The CIPS-SIGHAN CLP 2012 Chinese Word Segmentation on MicroBlog Corpora Bakeoff

Huiming Duan **Zhifang Sui** **Ye Tian** **Wenjie Li**
Key Laboratory of Computational Linguistics (Peking University)
Ministry of Education, CHINA
{duenhm, szf, ytian, lwj}@pku.edu.cn

Abstract

The CIPS-SIGHAN CLP 2012 Chinese Word Segmentation on MicroBlog Corpora Bakeoff was held in the autumn of 2012. This bake-off task of Chinese word segmentation is focused on the performance of Chinese word segmentation algorithms on MicroBlog corpora. 17 groups submitted 20 results, among which the best system has all the P, R and F values near 95%, and the average values of the 17 systems are 0.8931, 0.8981 and 0.8953, respectively.

1 Preface

After years of intensive researches, Chinese word segmentation has achieved a quite high precision. Five prior word segmentation bakeoffs, have been successfully conducted in 2003 (Sproat and Emerson, 2003), 2005 (Emerson, 2005), 2006 (Levow, 2006), 2007 (Jin and Chen, 2007) and 2012 (Zhao and Liu, 2010). These evaluations have established benchmarks for word segmentation with which researchers could evaluate their segmentation system.

However, the performance of segmentation is not so satisfying for the MicroBlog corpora. The corpus of a specific domain may have its characteristics in vocabulary, sentence pattern and style. MicroBlog makes no exception. The MicroBlog texts are much similar to oral expression, with a casual style and less deliberation in writing, resulting in a simple and comfortable style: the MicroBlog style. Like other domains, the vocabulary used in MicroBlog texts includes special "terms" and symbols, with which the authors may attract the reader's attention using simple and witty expressions. The MicroBlog style also indicates usage of words inconsistent with normative language, including homophonic word, character variants, word consisting of letters and misusage of punctuation.

In consideration of the characteristics described above, a successful word segmentation system on the MicroBlog corpora should take into consideration the special linguistic phenomena of the MicroBlog corpora and develop corresponding strategies, in addition to the techniques used for general-purpose word segmentation. This CIPS-SIGHAN-2012 bake-off task of Chinese word segmentation will focus on the performance of Chinese word segmentation algorithms on MicroBlog corpora.

2 Task Descriptions

This evaluation involves the following task: opened evaluation on simplified Chinese word segmentation task. This task provides no training set, and participants are free to use data learned or model trained from any resources.

Only a tiny amount of segmented data is given as a format reference of the segmentation systems, which consists of original data and segmented data. The standard of segmentation is in accord with the *Specification for Corpus Processing at Peking University*[1].

Most of the corpus used in this evaluation is selected from the randomly-collected large-scale MicroBlog corpora. Moreover, we manually added the MicroBlog corpus after new events to the corpora, in order to carry out new experiments of evaluation methods. The final corpora consist of 5000 sentences (or articles, strictly. For simplicity, we refer to the individual article as a sentence, since most of the MicroBlog articles consist of only one sentence.)

For evaluation, we adopt the evaluation method used in previous bake-off tasks, and use precision, recall and F-measure to measure the over-

[1] http://www.icl.pku.edu.cn/icl_groups/corpus/coprus-annotation.htm

Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 35–40,
Tianjin, China, 20-21 DEC. 2012

all performance of a system. Metrics used in this bake-off task are:

$$\text{Precision} = \frac{\text{Number of words correctly segmented}}{\text{Number of words segemented}} \times 100\%$$

$$\text{Recall} = \frac{\text{Number of words correctly segmented}}{\text{Number of words in the reference}} \times 100\%$$

