# The Second International Chinese Word Segmentation Bakeoff

**Thomas Emerson**
Basis Technology Corp.
150 CambridgePark Drive
Cambridge, MA 02140
tree@basistech.com

## Abstract

The second international Chinese word segmentation bakeoff was held in the summer of 2005 to evaluate the current state of the art in word segmentation. Twenty three groups submitted 130 result sets over two tracks and four different corpora. We found that the technology has improved over the intervening two years, though the out-of-vocabulary problem is still or paramount importance.

## 1 Introduction

Chinese is written without inter-word spaces, so finding word-boundaries is an essential first step in many natural language processing applications including mono- and cross-lingual information retrieval and text-to-speech systems. This word segmentation problem has been active area of research in computational linguistics for almost two decades and is a topic of active research around the world. As the very notion of "word-hood" in Chinese is hotly debated, so the determination of the correct division of a Chinese sentence into "words" can be very complex.

In 2003 SIGHAN, the Special Interest Group for Chinese Language Processing of the Association for Computational Linguistics (ACL) conducted the first International Chinese Word Segmentation Bakeoff (Sproat and Emerson, 2003). That competition was the first conducted outside of China and has become the benchmark with which researchers evaluate their segmentation systems. During the winter of 2004 it was decided to hold a second evaluation to determine how the latest research has affected segmentation technology.

## 2 Details of the Contest

### 2.1 The Corpora

Four corpora were used in the evaluation, two each using Simplified and Traditional Chinese characters.[1] The Simplified Chinese corpora were provided by Beijing University and Microsoft Research Beijing. The Traditional Chinese corpora were provided by Academia Sinica in Taiwan and the City University of Hong Kong. Each provider supplied separate training and truth data sets. Details on each corpus are provided in Table 1.

With one exception, all of the corpora were provided in a single character encoding. We decided to provide all of the data in both Unicode (UTF-8 encoding) and the standard encoding used in each locale. This would allow systems that use one or the other encoding to chose appropriately while ensuring consistent transcoding across all sites. This conversion was problematic in two cases:

1. The Academia Sinica corpus, provided in Unicode (UTF-16), contained characters found in Big Five Plus that are not found in Microsoft's CP950 or standard Big Five. It also contained compatibility characters that led to transcoding errors when converting from Unicode to Big Five Plus. A detailed description of these issues can be found on the Bakeoff 2005

[1] A fifth (Simplified Chinese) corpus was provided by the University of Pennsylvania, but for numerous technical reasons it was not used in the evaluation. However, it has been made available (both training and truth data) on the SIGHAN website along with the other corpora.

| Corpus | Abbrev. | Encodings | Training Size (Words/Types) | Test Size (Words/Types) |
|---|---|---|---|---|
| Academia Sinica (Taipei) | AS | Big Five Plus, Unicode | 5.45M / 141K | 122K / 19K |
| Beijing University | PK | CP936, Unicode | 1.1M / 55K | 104K / 13K |
| City University of Hong Kong | CityU | Big Five/HKSCS, Unicode | 1.46M / 69K | 41K / 9K |
| Microsoft Research (Beijing) | MSR | CP936, Unicode | 2.37M / 88K | 107K / 13K |

Table 1. Corpus Information

pages on the SIGHAN website. The data also included 11 instances of an invalid character that could not be converted to Big Five Plus.

2. The City University of Hong Kong data was initially supplied in Big Five/ HKSCS. We initially converted this to Unicode but found that there were characters appearing in Unicode Ideograph Extension B, which many systems are unable to handle. City University was gracious enough to provide Unicode versions for their files with all characters in the Unicode BMP. Specific details can be found on the Bakeoff 2005 pages of the SIGHAN website.

The truth data was provided in segmented and unsegmented form by all of the providers except Academia Sinica, who only provided the segmented truth files. These were converted to unsegmented form using a simple Perl script. Unfortunately this script also removed spaces separating non-Chinese (i.e., English) tokens. We had no expectation of correct segmentation on non-Chinese text, so the spaces were manually removed between non-Chinese text in the truth data prior to scoring.

