NetEase Automatic Chinese Word Segmentation

Li xin **Dai shuaixiang**
NETEASE INFORMATION TECHNOLOGY (BEIJING) CO., LTD.
SP Tower D, 26th Floor, Tsinghua Science Park Building 8, No.1 Zhongguancun East Road,
Haidian District Beijing, 100084, PRC.
lxin@corp.netease.com ddai@corp.netease.com

Abstract

This document analyses the bakeoff results from *NetEase Co*. in the SIGHAN5 Word Segmentation Task and Named Entity Recognition Task. The *NetEase* WS system is designed to facilitate research in natural language processing and information retrieval. It supports Chinese and English word segmentation, Chinese named entity recognition, Chinese part of speech tagging and phrase conglutination. Evaluation result shows our WS system has a passable precision in word segmentation except for the unknown words recognition.

1 Introduction

Automatic Chinese Word Segmentation (WS) is the fundamental task of Chinese information processing [Liu, 2000].Since there are lots of works depending on the automatic segmentation of Chinese words, different Chinese NLP-enabled applications may have different requirements that call for different granularities of word segmentation. The key to accurate automatic word identification in Chinese lies in the successful resolution of those ambiguities and a proper way to handle out-of-vocabulary (OOV) words (such as person names, place names and organization name etc.).

We have applied corpus-based method to extracting various language phenomena from real texts; and have combined statistical model with rules in Chinese word segmentation, which has increased the precision of segmentation by improving ambiguous phrase segmentation and out-of-vocabulary word recognition.

In the second section of this paper, we describe a Chinese word segmentation system developed by ***NetEase***. And we present our strategies on solving the problems of ambiguous phrase segmentation and identification of Chinese people names and place names. The third section is analysis of evaluation result.

2 Modern Chinese Automatic Segmentation System

2.1 System Structure

The WS system of NETEASE CO. supports Chinese and English word segmentation, Chinese named entity recognition, Chinese part of speech tagging and phrase conglutination. In ordering to processing mass data, it is designed as an efficient system. The whole system includes some processing steps: pre-processing, number/date/time recognition, unknown words recognition, segmenting, POS tagging and post-processing, as Fig 1 shows.

The *Prehandler* module performs the pre-processing, splits the text into sentences according to the punctuations.

Number/Data/Time recognition processes the number, date, time string and English words.

Unknown word recognition includes personal name recognition and place name recognition.

Segmenter component performs word-segmenting task, matches all the candidate words and processes ambiguous lexical.

POSTagger module performs part of speech tagging task and decides the optimal word segmentation using hierarchical hidden Markov model (HHMM) [Zhang, 2003].

Posthandler retrieves phrases with multi-granularities from segmentation result and detects new words automatically etc.

2.2 Ambiguous phrase segmentation

Assume that “AJB” are character strings and that W is a word list. In the field “AJB”, if “AJ”∈W,

Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pages 193–196,
Sydney, July 2006. ©2006 Association for Computational Linguistics

and “JB”∈W, then “AJB” is called ambiguous phrase of overlap type. For example, in the string "*当代表*", both "*当代*" and "*代表*" are words , so "*当代表*" is an ambiguous phrase of overlap type; and there is one ambiguous string.

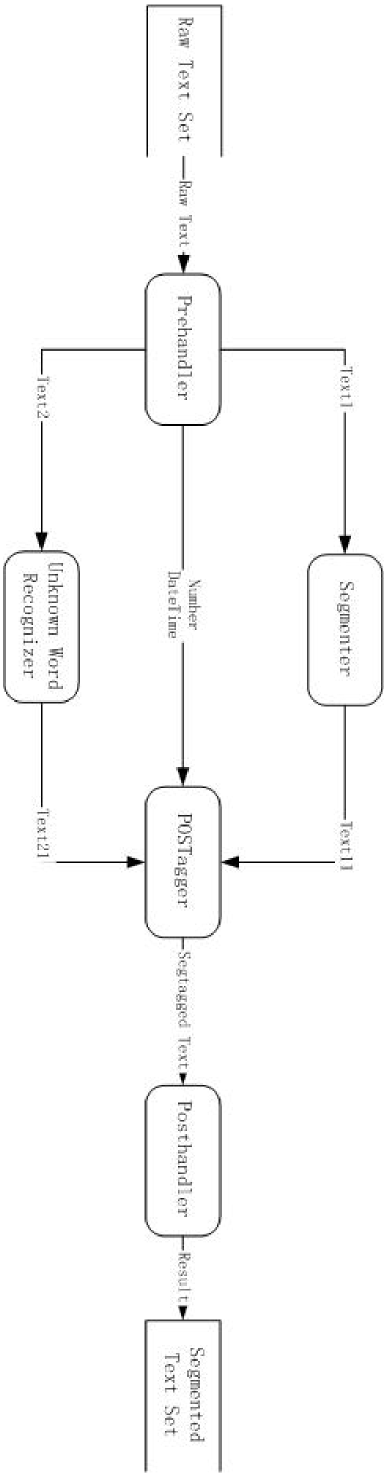

Fig 1 Structure and Components of WS

In the string “AB”, if “AB”∈W(word), “A”∈W, and “B”∈W, then the string “AB” is called ambiguous phrase of combination type. For example, in the string "*个人*", since "*个人*", "*个*" and "*人*"are all words, so the string "*个人*" is an ambiguous phrase of combination type.

We have built an ambiguous phrase lib of overlap and combination type from tagged corpus, which contains 200,000 phrases from 1-gram to 4-gram. For example: “*才/d 能/v 创造/v, 创造/vn 作/v 准备 vn*” If one ambiguous phrase found in raw text, the potential segmentation result will be found in the lib and submit to next module. If not found, *POS tagger* module will disambiguate it.