$$\text{F measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

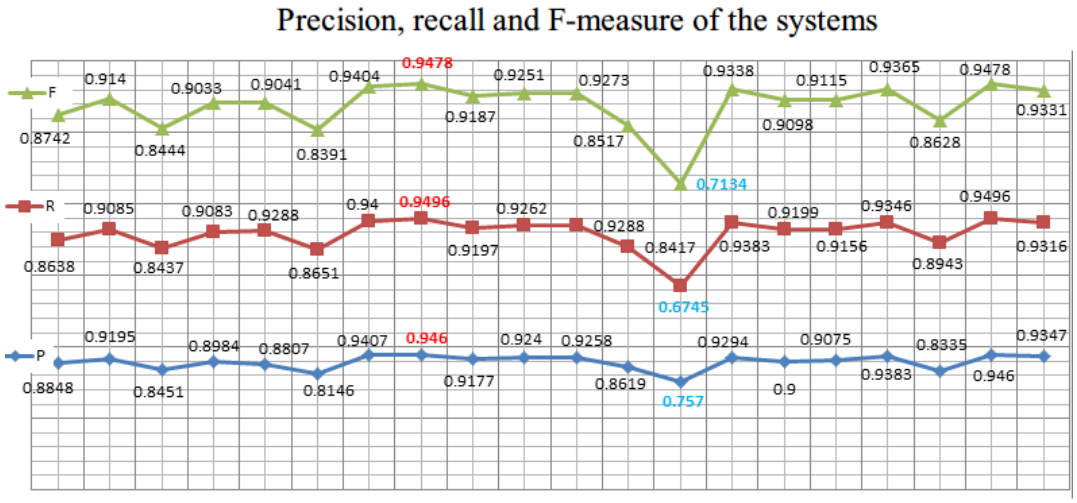

Figure 1 Precision, recall and F measure of the systems

3 Performance of the Contestants

Table 1 lists all the 17 groups of the bake-off task.

Site name	Contact
NLP group at the University of Macau	Longyue Wang (2 systems submitted)
Beijing Institute of Technology	Haizhao Lei
Beijing Information Science & Technology University	Chuan Xu
Beijing University of Posts and Telecommunications	Caixia Yuan
Dalian University of Technology	Jing Zhang
Fudan University	Xipeng Qiu
Individual	Kaixu Zhang
Harbin Institute of Technology	Yijia Liu
Harbin Institute of Technology at Weihai	Xiao Yang
Hefei University of Technology	Xiao Sun
Heilongjiang University	Heyu
Nanjing University	Bin Li (3 systems submitted)
Soochow University	Richen Xu
Zhengzhou University	Hongying Zan
Institute of Software, Chinese Academy of Sciences	Le Sun
Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences	Dan Tian
Institute of Automation, Chinese Academy of Sciences	Saike He

Table 1 List of contestants.

The maximal, minimal and average performances are listed as follows.

	Precision	Recall	F-measure	Number of correct sentences	Percentage of correct sentences
Max	0.946	0.9496	0.9478	2244	44.88%
Min	0.757	0.6745	0.7134	186	3.72%
Ave	0.8931	0.8981	0.8953	1370	27.396%

Table 2 Overall performance of the systems.

4 Results and Analysis

In addition to the traditional evaluation measures (precision, recall and F-measure), we added additional analyses and tests to gain a comprehensive view of the systems.

4.1 Performance of sentence segmentation

As indicated in Figure 2, the performances of sentence-level segmentation (the percentage of the correctly-segmented sentence) are uniformly lower than 50%, despite the fact that the precision, recall and F-measure of word-level segmentation of the systems reach 0.95. (Note: the arrangement of Figure 2 is different with figures above.)

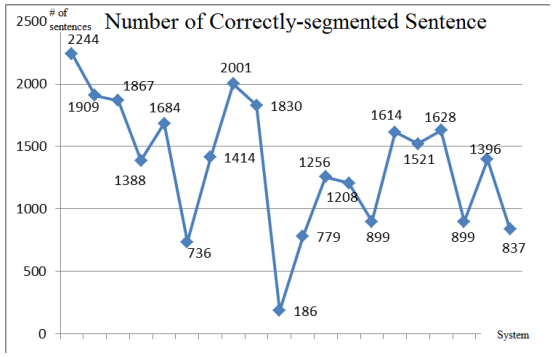

Figure 2 Number of correctly-segmented sentences of the systems.

