentence then Y; else N. |
| | ROLE | Grammatical role of the ANT: subject, object, or other. |
| | NUM | Single, plural, or unknown. |
| | GND | Male, female, neutral, or unknown. |
| | POS_HEAD | Part-of-Speech of the ANT head noun. |
| | NE | The ANT is a person name or an organization name. |
| | DEF | If the ANT is a definite noun phrase then Y; else N. |
| | EMB | If the ANT is a embedded noun phrase then Y; else N. |
| | CLASS | Semantic class of the ANT. |
| | SEN_DIST | The ANT occurs in the i-th clause of a sentence. |
| | OFFSET | Distance between the ANT and the ZA in terms of clauses. |

| | | |
|---|---|---|
| | CUR_ZA | If a ZA occurs in the current clause then Y; else N. |
| | RPT | If the ANT repeats more than once then Y; else N. |
| | CON_PAT | The conceptual pattern of a sentence in which the ANT occurs. |

### 4.2.4 Pattern Conceptualization

The ZA template contains ten features as well as its conceptual pattern. Conceptual patterns are utilized to measure the similarity between sentences in which zero anaphors occur. Each sentence is expressed as a pattern composed of semantic classes of nouns and grammatical categories of verbs. In the following examples, sentences (e10) and (e11), the corresponding conceptual patterns are represented. Each field bracketed by [ ] or ( ) indicates an item in the conceptual pattern.

(e10) $\phi$ 宣布(VE)明年(Nd)將(D)舉行(VC)大選(Na)。

((Councilor) announced that there will be an election next year.)

*Concept pattern representation*: $\phi$ (VE) [temporal] (VC) [events]

(e11) $\phi$ 決定(VE)明天(Nd)表決(VE)修正案(Na)。

((Chairman) decided that the amendment will be decided by vote tomorrow.)

*Conceptual pattern representation:* $\phi$ (VE) [temporal] (VE) [events]

Similarity between the input test sentence with a ZA and the case sentence with a ZA in the case base is described in Equation (7). For a given input test sentence *I* and a case sentence *C*, *CPSIM(I,C)* calculates the similarity value of sentences in which zero anaphors occur. For example, the similarity value of examples (e10) and (e11) is given as: $(2\times 4)/(5+5) = 0.8$.

$$CPSIM(I, C) = \frac{2 \times LENLCS(I, C)}{(LEN(I) + LEN(C))} \quad (7)$$

where

*I*: the input test sentence with a ZA

*C*: the case sentence with a ZA in case base

*LENLCS*(*I,C*): number of items in the longest common subsequence of *I* and *C*

*LEN*(*I*): number of items in *I*

*LEN*(*C*): number of items in *C*

### 4.2.5 ZA Detection

During the ZA detection phase, verbs in sentences are examined sequentially. If there is any omission of subjects or objects with respect to a verb, ZA detection will submit the sentence to reasoning module to decide whether there is a ZA. If there is any positive case retrieved, the ANT identification phase will be performed using the resolution template returned from the case base. If the retrieved case belongs to negative one, then the case is regarded as a non-anaphoric instance.

We must be mindful of the cataphora cases that may be mistakenly treated as a ZA. So we observe the following properties which can be utilized by our cataphora filter.

1. It often occurs after verbs tagged with VE.
2. There is no patient after the verb.
3. It occurs frequently in the first clause of a complex sentence.
4. The related verbs are followed by punctuation marks like “，” and “：”.
5. It often refers to the succeeding description rather than noun phrases.

The cataphora filter algorithm is shown as follows:

Step 1: For a ZA candidate, we define symbols as follows:

V = the verb preceding the ZA candidate;

VE = the set of reporting verbs;

W= the set of any words;

M=W-{nouns and verbs};

Step 2: If V∈{VE} and all the following conditions are satisfied, then return cataphora;

i. sentence pattern=[W*VM+，|W*VM+：];

ii. a ZA candidate occurs in the first clause of a complex sentence;

Step 3: For other cases, return ZA.

Moreover, it must be noted that the following conditions will not be considered while detecting ZAs around verbs [30] [55]. The conditions are described as follows:

(e12) “把” (Ba) sentence:

張三(Nb)已經(D)**把**(P)工作(Na)完成(VC)。

(Zhangsan has made the work done.)

(e13) “被” (Bei) sentence:

工作(Na)已經(D)**被**(P)張三(Nb)完成(VC)。

(The work has been finished by Zhangsan.)

(e14) In an adverbial case: when the verb functions as a part of an adverb, it is not the verb related to a ZA.

### 4.2.6 Antecedent Identification

Equation (9) is the similarity function used to compute the similarity between the input case and the stored case examples. The similarity computation concerns the similarity between ZA template features and conceptual patterns as described above. Subsequently, the ANT template with the highest similarity value is retrieved from the case base and used to identify the antecedent with respect to a given test case. For instance, to identify the antecedent of $\phi$ as shown in Table 4-4, the most similar case, shown in Table 4-3, is extracted by Equation (8). According to the ANT template in, "議員" (councilor) is selected as the antecedent because it matches the most features than other candidates such as "細節" (details).

We conduct a weighted k-nearest-neighbor algorithm in the case retrieval phase. The case retrieval phase captures the most similar case in the case base and employs the antecedent features to resolve the test case. The process is shown as follows:

1. Calculate the weight $w_{f_i}$ of each feature $f_i$ by Equation (8).

$$w_{f_i} = \inf(S) - \inf_{f_i}(S) \tag{8}$$

$$\inf(S) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

$$\inf_{f_i}(S) = \sum_{j=1}^{v}\frac{|S_j|}{|S|}\times\inf(S_j)$$

where

$f_i$: the $i^{th}$ feature with $v$ distinct values

$S$: the set of cases in the case base

$S_j$: the subset of $S$ for which feature $f_i$ has value $j$

$p$: the number of cases belonging to positive ones

$n$: the number of cases belonging to negative ones

2. Calculate the similarity $SIM(I,C)$ of the test case $I$ and each case $C$ in the case base by Equation (9).

$$SIM(I,C) = \frac{\sum_{i=1}^{|f|} w_{f_i} \times match(I_{f_i}, C_{f_i}) \times \alpha}{\sum_{i=1}^{|f|} w_{f_i}} + CPSIM(I,C) \times \beta \qquad (9)$$

where

$|f|$: the number of test case features $f$

$w_{f_i}$: the weight of the $i^{th}$ feature in $f$

$I_{f_i}$: the value of feature $f_i$ of the test case

$C_{f_i}$: the value of feature $f_i$ of the case in the case base

$match(I_{f_i}, C_{f_i})$: returns 1 if feature value of $I_{f_i}$ and $C_{f_i}$ are equal;

otherwise returns 0

$CPSIM(I,C)$: conceptual pattern similarity as shown in Equation (7)

$\alpha, \beta$: weighting factors where $\alpha + \beta = 1$

3. Retrieve k cases with the highest similarity value.

4. Let k retrieved cases vote on the antecedent features as a solution for the test case.

### 4.2.7 Centering Theory in ZA Resolution

Centering theory (CT) models local coherence and was initially developed to relate the focus of attention. In recent years, centering theory has been applied to the problem of anaphora resolution [10][21][55]. Two types of center are defined for each utterance $U_n$: a unique backward-looking center $C_b$ and a set of forward-looking center $C_f$. The most highly ranked entity in the resulting $C_f$ list is called the preferred center $C_p$. The ranking criterion is shown as follows:

Topic > Subject > Direct Object > Others

For comparison with our proposed method, a CT-based approach is implemented [55]. The basic idea is shown as follows:

1. $C_p$ of the previous clause $U_{i-1}$ is selected as the antecedent of the $ZA_i$ in the current clause $U_i$.
2. If there is one $ZA_{i-1}$ occurs in $U_{i-1}$ then select the antecedent of the $ZA_{i-1}$ in $U_{i-1}$ as the antecedent of $ZA_i$.
3. If more than one ZA occurs in $U_{i-1}$ then select the antecedent of the ZA in $U_{i-1}$ according to the forward-looking center ranking criterion.