2.3 Chinese Personal Name Recognition

At present we only consider the recognition of normal people name with both a family name and a first name. We got the statistical Character Set of Family Name and First Name data from corpus. And also consider the ability of character of constructing word. Some characters itself cannot be regarded as a word or composes a word with other characters, such as "*邓,聂,鑫*"; Some name characters which can compose word with other characters only, e.g. “*刘，张，英*” can construct words "*刘海儿，一张纸，英雄* ";Some name characters are also a common words themselves, e.g. “*汤，马*” .

The recognition procedure is as follows:

1) Find the potential Chinese personal names: Family name is the trigger. Whenever a family name is found in a text, its following word is taken as a first name word, or its following two characters as the head character and the tail character of a first name. Then the family name and its following make a potential people name, the probable largest length of which is 4 when it is composed of a double-character family name and a double-character first name.

2) Based on the constructing word rules and the protective rules, sift the potential people names for the first time. For example, when raw text is “*三张…,五周…*”, then the “*张,周*” were not family name. Because the “*三,五*” is number.

3) Compute the probabilities of the potential name and the threshold values of corresponding family names, then sift the people names again based on the personal name probability function and description rules.

4) According to the left-boundary rules and the right-boundary rules which base on title, for

example, “*总统,学员*”, and name frequent of context, determine the boundaries of people names.

5) Negate conflicting potential people names.

6) Output the result: The output contains every sentence in the processed text and the start and the end positions and the reliability values of all people names in it.

2.4 Chinese Place Name Recognition

By collecting a large scale of place names, For example, (1) The names of administrative regions superior to county; (2) The names of inhabitation areas; (3) The names of geographic entities, such as mountain, river, lake, sea, island etc.; (4) Other place names, e.g. monument, ruins, bridge and power station etc. building the place name dictionary.

Collecting words that can symbolize a place, e.g. “*地区*” ,“*城市*” ,”*乡*” etc.

Base on these knowledge we applied positive deduction mechanism. Its essence is that with reference to certain control strategies, a rule is selected; then examining whether the fact matches the condition of the rule, if it does, the rule will be triggered.

In addition, Those words that often concurrent with a place name are collected , including: “*在*” ,“*位于*” etc. And which often concurrent with a people name, such as “*同志*” ,“*说*” and so on, are also considered in NER.

WS system identifies all potential place names in texts by using place name base and gathers their context information; and through deduction, it utilizes rule set and knowledge base to confirm or negate a potential place name; hereupon, the remainders are recognized place name.

2.5 Multi-granularities of word segmentation

Whenever we deploy the *segmenter* for any application, we need to customize the output of the *segmenter* according to an application specific standard, which is not always explicitly defined. However, it is often implicitly defined in a given amount of application data (for example, Search engines log, Tagged corpus) from which the specific standard can be partially learned.

Most variability in word segmentation across different standards comes from those words that are not typically stored in the basic dictionary. To meet the applications of different levels, in our system, the standard adaptation is conducted by a post-processor which performs an ordered list of transformations on the output. For example: When input is “*国务院安全生产专家组*”, the output will be:

1. “*国务院/安全/ 生产/ 专家组* ”
2. “*国务院/安全生产/ 专家组*”
3. “*国务院/安全生产专家组*”

Result 1 is normal segmentation, also is minimum granularity of word. Result 2 and 3 is bigger granularity. Every application can select appropriate segmentation result according to its purpose.

3 Test results

The speed of *NetEase* WS system is about 1500KB/s--300KB/s in different algorithm and p4-2.8/512M computer. In SigHan5, the F-MEASURE of our word segmentation is 0.924, the IN Recall is 0.959, but OOV Recall Rate is only 0.656. This indicates that our unknown words recognition is poor; it makes a bad impact on the segmented result. It also shows our system should be improved largely in unknown words recognition. For example:

1. Name Entity Recognize: “*贝尔格勒, 中村植秀, 秋丽玲*” were falsely segment to “*贝/尔格/勒), 中村/植/秀, 秋/丽/玲*”.

2. Name Entity Ambiguous: “*向/ 瑞典/LO*C” are falsely recognized ”*向瑞典/PER*”.

3. Abbreviations of phrase: “*所国 (所罗门*)” was segment to “*所/国*”.

4. New Word: “*原著民, 宿求, 休职*”

5. Standard of Word: we think “*文化大革命*” and “*圣诞老人*” is one word, but criterion is “*文化/大/革命*”, “*圣诞/老人*” etc.

In evaluation, our system’s TOTAL INSERTIONS is 5292 and TOTAL DELETIONS is 2460. The result show: our WS usually segment out “shorter word”, for example, “*工商企业界*”, and “*血液循环*” is segmented to “*工商/企业/界*”, ”*血液/循环*” . But not every string is one word.

Much work needs to be done to evaluate this WS system more thoroughly. Refined pre-processing or post-processing steps could also help improve segmentation accuracy.

For example, pre-processing will split ASCII string and Chinese character, so "*DD6112H6 型, ic 卡, ｂｐ机*" will falsely segment "*DD6112H6/型, ic/卡, ｂｐ/机*"; In post-processing, by using consecutive single characters “*狭/持, 败/击*” to detect the valid out-of-vocabulary words ”*狭持*, *败击*” also is good idea.

References

Kaiying Liu. *Automatic Chinese Word Segmentation and POS Tagging.* Business Publishing House. Beijing. 2000.

Hua-Ping Zhang etc. *Chinese Lexical Analysis Using Hierarchical Hidden Markov Model.* Second SIGHAN workshop affiliated with 41th ACL, Sapporo Japan, July, 2003, pp. 63-70