Chinese Word Segmentation based on Mixing Model

Wei Jiang **Jian Zhao** **Yi Guan** **Zhiming Xu**
ITNLP, Harbin Institute of Technology, Heilongjiang Province, 150001 China
jiangwei@insun.hit.edu.cn

Abstract

This paper presents our recent work for participation in the Second International Chinese Word Segmentation Bakeoff. According to difficulties, we divide word segmentation into several sub-tasks, which are solved by mixed language models, so as to take advantage of each approach in addressing special problems. The experiment indicated that this system achieved 96.7% and 97.2% in F-measure in PKU and MSR open test respectively.

1 Introduction

Word is a logical semantic and syntactic unit in natural language. So word segmentation is the foundation of most Chinese NLP tasks. Though much progress has been made in the last two decades, there is no existing model that can solve all the problems perfectly at present. So we try to apply different language models to solve each special sub-task, due to "No Free Lunch Theorem" and "Ugly Duckling Theorem".

Our system participated in the Second International Chinese Word Segmentation Bakeoff (henceforce, the bakeoff) held in 2005. Recently, we have done more work in dealing with three main sub-tasks: (1) Segmentation disambiguation; (2) Named entities recognition; (3) New words[1] detection. We apply different approachs to solve above three problems, and all the modules are integrated into a pragmatic system (**ELUS**). Due to the limitation of available resource, some kinds of features, e.g. POS, have been erased in our participation system. This segmenter will be briefly described in this paper.

2 ELUS Segmenter

All the words are categorized into five types: Lexicon words (LW), Factoid words (FT), Morphologically derived words (MDW), Named entities (NE), and New words (NW). Accordingly, four main modules are included to identify each kind of words, as shown in Figure 1.

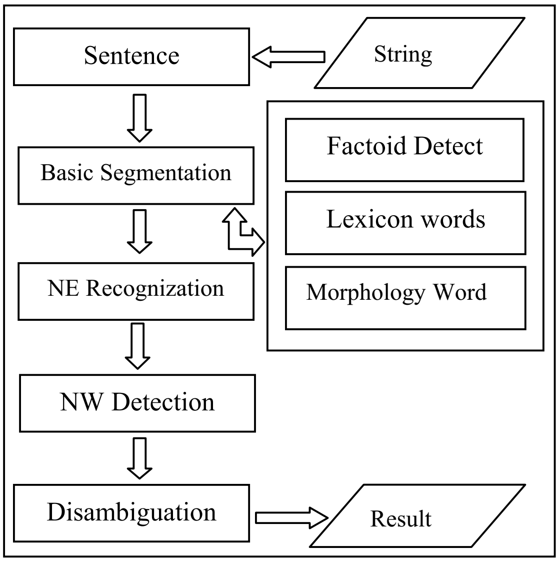

Figure 1 ELUS Segmenter

Class-based trigram model (Gao 2004) is adopted in the Basic Segmentation to convert the sentence into a word sequence. Let **w** = w_1 w_2 …w_n be a word class sequence, then the most likely word class sequence w* in trigram is:

$$\mathbf{w}^* = \arg\max_{w_1 w_2 \cdots w_n} \prod_{i=1}^{n} P(w_i \mid w_{i-2} w_{i-1}),$$

where let $P(w_0|w_{-2}\ w_{-1})$ be $P(w_0)$ and let $P(w_1|w_{-1}\ w_0)$ be $P(w_1|w_0)$. And w_i represents LW or a type of FT or MDW. Viterbi algorithm is used to search the best candidate. Absolute smoothing algorithm is applied to overcome the data sparseness. Here, LW, FT and MDW are idendified (Zhao Yan 2005). All the Factoid words can be represented as regular expressions. As a result, the detection of factoid words can be archieved by Finite State Machines.

[1] New words refer to this kind of out-of –vocabulary words that are neither recognized named entities or factoid words nor morphological words.

Four kinds of Named entities are detected, i.e. Chinese person name, foreign person name, location name and orgnization name. This is the most complicated module in ELUS.

