# Disambiguation of Chinese Polyphonic Characters

Hong Zhang, Jiangsheng Yu, Weidong Zhan, Shiwen Yu
The Institute of Computational Linguistics
Peking University, Beijing, 100871, P.R. China
http://www.icl.pku.edu.cn

**ABSTRACT** *In this paper,* [1] *we survey the Chinese polyphonic characters and introduce our solution to convert them into Pinyin. We focus on Polyphonic Mono-Character words, which we think are the key in converting Chinese text into Pinyin. Firstly, Some statistical data from a corpus are analyzed. Then we introduce the methods we used to judge the Pinyin of PMC words and the polyphones in undefined words in our automatic Chinese text-to-Pinyin system. We conclude that to utilize some Chinese context patterns, combined with priority files, is an effective way to solve the Chinese polyphonic characters.*

**KEYWORD** *Pinyin, Polyphonic Character, Polyphones, text-to-Pinyin,*

## I. Introduction

As an essential technique of Chinese natural language processing, the conversion of a Chinese text to a string of Pinyin, the Chinese phonetic symbol system, including disambiguation of Chinese polyphonic characters, is the first and an important stage of text-to-speech. Correctly converted Pinyin strings are prerequisite to later acoustic processing. But comprehensive studies of this issue are still not available in current literature. This paper firstly surveys a Chinese corpus about Chinese polyphonic characters, and then introduces our technical approach used in our automatic Chinese to Pinyin conversion system.

## II. Chinese Polyphonic Characters

The Chinese Polyphonic Characters, in this paper, refer to those Chinese characters with more than one pronunciation. For example, "背" can pronounces "*bei4*" as well as "*bei1*"; "朝"can pronounces "*zhao1*" as well as "*chao2*". They are called Chinese Polyphonic Characters. To make this term more clear, a formal definition of this term is given below.

**Definition** Let $C$ be the set of all Chinese characters and let $P$ be the set of all possible Chinese pronunciations. There exists a correspondence

$$f : C \to P$$

which maps each character $c$ to its all-possible pronunciations:

$$P_c = \{ p_1, p_2, \cdots, p_n \}.$$

If $| P_c | = 1$, then $c$ is called *monophonic character*, otherwise *polyphonic character*.

**Definition** Let $W$ be the set of all Chinese

[1] This work was supported by National 973 high-tech Project, National Science Foundation, and Peking University 985 Project.

words, then $f$ induces the restricted correspondence $f^w : C \cap W \to P$, which maps the word $c$ to its all-possible pronunciations $P_c^w = \{p_1^w, p_2^w, \cdots, p_m^w\}$. If $|P_c^w| = 1$, then $c$ is called *monophonic character word*, otherwise *polyphonic character word*.

**Example** $单 \in C$, $f(单) = \{dan1, shan4, chan2\}$, while $f^w(单) = \{dan1, shan4\}$( 单 sounds *chan2* only in 单于).

Because the phonetic notation is based on the well-done segmentation (i.e., the *Phoneticizing Unit* is word) and each element in the set of $W - C$ corresponds to a unique pronunciation except 53 words such as 朝阳 (*zhao1yang2* or *chao2yang2*), we could just focus on the disambiguation of multiple-pronunciation character word in a given sentence.

Polyphonic Characters here do not include those characters that are off-beating, ER-pronouncing and other tone modifications. For instance, “爸爸” is actually pronounced “ba4*ba5*”(off-beating) instead of “ba4*ba4*” in natural speech, though the Pinyin of “爸” is “ba4”. ER-pronouncing refers to such cases as “花儿” pronouncing “*huar1*” instead of “*hua1er2*”, combining the pronunciation of two characters into that of one character. As an example of tone modifications, the tone of “不是” is modified in natural speech into “*bu2*shi4”, instead of “*bu4*shi4”even though the Pinyin of “不” is “*bu4*”. All of these cases are ways to make speech more “natural”, not only to make it correct, which is the primary mission of conversion Chinese to Pinyin. Although to make Pinyin correct and natural is our ultimate goal, correctly converting characters to Pinyin, that is, disambiguating Polyphones, is the main task in this paper.

