ABSTRACT

The primary task of Chinese language processing is to establish efficient and accurate segmentation strategy. With the Chinese’s characteristics of been idea-phonetic language, the paper advances an automatic segmentation algorithm that is based on Chinese phoneme to realize disambiguation. First, the candidate tag set, which consists of ambiguous phrases that result from Chinese polyphones, is built up, and every possible segmentation result of each phrase compose the segmentation tag set, then, the calculation of posterior probability is transformed into solving optimization problem, and with genetic algorithm to get the optimal solution, furthermore, the approach also resolve the sparse data problem in HMM. The experiment shows that with this method to solve the ambiguity caused by polyphones is practicable and has a good effect.

Keywords: Chinese, Polyphones, Segmentation and Genetic Algorithm.

1. THE IMPORTANCE OF CHINESE PHONEME DISAMBIGUATION

In accordance with word’s morphology, the language can be divided into three categories: analytic language, agglutinative language and inflectional language. Chinese is one kind of analytic language with its characteristic is that: it hasn’t any affixation to show the grammatical meaning, it has few morphological changes and it has no space to act as break sign, so the main task of Chinese lexical analysis is to realize segmentation automatically, which means segment Chinese character strings into words. Automatic segmentation is the first step of modern Chinese syntactic analysis and is the basic of following analysis of syntax and semantic. It has been applied to many fields, such as: word frequency statistics, automatic indexing, automatic summarization, information extraction and machine translation. Thus, to establish efficient and accurate segmentation strategy is a necessary condition for high-quality Chinese processing. Since the early 1980s, many segmentation methods have been put forward. With the knowledge resources they used, they can be divided into three kinds of methods: rule-based, statistics-based and the combination of the former two. But all these methods have neglected one very important linguistic unit in Chinese, namely phoneme. The phoneme is composed of initial consonants, compound vowels and tones, and once one of them changes, the pronunciation changes also. The polyphones can be classified into two types: the character of variant pronunciations, the character of polyphonic and polysemous. The character of variant pronunciations means the pronunciations are different but the meaning is the same, while the character of polyphonic and polysemous means the pronunciations are different and the meanings are also different. Since the character of variant pronunciations has many-to-one relationship between the pronunciations and the meaning, and can’t cause segmentation ambiguity, therefore they are not been considered. However, the character of polyphonic and polysemous has many-to-many relationship between the pronunciations and the meanings, so the paper focuses on this kind of polyphones. There are many polyphones in Chinese. According to statistics, a total of 856 polyphones are in Modern Chinese Common Character Set (1998) and Chinese Coded Character Set for Information Interchange Basic Set, with many common characters included, such as: 了, 和, 得, 中, 长, 少, 行 etc. They are been used with very high frequency. In 1970’s, Chomsky also proposed that semantic interpretation must take phoneme into consideration, though this point of view can’t work with English and other kinds of inflectional language, but it is work with Chinese, because the polyphones’ pronunciation and meaning is closely related to each other in Chinese.

2. THE TYPES OF SEGMENTATION AMBIGUITY

The same Chinese phrase can be segment into different results for the reason of been lack of break signs, and thus lead to ambiguity. Segmentation ambiguity is one significant factor that impacts the correct rate of the whole segmentation system; it is also the toughest problem that encountered in the segmentation phase. There are three types of segmentation ambiguity:

1) Crossing ambiguity

The Chinese phrase $abc$ can be segment into not only $ab/c$, but also $a/bc$. For example:
①这块肉的确切得不错。
   (It can be segmented into:
   这块肉/的 dié/确切/得/不错/)。
②他的确切地址在这儿。
   (It can be segmented into:
   他/的 de/确切/地址/在/这儿/)。

(2) Combinatorial ambiguity
   The Chinese phrase $ab$ is a word, while $a$ and $b$ can also be words in the sentence. For example:
   ①他学会了解方程。
   (It can be segmented into:
   他/学会/了 le/解/方程/)。
   ②我对小华比较了解。
   (It can be segmented into:
   我/对/小华/比较/了 liǎo o/解/)。

