Analyzing Chinese Synthetic Words with Tree-based Information and a Survey on Chinese Morphologically Derived Words

Jia Lu
Nara Institute of
Science and Technology
jia-l@is.naist.jp

Masayuki Asahara
Nara Institute of
Science and Technology
masayu-a@is.naist.jp

Yuji Matsumoto
Nara Institute of
Science and Technology
matsu@is.naist.jp

Abstract

The lack of internal information of Chinese synthetic words has become a crucial problem for Chinese morphological analysis systems which will face various needs of segmentation standards for upper NLP applications in the future. In this paper, we first categorize Chinese synthetic words into several types according to their inside semantic and syntactic structure, and then propose a method to represent these inside information of word by applying a tree-based structure. Then we try to automatically identify the inner morphological structure of 3-character synthetic words by using a large corpus and try to add syntactic tags to their internal structure. We believe that this tree-based word internal information could be useful in specifying a Chinese synthetic word segmentation standard.

1 Introduction

Chinese word segmentation has always been a difficult and challenging task in Chinese language processing. 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 each research group has its own segmentation standard for their system, which means that there is no single segmentation standard for all tagged corpora which can be agreeable across different research groups. And we believe that this situation slows down the progress of Chinese NLP research.

Among all the differences of segmentation standards, the segmentation method for Chinese synthetic words is the most controversial part because Chinese synthetic words have a quite complex structure and should be represented by several segmention levels according to the needs of upper applications such as MT, IR and IME.

For instance, a long(upper level) segmentation unit may simplify syntactic analysis and IME application but a small(lower level) segmentation unit might be better for information retrieval or word-based statistical summarization. But for now, no Chinese morphological analysis system can do all kinds of these work with only one segmentation standard.

Furthermore, although every segmentation system has good performance, in the analysis of real world text, there are still many out-of-vocabulary words which could not be easily recognized because of the flexibility of Chinese synthetic word construction, especially proper names that could always appear as synthetic words.

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

In this paper, we first make the definition of Chi-

nese synthetic words and classify them into several categories in Section 2. In Section 3, two previous researches on Chinese synthetic words will be introduced. Then we propose a tree-based method for analyzing Chinese synthetic words and make a survey focused on 3-character morphological derived words to get the features for future machine learning process. In Section 4, we do an experiment by using SVM classifier to annotate 3-character morphologically derived words. Finally, Section 5 shows how this method could benefit Chinese morphological analysis and our future work.

2 Detailed study of Chinese synthetic words

2.1 Definition of Chinese words

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', because in Chinese a 'word' is by no means a clear and intuitive notion. But actually for native Chinese speakers, they know that words are those lexical entries which represent a complete concept and occur innately in the form of specific language rules based on the speaker's mental lexicon.

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

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

①one-character words:
人[human], 马[horse], 车[vehicle]

②one-morpheme words:
鹌鹑[quail], 翡翠[jadeite]
鸳鸯[mandarin duck]

③transliteration words:
比萨[pizza], 肯德基[Kentucky]
阿司匹林[aspirin]

The first kind is obvious single words in that an ordinary character in Chinese stands for an independent morpheme with one or several senses. The second kind shows those words which are composed of several characters and always used as a whole. For the last kind, as can be seen from the above examples, if we divide 肯德基[Kentucky] into '肯[can]', '德[moral]' and '基[base]', it definitely can not indicate the meaning of the well-known fried chicken restaurant chain from those three characters. So these three kinds of single-morpheme words should be segmented as one word in any morphological analysis systems.

However, it becomes much more complicated when dealing with synthetic words. 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 constituents. According to this definition, if we divide synthetic words into smaller parts, we could still somehow guess the original meaning from the meaning of internal parts despite the fact that it may not be a very precise one. For example the word 司机[driver]. If we don't know the meaning of '司机', but we do know the meaning of '司' is 'control' and the meaning of '机' is 'machine'. Then we can guess the meaning of '司机' may be connected with 'control' and 'machine', and actually the real meaning is the person who drives(controls) a car(machine).

In Chinese language, according to the encoding standard of GB2312, there are about 6,763 commonly used characters. And in our own system dictionary which has about 129,440 word entries, the number of one-character words is only 6,188 (about 4.78%). From these figures, we know that most Chinese words belong to synthetic words and a deep analysis for synthetic words is necessary for Chinese language processing.