Automatic word segmentation is known as the first step towards Chinese natural language processing. However, satisfactory results have not been yielded as far as the performance of sentence-level segmentation is concerned. Through investigating a series of test points, we can make further analysis and evaluation of the systems.

4.2 Test point evaluations

Test points are set to measure the relative strengths and weaknesses of the systems and to provide reference for further evaluation and improvement of segmentation systems, even if the test point evaluations are not fully convincing.

4.2.1 Settings

We chose 10 test points for this bake-off task: general term, MicroBlog term, symbols and emoticons, new word (unregistered word), location name, person name, proper name, combination ambiguity, overlapping ambiguity and rule-based combination of words.

test point 0	general term	坐班、做梦、不幸
test point 1	MicroBlog term	肿摸办、 咋米、肿么、 娘的、介个、下五
test point 2	symbols and emoticons	>_<、~~~~(>_<)~~~~
test point 3	new word	足管、住总、刑辩[abbrev]、叽里咕噜、官二代
test point 4	location name	迦错拉、渣滓洞、南市区
test point 5	person name	菅直人、仲井真弘多、郎咸平
test point 6	proper name	正大、粤来粤好、壳牌
test point 7	combination ambiguity	在外、再见、接下来
test point 8	overlapping ambiguity	真经典、在职场上、在手机上面
test point 9	rule-based combination of words	一串串、迷迷糊糊[duplication]、可信度[prefix]、暧昧感、装饰品[suffix]、昨儿[Erhua]

Table 3 Settings of test points

It remains dubious whether such classification is comprehensive, and various opinions exist towards the specific classification of each individual word. We leave such issues to further discussion.

4.2.2 Distribution of the test points

Some of the sentences in our evaluation corpus are easier, containing no test points. This evaluation contains 2147 test points in total, which are distributed in 1639 sentences, composing 32.78% of the sentences. Several sentences contain multiple test points.

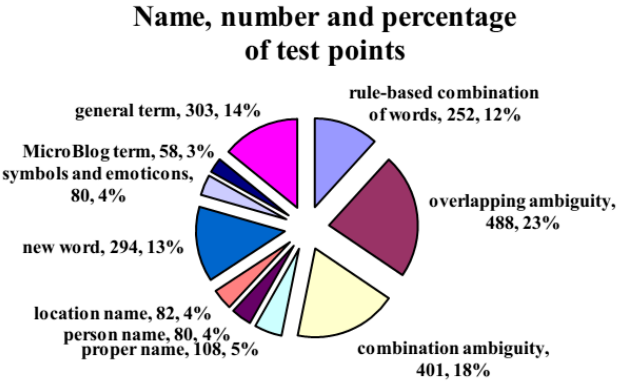

Figure 3 Distribution of test points

In a further merge, we combine combination ambiguity and overlapping ambiguity as ambiguity, combine location name, person name and other proper names as proper names, and combine MicroBlog term, symbols and emoticons as MicroBlog. The distribution of merged test points is illustrated in Figure 4.

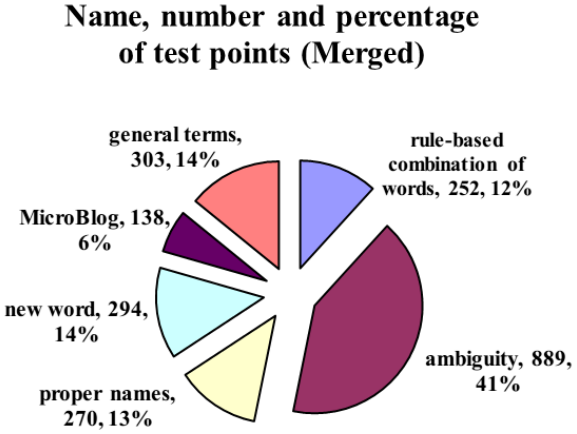

Figure 4 Distribution of merged test points

4.2.3 Evaluation results of test points

Figure 5 demonstrates the respective total number of the 10 test points and the comparison of the maximal segmentation performance of the system in these test points.