(3) Mixed ambiguity
   This type of ambiguity is cause by crossing ambiguity and combinatorial ambiguity using self-nested and cross combination. For example:
   ①他不了解答题的方法。
   (It can be segmented into:
   他/不/了 liǎo o/解/答题/的方法/)。
   ②他学会了解方程。
   (It can be segmented into:
   他/学会/了 le/解/方程/)。
   ③我们都了解他。
   (It can be segmented into:
   我/都/了 liǎo o/解/他/)。
   Here, “了解答” is of crossing ambiguity, while “了解” is of combinatorial ambiguity.

At present, there are three kinds of segmentation disambiguation methods respectively based on rule, statistics and the combination of the two former. Rule-based method mainly uses the information of syntax, semantic and pragmatic about the ambiguous phrase and the adjacent words to realize disambiguation. When with syntax rules, it can’t correctly segment the phrase, the semantic information is considered, and if the phrase still can’t be correctly segmented, the pragmatic information will be taken into account. However, statistics-based method achieves disambiguation by designing algorithm with the statistical information of words. This paper advances an automatic segmentation algorithm that is based on Chinese phoneme to realize disambiguation.

First, the candidate tag set, which consists of the phrases with segmentation ambiguity that result from Chinese polyphones, is built up. Then, every possible segmentation result of each phrase composes the segmentation tag set. Subsequently, HMM is used to construct the possibility of each segmentation result and with it to define the fitness function of genetic algorithm. Finally, simple variation is chosen to obtain the optimal segmentation results.

### 3. DESCRIPTION OF AUTOMATIC SEGMENTATION ALGORITHM BASED ON CHINESE PHONEME

#### 3.1 The selection of polyphones

From the statistics above, we get 856 polyphones, and 749 of them have two phonetic items, which account for 87.4%[1]. On the other hand, although the number of polyphones is large, their frequency of use is varying. After been sorted according to their frequency of use, we can see the first 100 commonly used polyphones account for more than 90%, and the first 180 commonly used polyphones account for more than 95%, so if the segmentation disambiguation of these 180 polyphones can be achieved, the segmentation disambiguation of polyphones is basically solved[2]. Here, 76 most common polyphones with two phonetic items is selected from them[3], as is listed in the table 1.

<table>
<thead>
<tr>
<th>扳</th>
<th>背</th>
<th>奔</th>
<th>篁</th>
<th>朝</th>
<th>冲</th>
<th>长</th>
<th>处</th>
<th>打</th>
<th>担</th>
</tr>
</thead>
<tbody>
<tr>
<td>弹</td>
<td>地</td>
<td>都</td>
<td>倒</td>
<td>都</td>
<td>发</td>
<td>房</td>
<td>缝</td>
<td>夫</td>
<td>干</td>
</tr>
<tr>
<td>更</td>
<td>冠</td>
<td>好</td>
<td>号</td>
<td>还</td>
<td>荷</td>
<td>间</td>
<td>隆</td>
<td>觉</td>
<td>假</td>
</tr>
<tr>
<td>看</td>
<td>将</td>
<td>卷</td>
<td>卡</td>
<td>空</td>
<td>乐</td>
<td>景</td>
<td>笼</td>
<td>宁</td>
<td>磨</td>
</tr>
<tr>
<td>屏</td>
<td>铺</td>
<td>强</td>
<td>奇</td>
<td>切</td>
<td>曲</td>
<td>圈</td>
<td>散</td>
<td>宿</td>
<td>供</td>
</tr>
<tr>
<td>少</td>
<td>箱</td>
<td>率</td>
<td>省</td>
<td>盛</td>
<td>越</td>
<td>营</td>
<td>为</td>
<td>相</td>
<td>行</td>
</tr>
<tr>
<td>应</td>
<td>肮</td>
<td>曾</td>
<td>粘</td>
<td>只</td>
<td>中</td>
<td>种</td>
<td>重</td>
<td>转</td>
<td>钻</td>
</tr>
<tr>
<td>折</td>
<td>壁</td>
<td>传</td>
<td>钉</td>
<td>斗</td>
<td>混</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### 3.2 The recognition of ambiguous phrases

