Chinese Synthetic Word Analysis with Tree-based Structure Information

Jia Lu Masayuki Asahara Yuji Matsumoto
Nara Institute of Science and Technology Graduate School of Information Science
{jia-l, masayu-a, matsu}@is.naist.jp

Abstract

Chinese word segmentation has always been a difficult and challenging task in Chinese language processing. Although there are several morphological analysis systems which have been developed until now, there is no single segmentation standard for all tagged corpora that is agreeable across different research groups. We found that most of the disagreements in these standards come from the segmentation of Chinese synthetic words. Furthermore, we found that in our own Chinese morphological analysis system, a large part of the out-of-vocabulary words are compound words or morphologically derived words, which also belong to Chinese synthetic words. In order to improve the performance of our Chinese morphological analysis system, we first try to solve the Chinese synthetic word problem by fertilizing our system dictionary with some synthetic word information. In this paper, we first survey Chinese synthetic words in our own Chinese word dictionary and then try to categorize them according to their inside semantic and syntactic structure. By doing this, we propose a method to save these inside information of word structure into our dictionary by applying a tree-based structure. We believe that this tree-based information could be useful in specifying a Chinese synthetic word segmentation standard.

1 Introduction

Several Chinese morphological analysis systems have been developed by different research groups and they all have quite good performance when doing segmentation of written Chinese. But there still remain some problems. The biggest one is that every research group has its own segmentation standard for its system, which means there is no agreement on the segmentation standard for Chinese language. Actually, most of these disagreements are the way to deal with synthetic words. Because different NLP applications, such as MT, IR and IME, need different ways of representing synthetic words, there is no Chinese morphological analysis system for now that could do all kinds of these work with only one segmentation standard.

Furthermore, although every segmentation system has good performance, there are still many out-of-vocabulary words, which could always be seen as synthetic words, they could not be easily recognized because of the flexibility of Chinese synthetic word construction process.

In order to make our Chinese morphological analysis system to recognize more out-of-vocabulary words and fitting different kinds of NLP applications, we try to analyze the structure of the inside information of Chinese synthetic word and store these information into a synthetic word dictionary by representing them with a kind of tree-based structure based on our system dictionary.

In this paper, we first make the definition of Chinese synthetic words and classify them into several different categories in section 2. In section 3, two previous researches on Chinese synthetic word will be first introduced. Then we propose a tree-based method for analyzing Chinese synthetic word with a little investigation on 3- character words. Finally, section 4 shows how this method could benefit Chinese morphological analysis and our future work.

2 A detailed study of Chinese synthetic word

2.1 Definition of Chinese word

There has always been a common belief that Chinese 'doesn't have words' but instead has 'characters', or that Chinese 'has no morphology' and so is 'morphologically impoverished'. But actually for native Chinese speakers, they know that words are those lexicons which represent a whole concept and occur innately in the form of specific language rules in the brain.

Though there are a lot of ways to classify Chinese word, we believe that Chinese word should be first divided into single-morpheme word and synthetic word according to their internal connection of different parts.

- ➢ Single-morpheme word
- ➢ Synthetic word

Single-morpheme words are those that could not be divided into smaller parts when representing as a whole concept. In other words, if we divide single-morpheme word into characters or parts, the meaning of different parts does not have any connection with the meaning of the original word. Following are the three different types of single-morpheme word.

- one-character word:
  人(human), 马(horse), 车(vehicle)
- one-morpheme word:
  鹌鹑(quail), 鸳鸯(mandarin duck), 翡翠(jadeite)
- transliteration word:
  比萨(pizza), 肯德基(Kentucky),
  阿司匹林(aspirin)

As you can see from the above examples, if we divide 肯德基(Kentucky) into '肯', '德' and '基', it definitely can not indicate the meaning of the famous fried chicken restaurant chain from those three characters. So these three kinds of single-morpheme word should be segmented as one word in any morphological systems.

However, it becomes much more complicated when dealing with synthetic word. Generally, synthetic words are the type of words which are composed of single-morpheme words and represent a new entity or meaning which can be indicated from the internal parts. According to this definition, if we divide synthetic words into smaller parts, we still could somehow guess the original meaning from the meaning of internal parts. For example:

国家(country), 毕业(graduate), 司机(driver)
游泳池(swimming pool)

Actually, in single-morpheme words, except a large number of transliteration words, there is only a little number of commonly used characters in Chinese. According to the encoding standard of GB2312, there are about 6,763 commonly used characters in Chinese language. In our own system dictionary which has about 129,440 word entries, there are only 6,188 (about 4.78%) one-character words, which means that most Chinese words belong to synthetic words.

2.2 Classification of Chinese synthetic word

Classification of Chinese synthetic word is a difficult task because sometimes even the native speakers could not determine which category a word should belong to. However, until now there have been a lot of researches on classification of Chinese synthetic word from both linguistic and computational points of view. Each of them has divided synthetic word into different categories according to their own criteria. In our research, based on our experience on Chinese morphological analysis and unknown word detection, we divide Chinese synthetic word into the categories in Figure 1.

