Project B

Chinese Word Segmentation

Introduction

Chinese is a language without word boundary. Therefore, Chinese word segmentation is a fundamental task for Chinese Natural Language Processing (Chinese NLP). Chinese text is different from an alphabetic text, e.g., an English text, in that it is composed of character strings from a large character set. Word segmentation aims to promote this advancement to the word level. Only with words being segmented, can other NLP tasks such as part-of-speech (POS) tagging, syntactic analysis, semantic analysis are made possible.

Objective and Background

This project is to provide a chance for students to learn how to resolve the problem of Chinese word segmentation. As the first step of Chinese NLP, its techniques are very important to guarantee the quality of word processing. However, Chinese word segmentation is also one of the difficult problems in Chinese NLP.

There are still a number of problems to be solved in Chinese word segmentation. One is the identification for the boundary of Chinese words, such as, the word boundary identification between single-character word and morpheme as well as between word and phrase. For example, 汉语 (Han Language or Chinese) or 汉 (Han or Chinese) | 语 (Language). Some phrases can be segmented in different ways. Furthermore, the principle for segmentation depends on the domain or application, such as 塑胶制品业 (Plasthetics Industry), 塑胶 (Plastic) | 制品业 (Products Industry), or 塑胶制品 (Plasthetics) | 业 (Industry).

Fortunately, in order to maintain a unified and consistent method for Chinese word segmentation, The Ministry of Mechanics and Electronics Industry of China released a national standard (信息处理用现代汉语分词规范，GB13715) for automatic word segmentation in 1992.

The existing Chinese word segmentation algorithms can be divided into three categories:

- String based matching method
- Understanding based word segmentation method
- Statistics based word segmentation method

The first method is also known as mechanical word segmentation method. It is based on a certain strategy to match analyzed Chinese character string with a "sufficiently large" machine lexicon. If the matching is successful, it means that this string is an actual Chinese word.

The second method is to let computer to simulate human’s understanding for sentences and then to identify the effect of the words. The basic idea is to do syntactic and semantic analysis at the same time during word segmentation, the syntactic and semantic information can be used for dealing

with ambiguity of segmentation. This method requires a large amount of language knowledge and information. Because of the complexity for the knowledge of Chinese language, it is difficult to organize all kinds of linguistic knowledge directly transformed into machine-readable form.

Formally, a word is stable combination of characters. Therefore, in the context of adjacent characters, the higher frequency their occurrences have, the more likely they form a word. We can do statistics for co-occurrence frequency of adjacent characters in a corpus and compute their mutual information (MI), which reflects the tightness degree of combination relationship between Chinese characters. When tightness degree is higher than a certain threshold, they can be thought as a word. The third method only considers doing statistics frequency of characters in a corpus, don't need segmentation lexicon. However, this method also has certain limitations. Sometimes it identifies some words with high co-occurrence frequency, but the words are not commonly used words such as "这一", "之一", "有的", "我的", "许多的" and so on. In addition, the recognition accuracy for commonly used words is poor, and the overhead is large. Practical statistics based word segmentation system has to utilize a basic word segmentation lexicon (common word lexicon) for string matching, and simultaneously uses statistical methods to identify some new words. Thus it not only does word segmentation with high efficiency, but also it doesn't need a lexicon to identify new words only according to the context and automatically disambiguates.

Existing Problems

Ambiguity problem: It means that a sentence may have **two or more** segmentations. There are two main ambiguities:

- overlapping ambiguity (交集型歧义)
- combination ambiguity (组合型歧义)

For example: "表面的", because the "表面" and "面的" are all the word, this phrase can be divided into "表面 | 的" and "表 | 面的". This is called overlapping ambiguity (also cross-ambiguity). Because of the absence of human knowledge, it is difficult for computer to determine which one is correct.

The combination ambiguity is more difficult to handle than overlapping ambiguity. For instance, in the sentence "这个门把手坏了", "把手" is a word; but in the another sentence "请把手拿开", "把手" is not a word. How to identify such words for computer?

Even though the above two types of ambiguity can be resolved, there is still a difficult problem in the ambiguity is called real ambiguity (真歧义). It means that even though people judge such words, they can also not know which should be word, which should not be word. For example, in the sentence "乒乓球拍卖完了", it can be segmented into "乒乓 | 球拍 | 卖 | 完 | 了", and can be also segmented into "乒乓球 | 拍卖 | 完 | 了". Without the context, no one knows whether "拍卖" here is a word or not.

Unknown word problem: Named entities (that is, personal names, place names, organization names etc.), new words, technical terms are called unknown words. That is, those in the lexicon are not included, but those words really can be called words. For example, in the sentence "王军虎去广州了", "王军虎" is a word. Because it is a personal name, it is difficult for computer to

identify it. If we save the all personal names in a lexicon, it is a huge project. Even though this work can be completed, there exist still problems. For instance, in the sentence “王军虎头虎脑的”, “王军虎” can be a word (personal name) or not?

