Detecting Segmentation Errors in Chinese Annotated Corpus

Chengjie Sun[*] Harbin Institute of Technology, Harbin, 150001, China

Chang-Ning Huang Microsoft Research, Asia, Beijing, 100080, China

Xiaolong Wang Harbin Institute of Technology, Harbin, 150001, China

Mu Li Microsoft Research, Asia, Beijing, 100080, China

{cjsun, wangxl}@insun.hit.edu.cn cnhuang@msrchina.research.microsoft.com muli@microsoft.com

Abstract

This paper proposes a semi-automatic method to detect segmentation errors in a manually annotated Chinese corpus in order to improve its quality further. A particular Chinese character string occurring more than once in a corpus may be assigned different segmentations during a segmentation process. Based on these differences our approach outputs the segmentation error candidates found in a segmented corpus and then on which the segmentation errors are identified manually. Segmentation error rate of a gold standard corpus can be given using our method. In Peking University (PK) and Academic Sinica (AS) test corpora of Special Interest Group for Chinese Language Processing (SIGHAN) Bakeoff1, 1.29% and 2.26% segmentation error rates are detected by our method. These errors decrease the F-measure of SIGHAN Bakeoff1 baseline test by 1.36% in PK test data and 1.93% in AS test data respectively.

1 Introduction

SIGHAN Bakeoff1[1] proposed an automatic method to evaluate the performance of different Chinese word segmentation systems on four distinct data sets. This method makes the performance of different Chinese word segmentation systems comparable and greatly promotes the technology of Chinese Word Segmentation. However, the quality of the reference corpora in the evaluation should be paid more attention because they provide training material for participants and they serve as a gold standard for evaluating the performance of participant systems.

This paper presents a semi-automatic method to detect segmentation errors in a manually annotated Chinese corpus in order to improve its quality further. Especially a segmentation error rate of a gold standard corpus could be obtained with our approach. As we know a particular Chinese character string occurring more than once in a corpus may be assigned different segmentations. Those differences are considered as segmentation inconsistencies by some researchers (Wu, 2003; Chen, 2003). Segmentation consistency is also considered as one of the quality criteria of an annotated Chinese corpus (Sun, 1999). But in order to provide a more clearer description of those segmentation differences we define a new

[*] This work was done while Chengjie Sun was visiting Microsoft Research Asia.

[1] http://www.sighan.org/bakeoff2003/

term, segmentation variation, to replace the original one, segmentation inconsistency.

Our approach of spotting segmentation errors within an annotated corpus consists of two steps: (1) automatically listing the segmentation error candidates with segmentation variations found in an annotated corpus, (2) spotting segmentation errors within those candidates manually. The target of this approach is to count the number of error tokens in the corpus and give the segmentation error rate of the corpus, which is not given for any gold standard corpus in Bakeoff1.

The remainder of this paper is structured as follows. In section 2, we discriminate the kinds of segmentation inconsistencies in test sets of SIGHAN Bakeoff1. In section 3, segmentation variation is defined and our approach to detect segmentation errors in a manually annotated corpus is proposed. In section 4 we conduct baseline experiments of PK and AS corpora with revised test sets in order to show exactly the impact of segmentation errors in the test sets of Bakeoff1. Section 5 is a brief conclusion.

2 Segmentation inconsistency

In the close test of Bakeoff1, participants could only use training material from the training data for the particular corpus being testing on. No other material was allowed (Sproat and Emerson, 2003). As we know that the test data should be consistent with the training data based on a general definition of Chinese words. That is if we collect all words seen in the training data and store them into a lexicon, then each word in a test set is either a lexicon word or an OOV (out of vocabulary) word (Gao et al., 2003). In another word, if a character string has been treated as one word, i.e. a lexicon word, in the training data, the same occurrence should be taken in the corresponding test data unless it is a CAS (combination ambiguity string) and vice versa.

As we all know that a CAS like "才能 [cai2-neng2]" may be segmented into one word or two words depending on different contexts. Thus segmentation inconsistency like "才能" (talent) and "才 能" (only can) could both be correct segmentations in a text. Therefore “segmentation inconsistency” should not be regarded as incorrect segmentations in general and should be clarified further. If one wants to discuss the segmentation errors based on segmentation inconsistencies, then from which those CAS instances should be excluded.