## 4.3 Experiments

Our resolution is justified by 382 narrative report articles selected from ASBC corpus. In experimental evaluation, five-fold cross-validation was conducted over the

selected data set. The positive and negative zero anaphora cases were annotated manually by domain experts. There are 5255 cases in total, which contains 3217 anaphoric cases and 2038 non-anaphoric cases respectively. During the testing phase, CKIP Chinese word segmentation system[9] is utilized for tagging POS.

Table 4-6 lists the statistical data regarding both the training and the testing corpora. Table 4-7 lists the results in terms of precision and recall at various matching thresholds. It is observed that optimal performance (in terms of F-score) is achieved when the $\alpha$ and $\beta$ values are 0.7 and 0.3, respectively. Moreover, we employ the presented weighted k-nearest-neighbor algorithm during resolution. According the result shown in Figure 4-2, the value of k is set to be 3 in our experiments.

Table 4-6. Statistical information of evaluation data.

| | Data set |
|---|---|
| Articles | 382 |
| Sentences | 12,775 |
| Words | 126,119 |
| Zero Anaphors | 5,255 |

Table 4-7. Performance at various thresholds.

| Threshold $\alpha$ | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | **0.7** | 0.8 | 0.9 | 1.0 |
|---|---|---|---|---|---|---|---|---|---|---|
| Recall | 0.42 | 0.48 | 0.55 | 0.59 | 0.68 | 0.74 | 0.78 | 0.75 | 0.71 | 0.66 |
| Precision | 0.41 | 0.47 | 0.49 | 0.61 | 0.69 | 0.75 | 0.79 | 0.78 | 0.72 | 0.67 |

[9] CKIP Chinese word segmentation system is available at http://ckipsvr.iis.sinica.edu.tw/

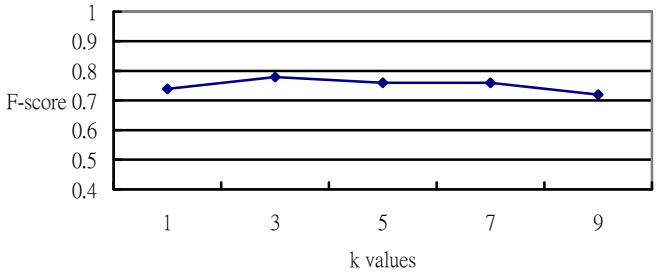



Figure 4-2. F-score over different k values.

In order to verify the impact of the extracted features, a baseline model is built in such a way that only grammatical features are used in ANT identification. Figure 4-3 shows that the highest F-score is obtained when all the ZA template features and conceptual patterns (denoted as “ALL”) are concerned and the baseline yields the worst performance by comparison. Additionally, the resolution performance can be enhanced significantly by applying semantic class features (denoted as “A”) and conceptual pattern mapping (denoted as “B”). We verify the sensitivity of training case size in our presented CBR approach for resolving zero anaphora. It is found from Figure 4-4 that feasible performance results can be obtained when the training corpus is two times the size of the testing corpus. If the training case size is half of the testing case size, performance may decrease by 25%. Besides, centering theory (CT) is applied as the frame work to resolve zero anaphora for comparison. Since only grammatical roles and constraints are major criteria used for resolution, the performance is not satisfactory. In addition, numerous errors are caused due to misjudgment of verbal nominalization (VN) and lack of cataphora filter (CF). Table 4-8 illustrates that the performance is indeed improved if VN and CF are incorporated in resolution. The result indicates that our proposed method significantly outperforms the CT approach by 13%.

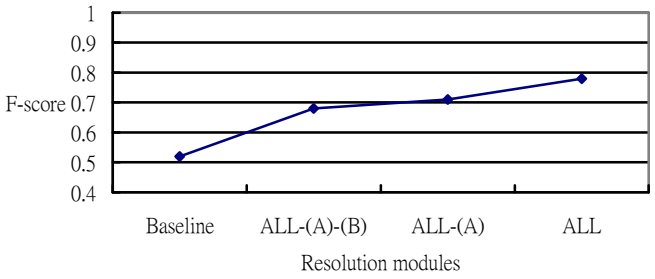



Figure 4-3. F-score after applying resolution modules.

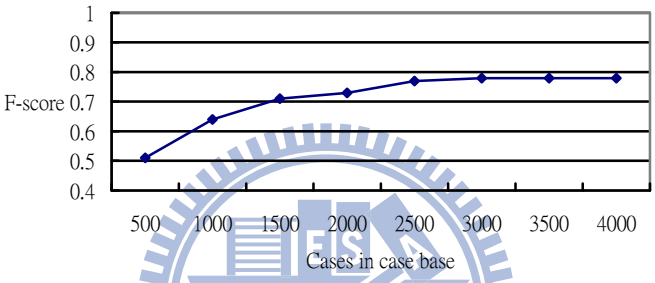



Figure 4-4. F-score over different case base scale.

Table 4-8. Performance evaluation with different methods.

| Method | F-score |
|---|---|
| CT | 66% |
| CT+VN | 69% |
| CT+VN+CF | 71% |
| Our method | 79% |

## 4.4 Analysis and Summary

A summary of errors of our proposed method is listed in Table 4-9. Seven types of errors are listed and the proportion of each error is calculated. Besides preprocessing errors, semantic class mismatch and inappropriate ANT template are two major errors. Semantic class mismatch denote that the system cannot identify the semantic class of antecedent candidates, for example, new words or words which are not included in lexicon. To overcome this kind of shortcoming, it takes more sophisticated method for word sense disambiguation. The error of inappropriate ANT temple comes from the property of case-based reasoning. Once no similar cases can be retrieved from case base, it may lead to error candidate selection.

Table 4-9. Error analysis of ZA.

| Error types | Ratio |
|---|---|
| POS tagging/chunking error | 20% |
| Semantic class mismatch | 16% |
| Inappropriate ANT template | 14% |
| Exceeding window size | 13% |
| Number mismatch | 11% |
| Gender mismatch | 10% |
| Multiple antecedents | 9% |
| Others | 7% |
| Total | 100% |

In this thesis, we present a case-based reasoning approach to Chinese zero anaphora resolution. Compared with rule-based resolution methods, the presented

approach turns out to be promising for dealing with both intra-sentential and inter-sentential zero anaphora. The contributions of our work are revealed from two aspects. First, a case-based reasoning approach with weighted KNN retrieval is demonstrated to be an effective method in comparison with the state-of-the-art rule-based approach. Second, we introduced two new features, semantic classes acquired from outer resources and conceptual patterns, for both ZA detection and ANT identification. Experimental results show that these two features can improve overall resolution performance by 11%. The drawback to this approach is that a case base must be constructed in advance. However, our experimental analysis shows that feasible performance results can be obtained when the training corpus is two times the size of the testing corpus.

