Hybrid Tag-set for Natural Language Processing

梁瑋洸
Leung Wai Kwong

Department of Systems Engineering & Engineering Management

香港中文大學
THE CHINESE UNIVERSITY OF HONG KONG

A Thesis
Submitted in Partial Fulfilment of the Requirements for the Degree of
Master of Philosophy
in
Systems Engineering and Engineering Management

July 1999

Abstract

In natural language processing (NLP), words are usually associated with part-of-speech tags. However, merely syntactic information cannot cope with most modern NLP applications such as parsing or indexing. In Chinese noun phrase parsing, intensive syntactic, semantic and contextual knowledges are required to analyze the textual data. It is observed that some information is missing in parsing, causing parsing ambiguities and thus, lowering the efficiency of parsing. Semantic consideration is required to overcome this problem. We propose a hybrid tag-set for Chinese natural language processing. The objective is to enhance the performance of syntactic parsing. This new tag combines the syntactic and semantic information and associated them with each other in a single representation. The semantic class from Cilin (同義詞詞林) [13] is employed to provide the semantic information. An algorithm for automatic assignment of hybrid tag to Chinese word is proposed. Apart from forming the hybrid tag-set, this algorithm is significant

in the way that it can be applied to enrich the Cilin for mordern linguistic reference. Unknown i.e. words, words that are not accounted for in the Cilin, can be automatically assigned with suitable semantic classes. This stratgey offers a correctness of 91.73%. Experiments have shown that the hybrid tag-set is effective. The CNP3 Chinese noun phrase partial parser was used for evaluation. Its correctness was improved by 12.15% with the hybrid tags.

摘要

在自然語言處理之中，大多數文本都標有詞的詞類(Part-of-Speech)。但是，單純應用詞類信息來處理資料, 是難以應付好像語法分析和索引等現代自然語言處理系統的需求。就以中文多詞短詞分析作例，要正確分析一段文字，是需要對詞類，詞義及前文後理有深入的認識。所以，只用詞類來表達文字，是不能提供足夠的語言信息的。這樣不但造成許多語法分析上的含糊不清，更降低了語法分析的整體效益。因此，要解決這樣的問題，詞義上的考慮是十分重要的。這項研究出一套 hybrid tag-set 以作中文自然語言處理提之用。其主要優勢是能提高句法分析的效率。這一套新的 tag-set，是在單一的表達方式上，結合了詞類及詞義的信息。這裡所指的詞義，是應用了《同義詞詞林》中的詞義。我們更提出一個自動中文字分配 hybrid tag標注的方法。它的設計完全是針對《同義詞詞林》的弱點。除了建立一套 hybrid tag-set 的功能，它還可以被應用到增加《同義詞詞林》作現代語言學上參考的功用。在同義詞詞林中未曾收錄的新詞，是可以透過它，自動成功分配一組對應的

詞義。這策略是可以提供 83.47% 和 91.73% 的召回率和精確率。概括而言，實驗結果顯示這一套 hybrid tag-set 是有效的，我們以 CNP3 這種中文短語句法分析器作試驗對象, Hybrid tag-set 中所提供的詞義資料的應用成功地把它的精確率提高了 12.15%。

Acknowledgements

I would like to express my most sincere gratitude to my supervisors Prof K.F. Wong and Prof. Edward Ho for their continual guidance and invaluable comments throughout my research work. I would also like to thank Prof. Helen Meng and Prof. B.T. Low for giving me helpful suggestions. Credits also go to the administrative staffs in the department office and all my colleagues in the I.S. Lab. Mandy and Iris are nice and responsible. Wai-Ip, Timothy and Edmund are always glad to help me whenever I encountered technical problems. Benson, Dong, Kin and Pao gave me enthusiastic support whenever I felt frustrated. They shared the tears and joys with me during my research days. Moreover, I would like to thanks my family for their unfailing support. My parents never complaint for the disturbs I created when I worked overnight and I would not forget the encouragement from my sisters, Yan and Foon. Finally, the greatest thanks will be dedicated to my beloved Emily. Without her, this thesis would never exist.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter elucidates the motivation of this research work. The objective is also stated, which is to enhance the performance of syntactic parsing with the hybrid tag-set proposed. Finally, the organization of this thesis is given.

1.1 Motivation

Today is the era of information technology. Noticeably, rapid development in information technology is observed in the East Asia. Large volume of information is being handled. The most common form of information is textual data. Many applications require automatic understanding of this textual data in a human-like fashion. As Chinese is an important language shared by many countries in the region, Chinese natural language process-

ing (NLP) is thus critical for information technology development in these countries. Parsing is a major stage in NLP. It aims to analyze a piece of text in natural language. A sentence is the basic processing unit. Parsers take a sentence as input and output a parse tree, which is the syntactic structure of the sentence. Accurate parsing will benefits many aspects of NLP, such as information retrieval and machine translation. Therefore, parsing can be regarded as the "preprocessing" to most NLP applications. Parsing requires syntactic, semantic and contextual analysis. Comprehensive linguistic knowledge and robust algorithms are also essential. Moreover, natural sentences are composed of various complex structure such as phrases or clauses. The complexity is much higher in Chinese sentences. It is impossible to define a complete set of grammar rules to cater for all the cases in Chinese. This caused the low correctness for parsing a complete sentence. Noun phrases are the basic building blocks of most sentences. The scope of parsing can be narrowed down to noun phrases instead of the whole sentence. It is hoped that the computation complexity will be reduced. However, most NLP applications apply the syntactic behaviors (part-of-speech tags) to represent Chinese words. For Chinese, syntactic representations lead to parsing ambiguities. Ambiguous noun phrases lower the performance of syntactic parsing. It is revealed that merely the syntactic representation for Chinese words are inadequate. For improvement, it is necessary to employ semantic informa-

tion in parsing. We represent a Chinese word in a single unified format with both syntactic and semantic information.

1.2 Objective

In this thesis, we propose a new tag-set, for enhancing the performance of syntactic parsing and prove that it is effective. The problems in establishment of the hybrid tag-set is discussed and solutions proposed.

1.3 Organization of thesis

The rest of this thesis is structured as follow: in Chapter two , different Chinese noun phrases categories are discussed and common structurally ambiguous noun phrases are outlined. Observations made over these ambiguous noun phrases suggested that semantic infromation should be incorporated for effective parsing.

Chapter three proposes a new tag-set and the problems encountered in establising the tag-set are also discussed. In particular, an algorithm for automatic semantic assignment of unknown words is proposed in Chapter four. The Word Sense Disambiguation (WSD) algorithm is proposed in Chapter five. WSD was designed to solve the problem of multiple semantic classes

associated with a single word.

Chapter six, presents a series of experiment for evaluating the hybrid tag-set for noun phrase parsing. Finally, in Chapter seven, summary of the research work and its major contirbution are given and possible extensions are given.

Chapter 2

Background

In this chapter, different Chinese noun phrase categories are discussed. With respect to Chinese noun phrases parsing, three structurally ambiguous noun phrases are identified. Observations made over these ambiguous noun phrases suggested that semantic infromation should be incorporated for effective parsing.

2.1 Chinese Noun Phrases Parsing

Parsing is the technique for understanding natural language text [7, 21, 43]. Most parsers interpret an entire sentence as a single unit. The result of the analysis is a parse tree, depicting the syntactic relationships between the words. A natural language sentence is composed of different grammatical

units such as phrases and/or clauses. Each of them can in turn be made up of other complex structures. To parse a sentence, intensive syntactic, semantic and contextual information are usually required. The computation complexity is large and comprehensive linguistic knowlege has to be incorporated. These become the barrier for natural language parsing. Therefore, to achieve better performance, most research works narrow down the scope to noun phrases instead of the complete sentence. Noun phrases are the constituent components of natural language sentences. If noun phrases can be parsed successfully, the complexity of full sentence parsing can be significantly reduced and moreover, the same algorithm may be extended to parsing in the sentence level [37]. Therefore, in our research, we focus on improving the performance of Chinese noun phrase parsing.

2.2 Chinese Noun Phrases

Based on syntactic classification, there are six categories of Chinese noun phrases [37]. They are listed below.

- Simple Noun Phrases
  - Syntactic Structure: modifier + single noun
  - Example: 白雲(white cloud)

- Classifier / Measure Noun Phrases
  - Syntactic Structure: numeral + measure word + noun phrases
  - Example: 五 個 萍果(five apples)
- Nominal Compound Noun Group
  - Subordinate Noun Phrases
    - Syntactic Structure: noun phrases as modifier + noun phrases
    - Example: 學生 宿舍(student hostel)
  - Apposition Noun Phrases
    - Syntactic Structure: noun phrase + noun phrase
    - Example: 醫生 護士(doctors and nurse)
  - Compound Noun Phrases
    - Syntactic Structure: [noun phrase] + verb + noun phrase
    - Example: 慶祝 活動(celebration activities)
- Associative Noun Phrases
  - Syntactic Structure: noun phrase + "的" + noun phrase
  - Example: 大學 的 設施(facilities of a university)
- Co-ordinate Noun Phrases

- Syntactic Structure: noun phrase + conjunction + noun phrases
- Example: 愛 與 誠(love and sincerity)

• Modifying Noun Phrases

- Noun Phrases with Relative Clause
  * Syntactic Structure: relative clause + "的" + noun phrase
  * Example: 唱歌 的 人(The one who sings ...)
- Attributive Noun Phrases
  * Syntactic Structure: adjective phrase + "的" + noun phrase
  * Example: 勤力 的 農夫(an industrious farmer)
- Genitive / Possessive Noun Phrases
  * Syntactic Structure: pronoun + "的" + noun phrase
  * Example: 我 的 電腦(my computer)
- Prepositional Noun Phrases
  * Syntactic Structure: prepositional + noun phrases + "的" + noun phrase
  * Example: 在 桌上 的 書(the book on the desk)

According to Li [23], a common noun phrase can be one of the following

- Pronoun

  Example: 我(I), 我們(we)

- Noun

  Example: 馬(horse), 萍果(apple)

- Noun with other elements.

  - Classifier phrases / measure phrases

    A classifier is a word that must occur with a number and/or a demonstrative, or certain quantifiers.

    Example: 三個人(three persons), 幾件衣服(a few garment)

  - Associative phrases

    It is a type of modification where two noun phrases are linked by the particle "的".

    Example: 我的家(my home), 中國的人口(population in China)

  - Modifying phrases

    It can be either a relative clause or an attributive adjective.

    Example: 騎自行車的人(the person who ride a bicycle)

- Compound noun phrases that consist of more than one of the above structure(s).

  Example: 三個騎自行車的人(three men who are riding bicycles)

Pun [32] also identified some forms of Chinese noun phrases. They are as below.

- Noun phrase + Noun phrase

  Example: 中國人民(people of China)

- Adjective + Noun phrase

  Example: 小孩子(little child)

- Noun phrase + "的" + Noun phrase

  Example: 我的老師(my teacher)

- Relative clause + Noun phrase

  Example: 喜歡中國的女孩(a girl who loves China)

- Nominal noun

  Example: 他吃的((the food) that he eat)

- Appositive clause + Noun phrase

  Example: 他生意失敗的事(the matter of the failure of his business)

For each of the above forms, combination with measure phrases is allowed with the form "(specifier) + (number) + (unit)". For example, 我的兩位老師(two teachers of mine).

In general, Chinese noun phrases are grouped in a different categories by different researchers. Among the three, Tse's [37] is a more complete definition and our work will be based on it.

2.3 Problems with Syntactic Parsing

Parsing ambiguities occur with the syntactic approach. Multiple parse trees may be produced for the same noun phrase. The three most ambiguous noun phrases are discussed below with illustrative examples.

2.3.1 Conjunctive Noun Phrases

Consider the following examples:

- 勤動的工人和農夫 (The hardworking worker and farmer)
  - Correct parsed structure:
    - [DENP [AP 勤動] 的 [CONJ-NP 工人和農夫]]
- 勤動的農夫和小狗 (The hardworking farmer and dog)
  - Correct parsed structure:
    - [CONJ-NP [DENP 勤動的農夫] 和 [NP 小狗]]

In the first noun phrase, the modifier is "勤動" (hardworking). It modifies both "工人" (worker) and "農夫" (farmer). In the second noun phrase, the modifier "勤動" (hardworking) only apply to "農夫" (farmer), but not "小狗" (dog). These two noun phrases have different parse trees, but bear the same syntactic structure (i.e. part-of-speech tag sequences). Therefore, ambiguities occur, as one could not distinguish the correct interpretation with syntactic information only. This implies further analysis, e.g. incorporation of semantic information is required. For example, If there was a clue that the adjective "勤動" (hardworking) can only be applied to a word, which described human (e.g. "農夫" (farmer), "工人" (worker)), and not to an animals ("小狗" (dog)), the correct parse tree could then be selected without ambiguity.

2.3.2 De-de Noun Phrases

De-de noun phrases are those noun phrases that are formed by the combination of noun phrases linked with "的" (i.e. of). Consider the following example:

- 去年第四季度 的 百分之一點一 的 水平
  - Correct parsed structure:
    - [DENP [NP 去年第四季度] 的 [DENP 百分之一點一 的

水平]]

- 包括廣大知識分子在內 的 工人階級內部 的 團結
  - Correct parsed structure:
    * [DENP [DENP 包括廣大知識分子在內 的 工人階級內部] 的 [NP 團結]]

These two noun phrases have the same syntactic structure, but their parse trees are different and present different interpretations. The problem is similar to that of conjunctive noun phrases and is best resolved semantically.

2.3.3 Compound Noun Phrases

According to Tse [37], nominal compound noun phrases account for most errors in compound noun phrases parsing. There can be infinite combinations in their formations, which makes it impossible to define a set of grammar rules to deal with them. Most syntactic parsers simply treat nominal compound noun phrases as separate grammatical unit and ignore their internal structures. Consider the following examples:

- 廣大勞動人民
  - Correct parsed structure:
    * [廣大 [勞動 [人民]]]

- 刺激銷售措施
  - Correct parsed structure:
    * [[刺激銷售] 措施]

They have the same syntactic representation, but their internal structures are different. In the first phrase, "勞動" (working) modifies "人民" (people) while "廣大" (all) modifies "勞動人民" (working people). However, in the second phrase, "刺激" (stimulate) modifies "銷售" (sales) and "刺激銷售" (stimulate sales) modifies "措施" (approach). It is inadequate to identify their internal modifying scope based on pure syntactic consideration. Once again, semantic information is required to reveal the internal structure of the compound noun phrases.