If we exclude CAS words in our investigation temporary then for a non-CAS character string, there are four kinds of situations violating the general definition of Chinese word, also called lexicon driven principle in automatic word segmentation technology:

S1. A character string is segmented inconsistently within a training data;

S2. A character string is segmented inconsistently within a test data;

S3. A character string is segmented inconsistently between a test data and its training data. This situation could be divided into the following two cases further:

S3.1 A word identified in a training data has been segmented into multiple words in corresponding test data;

S3.2 A word identified in a test data has been segmented into multiple words in corresponding training data.

Chen (2003) describes inconsistency problem found in cases S1, S2 and S3.1 of PK corpora. For example, he gives the amount of unique text fragments that have two or more segmentations within PK training data, within PK test data and also between PK training data and PK test data.

But those CAS words have not been excluded in his description. Ignoring the content of inconsistencies the influence about the number of segmentation inconsistencies of a particular corpus will be exaggerated greatly. In addition, Chen didn't consider the case of S3.2 which could also affect the evaluation significantly according to the lexicon driven principle. 53 word types found in case 3.2 (refer to Appendix part 2) were totally treated as OOV words in Bakeoff1 which impacts the identification of those authentic new words in the task. So the issue of segmentation inconsistency in reference corpora needs further investigation.

As mentioned before, in common knowledge "segmentation inconsistency" is a derogatory term. But our investigation shows that most of segmentation inconsistencies found in an annotated corpus turned out to be correct segmentations of CASs. Therefore it is not an appropriate technique term to assess the quality of an annotated corpus. Besides, with the concept of "segmentation inconsistency" it is hard to distinguish the different inconsistent components within an annotated corpus and finally count up the number of segmentation errors exactly. In the next section we propose a new term "segmentation variation" to replace the original one, "segmentation inconsistency".

3 Segmentation variation

3.1 Definition

Definition 1: In annotated corpora *C*, a set of f(*W*, *C*) is defined as: f(*W*, *C*) = {all possible segmentations that word *W* has in corpora *C*}.

Definition 2: *W* is a **segmentation variation type (segmentation variation** in short, hereafter) with respect to *C* iff |f(*W*, *C*)|>1.

Definition 3: An instance of element in f(*W*, *C*) is called a **variation instance**. Thus a segmentation variation (type) consists of more than one variation instances in corpora *C*. And a variation instance may include one or more than one **tokens**.

Definition 4: If a variation instance is an incorrect segmentation, it is called an **error instance** (**EI**).

The definitions of segmentation variation, variation instance and error instance (EI) clearly distinguish those inconsistent components, so we can count the number of segmentation errors (in tokens) exactly.

The term variation is also used to express other annotation inconsistency in a corpus by other researchers. For example, Dickinson and Meurers (2003) used variation to describe POS (Part-of-Speech) inconsistency in an annotated corpus.

Example 1: Segmentation variations (Bakeoff1 PK corpus):
Word "等同[deng3-tong2]" is segmented as "等同" (equal) and "等 同" (et al. with).
Word "黄金周[huang2-jin1-zhou1]" is segmented as "黄金周" (golden week) and "黄金 周" (gold week).
Word "冰清玉洁[bing1-qing1-yu4-jie2]" is segmented as "冰清玉洁" (pure and noble) and "冰 清 玉 洁" (ice clear jade clean).

In example 1, Words like "等同", "黄金周" and "冰清玉洁" are segmentation variation types. Segmentations "等同" and "等 同" are two variation instances of segmentation variation "等同". Besides, the variation instance "等 同" consists of two tokens "等" and "同". While the variation instance "冰 清 玉 洁" consists of four tokens "冰", "清", "玉" and "洁".

The existence of segmentation variations in corpora lies in two reasons: 1) ambiguity: variation type *W* has multiple possible segmentations in different contexts, or 2) error: *W* has been wrongly segmented which could be judged by a given lexicon.

