# NMSU Chinese Segmenter

**Wanying Jin**
Computing Research Laboratory
New Mexico State University
U.S.A.
wanying@crl.nmsu.edu

## 1. Introduction[1]

This presentation outlines an approach to identify words in Chinese corpora. The knowledge used in the system includes a word list; a list of affixes, classifiers and numbers; a list of Chinese surnames; a list of characters used in foreign names and the iteration patterns. The system integrates a heuristic method with a constrained statistical method. It contains five phases:

1) dictionary look-up;
2) iterative and fixed expression;
3) derivation generation;
4) foreign name recognation;
5) Chinese name recognition.

The result is a segmented text in which words are separated by delimiters. In the case of ambiguity a notation is given with an ID number as an index in the supplemental text providing multiple readings as a reference. The following is a sample Chinese text.

１９９３年重点支持了船舶、飞机、卫星发射、大型成套设备等资本性商品出口，共办理出口卖方信贷项目２３个，提供出口卖方信贷人民币近３４亿元，共支持出口成交９·１３亿美元。

The result from the segmenter is

¦１９９３¦年¦重点¦支持¦了¦船舶¦、¦飞机¦、¦卫星¦发射¦、¦<ambi id=4>¦大型¦成套设备¦等¦资本性¦商品¦出口¦</ambi>¦，¦共¦办理¦出口¦卖方¦信贷¦项目¦２３¦个¦，¦提供¦出口¦卖方¦信贷¦人民币¦近¦３４亿¦元¦，¦共¦支持¦出口¦成交¦９·１３亿¦美元¦。¦

where <ambi id=number> indicates the location of each ambiguity. Below is the sample of supplemental text.

<ambi id=4>¦
¦大型¦成套设备¦等¦资本¦性¦商品¦出口¦
¦大型¦成套设备¦等¦资¦本性¦商品¦出口¦
</ambi>¦

## 2. Chinese segmentation[2]

The system includes five phases. Each phase takes as input the output of previous phase and produces new output to the following phase as input until the final result is reached.

### 2.1 Dictionary look-up

The first phase finds words from the dictionary and produce a partially segmented text. It includes the following four steps:

1) find a string within a pair of punctuations;

[1]This project is supported by the Department of Defense, U.S.A. under the contract No. MDA904-94-c-E086

[2]For detailed description see "Identifying Unknown Words in Chinese Corpora" in the proceeding of NLPRS-95.

2) for each character in the string, find all the words beginning with this character. The result is a character-length table listing all the words and their lengths (in byte).

我忘了解放大街在哪里

will be processed into the following table:

| | | | |
|---|---|---|---|
| 我 2 | 忘 2 | 了 2 4 | 解 2 4 |
| 放 2 4 | 大 2 4 | 街 2 | 在 2 |
| 哪 2 4 | 里 2 | | |

3) concatenate all the words in the table from left to right to construct a sequence of words with delimiters inserted. The result is
¦我¦忘¦了¦解放¦大街¦在¦哪里¦
(I forgot where the Liberal street is)

4) in the case of ambiguity, produce multiple readings based on the above table.

¦我¦忘¦了解¦放大¦街¦在¦哪里¦
¦我¦忘¦了解¦放¦大街¦在¦哪里¦
¦我¦忘¦了¦解放¦大街¦在¦哪里¦

### 2.2 Iterations and fixed expressions

Iterative and fixed expressions are words composed of two to four characters with duplication such as 常常, 雄纠纠, 刮刮叫, 高高兴兴, 研究研究. Fixed expressions are words following certain structural patterns such as 想一想, 动不动, 一臂之力, 一窍不通, 忽冷忽热. The system applies the patterns over unsegmented character sequence to construct the expressions such as for the text

¦完¦完全¦全¦忘¦了¦解放¦大街¦在¦哪里¦

the second phase refines the segmentation as
¦完完全全¦忘¦了¦解放¦大街¦在¦哪里¦

### 2.3 Recognizing derivations

Derived words are formed from stem + affix/affix-like structure. A set of prefixes and suffixes are encoded in the knowledge base to generate derivations. Quantity expressions such as date expressions, time expressions, monetary expressions, percentages and generic number expressions are all treated as derivations.