The set of Chinese Polyphonic Characters, in this paper, is produced according to “Modern Chinese Dictionary”, because of its authority in Chinese. We draw out 809 polyphonic characters from “the Grammatical Knowledge-base of Contemporary Chinese” (GKCC Dictionary), which is annotated Pinyin according to “modern Chinese Dictionary”. This set still has a small part of off-beating and tone modification. It includes “guo5”, which is off-beating, as one of the two kinds of pronunciation of “过”. It also includes “bu2”, which is tone modification, as one the two kinds of pronunciation of “不”. But this paper still takes them into consideration in dealing with polyphones, since they are denoted in “modern Chinese Dictionary”.

We divided the polyphonic characters into three categories based on the disambiguation approaches available. The first one includes those characters which seldom occur in texts, such as “陂”, “嗌”, and “吡”. It also includes those that occur frequently but seldom has other pronunciations. For example, “厂” always pronounces “chang3” and has few cases with pronunciation of “an1”; “并”, usually pronounces “bing4”, seldom pronounces “bing1”; and “采”, usually pronounces “cai3”, seldom pronounces “cai4”. The category, which includes about 537 Chinese characters, is called Category A in this

paper. As the second category, some characters always have one kind of pronunciation when each of them acts as a word in text, although they have more than one pronunciation as a whole. That is, they only have other kinds of pronunciation when they combine with other characters to form words. For example, “否” pronounces “fou3” when it acts as a word itself, but pronounces “pi3” when it combines with other characters into words such as “臧否” and “否极泰来”; “大” always pronounces “da4” when it acts as a word itself, but pronounces “dai4” when it is in the words of “大夫” (it sometimes pronounces “da4” in this word)and “大王”. We can list all the exceptions in the form of word in a dictionary, and get these exceptional pronunciations from the words’ Pinyin, and set the normal pronunciations as the choices when they act as word by themselves. This category is called as B in this paper.

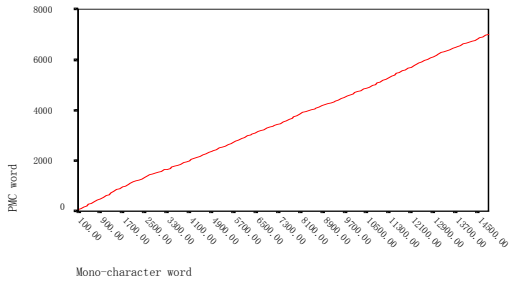
We called other polyphones C category. Let’s look at this category in the granularity of word. All of them have more than one pronunciation even when each of them acts as a word in a sentence. For example, “把” can pronounces “ba3” and “ba4” when it is a word itself; similarly, “单” can pronounces “dan1” and “shan4” as a word, though it pronounces “chan2” only when it is in the word “单于”。 This paper called these words as polyphonic mono-character words (PMC word). Most of polyphonic characters have definite pronunciation when they combine with other characters to words. But there are a few exceptions. For example, “朝阳” can pronounces “zhao1yang2” or “chao2yang2”; “倒车” can pronounces “dao4che1” or “dao3che1”; “倾倒” can pronounces “qing1dao4” or “qing1dao3”. In the GKCC Dictionary, there are 53 such words. This paper called these words as polyphonic poly-character words (PPC word).

Before talking about our approach to convert Chinese text to Pinyin, we first take a look at some data related to Chinese polyphones from our corpus.

## III. Corpus Data

We surveyed the 809 Chinese polyphonic characters in a corpus with 1 million Chinese characters selected from Ren-Ming Daily, and found that there are more than 25% characters are polyphones. We also found that many of the mono-character words are PMC words. The following section includes some data based on them.

Figure 1:




### 3.1 Survey on PMC word

Figure 1 describes the proportion of PMC words in all the mono-character words. It shows that there are above 50% of mono-character words (word tokens) are PMC words. We can see that PMC words account for a large part in all of the mono-character ones in

figure 1.