Example 2: A segmentation variation caused by ambiguity (Bakeoff1 PK corpus):
Segmentation variation: "国都[guo2-du1]"
Variation instances: "国都" (capital) and "国[guo2] 都[dou1]" (countries all). They are both correct segmentations in following sentences:

君士坦丁堡 成为 拜占庭 国都 。
(Constantinople became the capital of Byzantium.)

两 国 都 主张 通过 对话 和 谈判 来 解决 分歧 。
(Both countries all advocate solving disagreements by conversation and negotiation.)

Example 3: Segmentation variations caused by error (Bakeoff1 PK corpus):
Segmentation variation: "尽管如此 [jin4-guan3-ru2-ci4]"
Variation instances: "尽管如此" (still) and "尽管 如此" (despite so).
Segmentation variation: “冰清玉洁”
Variation instances: “冰清玉洁” and “冰 清 玉 洁” .

In the rest of the paper, a segmentation variation caused by ambiguity is called a CAS variation and a segmentation variation caused by error is called a non-CAS variation. Each kind of segmentation variations may include error instances (EIs).

Situation	Within test data	Between: One-to-Mult	Between: Mult-to-One
# of variation type	21	92	228
# of variation instances	87	129	506
# of EIs*	12(3)	68(4)	77
# of error tokens*	28(6)	142(8)	77

*: The number in the bracket is the amount caused by CAS.

Table 1 segmentation variations types, instances and EIs in PK test data

3.2 Finding error instances (EIs)

How to find the segmentation variations in corpora? Following is the algorithm of finding segmentation variations. According to our definition, the algorithm is quite straightforward. It takes two segmented Chinese corpora (reference corpus and corpus to be checked) and outputs a list of segmentation variation instances between the two corpora[2].

Algorithm steps:

1. Extract all the multi-character words in reference corpus and store their positions in reference corpus respectively;
2. Find the words that be segmented into N parts (N is from 2 to the length of current word) in the corpus to be checked. Store the positions of those segmentations found in the corpus to be checked;
3. Output a list of variation instances with their contexts between two corpora.

We use “AutoCheck” to stand for the processing using the algorithm above. In order to find the segmentation variations within one corpus, we can also make the reference corpus and the corpus to be checked be the same corpus. Data in Table 1 are obtained through “AutoCheck + manual

[2] These two corpora could be also regarded as one unique corpus: the corpus to be checked. A large scale reference corpus is always helpful in spotting more variations in the corpus to be checked.

checking”. That is firstly running “AutoCheck” 3 times as shown in Table 2 to get the list of variation types and instances in each situation respectively, and then EIs are found through manual checking.

In Table 1, situations “within test data”, “Between: One-to-Mult” and “Between: Mult-to-One” correspond to the Situations S2, S3.1 and S3.2 described in Section 2. Here we still include CAS segmentations in order to take a close look at the distribution of EIs in each kind of segmentation variation. We can see that in situation “Between: One-to-Mult”, there are only 4 EIs caused by CAS among 68 EIs. It is a very small fraction, so most of CAS variation instances are correct segmentations in a manually checked corpus.

Situation	Reference corpus	Corpus to be checked
Within test data	PK test data	PK test data
Between: One-to-Mult	PK training data	PK test data
Between: Mult-to-One	PK test data	PK training data

Table 2 Inputs of different AutoCheck runs

Except “黄金　周” (gold week) most of the EIs in S2: “within test data” are also found in S3.1: “Between one-to-mult”[3]. This is because in S3.1 the size of the reference corpus (training set) is much greater than the corpus to be checked (test set) so variations found in this case almost cover all of those found in S2 (test set only). EIs in S3.2: “between: mult-to-one” are such strings that they are never considered as one word in PK training data while always identified as one word in PK test data. For example, the segmentation variation (type) "上图 [shang4-tu2]" occurs four times as one word “上图” (above picture) in test data, but three of its variation instances "上　图" (upper picture) have been found in the training set. Thus, variation type "上图" should be identified as a segmentation error rather than an OOV word as in Bakeoff1. From Table 1, we can find 221 error tokens in all error instances (EIs) after removing the 26 redundant ones in PK test data (17194 tokens). So, the error rate of PK test data is 1.29%.