# Chapter 5

# Definite Anaphora Resolution

In this chapter, a novel approach using two strategies is presented to resolve Chinese definite anaphors in written texts. One is an adaptive weight salience measurement for antecedent identification. A weighted ranker is utilized to estimate the entire set of candidates simultaneously. Another is a Web-based knowledge acquisition model to extract useful lexical knowledge, such as gender, number, and semantic compatibility. The experimental results show that our proposed approach yields 72.5% success rate on 426 anaphoric instances, enhancing 4.7% improvement while compared with the result conducted by a conventional classifier.

The subsequence of the chapter is organized as follows. Section 5.1 introduces the commonly-seen definite anaphora instances in Chinese texts. Section 5.2 describes the proposed method by using feature weight learning and lexical knowledge acquisition in detail. Section 5.3 describes the experimental results and analysis. Section 5.4 presents the final summary.

## 5.1 Chinese Definite Anaphora

Definite noun phrase anaphora occurs in the situation that the antecedent is referred by a general concept entity. The general concept entity can be a semantically close phrase such as synonyms or hypernyms of the antecedent [33]. A definite noun phrase (the man, the car, the speaker) indicates that it is possible for the reader to uniquely identify the referent of the noun phrase, whereas an indefinite noun phrase (a man, a friend of mine, someone) indicates that the referent is not uniquely identifiable. The difference between “the animal” and “an animal” is a difference in definiteness. A

definite noun phrase is used when a writer expects readers to be able to pick out the referent for the noun phrase.

In Chinese definite anaphora (DA), an antecedent can be mentioned by a definite noun phrase preceded by demonstratives like “這” (this), “此” (this), “那” (that), “其” (that). Similarly, an English definite noun phrase is introduced by a definite article “the”. In this thesis, we tackle Chinese definite anaphor with the pattern like “[這 (this)] + [量詞 (quantifier)] + [實體名詞片語 (physical noun phrase)]”. Grammatically, definiteness is a feature of noun phrases, indicating entities which are specific and identifiable in a given context. The type of DA may be partial overlap relation as in example (1), synonymous relation as in example (2), or hyponymy-hypernymy relation as in example (3).

(1). Partial overlap relation:

波灣戰爭$_1$於 1991 年由美國主導進軍伊拉克，這場戰爭$_1$對國際政治與經濟造成巨大的影響。

Gulf War$_1$ was led by the United States to attack Iraq in 1991. The war$_1$ caused huge impact on international politics and economics.

(2). Synonymous relation:

上周鄰居家遭到小偷$_2$侵入，警方推測這竊賊$_2$可能是經由窗戶進入屋內。

Thieves$_2$ intruded into neighbor's house last week. The police thought that the burglars$_2$ probably enter the house through the windows.

(3). Hyponymy-hypernymy relation:

帕金森氏症$_3$是由於腦神經細胞退化所造成，在醫學上相信這種疾病$_3$與多巴胺有密切關係。

Parkinson's disease$_3$ is caused by the degradation of brain cells. Medically, it is believed that this disease$_3$ is closely related to dopamine.

Anaphoric relation in (1) can be resolved by matching the head nouns of noun phrases explicitly. As to the other two cases, surface features are no longer adequate to identify the correct antecedents. Most previous studies rely on pre-constructed lexicons as knowledge sources. However, it suffers from the problem of coverage. Besides, no sophisticated lexicon is available yet for identifying relation between Chinese expressions as shown in (3). Thus, we utilize a Web-based approach for exploiting semantic relationships such as hyponymy and hypernymy that are not included in lexicon resources.

In addition, we investigate the positional distribution of 618 anaphor-antecedent pairs in our training data. Table 5-1 shows that 93% of antecedents are in two sentences ahead of the definite anaphors.

Table 5-1. The positional distribution of anaphor-antecedent pairs.

| Relative Position* | (a) | (b) | (c) | (d) |
|---|---|---|---|---|
| Number of pairs | 223 | 433 | 575 | 585 |
| Ratio | 36.0% | 70.0% | 93.0% | 94.6% |

* Relative Position:

(a) Antecedents are in the same sentence.

(b) Antecedents are in the previous sentence.

(c) Antecedents are in the two previous sentences.

(d) Antecedents are in the same paragraph.

## 5.2 DA Resolution Framework

Figure 5-1 illustrates the presented definite anaphora resolution which is incorporated with three external resources, namely Web search results, CKIP lexicon,

and Tongyici Cilin[10]. The resolution is implemented in the training phase and the testing phase. The training phase involves feature weight learning and lexical knowledge acquisition. Three kinds of lexical knowledge are addressed, namely, gender, number, and semantic compatibility. In feature weight learning, an entropy-based approach is employed. The testing phase concerns text preprocessing, antecedent candidate identification, feature extraction, and antecedent identification. The following subsections describe each component and the resolution procedure.

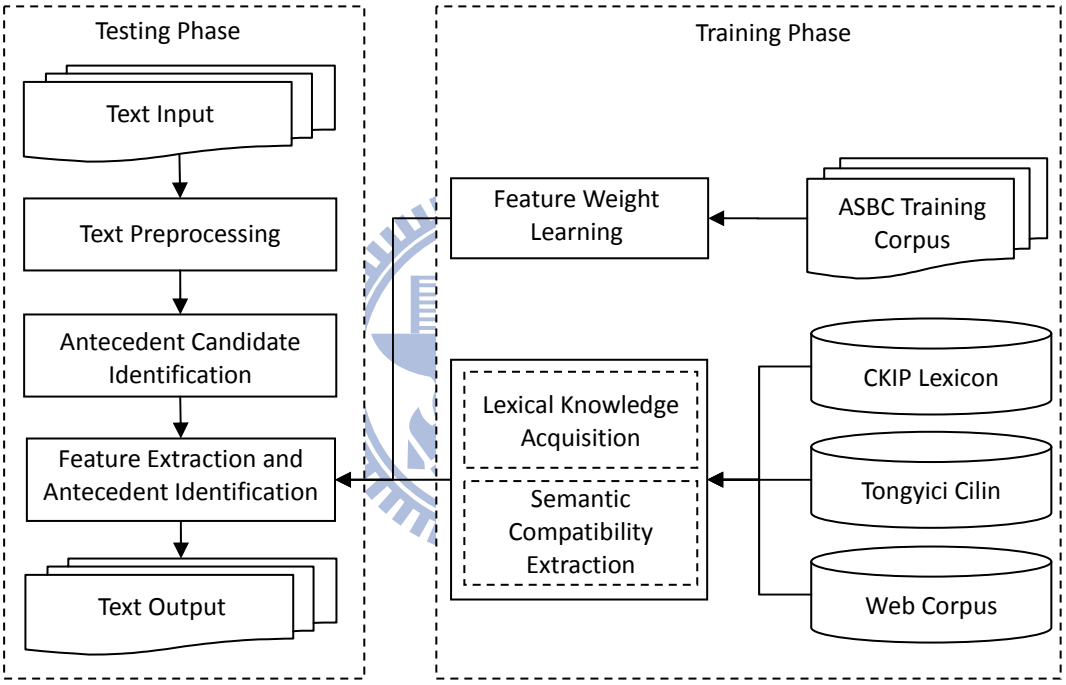



Figure 5-1. The system architecture.

