The CIPS-SIGHAN CLP 2014 Chinese Word Segmentation Bake-off

Huiming Duan
Key Laboratory of Computational Linguistics, Peking University, Beijing 100871, China
duenhm@water.pku.edu.cn

Zhifang Sui
Key Laboratory of Computational Linguistics, Peking University, Beijing 100871, China
szf@pku.edu.cn

Tao Ge
Key Laboratory of Computational Linguistics, Peking University, Beijing 100871, China
getao@pku.edu.cn

Abstract

This paper summarizes the SIGHAN 2014 Chinese Word Segmentation bake-off in several aspects such as dataset, evaluation results. In addition, we analyze errors of segmentation by instance and make a suggestion for improving segmentation systems.

1 Goal of the Chinese word segmentation bake-off

Chinese Word Segmentation is the preliminary step for Chinese information processing, which is extremely important and never neglected. Due to the properties of Chinese, the performance of Chinese word segmentation has an effect on the following analysis of Chinese text. As the organizer of the bake-off in Chinese word segmentation, not only do we show the performance of all participated systems, but also try to find out the weak point of these systems. In this way, participants are able to learn advantages of their systems and realize the problems which they did not pay attention to so that they could improve their system according to our feedbacks, which turns out to promote the study of Chinese word segmentation.

2 Dataset

2.1 Size of dataset

The dataset used in the SIGHAN2014 Chinese word segmentation bake-off is formed by sampling instances which are difficult to segment from approximately 1.3T Chinese corpus. This is a huge challenge for us. While sampling instances, we found that the distribution of sentences which are hard to segment does not depend on domains, in other words, these sentences appear in every domain.

2.2 Domains of dataset

Compared with the SIGHAN 2012 Chinese word segmentation bake-off which only focuses on the microblog domain, the dataset used in the shared task in SIGHAN2014 is formed by sampling sentences from a variety of domains. The dataset involves many subjects in both social sciences and natural sciences, and genres involved in the dataset are also taken into consideration. In this way, we can more clearly evaluate if current segmentation techniques can perform well in a wide range of domains.

2.3 Makeup of dataset

The SIGHAN2014 Chinese word segmentation bake-off mainly uses single sentences and paragraphs for evaluations. Additionally, discourses are also included.

As is known to all, there are two kinds of ambiguities in Chinese word segmentation – overlapping ambiguity and combinatorial ambiguity, which are difficult to deal with. In addition, OOV (out of vocabulary), which includes neologisms, abbreviations and uncommon terminology, is a challenge for Chinese word segmentation as well.

First, we show why the ambiguity of segmentation arises.

Segmentation ambiguity:

(1) Combinatorial ambiguity

It is not uncommon to see these words in Chinese: 树木、应对、根据地、正在、一道、一起、一块、一口气……

① 树木

树木自己要学会在土地里找水源，

——Here,树木 is a noun.

一年之计，莫如树谷；十年之计，莫如

Proceedings of the Third CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 90–95,
Wuhan, China, 20-21 October 2014

树木；终身之计，莫如树人。

——Here, 树木 is not a noun.树 is a verb rather than a noun.

② 应对

此时人们将无法正常地应对现实世界。

——Here,应对 is a verb.

在治疗前应对患者病变的部位(神经根定位)有明确的认识,

——Here, 应对 is two words。

③ 根据地

杨洁篪说，该报告毫无根据地攻击中国国防现代化，

——毫无根据地 should be segmented as 毫无 根据 地

(2) Overlapping ambiguity: 词语首尾的可成词性

There are many overlapping ambiguities in the dataset. For example:

塑造成：塑造+造成

心理学工作者：心理+理学+学工+工作+作者

司机会：司机 + 机会

心中立起：心中+中立+立起

正在家中看：正在+在家+家中+中看

在行军中：在行+行军+军中

以下划线：以下+下划线 (* All systems make a mistake segmenting this sequence)

在场论：在场+场论 (* “场论” is a word used in only a few domains)

享有的：享有+有的

We mainly test the performance of disambiguation of systems. Given that some ambiguous sequences of characters often appear in different context, we sometimes use multiple sentences to evaluate a sequence of characters. It is notable that some sentences’ context can provide helpful information while some sentences do not have such information. We want to see the capability of systems to use context to solve overlapping ambiguities. For example：

“无数学”

因有**无数学**子从这里走出去

将有**无数学**子背负着青春的理想

自然会有**无数学**者谈论

无数学过的占卜、巫术

仍有**无数学**者在对其进行着不断的研究。

都有**无数学**生在学校里轮流读着已知的二战死难者名单

有**无数学**者分析过

不能以有**无数学**公式及其推导来衡量文章的水平高低。

动物有**无数学**头脑

诺贝尔奖有**无数学**奖

心中虽有**无数学**识

也有**无数学**不尽的知识

“在行”