In addition, noun phrases which are combination of conjunctive noun phrases, De-de noun phrases and compound noun phrases are also ambiguous. Consider the following examples:

- 我 國 經濟 體制 改革 和 科技 體制 改革 的 推動
  - Wrongly parsed structure

    [DENP [CONJ-N [NP 我 國 經濟 體制 改革] 和 [NP 科技 體制 改革]] 的 推動]

No. of articles	No. of noun phrases	No. of conjuncitve noun phrases	No. of De-de noun phrases	No. of compound noun phrases
112	10886	390	2376	326

Table 2.1: *Distribution of Ambiguous Noun Phrases in corpus*

– Correct structure

[DENP [NP [NP 我 國] [CONJ-NP [NP 經濟 體制 改革] 和 [NP 科技 體制 改革]]] 的 推動]

where NP, CONJ-NP and DENP stand for noun phrase [1], conjunctive noun phrase and noun phrase with "的" ,respectively.

In this example, two nominal compound noun phrases (經濟 體制 改革, 科技 體制 改革) are connected together with "和" and "的". If semantic information indicates that "經濟 體制 改革"' and "科技 體制 改革" are nominal compound noun phrases, the conjunction "和" first joins the two compound noun phrases and then joins with "我國" to form a noun phrase.

2.4 Observations

The distribution of ambiguous conjunctive noun phrases and De-de noun phrases are shown in Table 2.1.

[1]See Appendix A for a complete list of POS tags.

2.4.1 Inadequacy in Part-of-Speech Categorization for Chinese NLP

There is no standardization in part-of-speech (POS) categorization. Most Chinese natural language researches make use of the POS tags developed by the Tsinghua University, which contains 109 POS tags (See Appendix A). However, one of the causes for parsing ambiguities is that although two sentences have different wordings, interpretation and parse trees, their representations in POS tags are the same (see before). Consider the following examples:

- 提高#vgn 速度#ng 的#usde 方法#ng

  (The method of increasing speed)

- 提高#vgn 汽車#ng 的#usde 速度#ng

  (Increase the speed of a car)

The symbols behind "#" are the POS tags. The first phrase is a noun phrase while the second one, which has the same syntactic structure is a relative clause. This complicates the parsing process. Therefore, most parsers equiped with syntactic information only, cannot distinguish between these two sentences and hence, treat them as the same pattern. This problem could be solved if a new tag categorization comprising both syntactic and

semantic information is used.

2.4.2 The Need of Semantic in Noun Phrase Parsing

It is observed that nominal compound noun phrases are the most problematic structures. They have numerous combination of formation. Thus, it is impossible to define a set of generic grammar rules to identify them accurately from a sentence. The usual practice in most parsers is to treat all nominal compound noun phrases as separate grammatical units without further revealing their internal structures [37]. As mentioned above, ignorance of compound noun phrases can lead to severe errors in parsing. In fact, the internal structures within compound noun phrases are useful for noun phrase indexing [19]. To solve this problem, semantics of individual components of a compound noun phrase must be considered. Our strategy is to first identify the compound noun phrases out of a sentence and then investigate into its internal structures.

2.5 Summary

In this chapter, Chinese noun phrase parsing is first discussed briefly. Several types of Chinese noun phrase formations are also described. First, it is the six syntactic categories from Tse [37]: simple NPs, classifier/measure NPs,

nominal compound noun groups, associative NPs, coordinate NPs and modifying NPs. This is follwoed by four categories from Li [23]: pronoun, noun, noun with other element and compound noun. Lastly, the six syntactic categories from Pun [32] is introduced. In addition, the parsing problems with syntactic tags are identified. Three ambiguous noun phrases are discussed in details. They are conjuntive NPs, de-de NPs and the nominal compound NPs. Two observations are concluded. It is observed that there is an inadequacy in part-of-speech (POS) categorization for Chinese words. Semantic information is also required to resolve the ambiguities caused by using POS alone. In the next chapter, a new tag-set, the hybrid tag-set will be proposed to improve the performance for Chinese noun phrase parsing. It is a combination of the syntactic (POS) and semantic tags. Problems concerning the establishment of the hybrid tag-set will be discussed and an overview for its implementation will be announced.

Chapter 3

Hybrid Tag-set

As discussed in the previous chapter, most parsing ambiguities arise due to the lack of semantic information. Therefore, in this chapter, a new tag-set is proposed. The problems encountered in establishing the tag-set are also discussed.

3.1 Objectives

3.1.1 Resolving Parsing Ambiguities

The new tag-set provide the semantic information required to resolve parsing ambiguities in syntactic parsers. Part-of-Speech (POS) tags are still used for parsing as most grammar rules are defined with POS tags. However, if ambiguous noun phrases such as conjunctive noun phrases or De-de noun

phrases are encountered, the semantic tag will then be applied. The goal is to improve the performance of existing syntactic parser without re-defining a new set of grammar rules.

3.1.2 Investigation of Nominal Compound Noun Phrases

Nominal compound noun phrases are most ambiguous. Most Chinese research works treat them as separate grammatical units. If semantic information is provided, it is possible to isolate them from a sentence, This reduces parsing errors and enhances the efficiency of noun phrase indexing in Chinese information retrieval [19].

3.2 Definition of Hybrid Tag-set

A hybrid tag consists of two parts. The first part is the syntactic tag and the second part is the semantic tag. The syntactic tag adopted is the Part-of-Speech (POS) tags developed by the Tsinghua University and the semantic classes in Cilin [13] are used for the semantic tags.

The following is the formal representation for a hybrid tag:

<POS | SC>

where POS is the part-of-speech tag and SC is the semantic tag.

3.3 Introduction to Cilin

Cilin <<同義詞詞林>> [13] is a semantic dictionary (or thesaurus) for daily use. It provides the readers the semantic tags of 56,000 Chinese words. The semantic of a word is expressed either by (a) its synonyms or antonyms; or (b) phrases or sentences describing the meaning of the word. In addition to semantics, homonyms, polysemous words and part-of-speech tags may be provided for each word. K.T. Lua has studied this thesaurus in some detail[25, 27, 28, 29].

Cilin classifies about 70,000 Chinese words according to a three-level semantic tree structure. This hierarchical structure reflects the semantic relationship between words. It is defined by 12 major (top level), 95 medium (middle) and 1428 minor (bottom level) semantic classes. Figure 3.1 is an example from the thesaurus. In the figure, the hierarchy spans from right to left, i.e. A-L are the 12 major classes, Aa-An are examples of the middle classes and Aa01-Aa06 are examples of the minor classes. In practice, word may be associated with more than one major class. Each major class may have multiple middle classes; and each middle class may in turn have multiple minor classes. However, there are still two shortcomings concerning the use of Cilin as semantic reference.

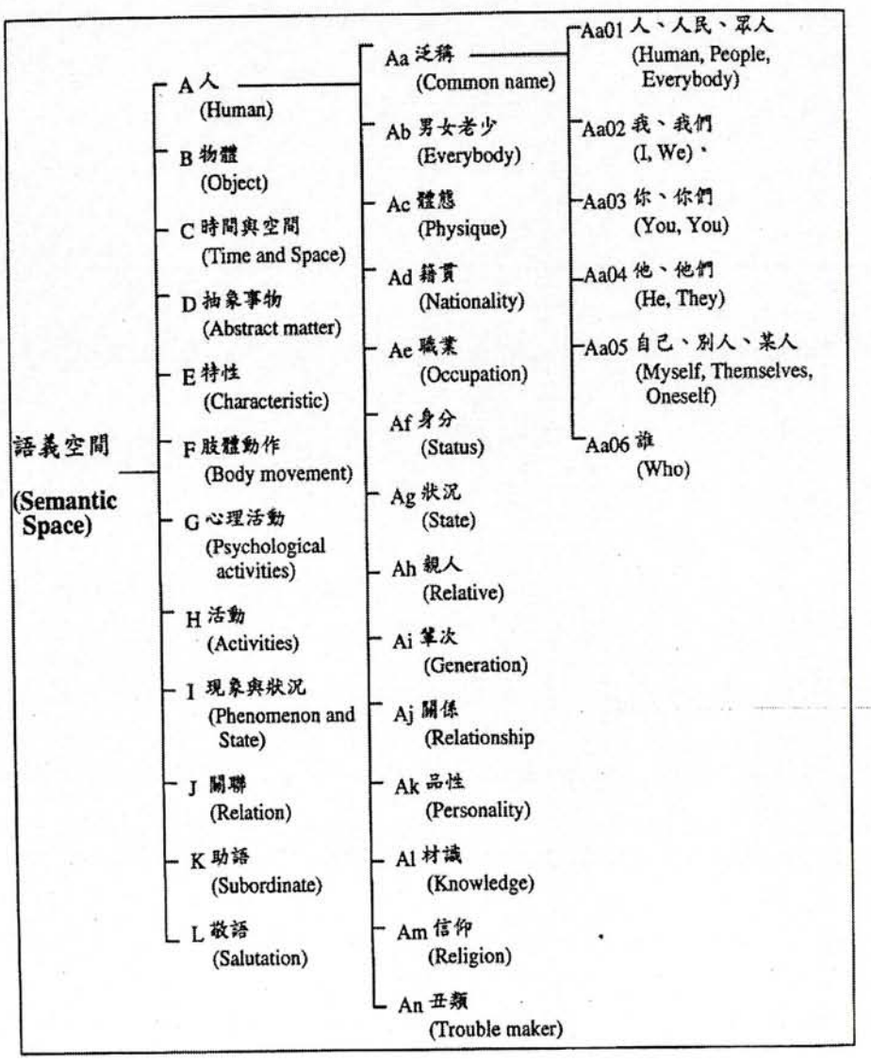

Figure 3.1: *An Example from the Cilin*

No. of Syllabic in Word	Word found in Cilin	WORD absence from Cilin
1	24103	3510
2	19013	10738
3	422	2100
4	210	671
5	5	372
6	0	94
7	23	79
8	0	28
9	0	35
10	0	28
More than 10	0	4

Table 3.1: *Distribution of unknown words in corpus*

3.4 Problems with Cilin

3.4.1 Unknown words

Cilin was developed in the last decade by Chinese lexicographers. Unlike English, Chinese characters have much flexibility to form new words by combination with characters or words. Due to culture development or technological innovations, many new words have been created as time passes. They are, however, not accounted for in the Cilin. These new words are referred to as unknown words in our research. To verify this, a corpus of 61435 Chinese words [1] is analyzed. The result is as below.

- The distribution of the number of syllabics in a Chinese word is uneven.

[1]The definition of a word depends on the word segmentation algorithm used. Theoretically, different thesaurus would be employed instead of the Cilin.

Most Chinese words are disyllabic. Focusing on the disyllabic words, 36.09% of words are absent from Cilin. The reason lying behind is due to the semantic transformation of Chinese word in their formation. This will be explained later.

- Most polysyllabic words are unknown words. This can be justified by the high flexibility of Chinese in word formation. A Chinese word is usually formed by the concatenation of two or more constituent words, which themselves can be one (monosyllabic), two (disyllabic) or more (poly-syllabic) characters in length.

- 29% of the Chinese words are missing from Cilin. This hinders the widespread adoption of Cilin for Chinese natural language processing. To alleviate this hindrance, one must assign a semantic class to each unknown word when it appears, e.g. in the middle of parsing. At present, this is done manually. This is error-prone and often leads to data inconsistency as different people may use different criterion in semantic class assignment. Hence, an automatic approach (SAUW) is proposed to improve this situation. The goal is to find a best matches of semantic classes for an unknown words.

- Chen Keh-Jiann and Chen Chao-Jan have also studied the formation of Chinese unknown words [17]. They conclude that unknown words

are usually numbers, proper nouns, abbreviations, compound nouns and words borrowed from foreign languages. Unknown words can be classified into two categories. They are the close set and the open set. Close set refers to words that are infinite in number, but all of them bear common regular expressions. For instance, compound nouns that are related to numbers such as or belong to this type. For open set, unknown words within this category are hard to be analyzed with regular expressions. It implies that some of the unknown words are of pre-defined structures. Morphological rules can be defined for semantic assignment to these unknown words.

3.4.2 Multiple Semantic Classes

Some words in Cilin have multiple semantic classes. For example, "人" has semantic classes Aa, Ab, Dd, De and Dn. The statistic about multiple semantic classes is depicted in Table 3.2.

From Table 3.2, Chinese words have an average of 2.45 entries in Cilin. It is observed the more syllabic a word contains, the clearer will be its meaming. Average number of semantic classes for monodisyllabic words is 3.73. However, the average number of semantic classes is 2.11 for each 5-syllabic

	No. of Semantic Class per Word				
No. of Syllabic per Word	1	2	3	4	5
1	51 (0.08%)	1820 (2.95%)	1607 (2.60%)	1264 (2.05%)	3267 (5.29%)
2	632 (1.02%)	28196 (45.63%)	7402 (11.98%)	1653 (2.68%)	479 (0.78%)
3	115 (0.19%)	5680 (9.19%)	427 (0.69%)	75 (0.12%)	8 (0.01%)
4	206 (0.33%)	6564 (10.59%)	1462 (2.37%)	219 (0.35%)	12 (0.02%)

Table 3.2: *Syllabic to Semantic Class*

word.

In order to choose a semantic class for the target word among multiple classes, the information from the other words within the sentence should be considered. This is regarded as word-sense disambiguation [3, 4, 18, 36, 35, 39]. This problem will be discussed later.