The Academia Sinica data separated tokens in both the training and truth data using a full-width space instead of one or more half-width (i.e., ASCII) spaces. The scoring script was modified to ignore the type of space used so that teams would not be penalized during scoring for using a different separator.

The segmentation standard used by each provider were made available to the participants, though late in the training period. These standards are either extremely terse (MSR), verbose but in Chinese only (PKU, AS), or are verbose and moderately bilingual. The PKU corpus uses a standard derived from GB 13715, the Chinese government standard for text segmentation in computer applications. Similarly AS uses a Taiwanese national standard for segmentation in computer applications. The CityU data was segmented using the LIVAC corpus standard, and the MSR data to Microsoft's internal standard. The standards are available on the bakeoff web site.

The PKU data was edited by the organizers to remove a numeric identifier from the start of each line. Unless otherwise noted in this paper no changes beyond transcoding were made to the data furnished by contributors.

## 2.2 Rules and Procedures

The bakeoff was run almost identically to the first described in Sproat and Emerson (2003): the detailed instructions provided to the participants are available on the bakeoff website at `http://www.sighan.org/bakeoff2005/` . Groups (or "sites" as they were also called) interested in participating in the competition registered on the SIGHAN website. Only the primary researcher for each group was asked to register. Registration was opened on June 1,

| ID | Site | Contact | Country | AS | PKU | CityU | MSR |
|---|---|---|---|---|---|---|---|
| 2 | ICL, Beijing University | Wuguang SHI | ZH |  |  |  | ◇ |
| 3 | Xiamen University | Xiaodong SHI | ZH | ◆◇ | ◆◇ | ◆◇ | ◆◇ |
| 4 | ITNLP Lab, Harbin Institute of Technology | Wei JIANG | ZH | ◆◇ | ◆◇ | ◆◇ | ◆◇ |
| 5 | France Telecom R&D Beijing | Heng LI | ZH | ◆◇ | ◆◇ | ◆◇ | ◆◇ |
| 6 | Information Retrieval Lab, Harbin Institute of Technology | Huipeng ZHANG | ZH |  | ◆◇ |  |  |
| 7 | Dept. of Linguistics, The University of Hong Kong | Guohong FU | HK | ◆◇ | ◆◇ | ◆◇ | ◆◇ |
| 8 | Computer Science Dept., Xiamen University | Hua-lin Zeng | ZH |  | ◆◇ |  | ◆◇ |
| 9 | Dept. of Linguistics, The Ohio State University | Xiaofei LU | US |  | ◆ |  |  |
| 12 | Dept. of Computer Science, The University of Sheffield | Yaoyong LI | GB | ◆◇ | ◆◇ | ◆◇ | ◆◇ |
| 13 | Nanjing University | Jiajun CHEN | ZH |  | ◆◇ |  | ◆◇ |
| 14 | Stanford NL Group | Huihsin TSENG | US | ◆ | ◆ | ◆ | ◆ |
| 15 | Nara Institute of Science and Technology | Masayuki ASAHARA | JP | ◆ | ◆ | ◆ | ◆ |
| 16 | Academia Sinica | Yu-Fang TSAI | TW |  | ◇ | ◇ |  |
| 19 | National University of Singapore | Hwee Tou NG | SG | ◇ | ◇ | ◇ | ◇ |
| 21 | Kookmin University | Seung-Shik KANG | KO |  | ◆ | ◆ | ◆ |
| 23 | US Dept. of Defense | Thomas Keenan | US |  | ◇ |  | ◇ |
| 24 | Dept. of Information Management, Tung Nan Institute of Technology | Jia-Lin TSAI | TW |  |  |  | ◆ |
| 26 | ICL, Peking University | Huiming DUAN | ZH |  |  |  | ◆◇ |
| 27 | Yahoo! Inc. | Aitao CHEN | US | ◆◇ | ◆◇ | ◆◇ | ◆◇ |
| 29 | The Chinese University of Hong Kong | Tak Pang LAU | HK |  | ◆ | ◆ | ◆ |
| 31 | City University of Hong Kong | Ka Po CHOW | HK | ◇ | ◇ |  |  |
| 33 | City University of Hong Kong | Chun Yu KIT | HK | ◆ | ◆ |  | ◆ |
| 34 | Institute of Computing Technology, Chinese Academy of Sciences | ShuangLong LI | ZH |  | ◆◇ |  | ◆◇ |