The table 1 below describes the distribution of PMC words (word types) in the mono-character words (word-types).

Table 1:

| Frequency | PMC word (word type) | Mono-character word (word type) | Percentage |
|---|---|---|---|
| > 2000 | 16 | 31 | 51.6% |
| 100 ~ 2000 | 76 | 300 | 25.3% |
| 10 ~ 100 | 162 | 886 | 18.3% |
| < 10 | 214 | 1422 | 15.0% |

It shows that most of the high frequently occurred mono-character word types are PMC ones. All of the data above shows the importance of PMC words in the issue of converting Chinese text to Pinyin string. Some of these PMC words always have one pronunciation, which belong to the category B. But the others are different. We can see it in the table below:

Table 2:

| Frequency | P1 | P2 | PMC word | Mono-character words |
|---|---|---|---|---|
| > 2000 | 2 | 6 | 8 | 10 |
| 100 ~ 2000 | 40 | 39 | 79 | 155 |
| 10 ~ 100 | 112 | 34 | 145 | 678 |
| < 10 | 88 | 15 | 93 | 685 |
| 总数 | 244 | 92 | 336 | 1528 |

P1: PMC words always with one pronunciation

P2: PMC words always with more than one pronunciations

It shows that in the high frequently occurred mono-character words, PMC words account for majority. The six PMC words with highest frequency mentioned in column P2 are “着”, “了”, “过”, “上”, “这”, and “不”. The two ones mentioned in column P1 with highest frequency are “把” and “的”.

We also can see that those 77 polyphones in column P2 should be more important than the others when we disambiguate polyphones.

**3.2 Polyphones in Poly-character Words**

There are about 45.5% words that contain polyphones in the 61537 poly-character words in GKCC Dictionary. Among the undefined words (all of them are poly-character words) extracted from the corpus, there are only 11.94% that contain polyphones. Its shows that dictionary can play an important role in solving polyphones in poly-character words,.

Besides, we also find that most of the poly-character words containing polyphones begin with the polyphones, just illustrated in the table below:

Table 3:

| Head | | Tail | | Middle | | Undefined Words containing polyphones | All words |
|---|---|---|---|---|---|---|---|
| Number | Percentage | Number | Percentage | Number | Percentage | | |
| 13189 | 47.08% | 9432 | 33.67% | 5393 | 19.25% | 28014 | 61537 |

There is similar distribution in the undefined words in the corpus.

**3.3 Polyphones in the undefined words**

Some polyphones in the undefined words always each have one pronunciation, and some others not, which is illustrated in Table 4.

Table 4:

| frequency | M1 | M2 |
|---|---|---|
| > 100 | 10 | 11 |
| 10 ~ 100 | 10 | 66 |
| 2 ~ 10 | 165 | 41 |
| 1 | 73 | 0 |
| Total | 336 | 118 |

M1: the polyphone which has one pronunciation in all of the undefined words

M2: the polyphone which has more than one pronunciation in all of the undefined words

The set of column M1 is different from the set of column P1 in table 2, although there is an intersection between them. It is indicated in the table below regarding to the above 77 polyphones in table 2.

Table 5:

| | Same | Different |
|---|---|---|
| Number of different polyphones | 34 | 43 |

Same: the polyphone (chosen from the above 77 polyphones) which has one pronunciation in all of the undefined words

Different: the polyphone (chosen from the 77 polyphones) which has more than one pronunciation in all of the undefined words.

**3.4 Conclusions of Corpus Data**

From above data, we can conclude that:

(1) Polyphones concentrate on high frequently occurred PMC words;

(2) Dictionary contains plenty of poly-character words containing polyphones;

(3) Some polyphones have one pronunciation as PMC words;

(4) Some Polyphones have one pronunciation in undefined poly-character words;

(5) Most of the poly-character words containing polyphones begin with a polyphone.