The system scans the results of the second phase. If stem + affix structure is found, the derivation is recognizeed. If number sequence is found, the system checks all number characters to construct quantity expressions. For the input text

¦经过¦一¦年¦多¦的¦政府¦倡导¦、¦工业¦
界¦策划¦和¦新闻¦界¦的¦宣传¦，¦美国¦“
¦信息¦高速公路¦”¦的¦宏伟¦前景¦已¦展现
¦在¦美国¦公众¦面前¦。¦把¦作为¦路面¦的
¦大¦容量¦光纤¦铺设¦到¦9¦5¦0¦0¦万
¦个¦美国¦家庭¦中¦，¦估计¦需¦耗资¦两¦
千¦亿¦美元¦。

the third phase produces a text

¦经过¦一¦年¦多¦的¦政府¦倡导¦、¦工业界
¦策划¦和¦新闻界¦的¦宣传¦，¦美国¦“¦
信息¦高速公路¦”¦的¦宏伟¦前景¦已¦展现
¦在¦美国¦公众¦面前¦。¦把¦作为¦路面¦的
¦大¦容量¦光纤¦铺设¦到¦9500万¦个¦美
国¦家庭¦中¦，估计¦需¦耗资¦两千亿¦美元¦
。

### 2.4 Recognizing foreign names

Foreign names listed in the dictionary will be caught at the first phase. For other foreign names, a set of character normally used to translate foreign names is used to recognize foreign names. The algorithm is similar to that used in quantity expression recognition. For characters used as both conjunction/adv and foreign name characters such as 也 or 及, etc., special heuristic is used to prevent these character to be used to generate incorrect foreign name.

¦莫¦依¦塞¦也¦夫¦总参谋长¦及¦前¦参谋总

长 ¦ 、 ¦ 现任 ¦ 总统顾问 ¦ 的 ¦ 阿 ¦ 夫 ¦ 若 ¦ 梅 ¦
也 ¦ 夫 ¦ 元帅 ¦ 。

are recognized as

¦ 莫依塞也夫 ¦ 总参谋长 ¦ 及 ¦ 前 ¦ 参谋总长 ¦ 、 ¦
现任 ¦ 总统顾问 ¦ 的 ¦ 阿夫若梅 也夫 ¦ 元帅 ¦ 。

### 2.5 Recognizing Chinese names

Chinese names for well-known people are listed in the dictionary. For others, Chinese surnames are used to constrain the n-gram. They are divided into three groups with different priorities.

1) the highest priority is given to those surnames frequently used without ambiguity such as 蔡, 曹, 崔, 董, 冯.

2) the second priority is given to the less frequently used surnames without ambiguity such as 樊, 傅, 葛, 龚.

3) the lowest priority is given to rarely used surnames such as 藏, 古, 谷, 简, 文 or frequently used surnames with high ambiguity such as 白, 时, 戴, Higher occurrence of these full names is required in order to confirm they are really Chinese names.

The algorithm uses N-gram along with the constraint of the surnames priority to make the best guess of Chinese names. The input text is scanned twice. First, the system scans a text to create a plausible Chinese name list. Next, it scans the text again to identify Chinese names by referring to the plausible Chinese name list. The detailed descriptions are:

1) use the surname list to constrain 2-gram and 3-gram analysis to get a potential name list with their frequencies such as

| | | | | | |
|---|---|---|---|---|---|
| 杨健 | 2 | 杨健著 | 1 | 黎利 | 11 |
| 文攻 | 1 | 文攻武 | 1 | 李坚持 | 6 |
| 白俄学 | 1 | 戴一付 | 1 | 范思 | 2 |
| 范思学 | 2 | 时间她 | 1 | 雷明 | 1 |
| 叶甫 | 1 | 叶甫图 | 1 | 毕汝 | 6 |
| 毕汝协 | 6 | 司马丽 | 23 | 赵一 | 32 |
| 赵一凡 | 27 | 华汝 | 1 | 郭路 | 17 |
| 郭路生 | 17 | | | | |