Using the same method, we also find out the 139 error instances (271 error tokens) in AS test data. The error rate of AS test data is 2.26% as shown in table 3.

Table 3 shows the error rate of AS test set is 2.26% and it is higher than PK test data which is 1.29%. So we believe that the reason why the evaluation result on AS corpus are higher than those on PK corpus of Bakeoff1 is not due to the segmentation quality of AS test data but because of the OOV rate (0.022) in AS test data which is much lower than PK test data (0.069).

data	PK test data	AS test data
Total tokens	17194	11985
Error tokens	221	271
Error rate	1.29%	2.26%

Table 3 Segmentation errors in PK and AS test data

“AutoCheck” outputs a list of all variation instances found in the corpus but it can not judge whether a variation instance is EI or not. Besides, the output of “AutoCheck” doesn’t include those segmentation errors which are not instances of any segmentation variation in a corpus. Two examples are given in Example 4. It

[3] “黄金　周” is considered as a segmentation error according to its variation instance “黄金周” (golden week).

means that “AutoCheck+manual checking”[4] can not spot all segmentation errors in a corpus. Despite of these disadvantages of “AutoCheck”, it is still a necessary assistant to find out almost all of the segmentation errors in an annotated corpus for its effective in finding segmentation error candidates.

Example 4: Segmentation errors which are not instances of any segmentation variation (Bakeoff1 PK corpus):

1）执政官 路易吉·马扎 和 马里诺·扎诺蒂 １６日 上午 举行 仪式

(Archon Marino Zanotti held a ceremony on the morning of 16th)

2）已 发展 成为 世界 上规模 最 大 的 代理 系统

(…has become the largest scale agency system in the world)

AutoCheck has been applied in preparing the MSRA (Microsoft Research Asia) annotated corpora of Chinese word segmentation (MS corpora, hereafter) that were submitted to SIGHAN Bakeoff2 as one of the data sets. "AutoCheck+manual checking" is applied as the principal way of quality control on MS corpora. Even only taking a manual check on those variations output by the AutoCheck could provide an approximate assessment about the quality of the annotated corpus. The lower the number of error instances (EIs) found in the output list the lower the segmentation error rate the annotated corpus reaches. For example, there are 37 variation instances output by AutoCheck in an annotated document #25 with 26K tokens in MS Corpora, in which no EIs has been found manually. Then the whole document was reviewed thoroughly by a person in which only two segmentation errors (shown in Example 5) have been found. Our practice shows that with the quality control method above the segmentation error rate of MS corpora reaches 0.1% in average at the worst cases.

Example 5: Segmentation errors in #25

Error 1: 已 有 三百六十多名 军师 职 领导干部 深入 部队 或 院校 。

(There are more than 360 leaders of corps and division working in grass roots of army and college.)

The string "军师[jun1-shi1]" (military counselor) should be corrected as "军 师" (corps and division).

Error 2: 它 似 一面 六 棱镜 ， 折射 出 时代 的 风貌 ，

(It is like a prism reflecting the style and features of the age)

The string "一面[yi1-mian4] (at the same time) should be corrected as "一 面" (a).

4 The impact to the evaluation caused by segmentation errors in corpora of Bakeoff1

In order to show the impact to the evaluation result caused by EIs existing in test data of Bakeoff1, we conduct the baseline close test with PK and AS corpora, i.e. we compile lexicons only containing words in their training data and then use the lexicons with a forward maximum matching algorithm to segment their test data respectively (Sproat and Emerson, 2003). Original and modified test data are used as gold standard in our baseline test.

In table 4, reference data PK1 and AS1 are the original PK test data and AS test data. Reference data PK2 and AS2 are obtained after correcting all segmentation errors found in their original data (Table 3).