### 5.2.1 Feature Set

There are fifteen features concerned as shown in Table 5-2. *can* denotes an antecedent candidate and *ana* denotes the definite anaphor. For each feature, we set its value to be 1 if an antecedent candidate satisfies the feature constraint; otherwise we

[10] Tongyici Cilin extended version is available at http://ir.hit.edu.cn/demo/ltp/Sharing_Plan.htm

set its value to be 0.

Table 5-2. Summary of features.

| Type | Feature | Description |
|---|---|---|
| Lexical | *Head_Match* | *can* and *ana* have the same head word. |
| | *Str_Overlap* | *can* and *ana* have overlapping words. |
| | *Non_Emb* | *can* is not an embedded noun phrase. |
| | *Definite* | *can* follows a determiner. |
| Grammatical | *Gender* | *can* and *ana* are the same gender. |
| | *Number* | *can* and *ana* are the same number. |
| | *Role* | *can* and *ana* are the same grammatical role. |
| Semantic | *Cilin_Syn* | *can* and *ana* are synonyms in Tongyici Cilin. |
| | *Animate* | *can* and *ana* are both animate entities. |
| | *Sem_Com* | The value of *Sem_Com*(*can, ana*) is maximum. |
| | *Same_SC* | *can* and *ana* are the same semantic class in CKIP lexicon. |
| Heuristic | *Coh_Cue* | *can* and *ana* are connected by coherence cue words. |
| | *Repeat* | *can* repeats more than once in the context. |
| | *Sent_Lead* | *can* is the first noun phrase in the sentence. |
| | *Fwd_Cent* | *can* is a forward looking center. |

### 5.2.2 Semantic Compatibility Extraction

To acquire semantic knowledge from the Web, we submit queries consisted of candidates and anaphors to the Google search engine. Queries are formed by patterns that structurally express the same semantic relationships. The co-occurrence statistics

of such patterns can then be used as a mechanism for detecting the hypernymy-hyponymy relation between the definite anaphor and its potential antecedents. In the case of a candidate "蘋果 (apple)" and the definite anaphor "這種水果 (this kind of fruit)", queries like < "蘋果是一種 (apple is a kind of )"+"水果 (fruit)" >, < "蘋果這種 (apple this kind of)"+"水果 (fruit)" >, and < "蘋果和其他 (apple and other)"+"水果 (fruit)" > are concerned and the implementation is shown as Figure 5-2.

**Algorithm 5.1. The semantic compatibility extraction algorithm for mining hypernymy and hyponymy relations**

**Input:** A candidate noun phrase *can,* a definite anaphora *ana*

**Output:** The value of Sem_Com for pair *can* and *ana*

**Procedure Sem_Com():**

Step 1: Identify the head noun of *can* as *m*
Identify the head noun of *ana* as *n*

Step 2: Submit query [*m*]+[是一種]+[*n*] to Google
Calculate the number of pages $cnt_{q1}$

Step 3: Submit query [*m*]+[這種]+[*n*] to Google
Calculate the number of pages $cnt_{q2}$

Step 4: Submit query [*m*]+[和其他]+[*n*] to Google
Calculate the number of pages $cnt_{q3}$

Step 5: Acquire the numer of pages $cnt_m$ by submitting *m* as query
Acquire the numer of pages $cnt_n$ by submitting *n* as query

Step 6: Calculate $Sem\_Com(can, ana) = log_2 \frac{p(cnt_{pair})}{p(cnt_m) \times p(cnt_n)}$ (10)

$$p(cnt_{pair}) = \frac{(cnt_{q1} + cnt_{q2} + cnt_{q3})}{cnt_{total}}$$

$$p(cnt_m) = \frac{cnt_m}{cnt_{total}}$$

$$p(cnt_n) = \frac{cnt_n}{cnt_{total}}$$

where $cnt_{total}$ is the number of Google pages

Step 7: Output the value of $Sem\_Com(can, ana)$

Figure 5-2. The semantic compatibility extraction algorithm.

### 5.2.3 Feature Weight Learning

The entropy value denotes the uncertainty associated with a random variable. In our case, a feature with lower entropy denotes that it can reduce uncertainty in selecting correct antecedents. Therefore, a feature with lower entropy is given a higher weight, and vice versa. In the training phase, 318 news documents containing 618 positive and 1077 negative pairs are used as training data. Figure 5-3 shows the entropy-based weight distribution of each feature.

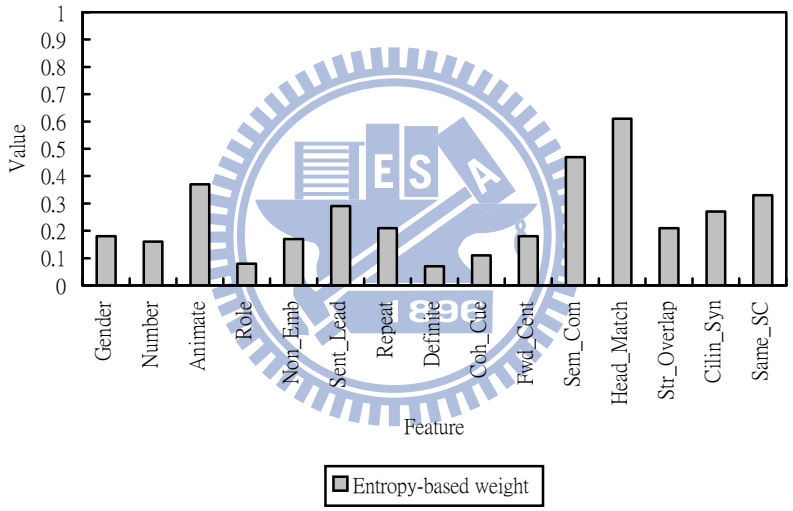



Figure 5-3. Entropy-based weight distribution.

### 5.2.4 Classification-based Module

Support Vector Machine (SVM) is a useful technique for data classification. It is widely used in the research of natural language processing problems. In anaphora resolution, SVM-based classifiers are commonly applied for identifying potential antecedents [4][26]. To compare with the performance of our proposed method, we

used SVM as a baseline model and utilize LIBSVM[11] as a classification tool.

## 5.3 Experiments

We extract 204 news documents from ASBC as our resolution corpus and from this corpus 426 anaphor-antecedent pairs are identified by experts. Table 5-3 lists the top 10 semantic class statistics in our corpus. To evaluate the performance of our proposed method, we implement three resolution strategies for comparison as shown in Table 5-4. The first model utilizes equal-weighted salience measures to identify antecedents. Namely, the weight of each feature is set to be 1. In the second model, a classification-based method is implemented by using SVM. In our proposed method, each feature is weighted by Equation (4). It is found that features with top five weights are *Head_Match*, *Sem_Com*, *Animate*, *Same_SC*, and *Sent_Lead*, respectively. This result indicates that *Head_Match*, *Animate* and *Same_SC* features are three dominant features for the characteristic of semantic agreement. In addition, the *Sem_Com* feature shows the significance of collocate compatibility in selecting antecedents. *Sent_Lead* justifies the fact that Chinese is a topic prominent language.