Three kinds of models are applied here. HMM model (one order) is described as:

$$\mathbf{T}^{\#} = \underset{T_1 T_2 \cdots T_n}{\arg\max} \prod_{i=1}^{n} P(W_i \mid T_i) P(T_i \mid T_{i-1}),$$

where T_i represents the tag of current word, Viterbi algorithm is used to search the best path. Another model is Maximum Entropy (Zhao Jian 2005, Hai Leong Chieu 2002). Take Chinese person name as example. Firstly, we combine HMM and Maximum Entropy (ME) model to lable the person name tag, e.g. "姚/CPB 铜/CPI 梅/CPI" (Tongmei Yao); Secondly, the tagged name is merged by combining ME Model and Support Vector Machine (SVM) and some aided rules, e.g. merged into "姚/铜梅" in PKU test.

Some complex features are added into ME model (described in Zhao Jian 2005), in addition, we also collect more than 110,000 person names, and acquire the statistic about common name characters, these kinds of features are also fused into the ME model to detect NE. The other kinds of NE recognition adopt similar method, except for individual features.

New Words is another important kind of OOV words, especially in closed test. Take PKU test as example, we collect NW suffixes, such as "市"(city),"灯"(lamp). Those usually construct new words, e.g. "景观灯"(sighting lamp).

A variance-based method is used to collect suffixes. And three points need to be considered:(1) It is tail of many words;(2) It has large variance in constructing word;(3) It is seldom used alone. We acquire about 25 common suffixes in PKU training corpus by above method.

We use Local Maximum Entropy model, e.g. "黄冈/1 市/1"(Huanggang city), i.e. only the nearer characters are judged before the suffix "市" (city). By our approach, the training corpus can be generated via given PKU corpus in the bakeoff. The features come from the nearer context, besides, common single words and punctuations are not regarded as a part of New Word.

The last module is Word Disambiugation. Word segmentation ambiguities are usually classified into two classes: overlapping ambiguity and combination ambiguity. By evaluating ELUS, the most segmentation errors are one segmentation errors (about 95%). i.e. the two words on both sides of current segmentation errors are right. These include LW ambiguities and FT ambiguities etc. Here, we adopt Maximum Entropy model. The same as other modules, it is defined over H×T in segmentation disambiguation, where H is the set of possible contexts around target word that will be tagged, and T is the set of allowable tags. Then the model's conditional probability is defined as

$$p(t \mid h) = \frac{p(h,t)}{\sum_{t' \in T} p(h,t')}, \text{ where}$$

$$p(h,t) = \pi\mu \prod_{j=1}^{k} \alpha_j^{f_j(h,t)}$$

where h is current context and t is one of the possible tags. The ambiguous words are mainly collected by evaluating our system.

In NE module and Word Disambiguation module, we introduce rough rule features, which are extracted by Rough Set (Wang Xiaolong 2004), e.g. "施展→才能"(display ability), "只有→才/能"(only→ can just), "记者+person+报道" (the reporter+person+report). Previous experiment had indicated word disambiguation could achieve better performance by applying Rough Set.

3 Performance and analysis

The performance of ELUS in the bakeoff is presented in Table 1 and Table 2 respectively, in terms of recall(R), precision(P) and F score in percentages.

Table 1 Closed test, in percentages (%)

Closed	**R**	**P**	**F**	**OOV**	**R_{oov}**	**R_{iv}**
PKU	95.4	92.7	94.1	5.8	51.8	98.1
MSR	97.3	94.5	95.9	2.6	32.3	99.1
CITYU	93.4	86.5	89.8	7.4	24.8	98.9
AS	94.3	89.5	91.8	4.3	13.7	97.9

Table 2 Open test, in percentages (%)

Open	**R**	**P**	**F**	**OOV**	**R_{oov}**	**R_{iv}**
PKU	96.8	96.6	96.7	5.8	82.6	97.7
MSR	98.0	96.5	97.2	2.6	59.0	99.0
CITYU	94.6	89.8	92.2	7.4	41.7	98.9
AS	95.2	92.0	93.6	4.3	35.4	97.9

Our system has good performance in terms of F-measure in simplified Chinese open test, including PKU and MSR open test. In addition, its IV word identification performance is remarkable, ranging from 97.7% to 99.1%, stands at the top or nearer the top in all the tests in which we have participated. This good performance owes to class-based trigram, absolute smoothing and word disambiguation module and rough rules.