由用户**在行**与行间选择要做这种计算的记录

尽管世行**在行**长提名权和任职条件上

其中结脉多因于气血凝滞，重**在行**气活血

并且**在行**文上有着程式性的规定

，**在行**业领先才能生存的前景下

在行唐县的推荐下，

应**在行**经前 3 天即开始服用

个个一专多能，吹、拉、弹、唱、舞样样**在行**，

不能担挑，拾柴却很在行，

As for names, we choose two lists of names as example:

Example 1：麦培东麦谢巧玲（女）麦耀堂严日初严建平严震铭苏开鹏苏西智苏丽珍（女）苏肖娟（女）苏泽光苏炤成苏晓鹏苏健康苏绮丽（女）苏耀华杜毅（女）杜耀明李乃尧李乃熹李大壮李子良李月华

Example 2：邓天生叶青纯田力普令狐安冯寿淼冯敏刚年福纯朱明国（黎族）朱保成刘玉亭刘亚洲刘建华（女）刘春良刘晓榕安立敏（女）许云昭许达哲孙忠同孙宝树孙思敬杜鹃（女）

3 Evaluation Results

Precision, recall and F-measure are used to evaluate participants’ systems, just as previous bakeoffs did. Since the number of participants is not large (6 institutes and 7 systems), we can analyze the systems in detail for finding the weak points of the systems, which would promote the study of Chinese word segmentation.

Precision, recall of F-measure of participants' systems

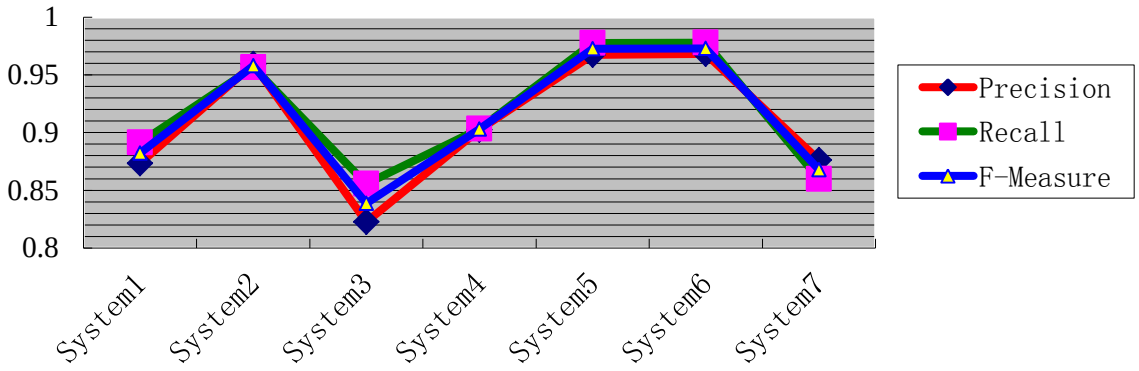

Table 1: Distribution of P,R,F of systems participating in this bake-off

3.1 Automatic Evaluation

For automatic evaluation, Precision, recall and F-measure are used to evaluate participants' systems.

The performance of 7 systems of 6 institutes participating in the bake-off is shown in Table1.

No.	Precision	Recall	F-Measure
System1	0.8734	0.8912	0.8822
System2	0.9592	0.9566	0.9579
System3	0.8226	0.8555	0.8387
System4	0.9025	0.9032	0.9029
System5	0.9673	0.9776	0.9724
System6	0.9681	0.9779	0.9730
System7	0.8760	0.8597	0.8678

Table 1: Precision, recall and F-measure of all systems participating in this bake-off

We compare the results in the bake-off with that in SIGHAN 2012

	Precision	Recall	F-Measure
2012	0.946	0.9496	0.9478
2014	0.9681	0.9779	0.9730

Table 2: The best systems in 2012 and 2014 bake-offs

	Precision	Recall	F-Measure
2012	0.9347	0.9316	0.9331
2014	0.9681	0.9779	0.9730

Table 3: Systems by the same institute in 2012 and 2014

	Precision	Recall	F-Measure
2012	0.1314	0.0845	0.1087
2014	0.1455	0.1224	0.1342

Table 4: Differences between the best system and the worst system in 2012 and 2014

3.2 Manual Inspection

3.2.1 Why manual inspection

In previous SIGHAN segmentation shared task, precision, recall and F-measure are only metric for evaluating systems. Although these metrics can reflect systems' performance to some extent, they cannot clearly show the specific weak point of the systems. It is likely that a system achieving high PRF does not deal with some details well and makes some silly mistakes. On the other hand, some systems whose PRF is not high can address some specific segmentation problems well. Of course, other factors such as the size of dictionary might also affect the results.