2.2 Classification of Chinese synthetic words

The Synthetic words may be understood as the result or 'output' of a word-formation rule in Chinese language. Classification of these Chinese synthetic words is a difficult task because the 'formation rule' is not so obvious and sometimes even a native speaker can not determine which category a word

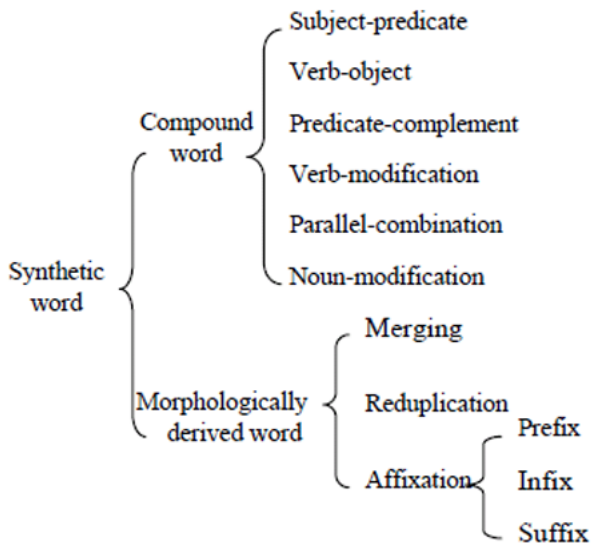

Figure 1: Types of Chinese synthetic words

should belong to. However, since it is quite important to understand the structure of Chinese words, there have been a lot of research on classification of Chinese synthetic words from both linguistic and computational points of view until now. Each of them has divided synthetic words 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 words into the categories as shown in Figure 1 to help us understand the inner constituents of them.

2.3 Compound words

Compound words, whose internal constituents have some syntactic relations with each other, can be divided into the following six kinds according to (Yuanjian He, 2004).

- Subject-predicate[主谓式]

  words that only have subject and predicate parts. This type is subdivided into two types: SV and VS.

  VS: 搬运/工[porter], 裁判/员[referee]

  SV: 胃/下垂[gastroptosia], 地/震[earthquake]

- Verb-object[动宾式]

  words that have verb and object parts, which contains two types: VO and OV.

  OV: 党/代表[representative of party]

  VO: 理/发[haircut], 反/政府[anti-government]

- Verb-modification[动词偏正式]

  words that have a verb part and an adjunct part which are neither subject nor object of the verb. The adjunct part always shows the property of the verb part or be the media of the verb's action.This type contains VX and XV.

  VX: 放大/器[amplifier], 冲印/店[print shop]

  XV: 自动/控制[automatic control]

  批/发[wholesale]

- Predicate-complement[述补式]

  words that have a verb part and a complement part, which shows the result, direction or aspect of the action. This type also have two kinds: VV and VA.

  VV: 跑/出来[running out of]

  打发/掉[get rid of]

  VA: 染/红[dyeing red]

- Parallel-combination[联合式]

  words that have a coordinate structure where the meanings of constituents are same, similar, related or opposite.

  Example: 开/关[switch], 学/习[learning]

  国/家[nation], 兄/弟[brother]

  中/日/韩[China, Japan and Korea]

- Noun-modification[名词偏正式]

  words that have a noun part which is the root of the word, and a modification part which shows the property of the noun part.

  Example: 电/脑[computer], 书/架[book shelf]

  汽车/站[bus stop]

Here we have made some compromises with these categories mentioned above based on the simplicity of machine learning process and our experience on tagging compund words. For example, we only

define SV and VS as the subject-predicate type because it will make machine learning process much easier when it comes to those words with a structure like SVO or SVX. Futhermore, in the case that a word with an internal constituent which has both NN and VV parts of speech, even human annotators can not easily tell which type this word should be categarized into between noun-modification and verb-modification. So we tagged all these words to verb-modification when they have an internal part with both NN and VV parts of speech.