3.5 Introduction to Chinese Word Formation

3.5.1 Disyllabic Word Formation

Wang has concluded six different approaches in Chinese Word formation [38]. They are as below:

- Pictorial (象形)
  - Words belonging to this type are created by simple concatenation of words.
  - Examples: 山水, 牛馬
- Indicative (指事)
  - It is to create words with combination of characters with appropriate meanings. Most of these words under semantic change, which means that the final meaning is normally unrelated to the meanings of the constituent characters.
  - Examples: 上海
- Semantic Aggregates (會意)
  - Words of this type are created by combining the smantics of its constituent characters.
  - Examples: 白雲, 藍天
- Pictophonetic (形聲)
  - This refers to words whose pronounctions of the constituent characters are used to describe the intended sounds.
  - Examples: 哇哇, 哈哈

- Derived, adaptive (轉注)
  - Words of this type have a semantic change of the meanings from the semantic of the constituent words
  - Examples: 高堂, 犬馬
- Borrow (假借)
  - These words are borrowed from foreign languages. There are also words that arise from technological advancement.
  - Examples: 可樂, 的士

3.5.2 Polysyllabic Word Formation

Most Chinese boun phrases are head-final [15] the same situation is found for Chinese polysyllabic words. The reason is that most polysyllabic words are formed by concatenation of two or more constituent words. The semantic of the whole word is usually the head of the polysyllabic word. The following examples depicts the observation and the words in < > are the head.

- 系統<工程>
- 消防<車>

3.5.3 Observation

Chinese characters themselves have high information to the word they form. When we decided to obtain the semantic of a newly encountered words, we can extract some useful meanings from the semantics of its constituent characters.

Considering disyllabic words, most of them are formed with the Pictorial or Semantic Aggregates approaches. There are four possible combinations in their formation [28]. These four types are as below:

1. X + X → X

    - It refers to two constituent characters of the same semantic formed a new word of the same meaning.
    - Examples: "士兵" and "兄弟"

2. X + Y → X

    - It refers to two constituent characters of the different semantic formed a new word of the same meaning as the first character.
    - Examples: "工程" and "利潤"

3. X + Y → Y

No. of words	Type 1	Type 2	Type 3	Type 4
33843	11132 (32.89%)	6059 (17.90%)	11792 (34.84%)	4860 (14.36%)

Table 3.3: *Distribution of Di-syllabic Words formations*

- It refers to two constituent characters of the different semantic formed a new word of the same meaning as the second character.
- Examples: "內戰" and "列車"

4. X + Y → Z

- It refers to two constituent characters form a new word with semantic unrelated to the constituent characters.
- Examples: "人馬", "千金" and "人煙"

In order to investigate the distribution of these four categories of di-syllabic words, 33,843 di-syllabic words from the Cilin were extracted. Since their semantics were known in advance, it was possible to classify them into these four categories. The result is shown in Table 3.3.

If we are going to derive the meaning of unknown words from the constituent chatacters, the algorithm should be able to decide whether the semantic of the first or second constituent characters is selected as the final meaning of the unknown word. On the other hand, words formed with the

fourth type of word formation belongs to Indicative, Derived or Borrow approaches. the semantics of them will be unrelated to their constituent characters or words. Especially for words with Dervied approaches, extra reference will be required to understand their meanings. Manual semantic assignment will be needed if an unknown word is of this type, but there should be some way to identify them out of a sentence in advance.

For polysyllabic unknown words, their meanings can be derived from their heads.

3.6 Automatic Assignment of Hybrid Tag to Chinese Word

The auotmatic assignment of a Hybrid Tag to a Chinese word is depicted in Figure 3.2. There are two main workflows in the whole process. The first one is syntactic tagging and the second is semantic tagging. The Part-of-Speech tagger developed by the Tsinghua University is adopted. It has an accurcy of up to over 90% and it will not be discussed in details. However, more effort is put on semantic tagging. The first process is process 2, which is to consult the Cilin directly. As mentioned in the previous section, there are many unknown words that are not found in the Cilin. Therefore, these

unknown words are assigned with possible semantic classes in process 3. This process will derive the semantics of the unknown words from their constituent characters or words. It also involves the manipulation of words that has undergone semantic changes.

We assume that that every noun phrase is segmented in advance. The definition of a word depends on which word segmentation algorithm is employed. That implies that although the Cilin is adpted, different thesaurus could be used for unknown word identification in our research. In our context, any word that does not appear in the Cilin is an unknown word. Therefore, an unknown word in the Cilin may not be unknown in the word segmentation dictionary. Effectively, Our research overcomes the inconsistency between the two linguistic resources. For example, consider a di-syllabic word AB results from word segmentation, i.e. it is a word found in the dictionary. it is an unknown word in the Cilin, a set of semantic classes will be assigned to AB through our semantic assignment process. The result is AB will be added to the Cilin. As a result, AB will no longer be an unknown word when re-appears.

After process 3, a set of best matches of semantic classes are assigned to each word. The next step is to select one semantic class out of them. It is known as word-sense disambiguation in process 4. Finally, the POS tag and the semantic class combine together to form the hybrid tag.

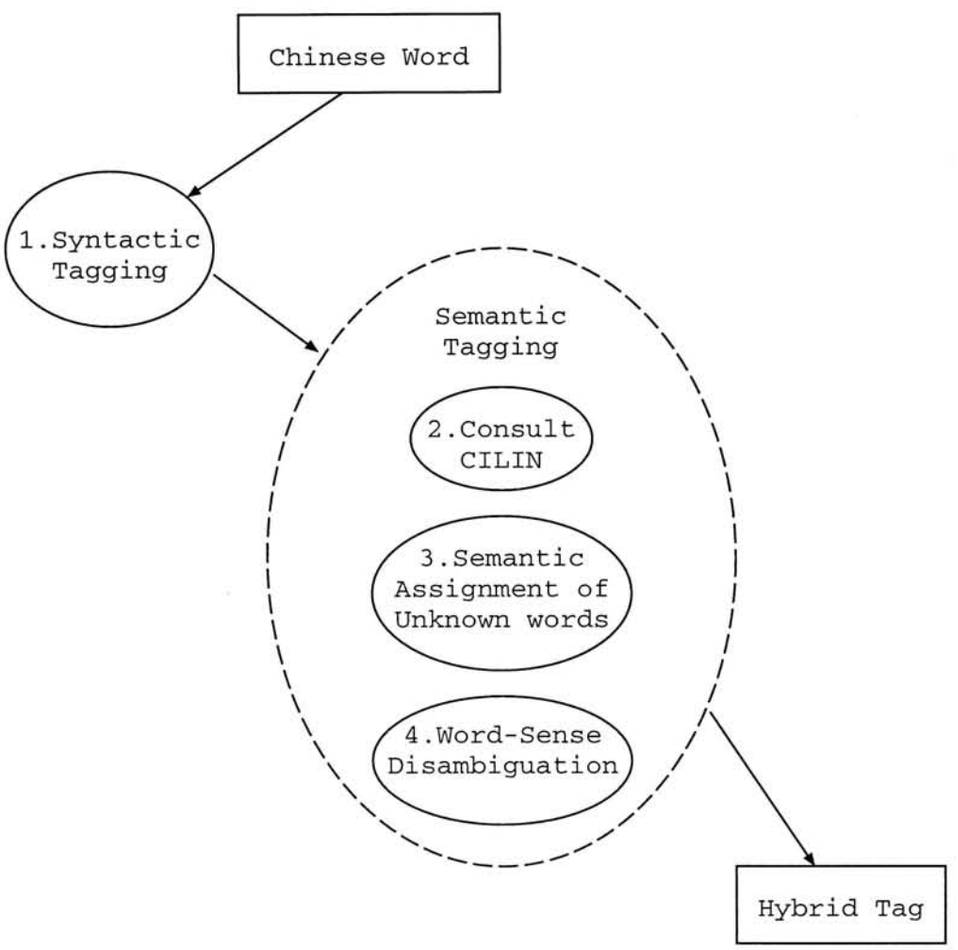

Figure 3.2: *Automatic assignment of Hybrid tag to Chinese word*

3.7 Summary

The hybrid tag-set is presented in this chapter. Each hybrid tag consists of two parts. The syntactic part is the part-of-speech, while the semantic part is the semantic class from the Cilin. Two problems concerning the use of the Cilin as semantic sources are pointed out. They are the unknown word problems and the problem of multiple semantic classes for a single word. The unknown word problem refers to the absence of entries from the Cilin. In an addition, the formations of Chinese words are also analyzed. It is shown that the semantic of an unknown word can be derived from the meaning of their constituent characters/words. Based on the above observations, a mechanism for automatic assignment of hybrid tag to Chinese word is proposed. There are two major processes in this mechanism. The frist one is "Syntactic Tagging", while the second "Semantic Tagging". The "Semantic Tagging" will be discussed in the next chapter in much details.

Chapter 4

Automatic Semantic Assignment

A hybrid tag consists of two parts. They are the syntactic tag and the semantic tag. In this chapter, semantic tagging is described. Cilin [13] is chosen as the semantic reference. Although it is a popular linguistic resource, its usage in natural language processing (NLP) has been seriously affected by the unknown word problem. To tackle this problem, an algorithm for automatic semantic assignment of unknown words (SAUW) is proposed [41]. This strategy makes use of both syntactic and semantic analysis to enrich the Cilin for handling modern NLP applications.

4.1 Previous Researches on Semantic Tagging

K.T. Lua has worked out an inductive unsupervised semantic tagger for Chinese words [26]. It is a statistical-based semantic tagger. The input corpus is first hand-tagged with part-of-speech (POS). Possible semantic tags are selected from the Cilin with respect to the POS and the conditional probability of the semantic from the POS. Finally, a semantic tag is assigned to each word by taking the semantic tags of its predecessor and successor into consideration. A large training corpus consisting of 340,000 words were prepared manually. This was very computationally expensive. Besides, inadequacy of word entries in Cilin introduced the major source of errors.

Chao-Jan Chen, Ming-hong Bai and Keh-Jiann Chen have proposed an algorithm for guessing part-of-speech (POS) categories for Chinese unknown words [6, 8, 17, 24]. In his context, unknown words refer to words that are not contained in the lexicon. They adopt a statistical method to predict the POS of unknown words. It is based on the prefix-category and suffix-category associations. In addition, the mutual information and dice metrics are employed to measure the association strength. Since most Chinese words are head-final, different entropy weightings between prefix-category and suffix-category are imposed for the different dsicrimination abilities on the unknown word categories. The algorithm has an accuracy of nearly 70%. However,

the tagging process takes no contextual information into consideration and it is quite difficult to determine the correct prefix and suffix.

4.2 SAUW - Automatic Semantic Assignment of Unknown Words

Figure 4.1 is the dataflow diagram[1] depicting the overall system architecture of the proposed automatic semantic assignment strategy (SAUM) for unknown words in Cilin [41]. It consists of four different functional stages, namely POS-SC Association, Morphology-based Deduction, Di-Syllabic Word Analysis and Polly-syllabic Word Deduction.

Input to SAUW is a collection of words, which do not exist in the Cilin. These words could be hand picked or automatically extracted from a corpus. In the latter case, they are words successfully determined by the word segmentation process but have no semantic class. The output from SAUW is the same word collection with the words assigned with the appropriate semantic classes.

[1]The DeMarco's dataflow diagram convnetion is used. Rectangular objects reresent external input and output. Circular objects are functional proceses and parallel bars represent data stores.

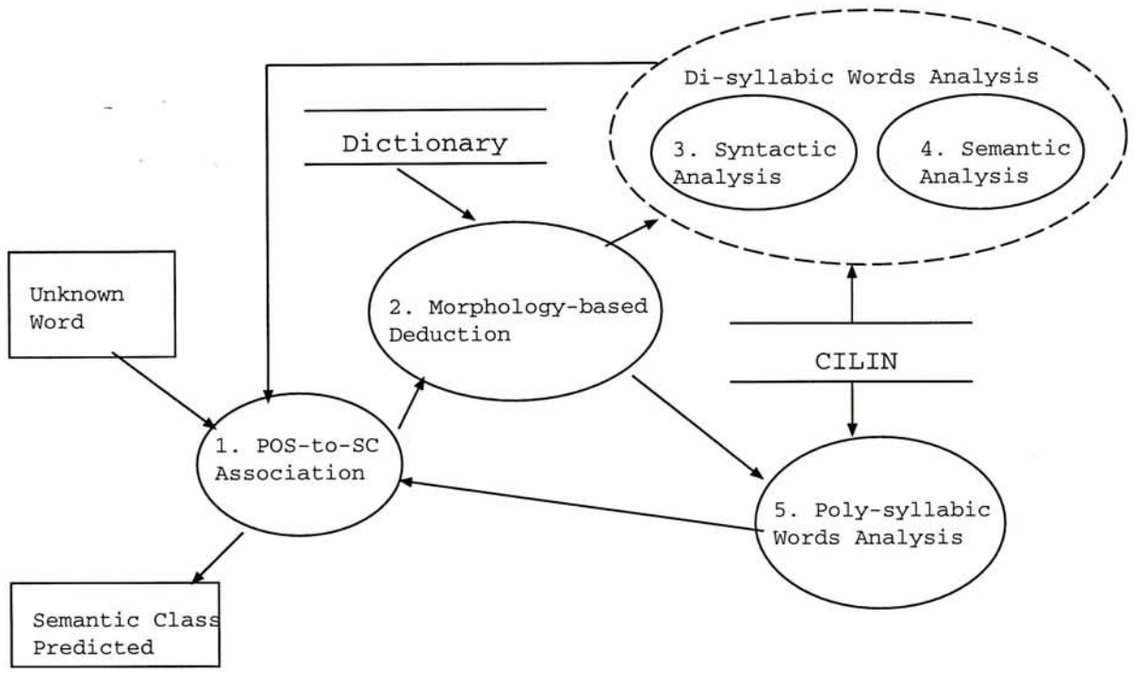

Figure 4.1: *The overall architecture of SAUW*

4.2.1 POS-to-SC Association (Process 1)

Each input word to SAUW is assigned with the appropriate syntactic, i.e. part-of-speech, (POS) tag(s)[2]. The POS tagger developed by Tsinghua University is used [2] (see also Appendix A). This tagger makes use of 109 POS tags. To deal with mulitple POS tags associated with one word, the POS tag with the highest occurence frequency is selected.