## IV. Approach of disambiguation

One of our approaches to disambiguate polyphones in Chinese text is to segment the text into words and get the Pinyin of poly-character words from a dictionary. According to our test with a corpus, the correct rate of converting text to Pinyin is 95.1% if simple using a priority file, not segmenting the text at first. But after we segment it using our segmentation and part-of-speech tagging software, the correct rate is increased to 97.1%, though there are some errors in segmentation and tags. If we have all the text correctly segmented and tagged, the correct rate of converting text to Pinyin, combined with priority file, is 98.7%. The correct rate here is computed using the formula below:

Correct rate = $\frac{N_{correct}}{N_{total}}$

$N_{correct}$ : Number of characters correctly converted to Pinyin

$N_{total}$ : Number of characters in text

Based on segmentation and tagging, we combine related rules with priority files, both of which obtained from the training corpus, to convert text to Pinyin.

The approach to disambiguate category A polyphones referred above is simple.

Considering that they seldom occur or seldom occur with another pronunciation in text, we can treat them as characters with only one pronunciation by setting priority to a specific pronunciation. For example, we can set the priority of "chang3" of "厂" to the highest priority and make the other pronunciation "an" always inactive.

As to a polyphone in category B, we can set one pronunciation active when it act as a word itself, and getting all the other pronunciation possibilities from a dictionary, by first segmenting the text to words.

Category C is much more complicated compared with these two categories, because we always need the information in syntax, semantics, even context to convert correctly the PMC word to Pinyin, although segmentation can help a lot in other cases. For example, when "地" is a word by itself, we have to judge its pronunciation according to its part-of-speech. Another PMC word "冲" also need part-of-speech information to be disambiguated. But in the sentences "这棵树长了三厘米" and "衣服长了三厘米", we have to depend on the semantic information of the word before the PMC word "长" to judge its pronunciation. More complicated, in a sentence like "他背着我去医院", we should refer to the context of the sentence to judge suitable pronunciation of the polyphone of "背". Besides in mono-character word, these polyphones sometimes occur in undefined words, surrounded by other characters, and thus we have to judge their pronunciation in these situations. Category C polyphones as character words and in undefined words will be discussed in detail below.

As to the PPC words, because they seldom occur in text and are similar to category C in our solution, they are not specially discussed in this paper.

**4.1 PMC words**

In category C, There are 15 PMC words expressing sentence mood, such as "啊", "唉", "吧", and "呀", etc.. We now simply choose their Pinyin according to the punctuations of the sentences where these words are, not considering the tones or context or other complex factors to judge it more precisely or naturally. There are also some PMC words that have specific pronunciations when they act as the family name of Chinese, such as "曾", "任", "区", and "朴", etc.. We have to depend to a name recognition tool to convert them correctly to Pinyin. Besides, there are some suffix characters of PMC words such as "边", "家", and "头", all with different pronunciations from their normal ones. We convert them also according to the result of segmentation and tagging.

To all the other ones in PMC words in category C, we utilize related context patterns and syntactic information from training text and cooperate them with priority files to judge their specific Pinyin in context. When we convert a PMC word to Pinyin after segmenting and tagging a text, we firstly check the rule of the word, if there is one. If the rule is satisfied, choose the Pinyin that the rule stipulates. Otherwise, choose the default one defined in the priority file for PMC words. To

be expedient and practical, we set the one that is most complicated and hard to be listed in rules to be the default Pinyin.

We take some examples with the most frequently occurred 4 PMC words illustrated in table 2, which are “着”, “了”, “过”, and “上”.

Between the pronunciations of “上”: “shang4” and “shang5”, we can differentiate them according to the part-of-speech tags immediately before and after the word. The condition expression of “shang5” in the corresponding rule is that the word before it is noun and the one after it is not noun. We set “shang4” as the default Pinyin.

