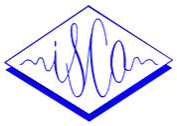


International Symposium on Chinese Spoken Language Processing (ISCSLP 2000)
Fragrant Hill Hotel, Beijing
October 13-15, 2000

RULE-BASED POST-PROCESSING OF PINYIN TO[1] CHINESE CHARACTERS CONVERSION SYSTEM

ZHANG Yan XU Bo ZONG Chengqing

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, 100080
E-Mail: {yzhang, xubo, cqzong}@nlpr.ia.ac.cn

ABSTRACT

Statistical method is a good way for pinyin to Chinese characters conversion and has gotten preferable conversion rate. However, there are still several percent words cannot be converted correctly with the method. This paper presents an error correction approach based on grammatical and semantic rules. According to the conversion results and neighboring information obtained from pinyin to Chinese characters using statistical method, we build a knowledge base consists of phrase rules, syntactic rules and semantic rules. By analyzing the syntactic structure of sentences, we check the semantic correction at some local part of speech node. This method is used for error correction as a post processing method under the assumption of localized error point at preliminary experiment. The experiments prove that the correct conversion rate is improved based on rule method.

Key words: Pinyin to Chinese characters conversion, error correction, knowledge base, post-processing

1. INTRODUCTION

A pinyin to Chinese characters conversion system is an expert system which can automatically convert pinyin strings to corresponding Chinese character string. The key of pinyin to Chinese characters conversion is to deal with homophone problem. In general, there are two approaches used to transfer pinyin to Chinese characters. One is statistical approach. The other is based on rules. Statistical approach needn't analyze Chinese grammatical and semantic structures and build complicated grammatical rules. It needs to build a very large corpus base and gets plenty of statistical information by sufficiently training large scale data. Furthermore the conversion rate is better and acceptable. Our system uses statistical method to convert pinyin to characters automatically. We select some sports news as test corpus and complete pinyin to characters conversion of these corpus. Then we observe three hundred sentences with errors and try to further correct these errors. While these errors are difficult to be corrected with statistical method, we need to use other method as a port-processing method to handle errors.

Rule method has some advantages which statistical method doesn't have. It applies lots of Chinese language knowledge, such as grammar, syntax, phrase and semantics, and is a good way to deal with ungrammatical and unmeaning semantic errors resulting from the statistical method due to uncertain long distance information or sparse training data. Here we try to apply rule-based method as a compensatory way of statistical method for error correction purpose..

In followed, section2 simply introduces statistical language model and the kinds of errors. Section3 presents the building of knowledge base including semantic lexicon and rule base. Section4 is the processing of rule method for error correction. Finally, We give the preliminary experiment result, prospect and conclusions.

2. STATISTICAL LANGUAGE MODEL

Since we process the results of pinyin to characters conversion based on statistical language model. It is necessary to briefly explain statistical method.

The target of pinyin to characters conversion system is to find a Chinese word string $W = w_1, w_2, ..., w_n$ satisfying

$$\hat{W} = \arg\max_{w} P(W / A) \qquad (1)$$

where A is the inputting pinyin string and $A = a_1, a_2, ...a_n$. W is the output characters corresponding A, $W = w_1, w_2, ...w_n$.

According to the Bayes' rule:

$$P(W / A) = \frac{P(A/W) \cdot P(W)}{P(A)} \qquad (2)$$

In the formula above, P(A) is a constant since the inputting A is finite, so we don't have to consider it. Therefore formula (2) can approximates:

[1] The research work described in this paper was supported by the National Key Fundamental Research Program (the 973 Program) of China under the grant G1998030504, the National Natural Science Foundation of China under the grant 69835030 and also the National 863 Hi-Tech Program under the grant 863-306-ZT03-02-2.

$$W = \arg\max_{w} P(A/W) \cdot P(W) \quad (3)$$

when we consider a Chinese sentence as a Markov source, that is to say, the occurring probability of one state is only related to the preceding (n-1) states and has nothing to do with other states. So we get the following formula:

$$P(W) = P(W_1 W_2 ... W_n)$$

$$= P(W_1) \prod_{i=2}^{i=n} P(W_1 W_2 ... W_{i-1}) \quad (4)$$

We usually adopt bigram and trigram model, respectively n=2 and n=3.