2) run constrained 2-gram analysis over sequences of single characters to find potential given names;

芒克 16 依群 14 是由 1 请在 1 或不 1

3) produce plausible name list from the potential name lists. The rules are:

a) remove those potential full names if their frequencies are lower than their surname thresholds such as 白俄学 1, 戴一付 1 and 时间她 1 because 白, 戴 and 时 are in the group with the lowest priority and their frequencies are low.

b) remove those 2-grams which overlap with 3-grams if their frequencies are equal to or lower than that of the 3-grams such as in the case of 范思 2 and 范思学 2, remove 范思.

c) remove those 2-grams which overlap with 3-grams if the frequencies of the 3-grams are much higher than the default surname threshold. For example, in the case of 赵一 32 and 赵一凡 27, remove 赵一 because the surname 赵 is in the group with the highest priority and the frequency of 赵一凡 is 27 that is much higher than the default threshold of surname 赵;

d) remove those potential given names if their frequencies are lower than the given name threshold, such as remove 是由, 请在, and 或不.

Combine the remaining 2-grams and 3-

grams to make a plausible name list as follows:

| | | | | | |
|---|---|---|---|---|---|
| 杨健 | 2 | 黎利 | 11 | 李坚持 | 6 |
| 范思学 | 2 | 雷明 | 1 | 叶甫图 | 1 |
| 毕汝协 | 6 | 司马丽 | 23 | 赵一凡 | 27 |
| 郭路生 | 17 | 芒克 | 16 | 依群 | 14 |

4) scan the text again and identify all names by matching them against the plausible name list.

## 3. Discussion

We discuss some problems we have experienced in the system and give some suggestions for further improvement.

1). Longer compound causes ambiguity and slow down the process:
If 中国航空器材公司 is in the dictionary, the phrase 在中国航空器材公 司寄售零备件 will be segmented as

¦ 在 ¦ *中国航空器材公司 ¦ 寄售 ¦ 零备件 ¦

If 航空器材公司 is in the dictionary, the segmentation will be

¦ 在 ¦ 中国 ¦ 航空器材公司 ¦ 寄售 ¦ 零备件 ¦

2). Words in the dictionary sometime causes error

¦ 花样 ¦ 游泳 ¦ 比赛 ¦ *中将 ¦ 分 ¦ 规定 ¦ 动作 ¦ 、
¦ 自选 ¦ 动作 ¦

3). Newly developed words are hard to recognize such as 热点 ,留成, 售汇, 结汇制, 攻关, etc.

4). The side effect of foreign name recognition:
Some foreign names have been included in the dictionary. During the foreign name process, some time the name found in the dictionary can be a part of another foreign name such as

¦ 吉尔 ¦ 吉 ¦ 斯 ¦
¦ 吉尔吉斯 ¦

Other time combine both names found in the dictionary cause error such as:

¦ 丹麦 ¦ 哥本哈根 ¦
¦ *丹麦哥本哈根 ¦

A constraint can be set up in case both names are in the dictionary to prevent the error. Also, unexpected foreign name can be produced such as

¦ 六百 ¦ *度高温 ¦ 的 ¦ 废气 ¦

We suggest do not use both Chinese name recognizer and foreign name recognizer unless it is necessary.

5). Place name is more difficult to recognize due to the lack of clue. Some place names are followed by location name such as 省, 市, 县, 地区, etc. which can be a clue, but it still difficult to decide the length of place name.

In conclusion, it has been noticed that a hybrid approach to Chinese segmentation, benefiting from both heuristic and statistical method, can achieve high quality results. Some problems mentioned above still need further improvement.

## 4. References

Jin, W., L. Chen (1995). Identify Unknown Words in Chinese Corpus, *Proc. of the Third Natural Language Processing Pacific-Rim Symposium* (NLPRS'95), Vol. 1, pp. 234-239. Seoul, Korea,

刘源, et al. (1994). 信息处理用现代汉语分词规范及自动分词方法, 清华大学出版社.