The other three PMC words have some similarities in that each of them can play as dynamic auxiliary, and pronounces differently then from in other cases. They also always appear in some special context patterns. Some context patterns are related to the partible verbs (for example, the word “*安心*” can be parted in “*安不了心*”), which are a special phenomenon in Chinese. The related patterns are listed in the table below:

Table 5:

| | Context Pattern | Polyphone | Pinyin | Example |
|---|---|---|---|---|
| 1 | V + 不 + 了 + N | 了 | Liao3 | 安不了心 |
| 2 | V + 得 + 了 + N | 了 | Liao3 | 安得了心 |
| 3 | V + 着 + N | 着 | Zhe5 | 理着发 |
| 4 | V + 了 + N | 了 | Le5 | 理了发 |
| 5 | V + 过 + N | 过 | Guo5 | 理过发 |

We can judge the proper Pinyin of these three PMC words by identifying these context patterns. We have discovered these partible verbs from the features in the GKCC Dictionary. There are also some other context patterns where these PMC words appear immediately after some special verbs. They are illustrated in table 6.

Table 6:

| | Context pattern | Polyphone | Pinyin | Example |
|---|---|---|---|---|
| 1 | V + 不了 | 了 | Liao3 | 解决不了 |
| 2 | V + 得了 | 了 | Liao3 | 解决得了 |
| 3 | V + 不着 | 着 | Zhao2 | 吃不着 |
| 4 | V + 得着 | 着 | Zhao2 | 吃得着 |
| 5 | 没 + V + 着 | 着 | Zhao2 | 没见着 |
| 6 | V + 着了 | 着/了 | Zhao2/le5 | 见着了 |
| 7 | V + 着 + NP + 了 | 着/了 | Zhao2/le5 | 吃着苹果了 |
| 8 | V + 了 + 一 + V | 了 | Le5 | 试了一试 |
| 9 | V + 了 + V | 了 | Le5 | 试了试 |
| 10 | V + 得过 | 过 | Guo4 | 信得过 |
| 11 | V + 过 | 过 | Guo4 | 翻过这座山 |
| 12 | V + 着/了/过 | 着/了/过 | Zhe5/le5/guo5 | 吃着/吃了/吃过 |

Most of transitive verbs are satisfactory to these patterns. We also can get this information from related features in GKCC Dictionary. Besides in verbs, there are similar patterns in adjectives, and the approach to apply them is the same.

But apparently, there are some patterns conflicting with other ones in above tables. In the table 6, the pattern 12 sometimes conflict with pattern 5, 6, 7, or 11. For example, the verb “穿” satisfies pattern 11 as well as pattern 12. So when judging the Pinyin of the polyphone “过” in the sentence “穿过这个山洞”, in which its Pinyin should be “guo4”, and the sentence “穿过这件衣服”, in which its

Pinyin should be “guo5”, there will be a confliction. According to our investigation, there are about 28 verbs that can have this confliction, and all of them are mono-character words. We have to depend on semantic or even context information to judge them, and by now they are still not solved in our system.

There are also conflicts related to “着”. For example, “吃” satisfies pattern 7 as well as pattern 12. But it should pronounce “zhao2” in the sentence of “吃着苹果了”, and pronounce “zhe5” in the sentence of “他正吃着苹果”. Our approach to disambiguate them is to set the pattern 12 the lowest priority when matching the patterns, and set “zhe5” as the default pronunciation.

We also utilize related context patterns to disambiguate other PMC words. We take an example of “长”, which is different from “着”, “了”, and “过” in that it can not act as an auxiliary. It has two Pinyin corresponding to the two meanings of “grow” as a verb and “long” as an adjective. We set “zhang3” as the default pronunciation, and stipulate the related patterns of “chang2” as below:

Table 7:

| 序号 | 模式 | 发音 | 例子 |
|---|---|---|---|
| 1 | “长” + “时间” | Chang2 | 长时间的努力 |
| 2 | “长” + “达” | Chang2 | 长达十年 |
| 3 | “长” + 长度单位 | Chang2 | 长三厘米 |
| 4 | 长度单位 + “长” | Chang2 | 三厘米长 |
| 5 | 程度副词 + “长” | Chang2 | 很长/最长/太长 |
| 6 | 数量词+“长”+名词 | Chang2 | 两根长木棍 |
| 7 | “以” + “见长” | Chang2 | 以能写文章见长 |
| 8 | V +“长”（述补动词） | Chang2 | 拉长/拖长 |
| 9 | “长” + “叹” | Chang2 | 长叹一声 |

These are not all of the patterns related to “chang2”, and the number of them should increase with learning from larger training corpus.