In addition to personal names, there are organization names, place names, product names, trade names, acronyms, abbreviation words etc. They are all difficult to identify. But they are exactly the words people often use. Therefore, the unknown word recognition is very important for word segmentation system. Currently, its accuracy evaluation has become an important feature for a Chinese word segmentation system.

Adopted Method and Algorithm

You will design a prototype system for Chinese word segmentation including user interface component, word segmentation component and lexicon management component (if it is needed). In this system, you can adopt different algorithm, such as string based matching method, understanding based word segmentation method, statistics based word segmentation method or other methods proposed by yourself. For each method, you can select an existing algorithm or new algorithms proposed by yourself. Note that you should use GB13715 guideline as golden standard for word segmentation.

In order to encourage students to study more difficult algorithms, if you accomplish a research task for more difficult algorithms, we will give you an additional premium.

Corpus Domain

On November 18, 1999, the Chinese government officially decided that Shanghai would bid for the 2010 World Expo. With support from home and abroad, Shanghai won the bid on December 3, 2002, at the 132nd General Assembly of the International Exhibitions Bureau.

The website www.expo2010.cn, managed by the Bureau of the Shanghai World Expo Coordination, is the only official Website of World Expo 2010 Shanghai China.

This website is the source of authoritative, accurate and timely information relating to the development of Expo 2010. It acts as an information center and a working platform for the running of the Expo in four versions: Simplified Chinese, Traditional Chinese, English and Japanese. The French version was also launched on July 13, 2008.

We utilize all of the texts on the website as our training corpus and test corpus. They can be categorized into Home, News, Encyclopedia, Pavilions, Activities, Forums, Services, Volunteer, Expo Online, Events etc. The texts can be directly obtained from the website by hand or extracted by automatically processing using computer. The following text is the example of news

上海世博会中国国家馆 12 月 1 日起续展半年

2010 年 11 月 25 日

世博网 11 月 25 日消息：为满足广大参观者的要求，让更多公众一睹“东方之冠”的风采，上海世博会中国国家馆将从 12 月 1 日起续展半年，即 2010 年 12 月 1 日至 2011 年 5 月 31 日。续展期间，中国国家馆普通票票价为每张 20 元，优惠票每张 15 元。

Concrete Requirements of Design and Implementation

User Interface Component: Design and implement a user interface component for prototype system using graphics module or TkInter module. Integrate the word segmentation and lexicon component developed by other team members with the user interface. The six sub-menus should be defined in the interface:

File: (1) open a file through choosing file name in hard disk or floppy disk; (2) exit from the system.

Edit: (1) edit a file; (2) input a sentence by hand.

Segmentation: (1) segment sentences of a text (input*.txt) by batch processing and output segmented result to another text (output*.txt); (2) segment a sentence into words.

Lexicon: (1) load the lexicon; (2) add a new word into the lexicon.

Rule: (1) load the rule library; (2) add a new rule into the rule library.

Help: (1) instructions about the system; (2) copyright information.

Word Segmentation Component: Design a word segmentation algorithm. It can effectively segment sentences and words. During the word segmentation process, you can utilize lexicons, rules (e.g., GB13715) and other resources (such as HowNet Knowledge Database) to help yourself do that. After the design, you will implement this component in the prototype system and debug it. The component should have high accuracy of segmentation. It should also have a good robust property. In addition, it can be added new words and rules through the user interface easily.

Lexicon Component: Design and implement a lexicon component. The lexicon includes word entry, index, other necessary information. The words in the lexicon can be modified. In addition, new words can be added to the lexicon.

After you accomplish prototype system integration, you should test the system using EXPO 2010 website texts. If word segmentation performance is not satisfied, you should improve your algorithm once more.

Input and Output Format

You can refer to the following input and output format for your user interface. But it is not only format.

Input:

上海世博会中国国家馆12月1日起续展半年

2010年11月25日

世博网11月25日消息：为满足广大参观者的要求，让更多公众一睹“东方之冠”的风采，上海世博会中国国家馆将从12月1日起续展半年，即2010年12月1日至2011年5月31日。续展期间，中国国家馆普通票票价为每张20元，优惠票每张15元。

Output1 - Sentence Segmentation:

Would you like to observe the result of sentence segmentation of the above text (yes = 1 / no = 0)?

Your input: 1

1. 上海世博会中国国家馆12月1日起续展半年

2. 2010年11月25日

3. 世博网11月25日消息：

4. 为满足广大参观者的要求，

5. 让更多公众一睹“东方之冠”的风采，

6. 上海世博会中国国家馆将从12月1日起续展半年，

7. 即2010年12月1日至2011年5月31日。

8. 续展期间，

9. 中国国家馆普通票票价为每张20元，

10. 优惠票每张15元。

Output2 - Word Segmentation:

Would you like to select one of the above sentences to look at the result of word segmentation? If yes, please input the indicated sentence no. If no, please input 0.

