Proceedings of the Second SIGHAN Workshop on Chinese
Language Processing, July 2003, pp. 152-155.

Chinese Word Segmentation at Peking University

Duan Huiming Bai Xiaojing Chang Baobao Yu Shiwen
Institute of Computational Linguistics, Peking University
{duenhm, baixj, chbb, yusw}@pku.edu.cn

Abstract

Word segmentation is the first step in Chinese information processing, and the performance of the segmenter, therefore, has a direct and great influence on the processing steps that follow. Different segmenters will give different results when handling issues like word boundary. And we will present in this paper that there is no need for an absolute definition of word boundary for all segmenters, and that different results of segmentation shall be acceptable if they can help to reach a correct syntactic analysis in the end.

Keyword: automatic Chinese word segmentation, word segmentation evaluation, corpus, natural language processing

1. Introduction

On behalf of the Institute of Computational Linguistics, Peking University, we would like to thank ACL-SIGHAN for sponsoring the First International Chinese Word Segmentation Bakeoff, which provides us an opportunity to present our achievement of the past decade.

We know for sure that it is very difficult to settle on a scientific and appropriate method of evaluation, and it might be even more difficult than word segmentation itself. We are also clear that each step in Chinese information processing requires great efforts, and a satisfactory result in word segmentation, though critical, does not necessarily guarantee good results in the following steps.

From the test results of this evaluation, we are very gratified to see that we have done a good job both as a test corpus provider and as a participant. According to the rule, we did not test on the corpus we provided, but it is quite encouraging that our supply tops the test corpus list to be elected by other participants.

Section 2 and Section 3 describes our work in the Bakeoff as the test corpus provider and the participant respectively.

2. The test corpus provider

2.1 Corpus

The corpus we provided to the sponsor includes:

- A training set from People's Daily (January, 1998)
- A test set from People's Daily (Page 4 of January 1, 1998)

Data from People's Daily features standard Chinese, little language error, a wide coverage of linguistic phenomenon and topics, which are required for statistic training. Meanwhile, the corpus we provided is a latest version manually validated, hence a high level of correctness and consistency.

2.2 Specification

When processing a corpus, we need a detailed and carefully designed specification for guidance. And when using the corpus for NLP evaluation, we also need such a specification to ensure a fair contest for different systems within a common framework.

We provided the latest version of our specification, which has been published in the *Journal of Chinese Information Processing*. Based on our experience of large-scale corpus processing in recent years, the current version gave us different perspectives in a consistent way, and we hope it will also help others in this field know better of our segmented and POS-tagged corpus.

3. The participant

3.1 Training and testing

Our research on word segmentation has been focusing on People's Daily. As we are one of the two providers of Chinese corpora in GB code in this Bakeoff, we had to test on the Penn Chinese treebank.

Not all the training and test corpus we got came from the Mainland China. Some were GB data converted from BIG5 texts of Taiwan. It is commonly known that in the Mainland, Hong Kong and Taiwai, the Chinese langauge is used diversely not only in the sense of different coding systems, but in respect of different wordings as well.

While training our segmenter, we studied the guidelines and training corpus of Penn Chinese treebank, tracing the differences and working on them. The main difference between the work of U. Penn and that of ours is notion of "word". For instance:

Differences of "Word"	U. Penn	PKU
Chinese name	刘卫东，彭少阳	刘　卫东, 彭　少阳
Number + "多|余……"	11.6 万余，八千七百多万	11.6 万　余，八千七百　多　万
Monosyllabic verb + complement	砌出，填上，读完	砌　出，填　上，读　完
Time word	十时三十分，９０年代	十时　三十分，９０　年代
Noun + suffix "们"	大学生们，企业家们	大学生　们，企业家　们
Disyllabic verb + "于"	领先于，受命于	领先　于，受命　于
… …		

These are different combinations in regard of words which follow certain patterns, and can therefore be handled easily by applying rules to the grogram. The real difficulty for us, however, is the following items:

U. Penn	PKU
有线　电视	有线电视
中央　军委	中央军委
中华　民族	中华民族
人民　日报	人民日报
知识　产权	知识产权
一　条　龙	一条龙
… …	… …

The Open Track allows us to use our own recourses, so we had to find the lexical correspondence to reduce the negtive effect caused by the difference between Penn Chinese treebank and our own corpus. However, as the training corpus is small, we could not remove all the negative effect, and the untackled problems remained to affect our test result.

Further, as we have been working on language data from the Mainland China, the lexicon of our segmenter does not contain words used in Taiwan. Such being the case, we added into our lexicon the entries that were not known (i.e., not found in the training set) and could not be handled by the rule-based makeshift either. But because we are not very familiar with the Chinese language used in Taiwan, we could not make a complete patch due to the limit of time.