Since SIGHAN 2012 Chinese word segmentation bake-off, we have attempted to introduce evaluations for some specific cases, which could inform participants of the approximate accuracy range of each case and allow them to learn the weak points of their systems.

By manual inspection, we found some typical mistakes which should have been corrected but were not solved by most systems.

3.2.2 Methods of manual inspection

We use different types of lines (a single line, double line or dash line) to indicate how to segment a sequence of Chinese characters.

事实上 ， 在 互联网 地图 日益 得到 广泛 应用 之 时 ， 一些 互联网 地图 服务 质量 不 高 ， 内容 不准 的 问题 也 广受 网友 诟病 。 国家 测绘 地理 信息局 有关 负责人 表示 ， 互联网 地图 服务 基本 纳入 法制化 、 规范 化 管理 的 轨道 ， 对 提 高 互联网 地图 服务 质量 、 方 便 社会 各界 更 好 享受 互联网 地图 服务 、 保障 国家 地理 信息 安全 将 起 到 良好 作用 。 .
事实上 ， 在 互联网 地图 日益 得到 广泛 应用 之 时 ， 一些 互联网 地图 服务 质量 不 高 ， 内容 不 准 的 问题 也 广 受 网友 诟病 。 国家 测绘 地理 信息局 有关 负责人 表示 ，
事实上 ， 在 互联网 地图 日益 得到 广泛 应用 之 时 ， 一些 互联网 地图 服务 质量 不 高 ， 内容 不 准的 问题 也 广 受 网友 诟病 。 国家 测绘 地理 信息 局 有关 负责人 表示 ，
事实上 ， 在 互联网 地图 日益 得到 广泛 应用 之 时 ， 一些 互联网 地图 服务 质量 不 高 ， 内容 不 准 的 问题 也 广 受 网友 诟病 。 国家 测绘 地理 信息 局 有关 负责人 表示 ，
事实上 ， 在 互联网 地图 日益 得到 广泛 应用 之 时 ， 一些 互联网 地图 服务 质量 不 高 ， 内容 不准 的 问题 也 广 受 网友 诟病 。 国家 测绘 地理 信息 局 有关 负责人 表示 ，
事实上 ， 在 互联网 地图 日益 得到 广泛 应用 之 时 ， 一些 互联网 地图 服务 质量 不 高 ， 内容 不 准的 问题 也 广 受 网友 诟病 。 国家 测绘 地理 信息 局 有关 负责人 表示 ，
事实 上 ， 在 互联网 地图 日益 得到 广泛 应用 之时 ， 一些 互联网 地图 服务 质量 不 高 ， 内容 不准 的 问题 也 广 受 网友 诟病 。 国家 测绘 地理 信息 局 有关 负责人 表示 ，

Table 5: Using different types of lines as indicators to conduct human inspection

Example 3：Merge

这 其实 我 根本 也 没有 做 主权 嘛

a single line indicates that the sequence should be merged as 做主权

Example 4：Segment

充电 时间的确 太 长

a double line indicates that the sequence should be segmented as：时间 的确

Example 5: Re-combine

其中 的 解决 方 案之一 就 是 ：

a dash line indicates that the sequence should be re-combined as 方案 之一

By using different types of lines as indicators, one can easily learn the mistakes made by each system, as table 5 shows.

As shown in table 5, only one system segments the sequence without any mistake. In contrast, one of the systems makes many mistakes when segmenting simple terms, which may arise from the problem of word-collection or some further problems.

4 Analysis of Results

4.1 Excessive word-collection may have an adverse effect

In table 6, only one system segments ‘对方’.

It can be verified by table 7 that this system did not include ‘对方’ in its dictionary.

As shown in table 6 and table 7, a system which includes ‘对方’ in its dictionary segments ‘对方’ correctly while others make a mistake here. We hope that the system actually pays attention to the detail rather than happen to segment it well. There are many similar cases such as ‘平等’ and ‘杜鹃’.

Example 6： 公司派张世平等一批技术骨干和管理人员到国外学习。

“杜鹃” in example 7 is a noun while it is a person’s name in example 2. Therefore, 杜鹃 should be segmented in example 2.