Table 2. Participating Groups (◆ = closed test, ◇ = open test)

2005 and allowed to continue through the time the training data was made available on July 11. When a site registered they selected which corpus or corpora there were interested in using, and whether they would take part in the open or closed tracks (described below.) On July 11 the training data was made available on the Bakeoff website for downloading: the same data was used regardless of the tracks the sites registered for. The web site did not allow a participant to

| Corpus | Word Count | R | P | F | OOV | Roov | Riv |
|---|---|---|---|---|---|---|---|
| AS | 122,610 | 0.909 | 0.857 | 0.882 | 0.043 | 0.004 | 0.950 |
| CityU | 40936 | 0.882 | 0.790 | 0.833 | 0.074 | 0.000 | 0.952 |
| MSR | 106,873 | 0.955 | 0.912 | 0.933 | 0.026 | 0.000 | 0.981 |
| PKU | 104,372 | 0.904 | 0.836 | 0.869 | 0.058 | 0.059 | 0.956 |

Table 3: Baseline scores generated via maximal matching using only words from the training data

| Corpus | Word Count | R | P | F | OOV | Roov | Riv |
|---|---|---|---|---|---|---|---|
| AS | 122,610 | 0.979 | 0.985 | 0.982 | 0.043 | 0.996 | 0.978 |
| CityU | 40,936 | 0.988 | 0.991 | 0.989 | 0.074 | 0.997 | 0.988 |
| MSR | 106,873 | 0.991 | 0.992 | 0.991 | 0.026 | 0.998 | 0.990 |
| PKU | 104,372 | 0.985 | 0.988 | 0.987 | 0.058 | 0.994 | 0.985 |

Table 4: Topline scores generated via maximal matching using only words from the testing data

add a corpus to the set they initially selected, though at least one asked us via email to add one and this was done manually. Groups were given until July 27 to train their systems, when the testing data was released on the web site. They then had two days to process the test corpora and return them to the organizer via email on Jul 29 for scoring. Each participant's results were posted to their section of the web site on August 6, and the summary results for all participants were made available to all groups on August 12.

Two tracks were available for each corpus, open and closed:

- In the open tests participants could use any external data in addition to the training corpus to train their system. This included, but was not limited to, external lexica, character set knowledge, part-of-speech information, etc. Sites participating in an open test were required to describe this external data in their system description.

- In closed tests, participants were only allowed to use information found in the training data. Absolutely no other data or information could be used beyond that in the training document. This included knowledge of character sets, punctuation characters, etc. These seemingly artificial restrictions (when compared to "real world" systems) were formulated to study exactly how far one can get without supplemental information.

Other obvious restrictions applied: groups could not participate using corpora that they or their organization provided or that they had used before or otherwise seen.

Sites were allowed submit multiple runs within a track, allowing them to compare various approaches.

Scoring was done automatically using a combination of Perl and shell scripts. Participants were asked to submit their data using very strict naming conventions to facilitate this: in only a couple of instances were these not followed and human intervention was required. After the scoring was done the script would mail the detailed results to the participant. The scripts used for scoring can be downloaded from the