There is almost the same IV performance between open test and closed test in MSR, CITYU and AS respectively, because we adopt the same Lexicon between open test and closed test respectively. While in open test of PKU, we adopt another Lexicon that comes from six-month corpora of Peoples' Daily (China) in 1998, which were also annotated by Peking University.

The OOV word identification performance seems uneven, compared with PKU, the other tests seem lower, due to the following reasons:

(1) Because of our resource limitation, NE training resource is six-month corpora of Peoples' Daily (China) in 1998, which came from Peking University, and some newspapers and web pages annotated by our laboratory;

(2) We have no traditional Chinese corpus, so the NE training resource for CITYU and AS is acquired via converting above corpora. Since these corpora are converted from simplified Chinese, they are not well suitable to traditional Chinese corpora;

(3) The different corpora have different criterions in NE detection, especially in location name and organization name, e.g. "崔村镇/香堂/猪场" (Cuicun Town Xiangtang Hogpen) in PKU and "崔村镇香堂猪场" in MSR criterion. Even if our system recognizes the "崔村镇/香/堂/猪场" as a orgnization name, we are not easily to regard "香堂" as one word in PKU, since "香堂" isn't a lexical word. However in MSR, that is easy, because its criterion regard the whole Orgnization as a word;

(4) We need do more to comply with the segmentation criterion, e.g. "露宿者"(outlier) in CITYU come from "露宿" + "者", while this kind of false segment is due to our bad understanding to CITYU criterion.

Though there are above problems, our system does well in regonization precision, since we adopt two steps in recognizing NE, especial in recognizing Chinese person name. And from the result of evalution in the bakeoff, we need to improve the NE recall in the future.

In order to make our New words comply with the criterion, we conservatively use New Word Detection module, in order to avoid having bad recognition result, since each corpus has its own New Word criterion.

4 Conclusion and Future work

We have briefly describled our system based on mixed models. Different approachs are adopted to solve each special sub-task, since there is "No Free Lunch Theorem". And mixed models are used in NE detection. This sytem has a good performance in the simplified Chinese in the bakeoff.

The future work is mainly concentrating on two directions: finding effective features and delicately adjusting internal relations among different modules, in order to improve segmentation performance.

References

Fu Fuohong. 2000. Research on Statistical Methods of Chinese Syntactic Disambiguation. Ph.D. Thesis. Harbin Institute of Technology, China.

Hai Leong Chieu, Hwee Tou Ng. Named Entity. Recognition: A Maximum Entropy Approach Using Global. Information. Proceedings of the 19th International Conference. on Computational Linguistics, 2002.

Hua-Ping Zhang, Qun Liu etc. 2003. Chinese Lexical Analysis Using Hierarchical Hidden Markov Model, Second SIGHAN workshop affiliated with 4th ACL, Sapporo Japan, pp.63-70, July 2003.

Jianfeng Gao, Mu Li et al. 2004. Chinese Word Segmentation: A Pragmatic Approach. MSR-TR-2004-123, November 2004.

Wang Xiaolong, Chen Qingcai, and Daniel S.Yeung. 2004. Mining PinYin-to-Character Conversion Rules From Large-Scale Corpus: A Rough Set Approach, IEEE TRANSACTION ON SYSTEMS, MAN. AND CYBERNETICS-PART B:CYBERNETICS. VOL. 34, NO.2, APRIL.

Zhao Jian, Wang Xiao-long et al. 2005. Comparing Features Combination with Features Fusion in Chinese Named Entity Recognition. Computer Application. China.

Zhao Yan. 2005. Research on Chinese Morpheme Analysis Based on Statistic Language Model. Ph.D. Thesis. Harbin Institute of Technology, China.