The recognition and collection of the ambiguity phrases that result from Chinese polyphones can be achieved by the methods of Minimum Matching and Reverse Maximum Matching, and the minimal matching unit is single character. With this method we can recognize not only crossing ambiguity but also combinatorial ambiguity[4]. The obtained ambiguous phrases constitute the candidate tag set, and is denoted as $W$, while each phrase is denoted as $w_i$, $f_i$ is the set of every segmentation result of $w_i$, and
\[ F_W = \{\{f_1\}, \{f_2\}, \ldots, \{f_i\}, \ldots\} \] is the segmentation tag set of the corresponding \( W \).

### 3.3 The optimal segmentation algorithm based on genetic algorithm

1) **Definition of fitness function**

Let \( w_i \) is the \( i \)-th phrase in the candidate tag set \( W \), and \( f_i \) is the set of every segmentation result of \( w_i \), \( w_{i,i+m} \) is composed of phrases from \( i \)-th to \( i+m \)-th, \( f_{i,i+m} \) is the corresponding sets of segmentation results, then with second-order conditions for HMM and with the hypothesis that the words are independent of each other, the optimal segmentation result is:

\[
\hat{f}_{i,n} = \arg_{f_i} \max P(f_i | w_{i,n})
\]

Here, \( P(w_i | f_i) \) is the ratio of the times of ambiguous phrase \( w_i \) is segmented into \( f_i \) to the times of \( f_i \) exist in training corpus, and \( P(f_i | f_{i-1}, f_{i-2}) \) is the ratio of the times of \( f_i \) exist in training corpus to the times of \( f_{i-1} \) and \( f_{i-2} \) exist together in training corpus.

Since the data in expression (1) is sparse, genetic algorithm is used to compute which segmentation result is best, and its fitness function is the logarithm of expression (1):

\[
\text{fitness}(f_{i,n} | w_{i,n}) = \log \prod_{i=1}^{n} P(w_i | f_i)P(f_i | f_{i-1}, f_{i-2})
\] (2)

With the fitness function, the occurrence frequency of each segmentation result of the phrase can be obtained.

2) **Chromosome coding**

Let the length of the chromosome is equal to the number of phrases in the candidate tag set \( W \), thus per chromosome locus is corresponding to the character in the phrases, and gene is represent the coding of the segmentation result of the phrases. According to limited choice of strategy, the chromosome gene can get its possible coding according to corresponding segmentation result, and the gene randomly selects one coding from the limited range, which is denoted as:

\[
\text{chrom}(f_{i,n}^0 | w_{i,n}) = \text{RandSelectF}(w_{i,n})
\] (3)

3) **Fitness evaluation**

Chromosome decoding is the inverse process of coding. With Chromosome coding, we can get the sequence of corresponding segmentation result, and after been substituted into the expression (2), we get the fitness value of the chromosome. When computing the fitness value, we use plus one law to smooth \( P(w_i | f_i) \) and \( P(f_i | f_{i-1}, f_{i-2}) \):

\[
P(w_i | f_i) = \frac{Q(w_i, f_i) + 1}{Q(f_i) + Y}
\] (4)

\[
P(f_i | f_{i-1}, f_{i-2}) = \frac{Q(f_{i-2}, f_{i-1}, f_i) + 1}{Q(f_{i-2}, f_{i-1}) + Z}
\] (5)

Here, \( Q(w_i, f_i) \) is the occurrence frequency of every segmentation result that is corresponding to \( w_i \), while \( Q(f_{i-2}, f_{i-1}, f_i) \) is the occurrence frequency of \( f_{i-2} \) followed by \( f_{i-1} \), \( Y \) is the average number of segmentation results of the phrase, \( Z \) is the average number of phrases before every binary group of segmentation result.