Table 5-4 shows that our method yields 72.5% success rate on 426 anaphoric instances by employing entropy-based weight scheme and web-based lexical knowledge. It improves about 4.7% success rate while compared with a classification-based model. To evaluate how useful the semantic compatibility feature is in our method, we conduct a leave-one-out evaluation in the third model. The success rate decreases 4.2% when feature *Sem_Com* is disabled. In addition, to find out the contribution of each type of features in our proposed method, we conduct a leave-group-out evaluation as shown in Table 5-5. Four types of features are

[11] The LIBSVM tool is available at http://www.csie.ntu.edu.tw/~cjlin/.

concerned, for example, lexical, grammatical, semantic, and heuristic. It shows that the type of semantic features plays the most important role since the success rate decreases significantly when this type of features is disable.

Table 5-3. Distribution of top 10 semantic classes.

| Semantic Class | Ratio |
|---|---|
| mankind | 22.0% |
| equipments | 8.9% |
| place | 6.1% |
| machines | 4.4% |
| organizations | 4.4% |
| buildings | 3.8% |
| fine_arts | 3.3% |
| nonhuman | 3.0% |
| solid | 3.0% |
| regions | 2.8% |

Table 5-4. Performance evaluation.

| Models | Success rate |
|---|---|
| Equal-weighted | 48.6% |
| Classification-based | 67.8% |
| Our method – *Sem_Com* | 68.3% |
| Our method | 72.5% |

Table 5-5. Performance of leave-group-out evaluation.

| Type | Success rate |
|---|---|
| Lexical | 62.6% |
| Grammatical | 66.8% |
| Semantic | 58.2% |
| Heuristic | 65.5% |

## 5.4 Analysis and Summary

Table 5-6 lists errors in definite anaphora resolution. Besides preprocessing errors, the major faults are attributed to semantic class features. Some error antecedent-anaphor pairs and their semantic classes are listed below. In example (1) and (2), definite anaphors are physical entities while the corresponding antecedents are non-physical entities. Example (3) shows the case that resolution could lead to ambiguity while semantic classes of anaphor and antecedent are unknown.

(1) [歐洲傳統武力裁減條約，這份文件] : [laws, document]

(2) [沈默寡言型，這種人] : [appearances, mankind]

(3) [象牙紅，這種樹花] : [not available, not available]

Table 5-6. Error analysis of NA.

| Error types | # of error instances | Ratio |
|---|---|---|
| Preprocessing error | 31 | 26% |
| Semantic class mismatch | 30 | 26% |
| Inappropriate salience | 24 | 20% |
| Exceeding window size | 13 | 11% |
| Multiple antecedents | 12 | 10% |
| Others | 7 | 6% |
| Total | 117 | 100% |

To our knowledge, our method represents the first attempt to use weight learning and Web-based knowledge acquisition for resolving definite anaphora in Chinese text. To overcome the drawback of common rule-based methods that employed manual

weights, an effective measurement is constructed on the basis of entropy-based weight to estimate the likelihood of antecedent candidates. Moreover, to cope with the difficulty of feature extraction in Chinese texts, a Web-based knowledge acquisition model is proposed to extract gender, number, and semantic compatibility from contextual information and Web resources. Our experimental results show that the method can achieve a significant increase in the success rate of around 4.7% when lexical knowledge learning and entropy-based weighting are utilized.

# Chapter 6

# Conclusions and Future Work

In this thesis, we present three methods to Chinese anaphora resolution based on weight learning and knowledge acquisition. Our contributions are that we proposed innovative methods for lexical knowledge acquisition and our study is the first one that utilizes entropy-based weight in anaphora resolution. To overcome the drawback of common rule-based methods that employed manual weights, an effective measurement is constructed on the basis of entropy-based weight to estimate the likelihood of antecedent candidates. Compared with the manual weight scheme, the presented entropy-based weight scheme is more capable to estimate the likelihood that a candidate turns out to be an antecedent. Moreover, the presented lexical knowledge acquisition such as collocate compatibility, conceptual patterns, and semantic compatibility are indeed able to acquire more semantic information from contexts and Web resources.

With the growing interest in natural language processing and its various applications, anaphora resolution is worth considering for further message understanding and the consistency of discourses. Although our proposed resolution methods to Chinese anaphora achieve better performance, there are still some problems left unsolved. Our future work will be directed into following studies.

(1) **Construct Chinese case frame ontology**

Case frames are useful knowledge resources for analyzing the relationship between nouns and verbs. In the application of anaphora resolution, case frames help to identify the antecedent by the verb of the anaphor. However, it is time-assuming and difficult to construct wide-coverage case frames

manually. Besides, verb sense ambiguity is another problem to be deal with since verbs with different meanings are classified as different cases frames. The integration of dictionaries and example phrases from large corpora such as web corpus could be helpful for constructing cases frames automatically [44].

(2) **Recognize semantic class of unknown words**

Our methods cannot recognize the semantic class of unknown words. Antecedent candidates or anaphors with unknown semantic class would cause errors during the antecedent selection process. For example, a semantic class often forms an inseparable connection between antecedents and definite anaphors. Therefore, word sense disambiguation should be considered as an important factor to enhance the overall performance.

**(3) Resolve the abstract anaphoric phenomenon**

In most studies of anaphora resolution, the default candidates are set to be noun phrases. However, instead of physical entities, antecedents can also be presented in the forms of clauses, sentences, or a group of sentences. In these cases, abstract anaphors often refer to expressions such as situations, facts, events, propositions, states, and actions. Thus, learning how to identify the boundary of utterance becomes an essential phase to resolve abstract anaphora.

(4) **Apply dependency structural feature**

Beside the features considered in our model, there are some more properties concerning dependency structure can be exploited. For example, Hobbs algorithm operates on parsed English sentences for pronominal anaphora resolution [18]. By applying the breadth-first searching strategy in Chinese

discourses, candidates that satisfy the constraints can be a feasible solution for antecedent selection.

# Bibliography


[1] Aamodt, A. and Plaza, E. “Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches,” *AI Communications*, IOS Press, 7(1), pp. 39-59, 1994.

[2] Baldwin, B. “CogNIAC: high precision coreference with limited knowledge and linguistic resources,” *Proceedings of the ACL'97/EACL'97 workshop on Operational factors in practical, robust anaphora resolution*, pp. 38-45, 1997.

[3] Barbu, C. and Mitkov, R. “Evaluation tool for rule-based anaphora resolution methods,” *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pp. 34-41, 2001.

[4] Bergsma, S. and Lin, D. “Bootstrapping Path-Based Pronoun Resolution,” *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pp. 33-40, 2006.

[5] Bunescu, R. “Associative Anaphora Resolution: A Web-Based Approach,” *Proceedings of the EACL Workshop on the Computational Treatment of Anaphora*, pp. 47-52, 2003.

[6] Chinese Knowledge Information Processing Group. “The content and illustration of Sinica corpus of Academia Sinica,” (Report No. 95-102), Institute of Information Science, Academia Sinica, 1995.

[7] Cardie, C. “Integrating case-based learning and cognitive biases for machine learning of natural language,” *Journal of Experimental and Theoretical Artificial Intelligence*, 11(3), pp. 297-337, 1999.

[8] Converse, S. P. “Resolving Pronominal References in Chinese with the Hobbs Algorithm,” *Proceedings of the 4th SIGHAN Workshop on Chinese Language*

*Processing*, pp. 116-122, 2005.

[9] Converse, S. P. "Pronominal Anaphora Resolution in Chinese," Dissertation, Computer and Information Science, University of Pennsylvania, 2006.

[10] Cui, Y. Z., Hu, Q., Pan, H., and Hu, J. "Zero Anaphora Resolution in Chinese Discourse," *Proceedings of the Computational Linguistics and Intelligent Text Processing*, pp. 242-248, 2006.