Syntactic tags give valuable clues in semantic class assignment. In practice, certain POS tags are always associated to the same semantic classes. For

[2]The same dictionary is used for word segmentation and POS tagging. Thus each word input to SAUW must have one or more or tag(s)

example, a time word is always semantically classified under "Ca". Table 4.1 summarizes the POS to semantic class (POS-to-SC) association [16, 23, 34]. On average, 37% of the unknown words can be marked with this association method. Using Table 4.1, the semantic class of an individual word can be easily looked up.

This POS-to-SC association is applied to unknown words twice in the whole mechanism. It is used to choose a single possible semantic class to the unknown word in the beginning. If there is no match in the POS entry or multiple entries in semantic classes are detected, the unknown word is fetched to the next process. Finally, this POS-to-SC association is re-applied to the unknown word. It is aimed to filter out the irrelevant semantic classes from the best matches of possible semantic classes, which are derived from our SAUW algorithm.

4.2.2 Morphology-based Deduction (Process 2)

Unknown words, which are unresolved from the previous stage, are checked against a set of morphological rules. Each rule represents a word pattern and depicts which part of the pattern contributes to the overall meaning of the word. For example, given the pattern "AABB" and an input word "安安定定", the semantic class of "安定" will also serve as the overall semantic

Part of Speech Tags	Semantic class in Cilin
nf, npf (Surname & Name of Person)	Aa
npu (Name of Organization)	Dm
m (Number of Words)	Dn
b, s (Name of place)	Cb
t (Time words)	Ca
i (Idioms)	Oi *
j (Abbreviations)	Oa *
f (Direction Words)	Cb
n (noun)	A, B, C, Da-Dm, Dn01-Dn03, Dn05-Dn07, E, F, Ga, Gb, H, I
r (pronoun)	A, B
v (verb)	F, G, H, I, J
d (adverb)	Ca, Cb, Dn05, Ka
p (preposition)	Kb
c (conjunction)	Kc
y, o, e (model word or interjection)	Ke, Kf
a (adjective)	E, Ga
u (particle)	Kd
x (Non-Chinese Words)	Ot *
* Oi, Oa and Ot are newly defined semantic classes	

Table 4.1: *Part-of-Speech to semantic class association.*

class of the word "安安定定". Notice that if "安定" itself is unresolved, the SAUW algorithm could be recursively applied.

Currently, there are 78 morphological rules (See Appendix B). They are divided into 4 categories. Following are some examples. Except for reduplicates, the pattern embraced in "<>" in a rule is the overall semantic class.

At the end, the remaining unresolved words will either be passed to the third or fourth stage depending on the length of the word.

4.2.3 Di-syllabic Word Analysis (Process 3 and 4)

Syntactic analysis is applied to the di-syllabic unknown words. But some words may belong to more than one syntactic structure. When this happens, semantic analysis will be applied.

Syntactic Analysis (Process 3)

Syntactic patterns are defined for short phrase formation [12, 16]. These patterns determine the structures of the phrases. When such a pattern is applied to di-syllabic word formation, it will take the form: <POS of the first character> <POS of the second character>. According to its general usage, each pattern is associated with a rule, which indicates whether one should select the first POS or the second as the core of the word. This, in turn, can help the system determine the overall semantic class of the di-syllabic

- Prefix
  - [上<X>,方位詞], [小<X>, 名詞], [大<X>, 名詞], [某<X>, 代詞], [本<X>, 代詞], [很<X>, 副詞], Etc.
  - For example, the word "大洲" contains "大" as the prefix and has the part-of-speech "名詞". The semantic body will be "洲".
- Suffix
  - [<X>子, 名詞], [<X>化, 名詞], [<X>性, 名詞], [<X>了, 助詞], [<X>上,方位詞], [<X>起, 動詞], Etc
  - For example, the word "提起" contains "起"as the suffix and has the part-of-speech "動詞". The semantic body will be "提".
- Negation
  - [不<X>], [非<X>], [未<X>], [不<A>不<B>], Etc
  - The negation character is segmented from the new words. For example, the word "不久" is segmented into "不" and "久".
- Reduplication
  - [AA, 形容詞, 量詞, 動詞, 副詞, 狀態詞, Semantic Body: A], [AABB, 形容詞, 動詞, 副詞, Semantic Body: AB], [又A又B, 副詞, Segment into A and B], Etc

Table 4.2: *Four categories of morphological rules*

qn(量名詞): q(量詞) + <ng(名詞)>	v(動詞): <v(動詞)> +vc(動補詞)
vg(動賓詞): <v(動詞)> + ng(名詞)	ng(名詞): a(形容詞) + <ng(名詞)>
qn(動量詞): v(動詞) + <q(量詞)>	ng(名詞): <ng(名詞)> + <ng(名詞)>
v(動詞): <v(動詞)> + ng(名詞)	v(動詞): d(副詞) + <v(動詞)>
a(形容詞): <a(形容詞)> + <a(形容詞)>	f(方位詞): ng(名詞) + <f(方位詞)>
ng(名詞): p(介詞) + <ng(名詞)>	v(動詞): <v(動詞)> + <v(動詞)>

Table 4.3: *Typical examples of syntactic rules*

unknown word. Altogether there are 140 word formation patterns/rules extracted from a training corpus (see Appendix C). Extraction was done manually. The following are some typical examples. The part-of-speech in "< >" is the semantic body of the whole word [1, 5, 11].

Example: Given the word "促成", which is missing from the Cilin. "促成" has the syntactical structure "vgv: vgv + vc". In the SAUW rule set, "vgv" will be selected; hence the first character "促" will be chosen for determining the overall semantic class of the word.

Another application of the syntactic patterns/rules is for detection of exceptional unknown word. As such, any di-syllabic word, whose formation pattern cannot be found in the SAUW rule set, is deemed to have undergone semantic changes, i.e. its semantic can no longer be deduced from its con-

stituent characters. For example, the semantic of "的士" is not related to "的" or "士". It belongs to the syntactical pattern "ng: usde + ng", which is not found in the current rule set.

However, ambiguities occur in some rules. For example, it is uncertain whether one should select the first or the second character if the syntactic pattern is "v: v + v". To improve this situation, the semantic of the constituent words are analyzed.

Semantic Analysis

Chinese characters have different semantic strength when they combine to form a new word. Some characters have a higher tendency to preserve their semantics over their partners in the new di-syllabic words. Given a di-syllabic word formed by two characters A and B, two semantic strength values are defined:

$$SS\_FRONT = \frac{\text{Frequencies character A as the overall semantic in a di-syllabic word AX}}{\text{Size of } SC_{AX}} \tag{4.1}$$

$$SS\_BACK = \frac{\text{Frequencies character B as the overall semantic in a di-syllabic word XB}}{\text{Size of } SC_{XB}} \tag{4.2}$$

SC_{AX} is the semantic class of all di-syllabic words starting with the A character; and SC_{XB} is the semantic class set with all di-syllabic words ending with the character B. Based on SS_BACK and SS_FRONT, the following algorithm is proposed:

```
If (SS_FRONT - SS_BACK) > Threshold then
        semantic-body = A
else
        semantic-body = B
end if
```

33,843 unique disyllabic words from Cilin were extracted for the calculation of the semantic strength. The example below depicts the procedure and the middle class in Cilin was used.

Given a character:	庭
SC_1:	Bn, Dm
[H] Disyllabic words beginning with 庭:	庭長, 庭訓, 庭院, 庭除, 庭燎, 庭闈
SC_{1X}:	Af, Df, Bn(2), Bp, Ah
SS_FRONT for 庭 =	0.333

Threshold Value	Precision
0	63.51%
-0.05	63.45%
-0.075	63.51%
-0.1	63.68%
-0.125	63.8%
-0.15	63.28%

Table 4.4: *Variation of threshold value to percentage of correctness*

Disyllabic words ending with 庭:	天庭, 法庭, 家庭, 訟庭, 開庭, 鯉庭, 椿庭
SC_{X1}:	Bk, Dm(2), Dp, Hm, Dk, Ah
SS_BACK for 庭 =	0.286

In order to decide the optimal value for the threshold, another 3475 disyllabic words from Cilin were analyzed. These words, whose semantics are known, were tested with the above algorithm. The results are shown in Table 4.4.

The low percentage of correctness was due to the fact that no morphology-based deduction and syntactic analysis was used in this test. Therefore, unknown words that undergo semantic change contributed signifcantly to the error. At present, the threshold value is set to -0.125, which is the best value.

4.2.4 Poly-syllabic Word Analysis (Process 5)

It was observed that most poly-syllabic words are head final [15]. For this reason, the semantic of a poly-syllabic word often appears in its nominal head, i.e. the last constituent word or character. This observation forms the basis of semantic assignment of poly-syllabic unknown words.

Consider a poly-syllabic word that does not appear in the Cilin. It will be re-partitioned for locating the constituent words. This is similar to un-grouping a long word into smaller word segments. A dictionary-based maximum matching approach was used[3]. Based on this approach, the word pattern containing the longest constituent word is selected. For example, the poly-syllabic word "辦工室女郎" will be re-partitioned to "辦工室" and "女郎" The nominal head "女郎" is then used as the overall semantic of the word.

4.3 Illustrative Examples

Example 1 Unknown words resolved by POS-to-SC association

- "美國", part-of-speech: "s" (name of place).

[3]In this context, the dictionary is Cilin.

- "1 9 5 2 年", part-of-speech: "t" (time words).

Example 2 Unknown words resolved by morphology-based deduction

- "提起" contains "起"as the suffix and has the part-of-speech "v" (動詞). The semantic body will be "提".

- "不久" is segmented into "不" & "久".

Example 3 Unknown words resolved by syntactic analysis

- "再度"is identified by "d(副詞): <d> + qnm(數詞)". The semantic body is "d" and it is "再".

Example 4 Unknown words resolved by semantic analysis

- "'魚類". SS_FRONT (魚) is 0.222 and SS_BACK (類) is 0.182. The difference between SS_FRONT and SS_BACK is 0.041, which is greater than the THRESHOLD (-0.125). Therefore, "魚" is selected as the semantic body.

Example 5 Poly-syllabic unknown word

- "自動系統" is re-segmented into "自動" and "系統". The head is "系統" and functions as the semantic body.

- "反坦克炮" is re-segmented into "反", "坦克" and "炮". The head is "炮" and functions as the semantic body.

4.4 Evaluation and Analysis

4.4.1 Experiments

Objective

The objective of this experiment is to evaluate the performance of the SAUW algorithm.

Procedure

The testing corpus is from the Hua Xia Wen Zhai (華夏文摘). There are 61,453 Chinese words and it is assumed to be correctly segmented. The middle classes (95 classes) in Cilin are used for the evaluation. The semantic strengths are gathered from 33,843 di-syllabic words in the Cilin.

	Unknown Characters	Di-syllabic Words	Poly-syllabic Words
Unknown Words	1380	8249	3223
Unique Unknown words	136	2154	1696
POS-to-SC Association	85	537	925
Morphology-based Deduction	NIL	235 (13)	144 (24)
Di-syllabic words analysis	NIL	774 (202)	NIL
Poly-syllabic Analysis	NIL	NIL	627(36)
Unknown word remained	51	608	0
Words handled	83.47%		
Correctness	91.73%		

Table 4.5: *Results for semantic assignment to new words*

Results

Table 4.5 gives the distribution of the new words from the corpus and the result of semantic assignment. The number of errors is embraced in "( ) ".

Total Errors	Errors from wrong POS tags	Errors from Syntactic Analysis	Errors from Semantic Analysis	Errors from Specific Domain
275	23 (8.36%)	130 (42.27%)	52 (19.05%)	70 (25.64%)

Table 4.6: *Error distribution for semantiic assignment to new words*

4.4.2 Error Analysis

There are several sources of errors. The error distribution is shown in Table 4.6. First, the words in the corpus were tagged with part-of-speech tags. Since the part-of-speech tagger was based on a probabilistic model, some errors are inevitable. This contributes much to the error in our algorithm as we rely heavily on the syntactic information. Besides, it was observed that many Chinese words tend to use homonyms to replace the correct constituent characters. They are sometimes mixed used of Feng Ti (繁體) and Jian Ti (簡體) characters. For example, "苦幹" was changed into "苦干". The semantic of "干" is much deviated from that of "幹". 369 unknown words are belong to this type. These words are manually rectified before the SAUW. They are not accounted for in the total errors. Lastly, some of the words were too specific and domain dependent, For examples, words like "核糖" , "液泡" or "電阻" . These also lowered the accuracy of our algorithm.

4.5 Summary

Many modern words are missing from the Cilin. Not unless they are catered for, the usability of Cilin will be affected. In this chapter, a new approach for automatic semantic class assignment to unknown words from the Cilin is proposed. First, the POS and semantic class association is considered. The morphological compositions of the unknown words are then analyzed. Lastly, syntactic and semantic analyses are performed. It was shown that the semantic classes of 80% of the 3,986 unknown words from a corpus of 61,453 could be successfully predicted. The missing 20% was mainly due to semantic changes in word formation and the use of homonyms for the original constituent characters . Now, each unknown word is assigned with semantic classes from the Cilin. Since multiple semantic classes may associate with a single unknown word, the next step is to select the best semantic class out of those assigned to the unknown word. This process is called word sense disambiguation and it will be discussed in the next chapter.

Chapter 5

Word Sense Disambiguation

In Chapter 3, it was pointed out that many Chinese words have more than one meaning. After the automatic semantic assignment of unknown words, the unknown words problem in the CILIN was solved. The next process is to solve the problem of multiple semantic classes associated with a single word. It is known as Word Sense Disambiguation (WSD). This determination of sense (semantic class) of ambiguous word (word with more than one semantic class) is already a major issue in Natural Language Processing. In this chapter, the problem of WSD will first be defined and an overview of some word-sense disambiguation scheme will be described. Our approach for WSD is then introduced. Finally, the proposed algorithm will be applied for WSD and the results will be discussed.

5.1 Introduction to Word Sense Disambiguation

Many Chinese words have more than one meaning. Words with multiple meanings are generally known as "Homonyms". Their existence in sentence parsing inevitably lead to word sense ambiguities. Consider the following example, the sentence is tagged with the hybrid tag-set with format in "<POS | SC>".