在 庭审 中 ， 对 方 律师 竟 对 中 方 托收 银行 寄 送 托受 文件 的 事实 全盘 否认 。
在 庭 审 中 ， 对方 律师 竟 对 中方 托收 银行 寄送 托 受 文件 的 事实 全盘 否认 。
在 庭审 中 ， 对方 律师 竟 对 中 方 托收 银行 寄 送 托 受 文件 的 事实 全盘 否认 。
在 庭审 中 ， 对方 律师 竟 对 中方 托收 银行 寄送 托 受 文件 的 事实 全盘 否认 。
在 庭审 中 ， 对方 律师 竟 对 中方 托收 银行 寄送 托 受 文件 的 事实 全盘 否认 。
在 庭审 中 ， 对方 律师 竟 对 中 方 托收 银行 寄 送 托 受 文件 的 事实 全盘 否认 。
在 庭审 中 ， 对方 律师 竟 对 中方 托收 银行 寄送 托受 文件 的 事实 全盘 否认 。

Table 6: Segmentation results of all systems for a sentence

这 也 与 学生 请 愿书 中 对 方 艳华 的 评价 相同 。
这 也 与 学生 请愿书 中 对方 艳华 的 评价 相同 。
这 也 与 学生 请愿 书 中 对方 艳 华 的 评价 相同 。
这 也 与 学生 请愿书 中 对方 艳华 的 评价 相同 。
这 也 与 学生 请愿书 中 对方 艳华 的 评价 相同 。
这 也 与 学生 请愿 书 中 对方 艳 华 的 评价 相同 。
这 也 与 学生 请愿书 中 对 方 艳华 的 评价 相同 。

Table 7: Segmentation results of all systems for another sentence

Example 7：在位于羊西线的西部花卉市场里，一排排水仙、菊花、杜鹃、郁金香等争奇斗妍、姹紫嫣红，前来赏花、买花的市民络绎不绝。

We can also give many other examples: 长江[江 can be surname], 孙子[孙 can be surname]，王[王 can be surname]子, 行李[尉健行李铁印] etc. To address these problems, an effective personal name recognition method is necessary.

4.2 A lack of attention to details

Example 8：进攻者比防御者更容易包围对方的全部军队以及切断它们的退路,因为防御者处于驻止状态,而进攻者是针对防御者的这种状态进行运动的。

Example 8 is an instance in test set. In this sentence, 进攻者 appears three times and 防御者 appears twice. Nonetheless, some systems cannot deal with these terms consistently. The cause of the phenomenon is that the systems do not exploit the context well.

进攻 者 比 防御 者 更 容易 包围 对方 的 全部 军队 以及 切断 它们 的 退路 ，因为 防御 者 处于 驻 止 状态 ，而 进攻者 是 针对 防御者 的 这种 状态 进行 运动 的 。

Example 9：于廿七号晚上出发，

In example 9, seldom has 廿七号 been used in written language in recent years. However, a good system is supposed to take into consideration these cases. Incorrect segmentations are shown as follows.

于 廿七 号 晚上 出发 ，
于 廿 七 号 晚上 出发 ，

5 Conclusion

Although languages have many properties in common, their unique characters do not allow researchers to directly use techniques for processing other languages to process Chinese.

In addition, when devoted to language study, one can find that Chinese has significant uniqueness and flexibility, which should be paid much attention to. Only by carefully analyzing unique properties of Chinese can researchers come up with a better solution to improving their systems. Even though Chinese is so flexible that one can-

not use a rule to describe the problems of Chinese word segmentation, researcher can try multiple rules to optimize their systems in multiple aspects and multiple levels, which requires them to be mindful of details.

As the organizers of this Chinese word segmentation bake-off, we may need to scrutinize details and make a standard which is detailed and easy to operate. For the bake-off, we are going to explore a better evaluation method which can show the results of systems more reasonably and objectively.

Acknowledgement: This paper is supported by National Key Basic Research Program of China 2014CB340504 and NSFC 61375074.

Reference

Shiwen Yu, Huiming Duan, Xuefeng Zhu, Bin Sun and Baobao Chang. 北大语料库加工规范：切分• 词性标注• 注音. 汉语语言与计算学报, 13(2), 121-158.

Hongmei Zhao and Qun Liu. The CIPS-SIGHAN CLP 2010 Chinese word segmentation bakeoff. In Proceedings of the First CPS-SIGHAN Joint Conference on Chinese Language Processing (pp. 199-209).

Duan, Huiming, Zhifang Sui, Ye Tian, and Wenjie Li. The CIPS-SIGHAN CLP 2012 Chinese Word Segmentation on Microblog Corpora Bakeoff. In Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing, Tianjin, China, pp. 35-40. 2012.