4) **Variation processing**

When the variation is been undertaken, a locus is randomly selected from the optimal chromosome, and coding the corresponding gene within the range of possible segmentation results. Keep repeating this process until 39 new chromosomes is obtained by variation, and together with the original chromosome whose adaptability is the strongest to constitute new population. The variation function is:

\[
\text{Mut(a}(chrom(f_{i,n}^0 )))
\]

5) **The process of genetic algorithm**

The process of computing optimal segmentation based on genetic algorithm is as follows:

**Generation** = 56, **Curfitness** = 0;

//Setting the number of generation and the optimal fitness value of the current chromosome

\[
\text{chrom}(f_{i,n}^0) = \text{RandSelectF}(w_{i,n})
\] (6)

//To generate the original chromosome with constrained strategy

While **i** ≤ **Generation** and

\[
\text{fitness}(f_{i,n}^0 | w_{i,n}) > \text{Curfitness}
\]

\[
\text{Curfitness} = \text{fitness}(f_{i,n}^0 | w_{i,n})
\] (7)

For **j** = 1 To 39 Step 1;

//Let the population quantity is 40

\[
\text{chrom}(f_{i,n}') = \text{Mut(a}(\text{chrom}(f_{i,n}^0)))
\] (8)

// The process of variation

\[
(f_{i,n}^0) = \arg_k \max \text{fitness}(f_{i,n}^k | w_{i,n})(0 \leq k \leq 39)
\]

//To decide which segmentation result is best
4. THE TEST OF AUTOMATIC SEGMENTATION

The resource of training corpus and test corpus is the text base of People's Daily (From 1993 to 1994, and from 1996 to 2000). We use CCRL to retrieve ambiguous example sentences with polyphones, and save them in specified format.

The test repeats for 8 times, and the result is compared to the result that is trained with the approach of SVM-based weight estimation\(^5\), the comparison result is shown in table 2.

Compare to SVM, the correct rate of high frequency characters with the algorithm based on Chinese Phoneme improves, while the correct rate of low frequency characters is decreases.

<table>
<thead>
<tr>
<th>The polyphone</th>
<th>The number of training ambiguous phrases</th>
<th>The number of testing ambiguous phrases</th>
<th>The correct rate of SVM-based weight estimation approach</th>
<th>The correct rate of the algorithm based on Chinese Phoneme</th>
</tr>
</thead>
<tbody>
<tr>
<td>长</td>
<td>2731</td>
<td>1879</td>
<td>88.12%</td>
<td>86.37%</td>
</tr>
<tr>
<td>地</td>
<td>3954</td>
<td>2556</td>
<td>62.62%</td>
<td>86.48%</td>
</tr>
<tr>
<td>量</td>
<td>1334</td>
<td>864</td>
<td>93.62%</td>
<td>93.55%</td>
</tr>
<tr>
<td>重</td>
<td>1487</td>
<td>1125</td>
<td>88.19%</td>
<td>86.03%</td>
</tr>
</tbody>
</table>

5. CONCLUSIONS

The exact and efficient automatic segmentation can ensure computer processing Chinese information with high quality, while the existence of polyphones in Chinese is a special linguistic phenomenon that differ from Indo-European family. Since polyphones have particular and complicated mapping relationship between pronunciations and meanings, so segmentation ambiguity caused by polyphones must be treat specially. On the other hand, because genetic algorithm has following characteristics: its fitness function can be defined flexible, it can easily adapt a lot of tags while the training data is relatively sparse, so the paper use it to automatically realize optimal segmentation about polyphones. Compare to SVM, the correct rate of high frequency characters improves, while the correct rate of low frequency characters is decreases, for the variation processing of genetic algorithm is not good at maintaining population differences of low frequency of chromosome.

ACKNOWLEDGEMENT

This dissertation is supported by senior professional research start-up fund of Jiangsu University (Grant No.: 07JDG046).

6. REFERENCES