2.4 Morphologically derived words

Morphologically derived words are those which have specific word formation. It can be categoried into the following types:

- 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
  [middle and preliminary school]

  上文+下文➔上下文[context]

  北京市+市长➔北京市长
  [mayer of Beijing city]

- 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[listen], 雄/赳赳[valiantly]

  研究/研究[research]

- Affixation

  words that are composed of a word and an affix(either a prefix, a suffix or an infix).

  Example: 副主席[vice president]
  总工程师[Executive Engineer]

  看不到[can't see]
  听得见[can hear]

  调查局[bureau of investigation]
  安全厅[security agency]

2.5 Exceptions

Apart from compound words and morphologically derived words, there still exist some types of words which need discussion about whether they belong to synthetic words or not. However, we can use some other methods like time expression extraction or named entity recognition to deal with these kinds of words.

- Abbreviations

  expressions that have a short appearance, but stand for a long term.

  Example: 中共➔中国共产党
  [Communist Party of China]

- Factoids

  expressions that indicate date, time, number, money, score or range. This kind of expressions have a large variation in their appearance.

  Example: 2007.1.30

  五点半[five thirty]

  三块五毛六[3.56 yuan]

- 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[sparsely visited]

  先天下之忧而忧
  [be the First to bear hardships]

3 Previous research and tree-based method

3.1 Previous research

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

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 words into different levels of word

components based on some pre-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[toaster]' into lower level such as '烤[bake] / 面包[bread] / 器[machine]', the translation of which will probably give us some information about the original meaning of the whole word. Although this system achieves higher score than other systems that do not have synthetic analysis, it only takes morphologically derived words into account, which means it does not contain information about internal syntactic relations.

The second one (C. Huang, 1997) is actually a proposal of segmentation standard rather than a detailed synthetic word analysis research. It is first used by Sinica when doing the tagging task of Chinese word segmentation. If the tagging object is a synthetic word, one tag among w0, w1 and w2, which stand for 'faithful', 'truthful' and 'graceful', will be selected for it. For example, if we have a synthetic word '北京市安全厅[security agency of Beijing city]', this tagging method will divide the word as follows:

<w2>
<w1><w0>北京</w0><w0>市</w0></w1>
<w1><w0>安全</w0><w0>厅</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 can not be used to 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 words in our current dictionary into the categories defined in section 2.2. Because intuitively most 2-character words, though they could have internal syntactic relations, are often used as single words by native speakers and have already been registered as lexical entries in our Chinese dictionary, we can first classify all 3-character words into those categories and link their internal components to 1-character words and 2-character words which are already in our dictionary.

After finishing the internal structure annotation for 3-character words, we can easily construct 4-character or 5-character words' structure by using 3-character and 2-character words' information and store these structure information into synthetic word dictionary.

Finally, when we get a long synthetic word, we can build a tree structure recursively like in Figure 2 by using the constituent words' internal structures, which have already been stored in our synthetic word dictionary.

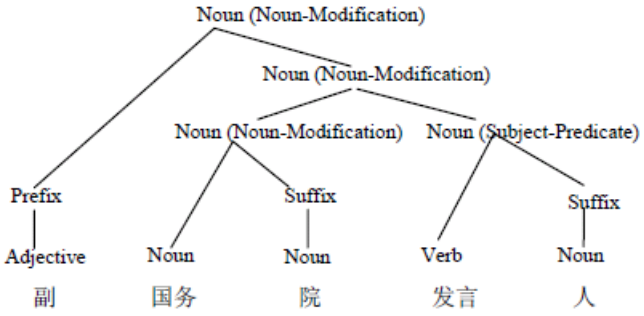

Figure 2: Synthetic word tree of '副国务院发言人[vice spokesman of State Department]'

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.

3.3 Annotation of morphologically derived words in system dictionary

Usually, in Chinese, 2-character words are thought and used as single words by native speakers. And words which have more than two characters are often synthetic words which can be categorized into

compound words or morphologically derived words. So during our work of analyzing Chinese synthetic words, we first choose 3-character words as our main target in that starting from 3-character words will give us a good chain effect when analyzing words which have more than three characters.