[11] Dagan, I. and Itai, A. "Automatic processing of large corpora for the resolution of anaphora references," *Proceedings of the 13th International Conference on Computational Linguistics*, pp. 330-332, 1990.

[12] Denber, M. "Automatic resolution of anaphora in English," Technical report, Eastman Kodak Co., 1998.

[13] Denis, P. and Baldridge, J. "A Ranking Approach to Pronoun Resolution," *Proceedings of IJCAI*, pp. 1588-1593, 2007.

[14] Ding, B. G., Huang, C. N., and Huang, D.G. "Chinese Main Verb Identification: From Specification to Realization," *International journal of Computational Linguistics and Chinese Language Processing*, 10(1), pp. 53-94, 2005.

[15] Gasperin, C. "Statistical anaphora resolution in biomedical texts," Technical report, Computer laboratory, University of Cambridge, 2009.

[16] Ge, N., Hale, J. and Charniak, E. "A Statistical Approach to Anaphora Resolution," *Proceedings of the Sixth Workshop on Very Large Corpora*, pp. 161-170, 1998.

[17] Guan, J., Zhou, Y. and He, H. "A sort Approach for Anaphora Resolution of Chinese Personal Pronoun Based on Machine Learning Method," *Proceedings of the Natural Language Processing and Knowledge Engineering*, pp. 293-300, 2007.

[18] Hobbs, J. “Resolving pronoun references,” *Lingua*, 44, pp. 311–338, 1978.

[19] Huang, Y. “Anaphora: A cross-linguistic study,” Oxford, England: Oxford University Press, 2000.

[20] Kennedy, C. and Boguraev, B. “Anaphora for everyone: Pronominal anaphora resolution without a parser,” *Proceedings of the 16th International Conference on Computational Linguistics*, pp. 113-118, 1996.

[21] Kong, F., Zhu, Q. M., Zhou, G. D., and Qian, P. D. “Coreference Resolution Based On Center Theory,” *Computer Science*, 36(6), pp. 219-222, 2009.

[22] Lappin, S. and Leass, H. “An Algorithm for Pronominal Anaphora Resolution,” *Computational Linguistics*, 20(4), pp. 535-561, 1994.

[23] Lee, C. L. “Zero anaphora in Chinese,” Crane publication company, Taipei, 2002.

[24] Li, G. C. and Luo, Y. F. “Chinese Pronominal Anaphora Resolution Via a Preference Selection Approach,” *Journal of Chinese Information Processing*, 19(4), pp. 24-30, 2005.

[25] Li, W. “Topic Chains in Chinese Discourse,” *Discourse Processes*, 37, pp. 25-45, 2004.

[26] Li, Y. C., Yang, Y., Zhou, G. D., and Zhu, Q. M. “Anaphora Resolution of Noun Phrase Based on SVM,” *Computer Engineering*, 35(3), pp. 199-204, 2009.

[27] Liang, T., Yeh, C.H., and Wu, D. S. “A Corpus-based Categorization for Chinese Proper Nouns,” *Proceedings of the National Computer Symposium*, pp. 434-443, 2003.

[28] Liang, T. and Wu, D. S. “Automatic Pronominal Anaphora Resolution in English Texts,” *International Journal of Computational Linguistics and Chinese Language Processing*, 9(1), pp. 21-40, 2004.

[29] Liu, D. R. and Ke, C. K. “Knowledge support for problem-solving in a production process: A hybrid of knowledge discovery and case-based reasoning,” *Expert Systems with Applications*, 33(1), pp. 147-161, 2007.

[30] Liu, Y. H., Pan, W. Y., and Gu, W. “Shiyong Xiandai Hanyu Yufa (Practical Modern Chinese Grammar),” The Commercial Press, 2002.

[31] Markert, K. and Nissim, M. “Comparing Knowledge Sources for Nominal Anaphora Resolution,” *Computational Linguistics*, 31(3), pp. 367-402, 2005.

[32] Mitkov, R. “Robust pronoun resolution with limited knowledge,” *Proceedings of the 18th International Conference on Computational Linguistics /ACL'98 Conference*, Montreal, Canada, pp. 869-875, 1998.

[33] Mitkov, R. “Anaphora resolution: the state of the art,” Working paper (Based on the COLING'98/ACL'98 tutorial on anaphora resolution), University of Wolverhampton, Wolverhampton, 1999.

[34] Mitkov, R., Evans, R., and Orasan, C. “A new fully automatic version of mitkov's knowledge-poor pronoun resolution method,” *Proceedings of the third International Conference on Computational Linguistics and Intelligent Text Processing*, pp. 168-186, 2002.

[35] Mitchell, T. M. “Machine Learning,” McGraw-Hall companies, 1997.

[36] Modjeska, N. N., Markert, K., and Nissim, M. “Using the Web in Machine Learning for Other-Anaphora Resolution,” *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 176-183, 2003.

[37] Muñoz, M., Saiz-Noeda, M., and Montoyo, A. “Semantic Information in Anaphora Resolution,” *Proceedings of the Third International Conference on Advances in Natural Language Processing*, pp. 63-70, 2002.

[38] Ng, V. “Machine learning for coreference resolution: From local classification to

global ranking," *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pp. 157-164, 2005.

[39] Ng, V. and Cardie, C. "Improving machine learning approaches to coreference resolution," *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 104-111, 2002.

[40] Poesio, M., Ishikawa, T., Walde, S. S., and Vieira, R. "Acquiring Lexical Knowledge for Anaphora Resolution," *Proceedings of the 3rd Conference on Language Resources and Evaluation*, pp. 1220-1224, 2002.

[41] Poesio, M. and Kabadjov, M. A. "A General-Purpose, off-the-shelf Anaphora Resolution Module: Implementation and Preliminary Evaluation," *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pp. 663-666, 2004.

[42] Qiu, L., Kan, M., and Chua, T. "A public reference implementation of the rap anaphora resolution algorithm," *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pp. 291-294, 2004.

[43] Saiz-Noeda, M. and Palomar, M. "Semantic Knowledge-Driven Method to Solve Pronominal Anaphora in Spanish Texts," *Proceedings of the Natural Language Processing*, pp. 204-211, 2000.

[44] Sasano, R., Kawahara, D., and Kurohashi, S. "Automatic Construction of Nominal Case Frames and its Application to Indirect Anaphora Resolution," *Proceedings of the 20th International Conference on Computational Linguistics*, pp. 1201-1207, 2004.

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

[46] Tao, L. and Healy, A. F. “Zero Anaphora: Transfer of Reference Tracking Strategies from Chinese to English,” *Journal of Psycholinguistic Research*, 34(2), pp. 99-131, 2005.

[47] Wang, H. F. and Mei, Z. “An Empirical Study on Pronoun Resolution in Chinese,” *Lecture Notes in Computer Science*, 2945, pp. 213-216, 2004.

[48] Wang, H. F. and Mei, Z. “Robust Pronominal Resolution within Chinese Text,” *Journal of Software*, 16, pp. 700-707, 2005.

[49] Wang, N., Yuan, C. F., Wang, K. F., and Li, W. J. “Anaphora Resolution in Chinese Financial News for Information Extraction,” *Proceedings of the 4th World Congress on Intelligent Control and Automation*, pp. 2422-2426, 2002.