| Participant | Run ID | Word Count | R | $C_r$ | P | $C_p$ | F | OOV | $R_{oov}$ | $R_{iv}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 15 | b | 122610 | 0.952 | ±0.00122 | 0.951 | ±0.00123 | 0.952 | 0.043 | 0.696 | 0.963 |
| 15 | a | 122610 | 0.955 | ±0.00118 | 0.939 | ±0.00137 | 0.947 | 0.043 | 0.606 | 0.971 |
| 14 |  | 122610 | 0.95 | ±0.00124 | 0.943 | ±0.00132 | 0.947 | 0.043 | 0.718 | 0.960 |
| 27 |  | 122610 | 0.955 | ±0.00118 | 0.934 | ±0.00142 | 0.945 | 0.043 | 0.468 | 0.978 |
| 12 |  | 122610 | 0.946 | ±0.00129 | 0.942 | ±0.00134 | 0.944 | 0.043 | 0.648 | 0.959 |
| 7 |  | 122610 | 0.947 | ±0.00128 | 0.934 | ±0.00142 | 0.94 | 0.043 | 0.523 | 0.966 |
| 15 | c | 122610 | 0.944 | ±0.00131 | 0.934 | ±0.00142 | 0.939 | 0.043 | 0.445 | 0.967 |
| 33 |  | 122610 | 0.944 | ±0.00131 | 0.902 | ±0.00170 | 0.923 | 0.043 | 0.234 | 0.976 |
| 5 |  | 122610 | 0.948 | ±0.00127 | 0.900 | ±0.00171 | 0.923 | 0.043 | 0.158 | 0.983 |
| 4 |  | 122610 | 0.943 | ±0.00132 | 0.895 | ±0.00175 | 0.918 | 0.043 | 0.137 | 0.979 |
| 3 |  | 122610 | 0.877 | ±0.00188 | 0.796 | ±0.00230 | *0.835* | 0.043 | 0.128 | 0.911 |

Table 5. Academia Sinica — Closed (italics indicate performance below baseline)

| Participant | Run ID | Word Count | R | $C_r$ | P | $C_p$ | F | OOV | $R_{oov}$ | $R_{iv}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 19 |  | 122610 | 0.962 | ±0.00109 | 0.95 | ±0.00124 | 0.956 | 0.043 | 0.684 | 0.975 |
| 27 |  | 122610 | 0.958 | ±0.00115 | 0.938 | ±0.00138 | 0.948 | 0.043 | 0.506 | 0.978 |
| 12 |  | 122610 | 0.949 | ±0.00126 | 0.947 | ±0.00128 | 0.948 | 0.043 | 0.686 | 0.961 |
| 7 |  | 122610 | 0.955 | ±0.00118 | 0.938 | ±0.00138 | 0.946 | 0.043 | 0.579 | 0.972 |
| 31 |  | 122610 | 0.943 | ±0.00132 | 0.931 | ±0.00145 | 0.937 | 0.043 | 0.531 | 0.962 |
| 4 |  | 122610 | 0.952 | ±0.00122 | 0.92 | ±0.00155 | 0.936 | 0.043 | 0.354 | 0.979 |
| 5 |  | 122610 | 0.952 | ±0.00122 | 0.919 | ±0.00156 | 0.935 | 0.043 | 0.311 | 0.981 |
| 3 |  | 122610 | 0.004 | ±0.00036 | 0.004 | ±0.00036 | *0.004* | 0.043 | 0.085 | 0 |

Table 6. Academia Sinica — Open (italics indicate performance below baseline)

Bakeoff 2005 web site. It was provided to the participants to aid in the their data analysis. As noted above, some of the training/truth data used a full-width space to separate tokens: the scoring script was modified to ignore the differences between full-width and half-width spaces. This is the *only* case where the half-width/full-width distinction was ignored: a system that converted tokens from full-width to half-width was penalized by the script.

### 2.3 Participating Sites

Thirty-six sites representing 10 countries initially signed up for the bakeoff. The People's Republic of China had the greatest number with 17, followed by the United States (6), Hong Kong (5), Taiwan (3), six others with one each. Of these, 23 submitted results for scoring and subsequently submitted a paper for these proceedings. A summary of participating groups and the tracks for which they submitted results can be found in Table 2 on the preceding page. All together 130 runs were submitted for scoring.