Statistical model has already been broadly applied to convert pinyin to Chinese characters. But some words are converted incorrectly. In general, there are following kinds of error result.

(1). Segmentation error: This is a common error. For example, in the sentence "他们有好的教练员", "有/好/的"is converted to "友好/的" because of segmentation error.

(2). Data sparseness problem: Some words still are not in the training set although we have trained large scale corpus for statistical model. For example: "出线" is often used in sports news but hardly appears in other field. So pinyin string "chu xian" is converted to "出现".

(3). Influence of long distance information: Bigram and trigram only express relations between neighboring words. But this relation is not enough. For example, in the sentence "这种对抗是个人显示能力的一场特殊方式", "是" and "方式" is long distance relation.

Above errors are difficult to be handled with statistical model, so we make use of rule method as a post-processing means to correct these errors.

3. KNOWLEDGE BASES

Our target is to correct errors with rule method. While rule method needs grammatical and semantic information, that is to say, it needs to build a knowledge base including a semantic lexicon and a rule base. We firstly explain the building of the semantic lexicon, then the rule base.

3.1 Design of semantic lexicon

Chinese word classification system is the base and key of rule approach to natural language understanding. Only Chinese words are classified to systematic grammatical and semantic categories, are Chinese words tagged with reasonable part-of-speech and semantic attributions. The semantic lexicon is built based on it.

In our approach, we classify Chinese words into seventeen categories based on part-of-speech as the first level category. Sometimes it is not enough to only denote part-of-speech attribution. It also needs semantic classification to express semantic information. We further classify each part-of-speech class into two level semantic subclasses. We refer to "Thesaurus Dictionary"[2] to get a comparison table. The tagged attributions are corresponding to semantic codes in this dictionary book. It is figured as follow.

Sequence Number	Part-Of-Speech	Brief Symbol	Semantic Class
1	Noun	N	A, B, C, D
2	Pronoun	P	Aa
3	Time Word	T	Ca
4	Place Word	W	Cb
5	Verb	V	F, G, H, I, J, L
6	Adjective	A	E, Ga
7	Numeral	Q	Dn
8	Quantifier	L	D
9	Adverb	D	Ka
10	Direction word	F	Cb
11	Preposition	R	Kb
12	Conjunction	C	Kc
13	Auxiliary	H	Kd
14	Sound Imitation Word	S	Kf
15	Mood Auxiliary Word	Y	Ke
16	Interjection	E	Ke
17	Idiom	I	Kd

Figure1 Comparison Table of Semantic Class

According to above word classification system, each word has three level attributions. First level denotes part-of-speech, and second and third level are both based on semantic attributions corresponding to semantic classification in "Thesaurus Dictionary". At the third level, there are ninety-four classes. Therefore, in semantic lexicon, each Chinese word has three parts which consist of Chinese character, pinyin string, semantic code. For example, 对 dui * NDK(LN,AEB,AED,VFA,VHI,RO).

3.2 Design of rule base

Rule base is built based on semantic lexicon. Rules generally express the relations among components in each Chinese sentence. we build rules according to above word classification system. Each rule is expressed with production formula. Since context-free grammar is appropriate and common to describe natural language, we describe the production rule with CFG.

In our system, we express grammatical components with following symbols:

- S and CS are starting symbols, respectively denotes a whole sentence and clause sentence.
- Non-terminal symbols are N, V, A, D…, NP,VP, AP,…and NBA, VFA…;

- Terminal symbols are Chinese words or phrase structures.
- The form of each rule is:

$$A_1 + A_2 + \ldots + A_n \rightarrow R \qquad (5)$$

By analyzing grammatical and semantic attributions, we get three types of rules including phrase, syntactic and semantic rules.

● Phrase Rules

We define eight forms of phrase structures, which consist of:

NP(Noun Phrase), VP(Verb Phrase), AP(Adjective Phrase), PP(Pronoun Phrase), DP(Adverb Phrase), RP(Preposition Phrase), QL(Numeral and Quantifier Phrase) and TP(Time Phrase).

According to Chinese grammar, we get some phrase rules. For example: R + V + N → RP

● Syntactic Rules

Phrase rules are the base of syntactic rules. We parse the correct sentences with parsing tree based on syntactic structure and each word is only considered its part-of-speech. There is a sentence:

这(P) 是(V) 不(D) 可(V) 更改(V) 的(H) 事实(N)

We get syntactic rules according to above sentence:

V + V → VP, D + VP → VP,

VP + H + N → NP, P + V + NP → S

● Semantic Rules

Semantic rules are got by analyzing semantic attribution of error words converted with statistical model, their homophones and neighboring words.