- Original sentence:

  漁政<ng | Oa Da Di> 管理<vg | Hc> 系統<ng | Db Dd> 的<usde | kd> 開發<vg | Hd He> 課題<ng | Dk>

The word "漁政", "系統" and "開發" are homonyms. Only one semantic class should be assigned to each of them in the hybrid tags. Word sense disambiguation is then performed. The goal of word sense disambiguation (WSD) is to assign the correct categories to a word from a range of possible senses given in a dictionary or thesaurus, which is most relevant to the context of the text. The disambiguation process not only involves one's understanding of the word, but also the contextual information in the sentence. Therefore, the senses of the neighboring words should also be taken into consideration The result for the above example is shown below:

- Resulting sentence:

  漁政<ng | Da> 管理<vg | Hc> 系統<ng | Dd> 的<usde | kd> 開發<vg | Hd> 課題<ng | Dk>

5.2 Previous Works on Word Sense Disambiguation

In this section, existing WSD algorithms are described [36]. They can be classified into 2 categories. They are linguistic-based approaches and corpus-based approaches. They have their own advantages and disadvantages. For linguistic-based approaches, they work on standard linguistic resources such as dictionary or thesaurus. No training is required. However, all linguistic sources are incomplete and it is impossible to define all the rules required. On the other hand, corpus-based approaches incorporate statistical data on the co-occurrence of the word senses. Sizeable corpus is required to achieve a high accuracy. It is quite computational expensive to build the co-occurrence statistics for the word senses.

5.2.1 Linguistic-based Approaches

The work of Y. Wilk's work [14] uses relatedness measure to determine word similarity for English. The relatedness measure of 2 words was in fact the co-occurrence statistics within the Longman Dictionary of Contemporary English (LDOCE). For instance, given the ambiguous word "ball", if "ball" and "spot" appeared together in the dictionary more frequently than "ball" and "dancing", the former word pair has a higher relatedness measure than the latter. This model has a accuracy up to 45%.

Y. Wilks and Mark Stevenson also proposed a WSD algorithm that performed unrestricted word disambiguation for English words [39]. They claimed that all NLP tasks, including WSD, are knowledge-dependent. They could not be solved in a modularized fashion, i.e. syntactic followed by semantic analysis. Different knowledge must be incorporated including semantic preferences, dictionary definitions and part-of-speech tags. The WSD algorithm was designed to assign sense tags (i.e. semantic categories) obtained from lexicon to words in general text. First, the input word was assigned a part-of-speech tag using the Brill's part-of-speech tagger. Secondly, the word was tagged with a set of suggested senses from a dictionary using dictionary word overlapping and simulating annealing. A hierarchical thesaurus was then used to provide semantic information. Afterward, a set of selection

rules was applied to restrict the choice of the sense. Each selection rule listed the possible sense expected for each syntactic type. Lastly, a decision list was trained to resolve the ambiguities and an accuracy of 92% was reported.

Eneko Agirre and German Rigau presented another WSD approach for English words using conceptual distance [3, 4]. The method relied on the use of the noun taxonomy of the WordNet (a hierarchical semantic net) and the notion of conceptual distance among concepts, which was captured by a conceptual density formula. The conceptual distance was to provide a basis for determining closeness in meaning among pairs of words. With it, the algorithm resolved lexical ambiguities of nouns by finding the combination of senses from a set of contiguous nouns that maximizes the total conceptual density among senses. A precision and recall of 71.2% and 61.4% were reported, respectively. However, contextual information, which could affect the accuracy of the algorithmwas, was ignored in the WSD process, .

Niwa and Nitta used distance vector to determine word similarity [42]. A distance vector measured the level of reference between two words. For example, given the definitions of the two words "Library" and "Book" from the dictionary.

- Library
  - A collection of books for reading and borrowing.
- Book

– A series of written or printed or plain sheets of paper fastened together at one edge and enclosed in a cover.

Since the definition of "Library" contains "Book", the distance vector between them is 1. Similarly, the distance vector between "Library" and "paper" is 2 because "Library" contains "Book" and "Book" contains "paper". They applied the distance vector approach to disambiguate the sense of 9 Japanese words. Niwa also applied the co-occurrence statistics approach to the same set of words. The results of linguistic-based (distance vector) and corpus-based (co-occurrence statistics) approaches are 60% and 100% in term of accuracy, which implies the co-occurrence statistics approach was more effective than the distance vector approach. However, since the test set was rather small, the experiment was hardly comprehensive and the results were not a good performance indicator.

5.2.2 Corpus-based Approaches

Harder modified Wilks technique on relatedness measure by using co-occurrence statistics to measure word similarity [20]. Syntactical information was also used in Harder's algorithm. Experiments on 4 English verbs showed a accuracy rate of 50%.

Shing-huan Liu, Keh-Jiann Chen, Li-ping Chang and Yeh-Hao Chin presented an automatic part-of-speech (POS) tagging algorithm for Chinese cor-

pora [24]. It was based on a hybrid approach. Disambiguation rules were applied to solve long distance dependency problems in most probabilistic POS tagger. The WSD component in the algorithm was a probabilistic-based model based on the relaxation labeling method. The bi-gram model was used in the relaxation labeling method. During each assignment to the Chinese words, relaxation technique selected the best label among several possible choices with respects to the local constraints between neighboring labels. The relaxation method gave an accuracy of about 70% for WSD and the tagger had an overall accuracy of more than 80%.

As discussed in the previous chapter, K.T. Lua proposed a semantic tagger [30]. The first process was to assign seven best matches of semantic classes to a word. The information used were conditional probabilities of semantics from POS and semantics from word. The WSD process designed to select the best semantic out of the matches. Tri-gram model is used which the tags before and after the current one were considered. Scores were assigned to each combination. It was the conditional probabilities of the semantic bi-grams among each combination, which were weighted by the conditional probabilities of the semantics from the word under consideration. The overall accuracy of this tagger was 91%. Unknown words from the CILIN contributed much to the errors. Moreover, only tags of the neighboring word(s) were considered in each run. Therefore, the long distance dependency problems also caused

some errors.

5.3 Our Approach

Consider a sentence T, which is composed of a sequence of n words:

$$\mathrm{T} = W_1\ W_2\ W_3\ \ldots\ W_{i-1}\ W_i\ W_{i+1}\ \ldots\ W_n$$

where W_i is the i-th word

Suppose W_{i-1}, W_i and W_{i+1} are homonyms (words with multiple semantic classes), which have l, m and n semantic classes, respectively.

$$\begin{aligned}
\text{Semantic classes of } W_{i-1} &= sc_{i-1,1}\ sc_{i-1,2}\ sc_{i-1,3}\ \ldots\ sc_{i-1,a}\ \ldots\ sc_{i-1,l} \\
\text{Semantic classes of } W_i &= sc_{i,1}\ sc_{i,2}\ sc_{i,3}\ \ldots\ sc_{i,b}\ \ldots\ sc_{i,m} \\
\text{Semantic classes of } W_{i-1} &= sc_{i+1,1}\ sc_{i+1,2}\ sc_{i+1,3}\ \ldots\ sc_{i+1,c}\ \ldots\ sc_{i+1,n}
\end{aligned}$$

where $sc_{i,j}$ is j-th semantic class of the i-th word

Our approach [1] is illustrated in Frgure 5.1. It is different from other existing methods [24, 28], whose sc selection of word is only based on the neighboring word(s). It selects an appropriate semantic class for a word with regard to the semantic classes of the complete sentence. It was designed to

[1] It makes use of the POS tagger developed by the Tsinghua University [2] (see also Appendix A)

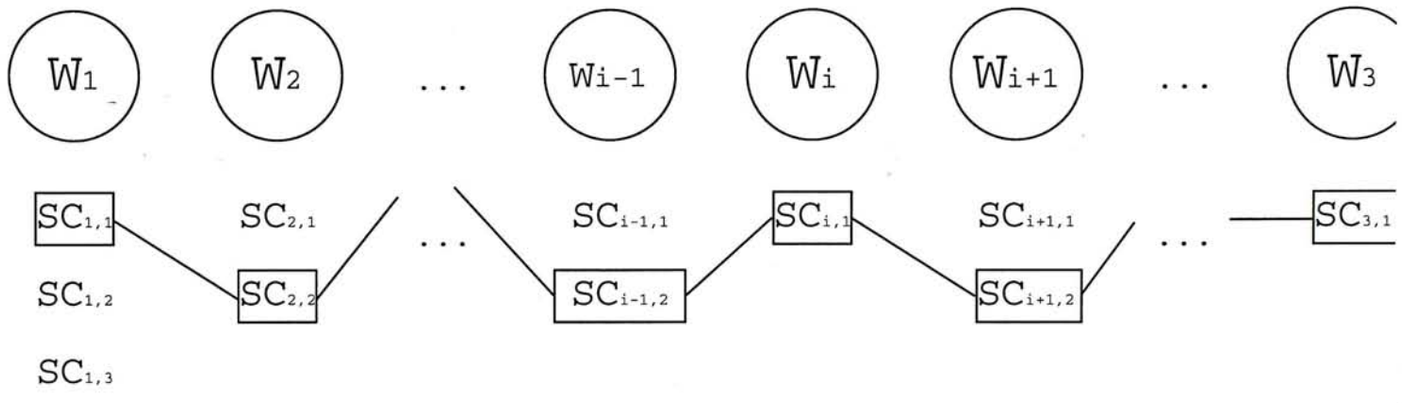

Figure 5.1: *Overview of the WSD*

find an optimal path to link a semantic class from the first word to that of the last word. The reason for this is to solve the long distance dependency problem. Consider the follow example:

- 從 上學 到 放學

The word "到" has six different semantic classes (Ed, Hf, Hj, Ie and Kb) from the CILIN, Hj is selected with a meaning of "arrive", but "到" should be related to "從" and should be tagged with Kb, which has the meaning of "until". Such an assignment cannot be obtained based on simply information from the neighboring words of "到".

Given a sentence with homonyms, each possible semantic class sequence from the first word to the last word is called a "path". All possible "path" are generated. The "path" with the maximum score value under our algorithm is selected and from that, obtain the appropriate semantic class for each word.

Since our algorithm is based on the corpus approach, co-occurence frequencies of the semantic classes should be gathered. We could employ either the bi-gram or the tri-gram model. The maximum score is thus the value obtained from maximizing the objective function illustrated in Equation 5.1 and 5.2. They will be described in details individually in the next subsection.

Bi-gram

$$\max \text{score} = \prod_{i=0}^{n-1} P(SC_i|SCi+1) \tag{5.1}$$

Tri-gram

$$\max \text{score} = \prod_{i=0}^{n-1} P(SC_i|SC_{i-1}SCi+1) \tag{5.2}$$

5.3.1 Bi-gram Co-occurrence Probabilities

The bi-gram conditional probability, $\mathrm{P}(sc_{i,b}|sc_{i+1,c})$ where $0 \leqslant b \leqslant m$ and $0 \leqslant c \leqslant n$, measures the co-occurrence probability that the b-th semantic class of W_i is followed by the c-th semantic class of W_{i+1}. Suppose there are k possible semantic classes in the CILIN. They are labeled from 1 to k as $SC_1, SC_2...SC_k$. A k $\times$ k two dimensional array, $A_{x,y}$ where $0 \leqslant x, y \leqslant k$, is used to record the bi-gram co-occurrence frequencies. An entry in $A_{x,y}$ is

the number of times the x-th semantic class, SC_x, is followed by the y-th semantic class SC_y in the CILIN. Assume $sc_{i,b}$ and $sc_{i+1,c}$ are equal to SC_α and SC_β respectively, the bi-gram conditional probability, $\mathrm{P}(sc_{i,b} \mid sc_{i+1,c})$, is equal to $\mathrm{P}(SC_\alpha \mid SC_\beta)$ and is obtained as below:

$$\begin{aligned} Freq(SC_\alpha) &= \text{The number of times } SC_\alpha \text{ appears in the corpus} \\ &= \sum_{y=1}^{k} A(x=\alpha, y) \end{aligned} \tag{5.3}$$

$$\begin{aligned} Freq(SC_\alpha, SC_\beta) &= \text{The number of times } SC_\alpha \text{ appears in the corpus} \\ &\quad \text{and it is followed by} SC_\beta \\ &= A(x=\alpha, y=\beta) \end{aligned} \tag{5.4}$$

$$\begin{aligned} P(SC_\alpha|SC_\beta) &= \text{Bi-gram probability for } SC_\alpha \text{ appears before } SC_\beta \\ &= \frac{Freq(SC_\alpha, SC_\beta)}{Freq(SC_\alpha)} \end{aligned} \tag{5.5}$$

5.3.2 Tri-gram Co-occurrence Probabilities

The tri-gram conditional probability, $\mathrm{P}(sc_{i,b} \mid sc_{i-1,a}\ sc_{i+1,c})$ where $0 \leqslant a \leqslant l, 0 \leqslant b \leqslant m$ and $0 \leqslant c \leqslant n$, measures the occurrence probability that the

b-th semantic class of W_i has the a-th semantic class of W_{i-1} as predecessor and the c-th semantic class of W_{i+1} as successor. A k × k × k three dimensional array, $B_{x,y,z}$ where $0 \leqslant x, y, z \leqslant k$, is used to record the occurrence frequencies. An entry $B_{x,y,z}$ is the number of times SC_x is before SC_y and SC_z is after SC_y. Assume $sc_{i-1,a}, sc_{i,b}$ and $sc_{i+1,c}$ are equal to SC_α, SC_β and SC_γ respectively, the tri-gram conditional probability, P($sc_{i,b}$ | $sc_{i-1,a}$ $sc_{i+1,c}$), is equal to P(SC_β | SC_α SC_γ) and is obtained as below:

$$
\begin{aligned}
Freq(SC_\alpha, SC_\gamma) &= \text{The number of times } SC_\alpha \text{ and } SC_\gamma \text{ appear in the corpus} \\
&= \sum_{y=1}^{k} B_{x=\alpha,y,z=\gamma} \qquad (5.6)
\end{aligned}
$$

$$
\begin{aligned}
Freq(SC_\alpha, SC_\beta, SC_\gamma) &= \text{The number of times } SC_\beta \\
&\quad \text{appears in the corpus and it is followed by } SC_\gamma \text{ in the corpus} \\
&\quad \text{and it follows } SC_\alpha \\
&= B_{x=\alpha,y=\beta,z=\gamma} \qquad (5.7)
\end{aligned}
$$

$$
\begin{aligned}
P(SC_\beta | SC_\alpha SC_\gamma) &= \text{Tri-gram probability for } SC_\beta \text{ appears to between } SC_\alpha \text{ and } SC_\gamma \\
&= \frac{Freq(SC_\alpha, SC_\beta, SC_\gamma)}{Freq(SC_\alpha, SC_\gamma)} \qquad (5.8)
\end{aligned}
$$

5.3.3 Design consideration

We adopted an empirical approach to decide whether the bi-gram or tri-gram model should be used in our WSD algorithm. The details are descirbed in this section.