## 3 Results

In order to provide hypothetical best and worst case results (i.e., we expect systems to do no worse than the base-line and to generally underperform the top-line), we used a simple left-to-right maximal matching algorithm implemented in Perl to generate “top-line” and “base-line”

| Participant | Run ID | Word Count | R | Cr | P | Cp | F | OOV | Roov | Riv |
|---|---|---|---|---|---|---|---|---|---|---|
| 14 |  | 40936 | 0.941 | ±0.00233 | 0.946 | ±0.00223 | 0.943 | 0.074 | 0.698 | 0.961 |
| 15 | a | 40936 | 0.942 | ±0.00231 | 0.941 | ±0.00233 | 0.942 | 0.074 | 0.629 | 0.967 |
| 15 | b | 40936 | 0.937 | ±0.00240 | 0.946 | ±0.00223 | 0.941 | 0.074 | 0.736 | 0.953 |
| 27 |  | 40936 | 0.949 | ±0.00217 | 0.931 | ±0.00251 | 0.94 | 0.074 | 0.561 | 0.98 |
| 7 |  | 40936 | 0.944 | ±0.00227 | 0.933 | ±0.00247 | 0.939 | 0.074 | 0.626 | 0.969 |
| 12 |  | 40936 | 0.931 | ±0.00251 | 0.941 | ±0.00233 | 0.936 | 0.074 | 0.657 | 0.953 |
| 29 | d | 40936 | 0.937 | ±0.00240 | 0.922 | ±0.00265 | 0.929 | 0.074 | 0.698 | 0.956 |
| 15 | c | 40936 | 0.915 | ±0.00276 | 0.94 | ±0.00235 | 0.928 | 0.074 | 0.598 | 0.94 |
| 29 | a | 40936 | 0.938 | ±0.00238 | 0.915 | ±0.00276 | 0.927 | 0.074 | 0.658 | 0.961 |
| 29 | b | 40936 | 0.936 | ±0.00242 | 0.913 | ±0.00279 | 0.925 | 0.074 | 0.656 | 0.959 |
| 21 |  | 40936 | 0.917 | ±0.00273 | 0.925 | ±0.00260 | 0.921 | 0.074 | 0.539 | 0.948 |
| 29 | c | 40936 | 0.925 | ±0.00260 | 0.896 | ±0.00302 | 0.91 | 0.074 | 0.639 | 0.948 |
| 4 |  | 40936 | 0.934 | ±0.00245 | 0.865 | ±0.00338 | 0.898 | 0.074 | 0.248 | 0.989 |
| 5 |  | 40936 | 0.932 | ±0.00249 | 0.862 | ±0.00341 | 0.895 | 0.074 | 0.215 | 0.989 |
| 3 |  | 40936 | 0.814 | ±0.00385 | 0.711 | ±0.00448 | *0.759* | 0.074 | 0.227 | 0.86 |

Table 7: City University of Hong Kong — Closed (italics indicate performance below baseline)

| Participant | Run ID | Word Count | R | Cr | P | Cp | F | OOV | Roov | Riv |
|---|---|---|---|---|---|---|---|---|---|---|
| 19 |  | 40936 | 0.967 | ±0.00177 | 0.956 | ±0.00203 | 0.962 | 0.074 | 0.806 | 0.98 |
| 16 |  | 40936 | 0.958 | ±0.00198 | 0.95 | ±0.00215 | 0.954 | 0.074 | 0.775 | 0.973 |
| 27 |  | 40936 | 0.952 | ±0.00211 | 0.937 | ±0.00240 | 0.945 | 0.074 | 0.608 | 0.98 |
| 7 |  | 40936 | 0.944 | ±0.00227 | 0.938 | ±0.00238 | 0.941 | 0.074 | 0.667 | 0.966 |
| 12 |  | 40936 | 0.933 | ±0.00247 | 0.94 | ±0.00235 | 0.936 | 0.074 | 0.653 | 0.955 |
| 4 |  | 40936 | 0.946 | ±0.00223 | 0.898 | ±0.00299 | 0.922 | 0.074 | 0.417 | 0.989 |
| 5 |  | 40936 | 0.94 | ±0.00235 | 0.901 | ±0.00295 | 0.92 | 0.074 | 0.41 | 0.982 |
| 3 |  | 40936 | 0.014 | ±0.00116 | 0.013 | ±0.00112 | *0.013* | 0.074 | 0.029 | 0.012 |

Table 8: City University of Hong Kong — Open (italics indicate performance below baseline)

numbers. This was done by generating word lists based only on the vocabulary in each truth (top-line) and training (bottom-line) corpus and segmenting the respective test corpora. These results are presented in Tables 3 and 4.