For example: 我可以感觉到海风的潮湿（超市）

海风 *NBF 的 *HS 潮湿 *AEB 超市 *NCB

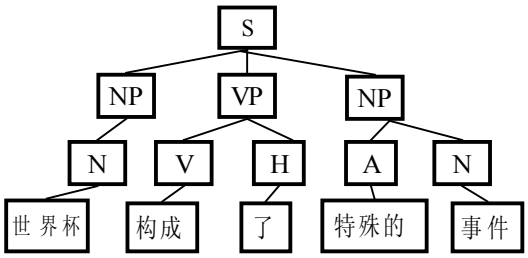
A semantic rule is produced: NBF + HS + AEB → NP

4. RULE APPROACH FOR ERROR CORRECTION

Our work is to correct errors with rule method. These errors exist in three hundreds sentences selected from test corpus. On the basis of knowledge base, we parse these sentences, analyze semantic attributions of error words and use rules matching to correct them.

This processing mainly consists of three steps. The first step is parsing the sentences and finding out error words based on syntactic structures. Then we compare the conversion results got from statistical model with original correct sentences and select the error words in all sentences. Their left and right neighbors are also obtained. Following figure illustrates this step. In which "事件" is converted to "时间". In our initial experiment, this step is finished manually.

The second step is to get the semantic codes of errors and their neighbors. When we get a wrong word, at the same time, we



search the pinyin sentence to get its corresponding pinyin string. We seek for Chinese words which are homophones according to the found pinyin string in the semantic lexicon. These candidate words with their semantic codes are kept in a stack. The left and right neighbors of the error are also kept. For example, pinyin string "jing ji" respectively corresponds to five Chinese words: 经济 *NDJ(VIF), 竞技 *VHH, 荆棘 *NDG, 静寂 *AEF, 惊悸 *AEG.

The final step is to find a word satisfying one of rules from the candidate words. Satisfying a rule means that the left and right neighbors can be connected with one of candidate word, that is to say, they have collocation relation. Because there are three kinds of rules, we select a rule base on following condition. First is to judge whether this type of collocation relation satisfies phrase rules. If only one rule is satisfied, stop searching other rules. If more than one rules are satisfied, we go on finding syntactic rules. According to the same way, we finally search semantic rules.

The processing of error correction is figured as figure2.

Let us give an example to illustrate the experiment.

Pinyin string: cong jiao lian dao qiu yuan dou shi qu le zhi hui

The result of statistical method: 从教练到球员都失去了**指挥**

We firstly get the error word "指挥" and its pinyin "zhi hui". By searching semantic lexicon, we get homophones " 智慧 *NDE(AEE)", "指挥*NAE(NAF,VHC)" and "只会 *VGB". We also get the neighboring word " 失去 *VJD". We go on to search rule base and find that a syntactic rule "V + N → VP" is satisfied by "指挥" and "智慧". While the semantic rule "VJD + NDE → VP" judges the only "智慧" is the correct word.

5. RESULT OF THE EXPERIMENT

We select sports news download from Internet as test corpus in which there are about two thousand sentences. We use statistical method to convert pinyin of these sentences to Chinese characters and find out the sentences including error results. Then we select some sentences only including one error but segmented correctly. In fact, it is very difficult to process segmentation errors although rule method is used. There are totally three hundred sentences to be post-processed. We analyze grammatical and syntactic structures of these sentences which are original sentences. And the homophones with error words and their neighboring words are also tagged with semantic attributions. However, only twenty percent of neighboring relations can produce semantic rules. We process these errors according to above processing of error correction. By experiment we get three kind of results. One is that