Experiment

Objective The objective of this experiment was to compare the performance of our WSD appraoch under the bi-gram and tri-gram model. A choice among these two models was made based on the experimental results.

Setup 15 Chinese articles from Hua Xia Wen Zhai (華夏文摘) were chosen randomly. 10 of them were used for training and the remainder for testing. These articles were word segmented and tagged with the POS tag-set from the Tsinghua University. They were then processed by the SAUW to resolve the unknown words. The errors in SAUW were then corrected manually.

Procedure Given a sentence with few embedded homonyms, all possible "paths" were generated and the one with maximum score value was selected.

Results The experimental results for the close and open tests are shown in Table 5.1 and 5.2, respectively.

Close Test			
		Bi-gram	Tri-gram
Articles	No. of noun phrases	Correctness	Correctness
1	17	76%	82%
2	26	73%	81%
3	20	85%	55%
4	17	88%	76%
5	23	86%	52%
6	27	81%	74%
7	31	74%	55%
8	29	62%	52%
9	34	82%	62%
10	25	68%	60%
Overall correctness		77.11%	63.05%

Table 5.1: *The Result of Close test for Word Sense Disambiguation*

Open Test			
		Bi-gram	Tri-gram
Articles	No. of noun phrases	Correctness	Correctness
11	16	19%	0%
12	18	28%	6%
13	14	21%	21%
14	21	54%	0%
15	16	50%	0%
Overall correctness		34.12%	4.71%

Table 5.2: *The Result of Open test for Word Sense Disambiguation*

It was shown that the bi-gram model performed better than the tri-gram model in this experiment. The situation was more significant in the open test. Chen [24] had similar result in his automatic part-of-speech tagger. He explained that such a phenomenon may be due to the strong bi-relationship between two consecutive Chinese words. More experiments are required to vertifiy this. Besides, our experiment was limited due to the small size of the corpus. Although the bi-gram model was selected for our WSD algorithm, further researches are required to compare the performance of the two models.

5.3.4 Error Analysis

Insufficient training corpus

Sparse distribution of statistical data is found in the result. This is because the training corpus we used was not large enough. Many semantic classes did not appear in the corpus and further many co-occurrence relationships were missing. It can be reflected from the fact that many values of the co-occurrence frequency are zero.

Feature of probabilistic approach

Selection of a semantic class is relative in nature. High-frequency semantic class are more likely to be selected. On the other hand, since we multiply

the co-occurrence probabilities to get the score for the semantic assignments, the value of the score is very small. It makes the comparison of the scores quite difficult. Weighting could be applied to enlarge the score.

5.4 Summary

Our approach for WSD makes use of the correlation relationship between two or three consecutive words. A maximization function is then applied to tackle the problem caused by long distance dependency. The major error is due to the lack of a pre-tagged corpus to provide abundant probabilistic distribution of each semantic class in the CILIN. Sparse distribution of data is resulted. Besides, high frequency semantic classes are more likely to be selected under the probabilistic approach, which is also a source of error. Despite all that, in close test, the WSD algorithm still gets accuracy rates of 77% and 63% for bi-gram and tri-gram, respectively. Up to this point, the whole mechanism for automatic assignment of hybrid tag to Chinese word has been discussed. The next chapter will outline the implementation and results of a series of experiments for evaluate of the performance of the hybrid tag-set in resolving ambiguities in Chinese noun phrase parsing.

Chapter 6

Hybrid Tag-set for Chinese Noun Phrase Parsing

This chapter describes the experiments for evaluating the hybrid tag-set in enhancing parsing performance. Ambiguous Chinese noun phrases, including conjunctive noun phrases, De-de noun phrases and compound noun phrases, are extracted from a corpus. These noun phrases are first parsed with a syntactic parsers using the syntactic information in the hybrid tags. The resulting noun phrases with erroneous parse trees are filtered out. Semantic information in the hybrid tags are then applied to resolve the parsing errors or ambiguities.

6.1 Resolving Ambiguous Noun Phrases

6.1.1 Experiment

Objective

The objective of this experiment is to evaluate the degree of parsing performance enhanced with semantic information from the hybrid tag-set.

Setup

10 Chinese articles from the Hua Xia Wen Zhai (華夏文摘) were prepared. The articles were first word segmented. [1] Each Chinese words were then allocated with their part-of-speech (POS) tags using the POS tagger and POS tag-set developed by the Tsinghua University (Appendix A). Afterwards, the noun phrase extraction algorithm (CNPext) proposed by Li [23] was adopted to extract noun phrases from the documents. Each noun phrase was embraced with the open, "[", and close, "]" boundaries. Maximal matching noun phrase extraction was used to pair the boundaries. In the training stage, probability of a noun phrase boundary occur between two consecutive words was calculated, which was followed by the pairing of the open and close boundaries using dynamic programming. 893 noun phrases were gathered and 340 ambiguous noun phrases were filtered out. These include conjunctive

[1] It is assumed that all the Chinese words are well segmented.

Articles	No. of np	No. of conj. np	No. of De-de np	No. of conj and De-de np
1	78	2	19	4
2	103	4	19	4
3	81	2	29	7
4	87	6	21	12
5	66	4	23	4
6	102	2	28	5
7	88	4	33	4
8	85	4	35	5
9	110	1	24	3
10	93	0	23	9
Total	893	29	254	57

Table 6.1: *Noun phrases extracted from the corpus*

noun phrases, De-de noun phrases and the combination of the two. They are summarized in Table 6.1.

Procedure

The CNP3 Chinese noun phrase partial parser [37] was employed to parse the 340 ambiguous noun phrases extracted. CNP3 was based on a hybrid approach in which syntactic parsing knowledge is acquired from linguistic

resources and by corpus training. Initially, 50 grammar rules from grammar books were defined. As the parsing proceeded, these sets of grammar rules were enhanced by corpus learning. Noun phrase patterns were extracted from the corpus. They were then generalized and merged with the grammar rules in the knowledge base. The generalized rule set consisted of 196 grammar rules. The CNP3 performed well, with a recall and precision of 72% and 89%, respectively.

Once the ambiguous noun phrases were parsed, their parsed structures were checked for validation. Those that were wrongly parsed were then tagged with the hybrid tags using the automatic hybrid tag assignment algorithm we proposed. Tagging errors were corrected manually, which included the assignment of semantic classes to words that undergone semantic changes (e.g.: "人馬" and "千金"). Afterwards, each wrongly parsed noun phrases were examined. The goal was to determine the parsing errors recovered by using the semantic information provided by the hybrid tags.

6.1.2 Results

The results for the CNP3 in parsing the ambiguous noun phrases is presented in Table 6.2. The distribution of those wrongly parsed ambiguous noun phrases is shown in Table 6.3.

Ambiguous Noun Phrases		
Articles	Correctly parsed	Wrongly parsed
1	20	5
2	20	7
3	30	8
4	26	13
5	18	13
6	27	8
7	32	9
8	35	9
9	23	5
10	21	11
Correctness	74%	

Table 6.2: *Result for noun phrases parsing with the CNP3*

Wrongly Parsed Noun Phrases			
Articles	No. of conj np	No. of De-de np	No. of conj and De-de np
1	1	0	4
2	1	2	4
3	0	1	7
4	0	1	12
5	1	8	4
6	0	3	5
7	0	5	4
8	1	3	5
9	0	2	3
10	0	2	9

Table 6.3: *Distribution of wrongly parsed noun phrases with CNP3*

It was shown that the CNP3 has an overall correctness of 74% in parsing the ambiguous noun phrases. Tse had concluded several problematic noun phrases that contribute to the errors. With analysis to these problematic noun phrases, it was shown that the correctness of the CNP3 could be enhanced with semantic consideration. Two processes were proposed. They are semantic aggregation and modifier identification. They are illustrated below and the result for the enhancement in parsing ambiguous noun phrases is shown in Table 6.4

Semantic Aggregation

Consider this sentence:

- 公 司#<ng|Dm> 總 部#<ng|Dm> 與#<cpw|Kd> 分#<b|cb> 公 司#<ng|Dm> 的#<cpw|Kd> 關係#<ng|Ie>

It should be parsed as:

- [DENP [CONJ-NP [NP 公司#<ng|Dm> 總部#<ng|Dm>] 與#<cpw|Kd> [NP 分#<b|cb> 公司#<ng|Dm>]] 的#<cpw|Kd> 關係#<ng|Ie>]

However, the CNP3 wrongly parsed this noun phrase as:

- [CONJ-NP [NP 公司#<ng|Dm> 總部#<ng|Dm>] 與#<cpw|Kd> [

DENP [NP 分#<b|cb> 公司#<ng|Dm>] 的#<cpw|Kd> 關係#<ng|Ie>

]]

This error is due to the fact that the CNP3 does not have the semantic information on parsing. When sub-noun phrases are linked with more than one conjunctions or "的", the CNP3 cannot not determine the exact scope of the conjunction and "的" just with part-of-speech tags. However, if hybrid tags are employed, similar internal pattern within a sentence can be detected. It is achieved by examining the semantic class of the words if the two sides of a conjunction is equal or not. In this example, it is revealed that "總部" (headquarters) and "公司" (company) have the same semantic class, Dm. They should be aggregated together and linked by the conjunction "與". The correct scope of the conjunctive noun phrase is then identified and the correct parse tree can be produced.

Most of the wrongly parsed noun phrases belonged to this category. The following are more examples of the same nature:

Example 1

- 按#<p|Kb> 中文#<ng|Dk> 習慣#<ng|Di> 表示#<vgn|Hi> 的#<usde|Kd> 時間#<ng|Ca> 和#<usde|Kd> 日期#<ng|Ca> 信息#<ng|Da> 等等#<x|Ot>

- Wrongly parsed structure from CNP3

- [CONJ-NP [PP 按#<p|Kb> 中文#<ng|Dk> 習慣#<ng|Di> 表示#<vgn|Hi> 的#<usde|Kd> 時間#<ng|Ca>] 和#<usde|Kd> 日期#<ng|Ca> 信息#<ng|Da> 等等#<x|Ot>]

- Solution

  "時間" and "日期" belongs to the same semantic class, Ca. They should be linked with the conjunction "和" to form a conjunctive noun phrase (CONJ-NP). Finally, this leads to the correct parsed structure.

- Correct parsed structure

  - [DENP [PP 按#<p|Kb> 中文#<ng|Dk> 習慣#<ng|Di> 表示#<vgn|Hi>] 的#<usde|Kd> [CONJ-NP 時間#<ng|Ca> 和#<usde|Kd> 日期#<ng|Ca> 信息#<ng|Da> 等等#<x|Ot>]]

Example 2

- 法國#<s|cb> 的#<usde|kd> 布爾#<npu|Dm> 和#<cpw|Kc> 意大利#<s|cb> 的#<usde|Kd> OLIVETTI#<xch|Ot> 公司#<ng|Dm>

- Wrongly parsed structure from CNP3

  - [CONJ-NP [DENP [DENP 法國#<s|cb> 的#<usde|kd> 布

爾#<npu|Dm>] 和#<cpw|Kc> 意大利#<s|cb>] 的#<usde|Kd>
OLIVETTI#<xch|Ot> 公司#<ng|Dm>]

- Solution

  "布爾" and "公司" belongs to the same semantic class, Dm. Therefore, "法國 的 布爾" and "意大利 的 OLIVETTI 公司" should be De-de noun phrases and linked together to form a conjunctive noun phrase.

- Correct parsed structure

  - [CONJ-NP [DENP 法國#<s|cb> 的#<usde|kd> 布爾#<npu|Dm>] 和#<cpw|Kc> [DENP 意大利#<s|cb> 的#<usde|Kd> OLIVETTI#<xch|Ot 公司#<ng|Dm>]]

Modifier Identification

Consider a sentence with structure: "高度#a 的#usde 創造性#ng 和#cpw 幹勁#ng". It should be parsed as. "[CONJ-NP [DENP 高度#a 的#usde 創造性#ng] 和#cpw 幹勁#ng]". However, the CNP3 wrongly parsed this noun phrase as: "[DENP 高度#a 的#usde [CONJ-NP 創造性#ng 和#cpw 幹勁#ng]]". It was observed that CNP3 produced incorrect result when there was a conjunctive noun phrase which connected a De-de noun phrase and a noun phrase. This is a typical error for syntactic parsers. These errors are due to the wrongly detection of the internal structures of a noun

Wrongly Parsed Noun Phrases recovered by hybrid Tags			
Articles	No. of conj np	No. of De-de np	No. of conj and De-de np
1	0	0	3
2	0	2	2
3	0	1	1
4	0	0	6
5	1	3	1
6	0	2	4
7	1	3	2
8	1	1	2
9	0	1	1
10	0	1	5
Total no. of np recovered		43	
Enhancement in correctness for parsing		12.65%	

Table 6.4: *Result for wrongly parsed noun phrases recovered by hybrid tags*

phrases. Semantic of the modifier can help to provide clues to determine the boundaries of the internal noun phrase structure. In this example, if semantic constraint showing that "高度" can only beapplied to "創造性", the correct parsed structure can be obtained.