Results in Table 4 are the output of Bakeoff1 evaluation program. Word count is the number of tokens in reference data and the change in word count is caused by our modification. Table 4 shows the impact of EIs in test data to the evaluation results. We

[4] “manual checking” is restricted on the output list only. Therefore it is a very effective way to assess approximately the quality of an annotated corpus.

can see the F measure increase to 0.879 (0.933) from 0.867 (0.915) and the OOV ratio is decrease to 0.065 (0.020) from 0.069 (0.022) when all EIs are corrected in PK (AS) test data.

Reference data	Word count	R	P	F	OOV	R_{OOV}	R_{IV}
PK1	17,194	0.909	0.829	0.867	0.069	0.050	0.972
PK2	17,200	0.920	0.841	0.879	0.065	0.053	0.980
AS1	11,985	0.917	0.912	0.915	0.022	0.000	0.938
AS2	11.886	0.939	0.926	0.933	0.020	0.000	0.958

Table 4 Baseline test results with original and revised PK and AS test data

5 Conclusion

A semi-automatic method to detect segmentation errors in a manually annotated Chinese corpus is presented in this paper. The main contributions of this research are:

- Offer an effective way to spot the segmentation errors in a manually annotated corpus and give the segmentation error rate of the corpus.
- Point out that segmentation inconsistency is not an appropriate technique term to assess the segmentation quality of an annotated corpus and define the concept of segmentation variation instead to get the segmentation error rate of a gold standard corpus.
- Show the influence to the evaluation result caused by the segmentation errors in a gold standard corpus. 1.29% error rate of PK test data and 2.26% error rate of AS test data decrease the F-measure of the SIGHAN Bakeoff1 baseline test by 1.36% and 1.93% respectively.

Acknowledgements

We would like to thank the members of the Natural Language Computing Group at Microsoft Research Asia, especially to acknowledge Jianfeng Gao, John Chen, and the two anonymous reviewers for their insightful comments and suggestions.

References

Aitao Chen. 2003. Chinese word segmentation using minimal linguistic knowledge. In Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, July 11-12, 2003, Sapporo, Japan.

Andi Wu. 2003. Chinese word segmentation in MSR-NLP. In Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, July 11-12, 2003, Sapporo, Japan.

Jianfeng Gao, Mu Li and Chang-Ning Huang. 2003. Improved source-channel models for Chinese word segmentation. In Proceedings of ACL-2003. July 7-12, 2003. Sapporo, Japan.

Markus Dickinson, W. Detmar Meurers. 2003. Detecting errors in Part-of-Speech annotation. In Proceedings of the 11th Conference of the European Chapter of the Association for

Computational Linguistics (EACL-03), 2003, Budapest, Hungary

Richard Sproat, Thomas Emerson. 2003. The first international Chinese word Segmentation Bakeoff. In Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, July 11-12, 2003, Sapporo, Japan.

Sun Maosong. 1999. On the consistency of word-segmented Chinese corpus. (In Chinese) Applied Linguistics, (2):88-91, 1999.

Appendix: Modified EIs in PK test data

1) EI found in CAS variations:

Original: 李 长春 、 罗 干 等同 首都 各界 人士

Modified: 李 长春 、 罗 干 等 同 首都 各界 人士

2) Some EIs in S3.2 which are considered as new words in Bakeoff1:

如果说 交通部门 问明 大世界 看着 全城 区内 换上 庆春节 没想到 踏上 文化村 老同志 迎春花市 守岛 喜迁 全过程 全天 双眼 双腿 族人 上图 共商 走入 领到 代为 一方 年味 岁岁 率兵 送往 前南 节节高 设在 驶过 工商部门 跳进 较大 翻看 惊呆 院团 外资金 下图 冰上 回老家 见到 县市区 极端分子 带回 这栋 办厂 小包 留在

3) Some EIs in S3.1 in which their variation types should be lexicon words

Original	Modified
科教 兴 国	科教兴国
火树银花 不 夜 天	火树银花不夜天
大 酒店	大酒店
集团 公司	集团公司
留学 人员	留学人员
尽管 如此	尽管如此
冰 清 玉 洁	冰清玉洁
羽毛 球队	羽毛球队
高等 学校	高等学校
直 指	直指
吐 翠	吐翠
报 春	报春
贺 岁	贺岁