**4.2 Undefined Poly-character words**

The undefined poly-character words include time terms, names of people, location, and organization, as well as some professional terms, and some verb phrases. Some of the polyphones in these words can be disambiguated by dividing into small words and get their Pinyin from a dictionary. For example, we can divide “无人过问” into “无”, “人”, and “过问”, and then get the Pinyin of “过问” from a dictionary to solve the polyphone of “过”; We also can divide “步行者” into “步行” and “者” to get the Pinyin of “步行” from a dictionary to solve the polyphone of “行”. Some other polyphones can be judged by identify the structure of the words such as suffix and prefix. For example, we can judge the polyphone of “率” in “产出率” by identify its role of suffix in the word, and we know that its Pinyin is “lv4” when it act as a suffix. The polyphones in all kind of names are most cases in the polyphones in undefined poly-character words. From the data in part 3, we can see that some polyphones always have one pronunciation when they are in the undefined words. So it’s a convenient and effective approach to create a special priority file aiming at polyphones in the undefined words.

## V. The structure of our system

The structure of our system of converting context to Pinyin is illustrated in figure 2 (at the end of the paper).

## VI. Conclusions and Future's Work

As it is showed in this paper, to convert Chinese text to Pinyin string involves identity techniques with other natural language processing ones, such as segmenting, tagging, semantic and context analysis. But it's different from these techniques in that it demands much higher precision, since its basic precision rate is above 95%. So it needs more precise processing techniques.

Our test shows that our system of converting Chinese text to Pinyin can reach the correct rate of above 98% or even 99%. Because the rules can be enriched and will be effective immediately, the correct rate is hopeful to be stable and increase. As a conclusion of our work and this paper, we think that to discover and utilize related Chinese context patterns to judge the Pinyin of PMC words is an important part in disambiguating the polyphones. It is also significant in other natural language processing techniques, such as segmenting and tagging. As a specific language characteristic of Chinese, many latent or manifest patterns frequently appear in Chinese sentences, such as "以…为", "为…所", and "所…的", which have characters as manifestation of the patterns, and "报了一个名", "澡也洗过了", and "长了…出来", which have latent information of the patterns, and can be identified with analyzing the features of the verbs, adjectives, and other kind of words around which the polyphones appear. We have discovered and utilized a few related patterns by now, and more and deeper discovery is anticipated. We also need to make the priority file for the polyphones in undefined words more perfectly. Besides these work, we need to explore in semantic information to solve the sentences like "衣服长了三厘米" and "树长了三厘米". To solve the problem such as "他背着我去医院" will be more challenging to us.