Figure 6 shows the percentage of correctly-segmented sentence and percentage of correct test points for each system.

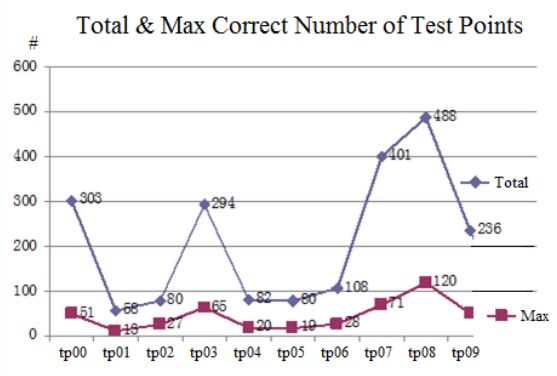

Figure 5 Comparison of the performance of the 10 test points

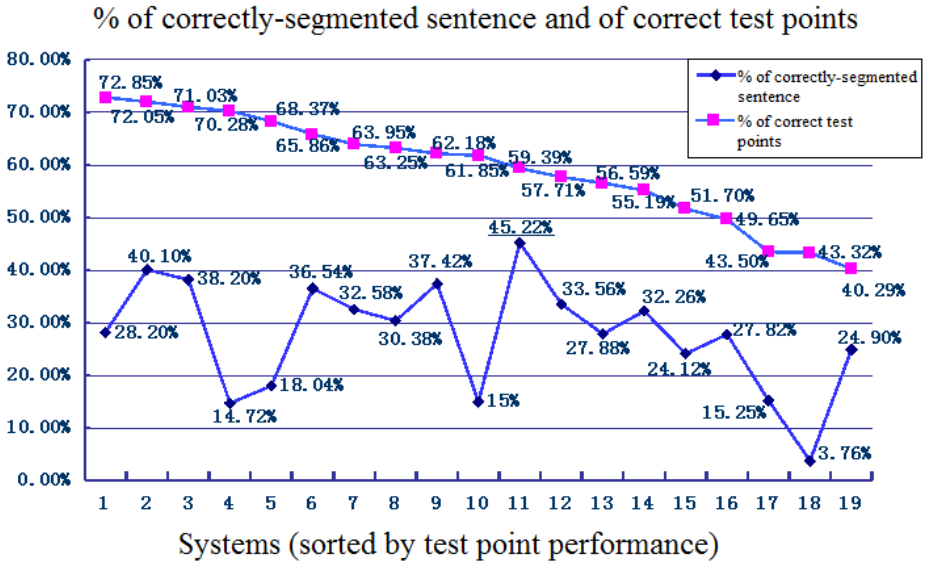

Figure 6 Number of correctly-segmented sentence and correct test points

It is shown in Figure 6 that the best system reaches a 73% precision in the test points, which proves that the bottleneck is almost broken through with more deliberation. We could also make further analysis and find out the weakness of each system. Figure 6 also shows that for systems that have a better performance in test points, they generally yield a low sentence-level performance. By making further development for the sentence-level tasks, such systems may further improve their overall performance.

4.2.4 Analysis

The final results of the systems generally outperform our expectations. However, space of improvement still exists in the critical issues, including ambiguity and proper names (refer to Figure 5).

For the sentences which contain neither ambiguity nor special terms, an optimal system may likely yield a satisfying result, but Figure 6 indicates that some systems perform quite well in sentence-level segmentation, but fail to handle with the test points well. Several possible explanations are as follows:

- Such systems may not deal with ambiguity, proper names and unregistered words well.
- Some systems tend to combine single characters to form complement structures or objective structures using its inbuilt "word formation" strategy.
- Contestants fail to combine some cases (e.g. year-month-date and family name-given name) due to misinterpretation of the task specification.

For the systems that perform well in test points, such issues have been paid more attention and are dealt with well.