At first, there is no other resource at hand except for the morphological analysis system dictionary with 129440 entries. Because a standard set with all category information of Chinese synthetic words is needed in further research, we first extracted 1000 3-character words from our dictionary and annotated them by hand according to the categories introduced in section 2. As the result of human annotation, Table 1 and 2 show the distribution of compound words and morphologically derived words. In Table 1, although about 6% of the 1000 words are single-morpheme words, we still can see that noun-modification words occupy the largest part (62.9%) in synthetic words from a syntactic point of view. Table 2 gives us the information that most synthetic words (83%) have an internal structure with a suffix.

subject-predicate	4.8%
verb-object	2.0%
verb-modification	21.3%
predicate-complement	3.2%
parallel-combination	0.2%
noun-modification	62.9%
single-morpheme word	5.4%

Table 1: Compound words in 1000 words

prefix	infix	suffix	merging	reduplication
9.0%	0.5%	83.0%	1.5%	0.2%

Table 2: Morphologically derived words in 1000 words

Since most 3-character Chinese words have the structure such as 'two+one' or 'one+two' character formation, it is obvious that we should first look at noun-modification words with frequently used suffixes 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 tend to have some fixed structures which makes them easy to recognize.

4 Experiment on Morphologically derived words

In order to apply our proposed tree-based analysis method, we first have to annotate all 3-character words in our dictionary with their internal parts linked to 2-character and 1-character words. Because most Chinese 3-character words have prefix or suffix structure, we assume that it will be much efficient for us to annotate 3-character words if we can classify them from the aspect of morphologically derived words.

Because we don't have any other useful resources except Chinese Gigaword(CGW), We first computed mutual information for all 3-character words in two ways by referencing the CGW. For example, if we have a word ABC and we assume that A, C, AB and BC are all independent entries in our dictionary, we compute the mutual information Mi-pre for A and BC, and the mutual information Mi-suf for AB and C.

Since A, C, AB and BC are all independent words, if the result shows Mi-pre$<$Mi-suf, we could say that the relation between A and BC is more independent than the relation between AB and C, which means it is more possible that A and some other 2-character word XY cound form the word AXY. So the possibility of A being a prefix is greater than the possibility of C being a suffix. Then we could conclude that the word ABC has a prefix internal structure. Otherwise, we could say it has suffix internal structure.

After this process, we got a Mi(Mi-pre or Mi-suf) and a possible internal structure(prefix or suffix) for every 3-character word in system dictionary. By comparing these results to the 1000 extracted words whose internal structure has been already known, we can easily get the correct ones whose possible internal structure is the same as the ones annotated by hand.

Except for the ones which have infix, merging or redplication structure, there are 920 words in the 1000 extracted words which are tagged as prefix or suffix structure by hand. After comparing, we got

676 out of 920 words(73.48%), which was divided correctly by only looking at the internal mutual information in a large corpus(Chinese Gigaword).

The above result shows that overall accuracy is quite low by only taking account the internal mutual information when classifying prefix and suffix structure. Some examples of wrongly classified words are shown in Table 3.

words	Mi-pre	Mi-suf	result
个体户	7.024e-07	9.981e-07	个 / 体户
水彩画	1.084e-06	1.171e-05	水 / 彩画
宝莲灯	1.384e-05	1.978e-05	宝 / 莲灯
交响曲	1.993e-06	2.440e-06	交 / 响曲
大批量	8.971e-08	6.762e-08	大批 / 量

Table 3: Examples of wrongly divided words by only using mutual information

As shown in Table 3, most uncorrect ones are words having suffix internal structure but wrongly classified to have prefix internal structure. This is because we only counted the frequencies of internal parts of words without considerring their properties such as parts of speech and the frequencies that the internal parts show out at a particular position, etc.

In order to improve the whole accuracy when recognizing prefix or suffix internal structure automatically, we used an SVM classifier with the following features(in the case of 3-character string ABC):

1.internal part: A, C, BC, AB, ABC

2.pos of each internal part:
pos(A), pos(C), pos(BC), pos(AB), pos(ABC)

3.frequency of each part in Chinese Gigaword:
fre(A), fre(C), fre(BC), fre(AB), fre(ABC)

4.mutual information of internal part:
Mi-pre(A-BC), Mi-suf(AB-C)

(In the actual classification process, we set the frequency range by 2000 and the mutual information range by $\log_{10}$)

After dividing 80% of 920 words into traning set and 20% into testing set, the accuracy of SVM classifier is 94.02%. The precision and recall are shown in the first row of Table 4. Because the above experiment(A) did not take the existence of 2-character words in system dictionary into account, we then add these features and run the SVM classifier again. Finally we get the result of experiment(B) shown in the second row of Table 4.