Your first input? 1

The result of word segmentation for the 1st sentence:

上海|世博会|中国|国家馆|12|月|1|日|起|续展|半|年|

Your next input? 5

The result of word segmentation for the 5th sentence:

让|更|多|公众|一|睹|“|东方之冠|”|的|风采|，|

Your next input? 0

Thank you for your testing, Goodbye!

Submission Instructions

Submit a program including user interface component, word segmentation component, and lexicon component. You also need to include a PDF file (*.pdf) that includes a final report (The format of the final report is shown below). These files should all be placed into one folder in a compressed file (e.g. 5120309000_XXX(team_projB_v3).rar) and submitted to the course representative of your class. The course representatives compress all of them with a *.rar file and upload it in the subdirectory team_projB.

The deadline is due the midnight on **Jan. 3, 2013**.

References

Steven Bird, Ewan Klein, Edward Loper. Natural Language Processing. Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 United States License (draft only). 2001-2008. http://nltk.googlecode.com/svn/trunk/doc/book/book.html.

Ronald Cole, editor. Survey of the State of the Art in Human Language Technology. Studies in Natural Language Processing. Cambridge University Press, 1997.

Steven Abney. Statistical methods and linguistics. In Judith Klavans and Philip Resnik, editors, The Balancing Act: Combining Symbolic and Statistical Approaches to Language. The MIT Press, 1996.

John M. Zelle. Python Programming: An Introduction to Computer Science (Chapter 5). Franklin, Beedle & Associates, 2004.

Fredrik Lundh. An Introduction to Tkinter. 1999.

Tkinter Documentation. 2010. http://wiki.python.org/moin/TkInter

John M. Zelle. Python Programming: An Introduction to Computer Science (Chapter 9). Franklin, Beedle & Associates, 2004.

The Python Tutorial. 2010. http://docs.python.org/release/2.6.6/tutorial/index.html.

The Ministry of Mechanics and Electronics Industry of China. 信息处理用现代汉语分词规范 (GB13715). 1992.

黄昌宁，赵海.中文分词十年回顾.《中文信息学报》21卷第三期. 2007年5月.

http://wenku.baidu.com/view/e711ae323968011ca3009124.html

Sun Maosong et.al. Chinese word segmentation without using lexicon and hand-crafted training data. In Proceeding of COLING '98. 1998. http://portal.acm.org/citation.cfm?id=980775

郑家恒等. 中文分词中歧义切分处理策略.山西大学学报(自然科学版), 30(2). 2007.

John M. Zelle. Python Programming: An Introduction to Computer Science (Chapter 11). Franklin, Beedle & Associates, 2004.

The Ministry of Mechanics and Electronics Industry of China. 信息处理用现代汉语分词规范 (GB13715). 1992.

Zhendong Dong and Qiang Dong. HowNet Knowledge Database. http://www.keenage.com/. 2011.

Appendix A: Final Report Format

Team Name: Team Leader:

Date:

1 Prototype System Introduction

1.1 Functions

1.2 Running Environment

Windows 7

1.3 Developing Environment

PyScripter 2.5.3

…

2 Task Allocation

The tasks for each team member.

3 System Architecture

3.1 User Interface Component

Graph and explanation.

3.2 Simulation Component

Graph and explanation.

3.3 Visualization Component

Graph and explanation.

4 Algorithm Description

4.1 User Interface Component

Flowchart and explanation.

4.2 Simulation Component

Flowchart and explanation.

4.3 Visualization Component

Flowchart and explanation.

5 Demo and Testing Result

5.1 Screenshots

User interface and every operation.

5.2 Testing Procedure, Data and Result

Explanation, testing data and result table, and result analysis

6 Conclusion

The discussion of your experiment and answers the questions listed above.

What is your research result? What are your experience and lesson on this project?

Appendix B: Input and Output for Chinese Sentence by Keyboard and Screen

```
list_ch = []
str_ch = input("Please input a Chinese string: ")
for i in range(len(str_ch)):
    list_ch.append(str_ch[i])
for i in range(len(list_ch)):
    print(list_ch[i], end = " ")
print()
```

Please input a Chinese string: 试验中文输入输出

试 验 中 文 输 入 输 出

Appendix C: Input and Output for Chinese Sentence by Files

```
list_ch = []
infile = open("input.txt", 'r')
outfile = open("output.txt", 'w')
for line in infile:
    for i in range(len(line)):
        list_ch.append(line[i])
infile.close()
for i in range(len(list_ch)):
    if list_ch[i] == "\n":
        outfile.write("\n")
    else:
        outfile.write(list_ch[i])
        outfile.write(" ")
outfile.close()
```

input.txt

上海交通大学

试验中文输入输出

output.txt

上 海 交 通 大 学

试 验 中 文 输 入 输 出