6.2 Summary

An experiment was performed for evaluating the effectiveness of using hybrid tags in partial noun phrases parsing. 10 articles were selected as the testing corpus and 340 ambiguous noun phrases were extracted. The CNP3, a syntactic parser based on hybrid approach, was chosen to parse these am-

biguous noun phrases. It was shown that the CNP3 offered a correctness value of 74%. Afterwards, those wrongly parsed noun phrases were tagged with the hybrid tags. In order to parse these noun phrases correctly, two approaches were proposed which made use of the semantic information from the hybrid tags. They are semantic aggregation and modifier identification. It was shown that the overall corectness of the CNP3 were enhanced by 12.65% if the semantic information provided by the hybrid tags are employed.

Chapter 7

Conclusion

In this chapter, a summary of the research work and its major contributions are given. Next, the difficulties encountered are described. Finally, several further extensions are proposed

7.1 Summary

This thesis describes a hybrid tag-set for Chinese natural language processing. An integrated algorithm is presented for tagging a word with the hybrid tag-set. There are two main processes in the algorithm. They are syntactic and semantic tagging. The part-of-speech tagger from the Tsinghua University is employed for the former. For semantic tagging, the semantic classes from the Cilin <<同義詞詞林>> are employed. However, two problems

are encountered. They are the unknown words and multiple semantic classes problems. Therefore, the semantic tagging process is further divided into two processes. The Semantic Assignment to Unknown Words (SAUW) module is designed to solve the unknown words problem, while the Word Sense Disambiguation (WSD) module is to tackle the multiple semantic classes problem.

The SAUW is a new approach for automatic semantic class assignment to unknown words from the Cilin. 3,890 unknown words are extracted from a corpus of 61,453 for testing. First, the part-of-speech and semantic class association is considered. 38.81% of the unknown words can be assigned with a semantic class with only part-of-speech consideration. The morphological compositions of the unknown words are then analyzed. The resulting unknown words are matched with more than 80 morphological rules, i.e. typical word patterns. 8.58% of unknown words can hence be resolved. Lastly, the remaining unknown words are separated into di-syllabic and poly-syllabic words. Syntactic and semantic analyses are performed for the di-syllabic unknown words. There are 140 syntactic rules, which are syntactic structure of di-syllabic word formation. It aims to apply these rules to identify words that undergo semantic changes and to locate the semantic bodies for the unknown words. The semantic analysis is to handle ambiguous cases where the semantic bodies cannot be identified with syntactic analysis. 14.35% of

unknown words are solved in this way. For the poly-syllabic unknown words, since most Chinse words are head-final, a heuristic to find the head within a word is applied and the semantic of the head is taken as the overall semantic of the word. This method further resolves 14.83% of the unknown words. In total, it was shown that the semantic classes of 80% of the unknown words could be successfully predicted. The missing 20% was mainly due to semantic changes in word formation and the use of homonyms for the original constituent characters .

Our WSD makes use of the correlation relationship among the semantic classes of neighboring words. A maximization function is applied to select the best semantic class for each word in a sentence. It aims to solve the problem caused by long distance dependency. However, there are two choices of co-occurence statistics, namely the bi-gram and tri-gram models. An empirical method was performed to decide which model performs better. 15 articles were prepared manually. 10 was used for the close test and 5 for the open test. It was shown that the bi-gram model was more effective than its tri-gram model counterpart. Therefore, the bi-gram model was chosen for our WSD algorithm.

Finally, an experiment was performed for evaluating the effectiveness of using hybrid tags in partial noun phrases parsing. 10 articles were selected as the testing corpus and 340 ambiguous noun phrases were extracted. The

CNP3, a syntactic parser based on hybrid approach, was chosen to parse these ambiguous noun phrases. It had been shown than the CNP3 offered a correctness value of 74%. Afterwards, those wrongly parsed noun phrases were tagged with the hybrid tags. In order to parse these noun phrases correctly, two approaches were proposed which made use of the semantic information from the hybrid tags. They were semantic aggregation and modifier identification. It was shown that the correctness of the CNP3 were enhanced by 12.15% if the hybrid tags were employed to provide the required semantic information.

7.2 Difficulties Encountered

Several difficulties were encountered during the research work. They are summarized as below:

7.2.1 Lack of Training Corpus

The main difficulty in this research work was the lack of semantically tagged corpus. Both of the semantic strength in the SAUW and the bi-gram or tri-gram co-occurrence statistics in the WSD required a large corpus to provide the probabilistic information. In the SAUW, the semantic strength could still be gathered from the 33,843 di-syllabic words from the Cilin, for their seman-

tic classes were known. However, a small corpus could only be prepared for manual training in WSD. The co-occurrence information was inadequate and some semantic classes were missing in the corpus. Although the validation of the WSD algorithm could be demonstrated, the accuracy of the word sense disambiguation was affected with an application point of view.

7.2.2 Features of Chinese word formation

First, di-syllabic words, which undergone semantic change, cause a problem. Their meanings were unrelated to their constituent characters and extra references were required to derive the semantics of these category of words. In our research, syntactic composition of unknown words were analyzed to identify words that has undergone semantic changes. The syntactic composition of an unknown word was matched with the set of syntactic rules governing the di-syllabic words formation. If it was not found in the rule set, the word was proven to be a word that undergone semantic changes. This approach was adequate for typical example like "的士" (taxi), as it had a special syntactic composition, "ng: usde + ng". However, it did not work for some others. For example, "人馬" (manpower) had a syntactic structure of "ng: ng + ng". It was the same as the syntactic structure of "糖果" (sweet), which was contained in the Cilin. This accounts for 47.27% of the total errors. More

accurate method is needed to identify these words. Besides, once these words were identified, manual process was involved in semantic classes assignment to these words in our research. Automatic approach would be desirable in this process. Second, it was observed that many Chinese words tended to use homonyms to replace the correct constituent characters in word formation. Many of them were also mixed with Feng Ti (繁體) and Jian Ti (簡體) characters. Examples were "苦干" (hardworking), "關系" (relationship) and "然后" (afterward). These characters were unrelated to the meanings of the whole words. They were similar to those words that has undergone semantic changes and efforts should be taken to handle them. In this research, 369 words belonging to this type were extracted from 3890 unknown words and they were rectified manually before the SAUW

7.2.3 Problems with linguistic sources

The Chinese dictionary and the Cilin were the two major linguistic sources used in this research. Since they were developed independently, inconsistency inevitably exists between them, i.e. a word found in the dictionary might not exist in the Cilin and vice versa. This caused problem in the SAUW. In the SAUW, given an unknown word, the part-of-speech of it and its constituent characters were required in the morphology-based deduction and syntactic

analysis. Similarly, the semantic classes of them were also required in the semantic analysis. This interrupted the automatic process as the users have to be involved to rectified the errors. Moreover, 25.64% of the errors in the SAUW came from technical words or other specific terms such as "核糖" or "液泡". Some of these words could not derive their meanings from its constituent characters. Domain specific dictionary must be added to enrich the linguistic sources.

7.3 Contributions

7.3.1 Enrichment to the Cilin

The hybrid tag proposed in this thesis consists of two parts, i.e. the syntactic and semantic parts. The semantic part is the semantic classes associated with the word from the Cilin. The unknown words referred in this thesis are those words that are absent from the Cilin. These missing words can seriously reduce the effectiveness of the Cilin in many NLP applications, e.g. parsing, automatic indexing, etc. In order to overcome the unknown words problem in the Cilin, we proposed an algorithm for automatic semantic class assignment to unknown word (SAUW). This is achieved by predicting the semantic class of the unknown words from the syntactic and/or semantic information of the

characters/words, which make up the word. 83.47% of the unknown words are handled and the correctness is 91.73%, respectively. It can broaden the coverage of the Cilin by assigning semantic classes to unknown words. The process is automatic and dynamic. In this way, modern words can be added into the Cilin and the burden of the lexicographers could be eased, saving much effort and time in compilation. This moduale is useful many other applications that use the Cilin as semantic references.

7.3.2 Enhancement in syntactic parsing

The hybrid tag-set were designed to enhance parsing performance of syntactic parsing. Semantic information was provided by the hybrid tags. Semantic constraints could be incorporated to Chinese noun phrases parsing. Two approaches are proposed. They are semantic aggregation and modifier identification. The CNP3, an in-house syntactic parser based on hybrid approach, was used for evaluation. It was shown that the correctness of the CNP3 was enhanced by 12.15% with the hybrid tags.

7.4 Further Researches

7.4.1 Investigation into words that undergo semantic changes

It was shown that solely the syntactic analysis was inadequate to identify this categories of words. In fact, syntactic information of the neighboring words are also useful. The relationship among the semantic classes of a word and the part-of-speech of its neighboring words could provide clues to identify such kind of word. Consider the following examples:

- ... 厲害#a 角色#ng ...

The word "角色" (role) was absent from the Cilin and it was a word that undergone semantic change. Our algorithm was not able to identify it with syntactic analysis. It was treated as ordinary unknown word and the semantic of "色" (color) was assigned to it. However, if it could be shown that the word "厲害" (tough), which is an adjective, could not be applied to "色" if it was semantically tagged as color or even as (Semantic of "角", which means "corner"). "角色" could be proven to be a word that undergone semantic change.

Besides, manual process was involved to assign semantic classes to these words in this research. It might be possible to assign a semantic class to

these words automatically based on the context information.

7.4.2 Incorporation of more information into the hybrid tag-set

Apart from the semantic information from the Cilin, other information can be incorporated into the hybrid tag. Thematic roles of words in Chinese is a good example [9]. The use of thematic knowledge can link up the semantic, conceptual and syntactic constituents of the word. It can also facilitate conflict resolution in parsing. The most common thematic roles are Agent, Goal, Source, Instrument, Theme, Beneficiary, Location, time, Quantity, Proposition, Manner, Cause and Result. "辭海" is another good source of linguistic information. The information could be added into the hybrid tag. It is a thesaurus for Chinese words matching (搭配詞典). There are 7781 word strings (詞條) with 770,000 matching examples (搭配實例). For each word string, there are several semantic items (義項) with a total of 12841. there are also a word category (詞類), semantic (釋義文本) and matching examples (搭配實例) for each semantic item. It is expected that this information may further improve the effectiveness of the hybrid tag-set in enhancing Chinese parsing.

§ END §

Bibliography

[1] 劉源, 譚強, and 沈旭昆. 信息處理用現代漢語分詞規范及自動分詞方法, 1994.

[2] 清華大學計算機科學與技術系. 英語詞性自動標注系統. Technical report, Tsinghua University, March 1992.

[3] Eneko Agirre and German Rigau. Word sense disambiguation using concpetual density. In *cmp-lg*, Jun 1996.

[4] Eneko Agirre and German Rigau. A proposal for word sense disambiguation using conceptual distance. In *cmp-lg*, Oct 95.

[5] Chao-Huang Chang and Cheng-Der Chen. A study on integrating chinese word segmentation and part-of-speech tagging. In *Communication of COLOPS*, pages 69–77, 1993.

[6] Ming-hong Bai chao-jab Chen and Keh-Jiann Chen. Category guessing for chinese unknown words. In *Proceedings of the Natural Language Processing Pacific Rim Symposium*, pages 35–40, 1997.

[7] Keh-Jiann Chen. A model for robust chinese parser. In *Computational Linguistics and Chinese Language Processing*, volume 1 no. 1, pages 183–204, Aug 1996.

[8] Keh-Jiann Chen and Ming-Hong Bai. Unknown word detection for chinese by a corpus-based learning method. In *Proceedings of ROCLING*, pages 159–174, Aug 1997.

[9] Tung-Bo Chen, Koong H.C. Lin, and Von-Wun Soo. Training recurrent neural network to learn lexical encoding and thematic role assignment in parsing mandarian chinese sentences. In *Neurocomputting*, volume 15, pages 383–408, 1997.

[10] Chun-Hung Cheng and Kun-Chung Timothy Chan Kam-Fai Wong. A multi-pararmetric approach for chinese noun phrase extraction. In *Proceedings of the Frist ACM Hong Kong Postgraduate Research Day*, pages 132–138, Oct 1998.

[11] Han Dezhi. Fifty patterns of modern chinese, 1993.

[12] Ji Donghong and Hunag Changning. A semantic composition model for chinese nouns and adjectives. In *cmp-lg*, Jun 1996.

[13] Mei et al., 梅家駒, 竺一鳴, 高琦, and 殷鴻翔. 同義詞詞林, 1983.

[14] Wilks Y. et al. Providing machine tractable dictionary tools, semantics and the lexicon, 1993.

[15] J. Fu. *On deriving Chinese derived nominals: Evidence for V-to-N raising.* PhD thesis, University of Massachusetts Amherst, September 1994.

[16] Wan Jiancheng, Tan Ming, and Wan Fang. Chinese word semantic association in syntax. In *Communications of COLIPS*, volume 4 no. 2, pages 103–111, December 1994.

[17] Chen Keh-Jiann and Chen Chao-Jan. A corpus-based study on computational morphology for mandarin. In *Quantitative and Computational Studies on the Chinese Language*, pages 283–300, 1998.

[18] Adam Kilgarriff. What is word sense disambiguation good for. In *cmp-lg*, Dec 1997.

[19] Pang Chun Kiu. A natural language based indexing technique for chinese information retrieval. Master's thesis, The Chinese University of Hong Kong, Jun 1997.

[20] Harder L.B. Sense disambiguation using on-line dictionaries, natural language processing: the plnlp approach, 1993.

[21] Hsi-Jian Lee and Pei-Rong Hsu. Parsing chinese sentences in a unification-based grammar. In *Computer Processing of Chinese and Oriental Languages*, Nov 1991.

[22] C.N. Li and S.A. Thompson. Mandarian chinese: A functional reference grammar, 1981.

[23] W. Li. *Automatic Noun Phrase Extraction from Full Chinese Text.* PhD thesis, The Chinese University of Hong Kong, September 1997.

[24] Shing-Huan Liu, Keh-Jiann Chen, Li-Ping Chang, and Yeh-Hao Chin. Automatic part-of-speech tagging for chinese corpora. In *Computer Processing of Chinese and Oriental Languages*, volume 9 no. 1, pages 31–47, June 1995.