[50] Wu, D. S. and Liang, T. “Chinese Pronominal Anaphora Resolution Using Lexical Knowledge and Entropy-based Weight,” *Journal of the American Society for Information Science and Technology*, 59(13), pp. 2138-2145, 2008.

[51] Wu, D. S. and Liang, T. “Zero Anaphora Resolution by Case-based Reasoning and Pattern Conceptualization,” *Expert Systems with Applications*, 36(4), pp. 7544-7551, 2009.

[52] Xu, J. J. “Anaphora in Chinese Texts,” China social science, Beijing, 2003.

[53] Yang, X. F., Su, J., and Tan, C. L. “Kernel-Based Pronoun Resolution with Structured Syntactic Knowledge,” *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pp. 41-48, 2006.

[54] Yang, X. F., Su, J., and Tan, C. L. “A Twin-Candidate Model for Learning-Based Anaphora Resolution,” *Computational Linguistics*, 34(3), pp. 327-356, 2008.

[55] Yeh, C. L. and Chen, Y. C. “Zero anaphora resolution in Chinese with shallow parsing,” *Journal of Chinese Language and Computing*, 2005.

[56] Yu, C. H. and Chen, H. H. “A Study of Chinese Information Extraction Construction and Coreference,” Unpublished master’s thesis, National Taiwan University, Taiwan, 2000.

[57] Zhao, S. and Ng, H. T. “Identification and Resolution of Chinese Zero Pronouns: A Machine Learning Approach,” *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 541–550, 2007.

# Appendix A- Tagged Data

## A1. PA Tagged Data

資料表 'test.' 中的資料 (在 'PA' 中) 於 '(local)'

| ana | anarange | ant | antrange | same_pro | per_pro | gender | number | animate | role | parallel | ne_per |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 他 | 8:1:1 | 丘成桐 | 1:10:10 | <NULL> | 1 | 1 | 1 | 1 | <NULL> | <NULL> | 1 |
| 他 | 4:3:3 | 楊振寧 | 1:5:5 | <NULL> | 1 | 1 | 1 | 1 | <NULL> | <NULL> | 1 |
| 他 | 10:1:1 | 楊先生 | 5:1:2 | <NULL> | 1 | 1 | <NULL> | 1 | <NULL> | <NULL> | <NULL> |
| 他 | 38:3:3 | 羅伯茲 | 37:2:2 | <NULL> | 1 | <NULL> | 1 | 1 | 1 | 1 | 1 |
| 她 | 81:9:9 | 周博士 | 76:1:2 | <NULL> | 1 | <NULL> | <NULL> | 1 | <NULL> | <NULL> | <NULL> |
| 她 | 84:1:1 | 她 | 81:9:9 | 1 | 1 | 1 | 1 | 1 | <NULL> | <NULL> | <NULL> |
| 她 | 85:4:4 | 她 | 84:1:1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | <NULL> |
| 她 | 87:2:2 | 她 | 85:4:4 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | <NULL> |
| 她 | 107:1:1 | 周博士 | 99:1:2 | <NULL> | 1 | <NULL> | <NULL> | 1 | <NULL> | <NULL> | <NULL> |
| 他 | 3:3:3 | 楊振寧 | 1:5:5 | <NULL> | 1 | 1 | 1 | 1 | <NULL> | <NULL> | 1 |
| 他 | 2:2:2 | 李遠哲先生 | 1:5:6 | <NULL> | 1 | 1 | <NULL> | 1 | <NULL> | <NULL> | <NULL> |
| 他 | 5:1:1 | 他 | 2:2:2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | <NULL> |
| 它 | 32:1:1 | 學者的社區 | 27:4:6 | <NULL> | <NULL> | <NULL> | <NULL> | <NULL> | <NULL> | <NULL> | <NULL> |
| 他們 | 53:1:1 | 優秀的心靈 | 47:6:8 | <NULL> | 1 | <NULL> | <NULL> | -1 | <NULL> | <NULL> | <NULL> |
| 它們 | 11:6:6 | 台邁南島語言 | 8:1:3 | <NULL> | <NULL> | <NULL> | <NULL> | <NULL> | <NULL> | <NULL> | <NULL> |
| 它們 | 24:2:2 | 台邁南島民族的語 | 23:1:6 | <NULL> | <NULL> | <NULL> | <NULL> | <NULL> | <NULL> | <NULL> | <NULL> |
| 他 | 7:1:1 | 柯適教授 | 3:1:2 | <NULL> | 1 | <NULL> | <NULL> | 1 | <NULL> | <NULL> | <NULL> |
| 他 | 8:1:1 | 他 | 7:1:1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | <NULL> |
| 他 | 2:1:1 | 簡泰安 | 1:9:9 | <NULL> | 1 | 1 | 1 | 1 | <NULL> | <NULL> | 1 |
| 他們 | 11:6:6 | 一對男女 | 10:1:3 | <NULL> | 1 | <NULL> | 1 | 1 | <NULL> | <NULL> | <NULL> |
| 他 | 37:1:1 | 那個男生 | 35:2:4 | <NULL> | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 她 | 27:6:6 | 朱美玲 | 24:1:1 | <NULL> | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 她 | 35:1:1 | 她 | 28:7:7 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | <NULL> |
| 她 | 28:7:7 | 她 | 27:6:6 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | <NULL> |
| 他 | 15:9:9 | 張振宇 | 15:1:1 | <NULL> | 1 | <NULL> | 1 | 1 | 1 | <NULL> | 1 |
| 他 | 18:9:9 | 張振宇 | 18:4:4 | <NULL> | 1 | <NULL> | 1 | 1 | <NULL> | <NULL> | 1 |
| 他 | 33:3:3 | 張振宇 | 32:1:1 | <NULL> | 1 | <NULL> | 1 | 1 | 1 | <NULL> | 1 |
| 他 | 41:1:1 | 張振宇 | 37:1:1 | <NULL> | 1 | <NULL> | 1 | 1 | 1 | 1 | 1 |
| 她們 | 50:21:21 | 藝術中心的義工們 | 50:1:5 | <NULL> | 1 | <NULL> | <NULL> | 1 | 1 | 1 | <NULL> |
| 它 | 30:3:3 | 這畫 | 27:5:6 | <NULL> | <NULL> | <NULL> | <NULL> | <NULL> | <NULL> | <NULL> | <NULL> |
| 它 | 34:2:2 | 它 | 30:3:3 | 1 | <NULL> | <NULL> | 1 | <NULL> | <NULL> | <NULL> | <NULL> |
| 它 | 35:2:2 | 它 | 34:2:2 | 1 | <NULL> | <NULL> | 1 | <NULL> | <NULL> | 1 | <NULL> |
| 它 | 36:2:2 | 它 | 35:2:2 | 1 | <NULL> | <NULL> | 1 | <NULL> | <NULL> | <NULL> | <NULL> |