Some issues are still under debate, including the definition of word, rules of word formation,

towards which there exists no uniform standard. It is not necessary to demand a uniform standard, but without which the evaluations are impossible to realize.

5 Suggestions

5.1 Further considerations in segmentation evaluation

Word segmentation, though a seemingly simple task, has been making no substantial progress despite the continuous research in recent years. As far as ambiguity is concerned, it involves lexical semantics, word formation and the size of vocabulary. Researchers have made enough efforts in expanding the scale of vocabularies, but the inner structure of words still requires further consideration in scale and depth. Words of ambiguity are prevailing and ubiquitous rather than a closed set. For example:

"总会" is a noun when treated as one word, but "adverb + auxiliary verb" when treated as two words.

Example: 游戏里每个人总会分到一些钱

Translation: Every one of the game always gets some money.

"看中" is a verb and is pronounced *kan4 zhong4* when treated as one word, but "verb + localizer" and *kan4 zhong1* when treated separately.

Example: 拿到书了，慢慢看中……

Translation: I have got the book and have been reading slowly…love creatures, love life.

Example: 我看中一只包就问服务员多少

Translation: I fancied a bag and asked the salesman how much it was.

"着手" is a verb and is pronounced *zhuo2 shou3* when treated as one word, but "particle + noun" and *zhe5 shou3* when treated separately.

Example: 从小处着手，大处着眼

Translation: Start small, and see the big picture.

Example: 看着手都抽筋啊

Translation: Even looking at it makes my hand cramp.

Example: 所有的同学都拿着手机在埋头苦忙 (overlapping ambiguity)

Translation: All the students are holding their cell phones and burying themselves, busied.

Furthermore, after a close investigation of the segmentation results, we found that for the systems trained by statistical data, rule-based post-processing is basically employed to increase recall and avert errors. Each of the systems has further space for improvement, which is easy to achieve as long as the researchers refine their systems.

5.2 Suggestions for future evaluations

Due to various factors and complication of the evaluations, we could only ensure relative fairness for each of the evaluation results. We expected the participants to conform to the segmentation standard proposed by Peking University, but we observed from the final results that some systems failed to take it into consideration, which resulted in unnecessary errors.

Is there any fairer method to evaluate the segmentation systems?

Is it possible to adopt a standardized core vocabulary?

From the technical specifications returned by the participants, we could we that the scale of vocabularies and the scope of domains vary from system to system, which influenced the evaluation results and may yield to difficulties in further analysis.

To make the evaluation results comparable, we should use a uniform standard to make evaluation (though standard of segmentation is specified for this bake-off task, it is possible that systems are not adjusted accordingly due to time limitations or just ignorance of the standard).

Above are our preliminary views towards this evaluation task. We wish to listen to the participants for their viewpoints and make the evaluation task play its due role.

Acknowledgements

This work is supported by NSFC Project 61075067 and National Key Technology R&D Program (No: 2011BAH10B04-03).

Reference

1. Richard Sproat and Thomas Emerson. 2003.The first international Chinese word segmentation bakeoff.In *The Second SIGHAN Workshop on Chinese Language Processing*, pages 133–143, Sapporo, Japan.
2. Thomas Emerson. 2005. The second international Chinese word segmentation bakeoff.In*Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 123–133, Jeju Island, Korea.
3. Gina-Anne Levow. 2006. The third international Chinese language processing bakeoff: Word segmentation and named entity recognition.In*Proceedings of the Fifth SIGHAN Work-*

shop on Chinese Language Processing, pages 108–117, Sydney, Australia, July. Association for Computational Linguistics.

4. Guangjin Jin and Xiao Chen. 2007. The Fourth International Chinese Language Processing Bakeoff: Chinese Word Segmentation,Named Entity Recognition and Chinese POS Tagging.In*The Sixth SIGHAN Workshop on Chinese Language Processing*, pages 69-81, Hyderabad,
5. Hongmei ZhaoandQun Liu, The CIPS-SIGHAN CLP 2010 Chinese Word Segmentation Bakeoff, *The first CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP2010)*, August 28-29, Beijing, China