Exp	Acc. (%)	F	Prefix(%)		Suffix(%)	
			Rec.	Pre.	Rec.	Pre.
A	94.02	0.56	38.89	100.0	100.0	93.79
B	94.57	0.67	55.56	83.33	98.80	95.35

Table 4: Results of Recall and Precision for words which have prefix or suffix structure

This result is quite unbalanced because there are only a few intances of prefixes both in training(72/736=9.78%) and testing(18/184=9.78%) sets. This is the reason of low recall in classifying instances of prefixes. The following words are the ones which were wrongly classified.

主色调，土坷垃，大批量，小卖部，山大王
市中心，菲军方，零备件，学联会，罗影剧

It turns out that these words contains mainly two types: the first type contains word like '大批量', a prefix structure word whose last character has a quite high probability to be a suffix. This makes it difficult for SVM to determine to which class the whole word should be classified; an example of the second type is suffix structure word like '罗影剧', whose front part '罗影' does not appear in the Chinese Gigaword independently, which in the end make it unsure for SVM to classify it into suffix structure word.

Though there are some words that were wrongly categorized, we still got a overall accuracy of 94.57% which would be much higher if we recursively use SVMs for classification. We believe that this method could classify morphologically derived words quite efficiently if we add some more rules for recognizing merging and reduplication words. And by applying the tree-based method, we could use this method on words with more than 3 characters in future.

5 Conclusion and future work

This paper proposed a tree-based method for analyzing Chinese synthetic words by constructing a Chinese synthetic word dictionary. This method is based on the classification of Chinese synthetic words both from syntactic and morphological ways. After annotating and investigating the distribution of one thousand 3-character words, we used frequencies and

mutual information as features from Chinese Gigaword for machine learning. With these features, we tried to classify morphologically derived words into prefix or suffix internal structure by using SVM classifier.

For future work, we have to take other morphological internal structures into account and try to classify all synthetic words into morphologically derived word categories. Then, we should also find some thesaurus that contain syntactic information of words or characters to help us analyze the compound words' internal structure. Finally, after gathering the information of Chinese synthetic words from both syntactic and morphological aspect, we will build a Chinese synthetic word dictionary and try to use it to improve the performance of our morphological analysis system and unknown word extraction.

References

[Andi Wu2003] 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. Huang1997] C. Huang, K. Chen and L. Chang 1997. *Segmentation standard for Chinese natural language processing.* International Journal of Computational Linguistics and Chinese Language Processing

[Chooi-Ling Goh2006] Chooi-Ling Goh, Jia Lu, Yuchang Cheng, Masayuki Asahara and Yuji Matsumoto 2006. *The Construction of a Dictionary for a Two-layer Chinese Morphological Analyzer.* PACLIC 2006

[Hiroshi Nakagawa, Hiroyuki Kojima, Akira Maeda2004] Hiroshi Nakagawa, Hiroyuki Kojima, Akira Maeda 2004. *Chinese Term Extraction from Web Pages Based on Compound word Productivity.* Third SIGHAN Workshop on Chinese Language Processing, ACL 2004

[Huihsin Tseng and Keh-Jiann Chen2002] Huihsin Tseng and Keh-Jiann Chen 2002. *Design of Chinese Morphological Analyzer.* First SIGHAN Workshop 2002

[Jerome L. Packard2000] Jerome L. Packard 2000. *The Morphology of Chinese-A Linguistic and Cognitive Approach.*

[Keh-Jiann Chen, Chao-jan Chen2000] Keh-Jiann Chen, Chao-jan Chen 2000. *Automatic Semantic Classification for Chinese Unknown Compound Nouns.* ACL 2000

[Shengfen Luo and Maosong Sun2003] Shengfen Luo and Maosong Sun 2003. *Two-character Chinese Word Extraction Based on Hybrid of Internal and Contextual Measures.* ACL 2003

[Yuanjian He2004] Yuanjian He 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. The Chinese University of Hong Kong