As you can see from Figure 1, we describe synthetic word either as compound word based on the internal syntactic rules or as morphological derived word based on the structure feature of internal word components. While the compound words are usually two-character words which are most frequently used in Chinese language, the morphological derived words often represents words that have more than 2 characters which could be easily recognized by the arrangement of word components. So there may be some overlap of these two kinds of synthetic words.

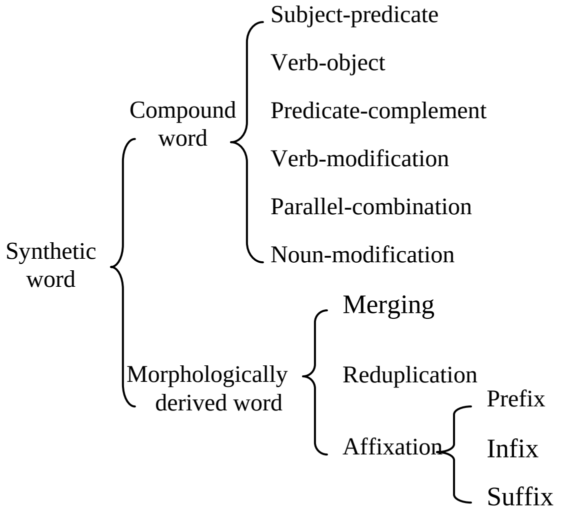

Figure 1. Classification of Chinese synthetic word

2.2.1 Compound word

Compound words are those whose internal word components have some syntactic relations with each other.

- Subject-predicate (主谓式)
  words that have subject and predicate parts. This type is subdivided into three types: SV, SVX and VS.
  Example: 地 / 震， 人 / 造 / 棉， 搬运 / 工
- Verb-object (动宾式)
  words that have verb and object parts. This type is subdivided into six types: VO, VOS, VOX; OV, OVS and OVX.
  Example: 理 / 发，投 / 资 / 商，度 / 假 / 村，
  瓶 / 塞，曲 / 作 / 者，数据 / 分析 / 仪
- Predicate-complement (述补式)
  words that have predicate part and complement part and show the result, direction or aspect of the action. This type is subdivided into two types: VV and VA.
  Example: 煽 / 动，说 / 明
- Verb-modification (动词偏正式)
  words that have verb part and modification part and show the property of the verb part or be the root of the verb part. This type is subdivided into three types: VX, XV and XVX.
  Example: 放大 / 器，激光 / 打印，
  自动 / 控制 / 器
- Parallel-combination (联合式)
  words that are composed of equal parts which could have the same, similar, related or opposite meaning.

Example: 开 / 关，学 / 习，国 / 家，兄 / 弟

- Noun-modification (名词偏正式)
  words that have noun part and modification part and show the property of the noun part or be the root of the noun part.
  Example: 笔 / 筒，书 / 架，汽车 / 站

2.2.2 Morphologically derived word

Morphologically derived words are those ones which have specific appearance of word formation.

- Merging
  words that are composed of two adjacent and semantically related words which have some characters in common. It could be seen as a kind of abbreviation.
  Example: 中学+小学→中小学
  上文+下文→上下文
  北京市+市长→北京市长
- Reduplication
  words that contain some reduplicated characters. There are eight main patterns of reduplication: AA, ABAB, AABB, AXA, AXAY, XAYA, AAB and ABB.
  Example: 听听，研究研究，雄赳赳
- Affixation
  words that are composed of a word and an affix(either a prefix, a suffix or an infix).
  Example: 副主席，总工程师
  看不到，听得见
  监督局，安全厅

2.3 Exceptions

Apart from compound word and morphologically derived word, however, there still exist some types of words which need discussion about whether they should belong to synthetic word or not.

- Abbreviations
  expressions that have a short appearance, but stand for a long term.
  Example: 中共→中国共产党
  国资委→国有资产监督管理委员会
- Factoids
  expressions that indicate date, time, number, money, score and range. This kind of expressions has a large variation of appearance.
  Example: 2007 年 1 月 30 日， 2007.1.30
  五点半，3 点 10 分
  一比二，三块五毛六
- Idioms, proverbs, sayings and poems
  expressions that usually consist of more than three characters and always have a special meaning.
  Example: 门可罗雀，明日黄花
  刀子嘴豆腐心
  先天下之忧而忧

3 Synthetic word analysis and experiment

3.1 Previous research

There is little specific research on Chinese synthetic word. However, every institution has its own way of dealing with synthetic word in their segmentation standard when doing Chinese morphological analysis. There are two main previous researches on the analysis of Chinese synthetic word.