3.2 Result analysis

From the test result that the sponsor provided, we can see our segmenter failed to score when the notion of "word" and the recognition of

unknown words are involved.

Example 1:
[U. Penn] 实施 初期 将 以 **除罪化** 与 小额 贸易 合法化 为主 ， 在 观察 小三通 对 **金 马** 地区 治安 与 产业 影响 后 ， 才 会 考虑 在 第二 阶段 开放 商业性 行为 ， 至于 大 三通 的 实施 ， 尚 没有 明确 的 时间表 。
[PKU] 实施 初期 将 以 **除 罪 化** 与 小额 贸易 合法化 为主 ， 在 观察 小三通 对 **金马** 地区 治安 与 产业 影响 后 ， 才 会 考虑 在 第二 阶段 开放 商业性 行为 ， 至于 大 三通 的 实施 ， 尚 没有 明确 的 时间表 。

Example 2:
[U. Penn] 一 家 **宽频** 公司 负责 管理 两 个 大型 鱼缸 的 赖 小姐 更 表示 ， 公司 受 景气 影响 ， 不免 人人 节衣缩食 ， 但 每 个 月 五 、 六千 元 的 鱼缸 清理费 可 不 曾 少 过 ， 更 别 提 不时 得 补充 鱼 、 饲料 等 费用 。
[PKU] 一 家 **宽 频** 公司 负责 管理 两 个 大型 鱼缸 的 赖 小姐 更 表示 ， 公司 受 景气 影响 ， 不免 人人 节衣缩食 ， 但 每 个 月 五 、 六千 元 的 鱼缸 清理费 可不 曾 少 过 ， 更 别提 不时 得 补充 鱼 、 饲料 等 费用 。

In addition, there are also cognitive differences concerning the objective world, which did come up to influence our fine score.

Example 3:
[U. Penn] 吴思华 则 以 英特尔 为 例 ， 说明 知识 「 点 矽 成 金 」 的 威力 ： 一 张 名片 大小 的 CPU ， 说穿 了 就 是 一 块 烧 有 复杂 电路图 的 **矽 晶 板** 。
[PKU] 吴思华 则 以 英特尔 为 例 ， 说明 知识 「 点 矽 成金 」 的 威力 ： 一 张 名片 大小 的 CPU ， 说穿 了 就 是 一 块 烧 有 复杂 电路图 的 **矽晶 板** 。

Example 4:
[U. Penn] 啊 ， 外太空人 入侵 ？ 别 紧张 ， 这 是 新 车 大展 中 ， 观众 戴着 特制 的 「 **虚拟 实 境** 」 头盔 ， 体会 一下 自己 驾驶 新型 车 的 超炫 快感 。
[PKU] 啊 ， 外 太空人 入侵 ？ 别 紧张 ， 这 是 新车 大展 中 ， 观众 戴着 特制 的 「 **虚拟 实境** 」 头盔 ， 体会 一下 自己 驾驶 新型 车 的 超 炫 快感 。

The recognition of unknown words has long been a bottleneck for word segmentation technique. So far we have not found a good solution, but we are confident about a progress in this respect in the near future.

4. Conclusion

Word segmentation is the first step yet a key step in Chinese information processing, but we have not found a perfect solution up till now. From an engineering perspective, we think there is no need for a unique result of segmentation. All roads lead to Rome. The approach you take, technical or non-technical, will be a good one if the expected result is achieved. And it would be more desirable if the processing program in each step can tolerate or even correct the errors made in the previous step.

We learn from our experience that the computer processing of natural language is a complex issue, which requires a solid fundamental research (on the language itself) to ensure a higher accuracy of automation. It is definitely hard to achieve an increase of one percent or even less in the accuracy of word segmentation, but we are still confident and will keep working in this respect.

Finally, we would like to thank Dr. Li Baoli and Dr. Bing SWEN for their great efforts on the maintenance of our segmentation program.

Reference

Yu, Shiwen, DUAN, Hui-ming, ZHU, Xue-feng, Bing SWEN. 2002. *The Specification of Basic Processing of Contemporary Chinese Corpus*. Journal of Chinese Information Processing, Issue 5 & Issue 6, 2002.

Yu, Shiwen, et al. 2002. The Grammatical Knowledge-base of Contemporary Chinese – A Complete *Specification (Second Version)*. Beijing: Tsinghua University Press.

Liu, Yuan, et al. 1994. *Specification and Automation of Word Segmentation of Contemporary Chinese for Information Processing*. Beijing: Tsinghua University Press.

Fie Xia. 2000. *The segmentation guidelines for the Penn Chinese tree bank (3.0)*. see http://www.cis.upenn.edu/~chinese/segguide.3rd.ch.pdf