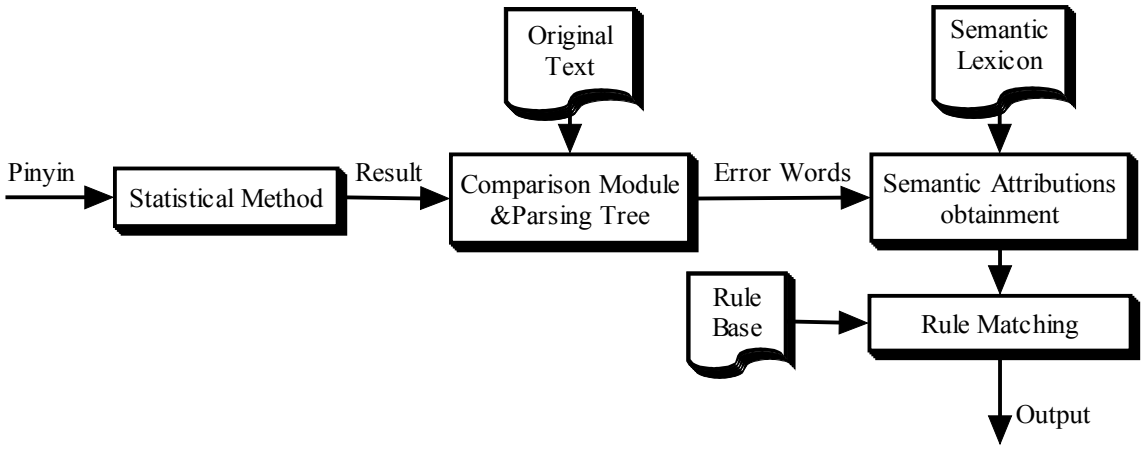


Figure2 Processing of Error Correction

some error words can be converted to the correct words. The conversion rate of error correction is 32.3%. Another result is that more than one candidate words satisfy rules. The probability of this case is 31%. For example, in the sentence "他们用自己的经历指挥球队", "经历" is converted to "精力". While these two words respectively satisfy semantic rule "NDA+VHC→CS" and "NDD+VHC→CS". The third is that the error words are not converted because their neighboring relation can not satisfy any rule. Figure 3 shows the result of experiment.

.

Three Cases	Number of Sentences	Conversion Rate
Accurate conversion	97	32.3%
Multi-candidate Result	93	31%
No Conversion	110	36.7%

Figure 3 Result of Experiment

6. CONCLUSION AND PROSPECT

From the result of experiment, we obtain following conclusions.

(1) Rule method is a feasible way to correct errors and can improve accurate rate. It has further research prospect. For example, automatic parsing of sentences, learning of rules is absolutely necessary. Furthermore grammatical and semantic knowledge is very important information to pinyin to Chinese characters conversion system.

(2) Semantic classification is a basic and key problem. If the classification is unreasonable or not detailed enough, it affects the building of rule base.

(3) The building of rule base needs plenty of Chinese knowledge, especially collocating relations. It is very difficult to build a complete rule base, especially semantic rules. However, it's possible to set up some rules specially handling the errors from statistical method.

(4) Candidate words and their neighboring words may match each other and satisfy more than one rule. Once it happens, Probabilities of these rules to find a best one.

In a word, in the process of error correction, rule method can solve some grammatical and semantic errors which are not solved by statistical method. The final object is combine statistical method with rule method together during coding.

7. REFERENCES

[1] Yang Kaicheng, He Kekang, Design and Implementation of Practical PinYin-Chinese Conversion system, ICCPOL'97, vol.Ⅱ, 1997.

[2] Mei Jiaju, YuanYiming, "Thesaurus Dictionary", Shanghai Dictionary Press, 1982

[3] Eugene Charniak, "Statistical Language Learning", the MIT press, Cambridge, Massachusetts, 1993.

[4] Zong Chengqing, "Research on the Transliteration from Pinyin to Chinese Characters and Normalization for Chinese Sentences", Ph.D. thesis, Institute of Computing Technology, The Chinese Academy of Sciences, 1997.

[5] Zhang Ruiqiang, Wang Zuoying, Zhang Jianping, Chinese Pinyin to Text Translation Technique with Error Correction Used for Continuous Speech Recognition, Journal of Tsinghua University (Sci & Tech), vol.37, 1997.

[6] Guan Yi, Wang Xiaolong, Zhang Kai, Automatic Rule Acquisition Based on Transformation for a Chinese Syllable to Character Conversion System, Journal of Computer Research & Development, vol.36, No.3, 1999.

[7] Zhang Yangsen, Automatic Lexicon Errors Detecting of Chinese Texts Based on the Orderly-neighborship, ICMI, 1999.