## A2. ZA Tagged Data

資料表 'ZA case' 中的資料 (在 'ZA' 中) 於 '(local)'

| DEF | EMB | CLASS | SEN_DIST_A | OFFSET | CUR_ZA | RPT | CON_PAT_A | ZA | ANT |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | events | 1 | 1 | 0 | 1 | [unknown](VC)[unknown] | 9:13:13 | 9:1:1 |
| 0 | 1 | unknown | 1 | 1 | 0 | 0 | [mankind](VC)[events] | 3:6:6 | 3:3:3 |
| 0 | 0 | mental | 1 | 1 | 0 | 0 | [unknown](VC)[events] | 2:50:50 | 2:44:46 |
| 0 | 0 | mankind | 1 | 1 | 0 | 1 | [atrifacts](VC)[events] | 5:9:9 | 5:1:3 |
| 1 | 0 | mankind | 1 | 2 | 0 | 1 | [atrifacts](VC)[events] | 5:21:21 | 5:12:14 |
| 0 | 0 | events | 1 | 1 | 0 | 0 | [unknown](VA)[events] | 6:5:5 | 5:21:21 |
| 0 | 1 | mankind | 1 | 1 | 0 | 1 | [mankind](VA)[events] | 7:1:1 | 5:1:3 |
| 0 | 0 | unknown | 1 | 1 | 0 | 1 | [unknown](VC)[events] | 9:3:3 | 7:4:6 |
| 0 | 0 | matter | 2 | 2 | 0 | 1 | [unknown](VA)[unknown] | 12:19:19 | 12:1:1 |
| 0 | 0 | unknown | 1 | 1 | 0 | 1 | [mankind](VC)[events] | 1:24:24 | 1:18:20 |
| 0 | 0 | mankind | 1 | 1 | 0 | 1 | [unknown](VC)[events] | 2:1:1 | 1:24:24 |
| 0 | 0 | mental | 1 | 1 | 0 | 1 | [unknown](VC)[events] | 5:20:20 | 5:1:1 |
| 0 | 0 | unknown | 1 | 1 | 0 | 1 | [matter](VC)[events] | 5:2:2 | 4:1:1 |
| 0 | 0 | temporal | 1 | 1 | 0 | 1 | [atrifacts](VD)[events] | 5:15:15 | 5:2:2 |
| 0 | 0 | mankind | 1 | 1 | 0 | 0 | [matter](VD)[unknown] | 7:1:1 | 6:1:3 |
| 0 | 0 | mankind | 1 | 1 | 0 | 1 | [unknown](VC)[events] | 8:1:1 | 7:1:1 |
| 0 | 0 | events | 3 | 1 | 0 | 0 | [unknown](VC)[place] | 2:6:6 | 2:1:1 |
| 0 | 0 | mankind | 1 | 1 | 0 | 1 | [atrifacts](VA)[temporal] | 5:8:8 | 5:2:2 |
| 0 | 0 | mankind | 1 | 1 | 0 | 1 | [unknown](VC)[events] | 7:19:19 | 7:3:6 |
| 0 | 0 | mental | 1 | 1 | 0 | 1 | [unknown](VC)[matter] | 11:1:1 | 10:18:18 |
| 0 | 0 | unknown | 1 | 1 | 0 | 1 | [mankind](VD)[events] | 13:2:2 | 12:20:22 |
| 1 | 0 | mankind | 1 | 1 | 0 | 1 | [unknown](VC)[events] | 18:8:8 | 18:1:1 |
| 0 | 0 | mankind | 1 | 1 | 0 | 1 | [atrifacts](VC)[unknown] | 23:25:25 | 23:12:15 |
| 0 | 0 | mankind | 1 | 1 | 0 | 0 | [mankind](VC)[events] | 26:19:19 | 26:13:17 |
| 0 | 1 | events | 3 | 1 | 0 | 1 | [mankind](VC)[mankind] | 31:47:47 | 31:33:34 |
| 0 | 0 | mankind | 1 | 1 | 0 | 0 | [mankind](VC)[events] | 38:13:13 | 38:3:4 |
| 0 | 0 | mankind | 1 | 1 | 0 | 1 | [matter](VD)[events] | 57:1:1 | 56:18:19 |
| 0 | 0 | mankind | 1 | 1 | 0 | 1 | [mankind](VC)[events] | 57:36:36 | 57:20:23 |
| 0 | 0 | mankind | 1 | 1 | 0 | 0 | [mankind](VE)[events] | 60:20:20 | 60:14:18 |
| 0 | 0 | mankind | 1 | 1 | 0 | 1 | [unknown](VC)[events] | 46:1:1 | 45:33:33 |
| 0 | 0 | mankind | 1 | 2 | 1 | 1 | [unknown](VC)[events] | 47:1:1 | 46:1:1 |
| 0 | 0 | mankind | 1 | 1 | 0 | 1 | [mankind](VE)[events] | 52:2:2 | 51:2:2 |
| 0 | 0 | mankind | 1 | 1 | 0 | 0 | [atrifacts](VE)[events] | 53:13:13 | 52:2:2 |

## A3. NA Tagged Data

資料表 'test1' 中的資料 (在 'NA' 中) 於 '(local)'

| Cno | NSnoW1W2 | CSnoW1W2 | Non_Emb | Gender | Number | Animate | Role | Sent_Lead | Repeat |
|---|---|---|---|---|---|---|---|---|---|
| 2040 | 這次比賽 | 會長盃國際棒球賽 | 1 | 1 | 0 | 1 | 0 | 1 | 0 |
| 2042 | 這項比賽 | 世界職業網球賽 | 1 | 1 | 0 | 1 | 0 | 0 | 0 |
| 2047 | 這個分數 | 高得分 | 1 | 1 | 0 | 1 | 0 | 0 | 1 |
| 2053 | 這份薪水 | 總幹事月薪 | 1 | 1 | 0 | 1 | 1 | 0 | 0 |
| 2066 | 這項公路賽 | 自由車公路賽 | 1 | 1 | 0 | 1 | 0 | 1 | 1 |
| 2083 | 這記觸身球 | 後援投手劉志昇投 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| 2085 | 這位官員 | 央行官員 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 2111 | 這個爛隊 | 沙加緬多帝王隊 | 1 | 1 | 0 | 1 | 1 | 0 | 1 |
| 2111 | 這一個球隊 | 沙加緬多帝王隊 | 1 | 1 | 0 | 1 | 1 | 0 | 1 |
| 2123 | 這一球 | 外野高飛球 | 1 | 1 | 0 | 1 | 0 | 0 | 1 |
| 2124 | 這項全民運動聯盟 | 推動的全民運動場 | 1 | 1 | 0 | 1 | 0 | 0 | 1 |
| 2128 | 這次雙打錦標賽 | 世界杯桌球雙打混 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 2137 | 這次大會 | 台灣省七十九年中 | 1 | 1 | 0 | 1 | 0 | 1 | 0 |
| 2137 | 這項殊榮 | 十二屆的冠軍寶座 | 1 | 1 | 0 | 1 | 1 | 0 | 0 |
| 2169 | 這場比賽 | 中。日職棒對抗混 | 1 | 1 | 0 | 1 | 1 | 1 | 0 |
| 2185 | 這項危機 | 波斯灣危機 | 1 | 1 | 0 | 1 | 1 | 0 | 1 |
| 2187 | 這枚人造衛星 | 一枚人造衛星 | 1 | 1 | 1 | 1 | 1 | 0 | 1 |
| 2192 | 這把尚方寶劍 | 尚方寶劍 | 1 | 1 | 0 | 1 | 0 | 0 | 1 |
| 2196 | 這位官員 | 匈牙利一位高級官 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2200 | 這艘船 | 沙國運輸艦 | 1 | 1 | 1 | 1 | 0 | 0 | 1 |
| 2218 | 這項措施 | 伊拉克動武的決議 | 1 | 1 | 0 | 1 | 1 | 0 | 0 |
| 2225 | 這項儀典 | 日本天皇的登基大 | 1 | 1 | 0 | 1 | 0 | 0 | 0 |