[25] K.T. Lua. Associative thinking as derived from semantics of chinese characters and chinese semantic field. In *Communications of COLIPS*, volume 3 no. 1, pages 11–30, 1992.

[26] K.T. Lua. An efficient inductive unsupervised semantic tagger. In *cmp-lg*, 1992.

[27] K.T. Lua. The number of syllabics in a chinese sentences. In *Computer Processing of Chinese and Oriental Languages*, volume 7 no. 1, pages 125–131, June 1993.

[28] K.T. Lua. A study of chinese word semantics. In *Computer Processing of Chinese and Oriental Languages*, volume 7 no. 1, pages 37–60, June 1993.

[29] K.T. Lua. A study of chinese word semantics and its prediction. In *Computer Processing of Chinese and Oriental Languages*, volume 7 no. 2, pages 167–189, December 1993.

[30] K.T. Lua. An efficient inductive unsupervised semantic tagger. In *cmp-lg*, 1996.

[31] C.K. Pang, K.F. Wong, B.T. Low, and V.Y. Lum. Structural and contextual index extraction for chinese documents. In *2nd International Workshop on Information Retrieval on Asian Languages, Tsukuba, Japan*, pages 51–67, October 1997.

[32] K.H. Pun and B. Lum. Resolving ambiguities of complex noun phrases in a chinese sentence by case grammar. In *Computer Processing of Chinese and Oriental Languages*, July 1989.

[33] Man-Tak Shing and Paul Ling. A knowledge engineering approach to natural language processing. In *Computer Processing of Chinese and Oriental Languages*, Jul 1989.

[34] Tang Siu-Lam. syntactic and semantic interplay during chinese text processing. Master's thesis, The Chinese University of Hong Kong, 1996.

[35] Ricjard Sproat, Chilin Shih, William Cale, and Nancy Chang. A stochastic finite-state word-segmentation algorithm for chinese. In *cmp-lg*, May 1994.

[36] Lam Sze-Sing. A new approach for extracting inter-word semantic relationship from a contemporary chinese thesaurus. Master's thesis, The Chinese University of Hong Kong, Jun 1995.

[37] A. Tse. Chinese noun phrase parsing with a hybrid approach. Master's thesis, The Chinese University of Hong Kong, 1996.

[38] Patrick S P Wang. The intelligent chinese characters. In *PRoceedings of 1987 International Conference on Chinese and Oriental Language computing*, pages 85–88, 1987.

[39] Yorick Wilks and Mark Stevenson. Word sense disambiguation using optimised combination of knowledge sources. In *cmp-lg*, Jun 1998.

[40] Kam-Fai Wong, Sze-Sing Lam, and Vincent Lum. Extracting the inter-word semantic relationship. In *Computer Processing of Oriental Languages*, volume 10 no. 3, 1997.

[41] Kam-Fai Wong and Wai-Kwong Leung. Automatic semantic assigment of unknown words according to cilin. In *Proceedings of International Symposium on Machine Translation and Computer Language Processing*, pages 229–236, Jun 1998.

[42] Niwa Y. and Nitta Y. Co-occurrence vectors from corpora vs distance

vectors from dictionaries. In *Proceedings of COLING's 94*, pages 304–309, 1994.

[43] Jingye Zhou and Shi-Kuo Chang. A methodology for deterministic chinese parsing. In *Computer Processing of Chinese and Oriental Languages*, May 1986.

Appendix A

POS Tag-set by Tsinghua University (清華大學)

1. a 形容詞
2. b 區別詞
3. c 連詞
   (a) cb 主從連詞後段 (e.g.: 所以,但是)
      i. cbc 連接分句，詞語
      ii. cbs 連接句子
   (b) cf 主從連詞前段 (因爲,雖然)
   (c) cp 幷列連詞 (和,與,並且)
      i. cpc 連接分句
      ii. cps 連接句子
      iii. cpv 連接詞語
4. d 副詞
5. e 嘆詞
6. f 方位詞
7. h 前綴
   (a) hm 數詞前綴
   (b) hn 名詞前綴
8. i 成語,習用語

9. j 簡稱略語
10. k 後綴
11. mark 標點符號
12. m 數詞
    (a) mb 倍數詞
    (b) mf 分數詞
    (c) mh 數詞“半”
    (d) mm 數量詞
    (e) mo 數詞“零”
    (f) mq 概數詞
    (g) mv 位數詞
    (h) mx 序數詞
13. n 名詞
    (a) nf 姓氏
    (b) ng 普通名詞
    (c) np 專有名詞
        i. npf 人名
        ii. npr 其他專有名詞
        iii. npu 組織,機構名
14. o 象聲詞
15. p 一般介詞
    (a) pba 介詞“把”,“將”
    (b) pbei 介詞“被”,“讓’，“叫”
    (c) pzai 介詞“在”
16. q 量詞
    (a) qn 名量詞
        i. qnc 集合量詞
        ii. qnf 量詞

iii. qng 名量詞“個”
iv. qnk 種類量詞
v. qnl 個體量詞
vi. qnm 度量詞
vii. qns 不定量詞
viii. qnt 臨時量詞
ix. qnv 容器量詞
x. qnz 準量詞

(b) qv 動量詞
i. qvn 借用名詞做動量詞
ii. qvp 專有動量詞

17. r 代詞
(a) rd 副詞性代詞
(b) rn 體詞性代詞
(c) rp 謂詞性代詞

18. s 處所詞人 (國名)

19. t 時間詞

20. u 助詞
(a) up “被”,“給”
(b) us 結構助詞
i. usde “的”
ii. usdf “得”
iii. usdi “地”
iv. ussb “不”
v. ussi “似的”
vi. ussu “所”
vii. uszh “之”
(c) ut 時態助詞
i. utg “過”
ii. utl “了”
iii. utz “著”

21. v 動詞
    - (a) va 助動詞
    - (b) vc 補語(趨向)動詞
    - (c) vg 非謂或帶動詞補語的動詞
        - i. vga 帶形容詞性賓語
        - ii. vgd 帶雙賓語
        - iii. vgj 帶兼語賓語
        - iv. vgn 帶體詞性賓語
        - v. vgo 不帶賓
        - vi. vgs 帶小句賓語
        - vii. vgv 帶動詞性賓語
    - (d) vh "有"動詞
    - (e) vi 系動詞
    - (f) vv "來","去"形成連詞
    - (g) vy "是"動詞
22. x 其他
    - (a) xch 非漢字字符串或數學公式
23. y 語氣詞人(了,嗎,吧)
24. z 狀態詞
25. Non-terminal
    - (a) AP (Adjective phrase) 形容詞短語
    - (b) CONJ-NP (Conjunctive phrase) 連詞短語
    - (c) DENP (De noun phrase) "的"名詞短語
    - (d) CPD-N (Compound noun) 複合名詞
    - (e) NP (Noun phrase) 名詞短語
    - (f) PP (Prepositional phrase) 前置詞短語
    - (g) NUM (Number) 數字
    - (h) RELCLS (Relative clause) 關係子句
    - (i) TIME (Time) 時間

Appendix B

Morphological Rules

The following are four categories of the morphological rules used in automatic semantic classes assignment to unknown words. The symbol "X", "A" or "B" stand for Chinese character. For some rules, the character embraced with "< >" contributes to the overall semantics while others involve decomposition to the words to identify the semantic bodies.

Prefix	
Chinese Words	Description
<X>兒	名詞
<X>子	名詞and X is not equal to小
<X>手	名詞
<X>性	名詞 副詞 區別詞 and X is動詞 形容詞 名詞
<X>化	名詞
<X>了	助詞
<X>過	助詞
<X>們	代詞
<X>地	助詞
<X>的	助詞
<X>頭	名詞
<X>於	帶體詞性賓語
<X>出	動詞
<X>起	動詞 副詞
<X>來	動詞
<X>式	名詞 動詞
<X>于	動詞 副詞 一般介詞 連接句子

Prefix	
Chinese Words	Description
<X>之	NIL
<X>著	動詞
<X>以	NIL
<X>開	動詞
<X>此	連接句子
<X>不	副詞
<X>到	副詞 帶體詞性賓語
<X>去	不帶賓
<X>上	名詞
<X>下	名詞

Negation	
Chinese Words	Description
沒<X>	Segment into沒and <X>
不<X>	Segment into不and <X>
非<X>	Segment into非and <X>
A不B	Segment into A and不B (A can't be equal to B)
莫<X>	segment into 莫 and <X>
無<X>	segment into 無 and <X>
反<X>	名詞segment into 反 and <X>

Suffix	
Chinese Words	Description
小<X>	名詞
老<X>	名詞
阿<X>	名詞
各<X>	代詞
每<X>	代詞
某<X>	代詞
本<X>	代詞 名詞
該<X>	代詞
此<X>	代詞
全<X>	代詞
很<X>	副詞
<X>	副詞
剛<X>	副詞
來<X>	帶動詞性賓語
大<X>	名詞
一<X>	名詞
之<X>	NIL
所<X>	區別詞
而<X>	帶動詞性賓語
上<X>	方向詞
前<X>	方向詞
後<X>	方向詞
中<X>	方向詞
東<X>	方向詞
南<X>	方向詞
西<X>	方向詞
北<X>	方向詞
內<X>	方向詞
外<X>	方向詞
有<X>	動詞
X<人>	名詞
X<軍>	名詞
X<者>	名詞
X<員>	名詞
以<X>	NIL
下<X>	方向詞 體詞性代詞

Reduplicates		
Chinese Words	Description	Semantic Bodies
AA	形容詞 量詞 動詞 副詞 狀態詞	A
AABB	形容詞 動詞 副詞 名詞	AB
ABB	形容詞 動詞 體詞性代詞	AB
AAB	形容詞 動詞	AB
ABAB	形容詞 動詞 名詞	AB
A里AB	形容詞	AB
A一A	動詞	A
A了A	動詞	A
A了一A	動詞	A
一A一B	形容詞	Segment into A and B
一A二B	形容詞	Segment into A and B
半A半B	形容詞	Segment into A and B
半A不B	形容詞	Segment into A and 不B
有A有B	形容詞	Segment into A and B
又A又B	副詞	Segment into A and B
不得不A	副詞	A
不能不A	副詞	A
越來越A	副詞	A
越A越B	副詞	B

Appendix C

Syntactic Rules for Di-syllabic Words Formation

The following are xxx syntactic rules defined for di-syllabic words formation. In each rule, there are three parts. The first part is the part-of-speech of the first character while the second one is the part-of-speech of the second character. The last part is the part-of-speech of the whole di-syllabic word.

Syntactic Rule for Di-syllabic Words Formation		
POS for the first character	POS for the second character	POS for the di-syllabic words
q	<n>	n
<v>	n	vg
<v>	qv	v
<v>	vc	v
a	<n>	n
n	n	n
<v>	n	v
d	<a>	a
d	<v>	v
a	a	a
n	<f>	f
<n>	a	n
<n>	v	n
r	<v>	v
r	<a>	a

Syntactic Rule for Di-syllabic Words Formation		
POS for the first character	POS for the second character	POS for the di-syllabic words
v	v	v
p	n	n
d	<a>	a
d	<ng>	vg / ng
p	<f>	f
b	<ng>	ng
a	<vgn>	vgn
a	<qnm>	ng
<d>	p	d
<p>	ng	ng
n	v	n
n	v	v
n	a	a
n	u	n
n	a	n
v	n	n
v	v	n
v	a	v
vc	n	vgo
v	p	v
v	vh	v
a	v	a
a	n	a
a	a	a
a	q	n
f	n	f
f	vv	vc
d	a	d
d	d	d
f	n	n
b	<qnm>	ng
<d>	qnm	d
<d>	qnk	ng
d	<vgn>	ng
<d>	vgn	a
<t>	ng	b/d

Syntactic Rule for Di-syllabic Words Formation		
POS for the first character	POS for the second character	POS for the di-syllabic words
t	<ng>	ng
mx	<ng>	f
mx	<ng>	d
mh	<ng>	ng
ng	<d>	d
ng	f	ng
ng	qni	ng
<a>	p	a
a	mx	ng
<a>	p	a
a	<vg>	ng
<a>	va	ng
a	<ng>	rn
<v>	mx	ng
<vg>	qni	vg
va	<vgo>	a
<vgn>	ng	d
vc	<qnc>	z
<vgn>	f	vgn
<vg>	qni	ng
<va>	ng	a
<vg>	p	a
<vg>	p	nvg
p	<ng>	a
<p>	utz	p
p	<d>	d
<p>	qni	d
<qni>	qvp	b
<qni>	qni	ng
<qnm>	ng	vgn
p	va	vg
p	p	p
p	f	ng
p	vgn	vgn
p	rn	p/d
pzai	<ng>	d

Syntactic Rule for Di-syllabic Words Formation		
POS for the first character	POS for the second character	POS for the di-syllabic words
p	vgn	a
p	d	a
p	rn	cf
p	p	vgn
p	ng	b
d	d	b
d	va	cf
<d>	f	ng
d	a	ng
d	t	d
d	<a>	b
d	<vgv>	b
b	vgn	ng
b	<vgn>	ng
b	vy	d
<f>	vg	f
f	ng	b
f	f	f
<rn>	ng	rn
rn	ng	ng
rn	qns	rn
rn	rn	b
rn	qv	rn
rn	qnm	a
mx	<qnm>	d
mx	<vgn>	d
mx	<b>	d
mx	qnm	qnm
t	vh	b
<cpw>	rn	cpw
qnk	qnc	ng
qni	vgj	ng
ng	qnk	ng
ng	ng	vg
ng	qnc	ng
ng	qni	ng

Syntactic Rule for Di-syllabic Words Formation		
POS for the first character	POS for the second character	POS for the di-syllabic words
a	vgn	vgo/vg
a	<vgo>	vgn
a	p	vgn
a	ng	vgn
a	ng	b
<vg>	usdf	vgn
vg	ng	b
vgn	ng	z
vgn	a	nvg
va	a	d/a
va	vg	b
vgn	vgn	d
vg	d	vgo
vgn	p	p
vc	mb	d
<vg>	mx	vg
<vc>	cbc	d