The first one is done by Microsoft [Andi Wu, 2003] by creating a customizable segmentation system of Chinese morphologically derived words. This system uses a parameter driven method which can divide synthetic word into different levels of word components based on some advanced defined rules, according to the needs of different NLP application. For instance, in machine translation, we will translate ‘烤面包器’ into ‘toaster’ if our system dictionary has this kind of information, but if we do not have this entry in our dictionary, we have to split ‘烤面包器’ into lower level such as‘烤 / 面包 / 器’ the translation of which will probably give us some information about the original meaning of the whole word. Although this system gains high score against other systems that do not have synthetic analysis, it only takes morphologically derived words into account, which means it does not contains information about internal syntactic relations.

The second one [C. Huang 1997] is actually a segmentation tagging standard rather than a detailed synthetic word analysis research. It is first used by Sinica when doing the tagging task of Chinese word segmentation, which will select one tag from w0, w1 and w2 which stand for ‘faithful’, ‘truthful’ and ‘graceful’ if the tagging object is a synthetic word. For example, if we have a synthetic word ‘副总工程师’, this tagging method will divide the word like this:

<w2>
<w1><w0>北京</w0><w0>市</w0></w1>
<w1><w0>矿物</w0><w0>局</w0></w1>
</w2>

Again, this kind of method does not take word internal syntactic relations into account either. Furthermore, it even does not have the POS information of different levels of word, thus could not construct a customizable system.

3.2 Synthetic word analysis with tree-based structure information

For specifying consistent Chinese segmentation standard for our morphological analysis system and fertilizing the information of our dictionary, we propose a synthetic word analysis method with tree-based structure information.

We assume that words which are already in our current system dictionary could be word components of other out-of-vocabulary synthetic words. So the first thing to do is to classify all synthetic word in our current dictionary into the categories based on section 2.2. Because intuitively most 2-character words have already become lexicons in Chinese, though they could have internal syntactic relations, we can first classify all 3-character words into those categories and from here, we can easily construct 4-character or 5-character words' structure by using 3-character and 2-character words' information. Finally, when we get a long synthetic word, we can build a tree structure (Figure 2) by using the constructed synthetic word dictionary.

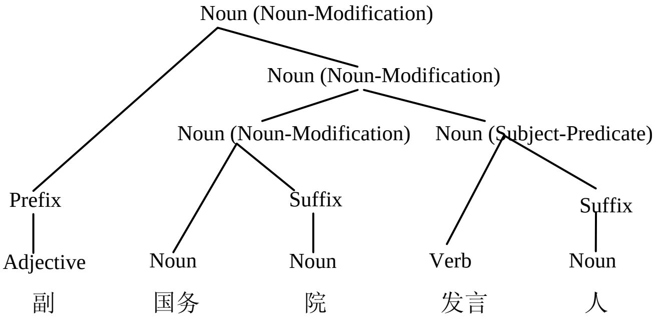

Figure 2. Synthetic word tree of '副国务院发言人'

When constructing this kind of tree, we can use some rules which have the following form:

A + B → Category

or A + B + C → Category

where A, B and C are parts of speech, affixation or other properties of word components.

In order to tagging all 3-character words in our dictionary, we first randomly selected 1000 3-character words and tagged them by hand. Table 1 and Table 2 show the result of this process.

From Table 1 which shows the distribution of compound words, we can see that noun-modification words are the largest part (62.7%) in synthetic words from a syntactic point of view. Table 2 which shows the distribution of morphologically derived words gives us the information that most synthetic words (84.3%) have an internal forming structure with a suffix part. Because most 3-character Chinese words have the structure such as 'two+one' or 'one+two' character formation, it is obvious that we should look at noun-modification words with frequently used suffixes first as the beginning of our analysis. We could get a list of characters of possible affixation from this process too. Furthermore, we also find that parallel-combination words and reduplication words intend to have some fixed structures which may make them easy to recognize.

subject-predicate	2.1%
verb-object	7.5%
verb-modification	16.9%
predicate-complement	3.1%
parallel-combination	0.4%
noun-modification	62.7%
single-morpheme word	7.2%

Table 1. Compound words in 1000 words

prefix	5.9%
infix	0.4%
suffix	84.3%
merging	1.3%
reduplication	0.4%

Table 2. Morphologically derived words in 1000 words

4 Conclusion and future work

This paper proposed a tree-based method for analyzing Chinese synthetic word for constructing a Chinese synthetic word dictionary. This method is based on the classification of Chinese synthetic word both from syntactic and morphological ways. After investigating the distribution of 3-character words, we will try to build a automatic synthetic word analysis system by using semi-supervised learning method in the near future and finally improve the performance of our morphological analysis system with the built Chinese synthetic word dictionary.

References

Andi Wu, 2003, Customizable Segmentation of Morphologically Derived Words in Chinese, Vol.8, No.1, February 2003, pp. 1-28, Computational Linguistics and Chinese Language Processing

C. Huang, K. Chen and L. Chang, Segmentation standard for Chinese natural language processing. International Journal of Computational Linguistics and Chinese Language Processing, 1997

何元建, 2004, 汉语真假复合词-从普遍语法原则看汉语复合词的语序、类型及结构, 香港中文大学

Jerome L. Packard, 2000, The Morphology of Chinese-A Linguistic and Cognitive Approach