All of the results comprise the following data: test *recall* (R), test *precision* (P), balanced *F score* (where $F = 2PR/(P + R)$), the *out-of-vocabulary* (OOV) rate on the test corpus, the *recall on OOV* words ($R_{oov}$), and the *recall on in-vocabulary words* ($R_{iv}$). We use the usual definition of out-of-vocabulary words as the set of words occurring in the test corpus that are not in the training corpus.

As in the previous evaluation, to test the confidence level that two trials are significantly different from each other we used the Central Limit Theorem for Bernoulli trials (Grinstead and Snell, 1997), assuming that the recall rates from the various trials represents the probability that a word will be successfully identified, and that a binomial distribution is appropriate for the experiment. We calculated these values at the 95% confidence interval with the formula *±2 √(p*

| Participant | Run ID | Word Count | R | Cr | P | Cp | F | OOV | Roov | Riv |
|---|---|---|---|---|---|---|---|---|---|---|
| 14 | | 106873 | 0.962 | ±0.00117 | 0.966 | ±0.00111 | 0.964 | 0.026 | 0.717 | 0.968 |
| 7 | | 106873 | 0.962 | ±0.00117 | 0.962 | ±0.00117 | 0.962 | 0.026 | 0.592 | 0.972 |
| 27 | a | 106873 | 0.969 | ±0.00106 | 0.952 | ±0.00131 | 0.960 | 0.026 | 0.379 | 0.985 |
| 27 | b | 106873 | 0.968 | ±0.00108 | 0.953 | ±0.00129 | 0.960 | 0.026 | 0.381 | 0.984 |
| 4 | | 106873 | 0.973 | ±0.00099 | 0.945 | ±0.00139 | 0.959 | 0.026 | 0.323 | 0.991 |
| 15 | b | 106873 | 0.952 | ±0.00131 | 0.964 | ±0.00114 | 0.958 | 0.026 | 0.718 | 0.958 |
| 5 | | 106873 | 0.974 | ±0.00097 | 0.940 | ±0.00145 | 0.957 | 0.026 | 0.21 | 0.995 |
| 13 | | 106873 | 0.959 | ±0.00121 | 0.956 | ±0.00125 | 0.957 | 0.026 | 0.496 | 0.972 |
| 12 | | 106873 | 0.952 | ±0.00131 | 0.960 | ±0.00120 | 0.956 | 0.026 | 0.673 | 0.96 |
| 24 | 6 | 106873 | 0.958 | ±0.00123 | 0.952 | ±0.00131 | 0.955 | 0.026 | 0.503 | 0.97 |
| 24 | 7 | 106873 | 0.958 | ±0.00123 | 0.952 | ±0.00131 | 0.955 | 0.026 | 0.504 | 0.97 |
| 24 | 4 | 106873 | 0.958 | ±0.00123 | 0.949 | ±0.00135 | 0.954 | 0.026 | 0.465 | 0.972 |
| 24 | 5 | 106873 | 0.958 | ±0.00123 | 0.951 | ±0.00132 | 0.954 | 0.026 | 0.493 | 0.971 |
| 24 | 3 | 106873 | 0.968 | ±0.00108 | 0.938 | ±0.00148 | 0.953 | 0.026 | 0.205 | 0.989 |
| 33 | | 106873 | 0.965 | ±0.00112 | 0.935 | ±0.00151 | 0.950 | 0.026 | 0.189 | 0.986 |
| 15 | a | 106873 | 0.955 | ±0.00127 | 0.942 | ±0.00143 | 0.949 | 0.026 | 0.378 | 0.971 |
| 21 | | 106873 | 0.945 | ±0.00139 | 0.949 | ±0.00135 | 0.947 | 0.026 | 0.576 | 0.955 |
| 24 | 0 | 106873 | 0.956 | ±0.00125 | 0.938 | ±0.00148 | 0.947 | 0.026 | 0.327 | 0.973 |
| 34 | | 106873 | 0.948 | ±0.00136 | 0.942 | ±0.00143 | 0.945 | 0.026 | 0.664 | 0.955 |
| 24 | 2 | 106873 | 0.964 | ±0.00114 | 0.924 | ±0.00162 | 0.944 | 0.026 | 0.025 | 0.989 |
| 15 | c | 106873 | 0.964 | ±0.00114 | 0.923 | ±0.00163 | 0.943 | 0.026 | 0.025 | 0.99 |
| 24 | 1 | 106873 | 0.963 | ±0.00115 | 0.924 | ±0.00162 | 0.943 | 0.026 | 0.025 | 0.989 |
| 29 | a | 106873 | 0.946 | ±0.00138 | 0.933 | ±0.00153 | 0.939 | 0.026 | 0.587 | 0.956 |
| 29 | b | 106873 | 0.941 | ±0.00144 | 0.932 | ±0.00154 | 0.937 | 0.026 | 0.624 | 0.95 |
| 8 | b | 106873 | 0.957 | ±0.00124 | 0.917 | ±0.00169 | 0.936 | 0.026 | 0.025 | 0.982 |
| 8 | c | 106873 | 0.955 | ±0.00127 | 0.915 | ±0.00171 | 0.935 | 0.026 | 0.025 | 0.98 |
| 26 | | 106873 | 0.937 | ±0.00149 | 0.928 | ±0.00158 | *0.932* | 0.026 | 0.457 | 0.95 |
| 3 | | 106873 | 0.908 | ±0.00177 | 0.927 | ±0.00159 | *0.917* | 0.026 | 0.247 | 0.926 |
| 8 | a | 106873 | 0.898 | ±0.00185 | 0.896 | ±0.00187 | *0.897* | 0.026 | 0.327 | 0.914 |