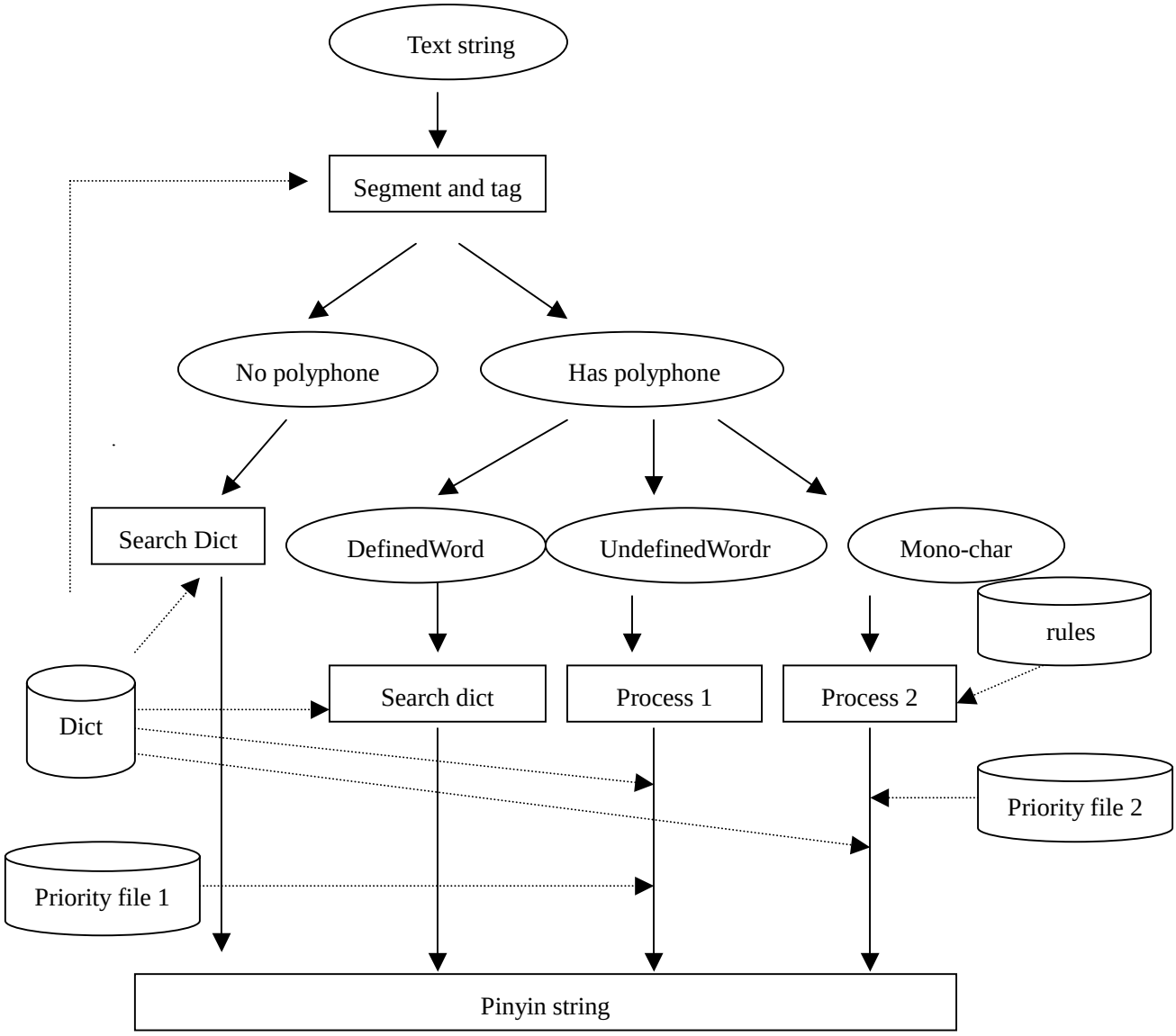
## Acknowledgements

We would like to deeply thank Prof. Xuefeng Zhu for her guidance and help in our work. Also, we appreciate Prof. Huiming Duan and Dr. Bin Sun, without their kindly suggestions on segmentation and POS tagging, our work will seem bleak. Lastly, thank all the researchers in our seminar of Natural Language Processing.

## Reference

1. 俞士汶等, 《现代汉语语法信息词典详解》, 清华大学出版社，1999.4
2. 中国社会科学院语言研究所词典编辑室,《现代汉语词典》修订本，1997
3. 俞士汶，朱学锋，段惠明，大规模现代汉语标注语料库的加工规范,《多语言信息处理国际会议 2000 ICMIP 论文集》, 2000.8
4. Bing Swen and Yu Shiwen, A Graded Approach for the Efficient Resolution of Chinese Word Segmentation Ambiguities, Proceedings of NLPRS 99, Beijing, Nov. 1999
5. 揭春雨，刘源，梁南元，论汉语自动分词方法，中文信息学报，vol. 3 No. 1., 1989
6. 王文俊等，破音字发音的预测方法，台湾第八届计算语言学会议论文集，1995

Figure 2


Text string
Segment and tag
No polyphone
Has polyphone
Search Dict
DefinedWord
UndefinedWordr
Mono-char
rules
Dict
Search dict
Process 1
Process 2
Priority file 2
Priority file 1
Pinyin string