Table 9: Microsoft Research — Closed (italics indicate performance below baseline)

$(1 - p)/n)$ where $n$ is the number of words. This value appears in subsequent tables under the column $c_r$. We also calculate the confidence that the a character string segmented as a word is actually a word by treating $p$ as the precision rates of each system. This is referred to as $c_p$ in the result tables. Two systems are then considered to be statistically different (at a 95% confidence level) if one of their $c_r$ or $c_p$ are different. Tables 5–12 contain the results for each corpus and track (groups are referenced by their ID as found in Table 2) ordered by F score.

| Participant | Run ID | Word Count | R | Cr | P | Cp | F | OOV | Roov | Riv |
|---|---|---|---|---|---|---|---|---|---|---|
| 4 | | 106873 | 0.98 | ±0.00086 | 0.965 | ±0.00112 | 0.972 | 0.026 | 0.59 | 0.99 |
| 19 | | 106873 | 0.969 | ±0.00106 | 0.968 | ±0.00108 | 0.968 | 0.026 | 0.736 | 0.975 |
| 7 | | 106873 | 0.969 | ±0.00106 | 0.966 | ±0.00111 | 0.967 | 0.026 | 0.612 | 0.979 |
| 27 | b | 106873 | 0.971 | ±0.00103 | 0.961 | ±0.00118 | 0.966 | 0.026 | 0.512 | 0.983 |
| 5 | | 106873 | 0.975 | ±0.00096 | 0.957 | ±0.00124 | 0.966 | 0.026 | 0.453 | 0.989 |
| 13 | | 106873 | 0.959 | ±0.00121 | 0.971 | ±0.00103 | 0.965 | 0.026 | 0.785 | 0.964 |
| 27 | a | 106873 | 0.97 | ±0.00104 | 0.957 | ±0.00124 | 0.963 | 0.026 | 0.466 | 0.984 |
| 12 | | 106873 | 0.95 | ±0.00133 | 0.958 | ±0.00123 | 0.954 | 0.026 | 0.648 | 0.958 |
| 26 | | 106873 | 0.925 | ±0.00161 | 0.936 | ±0.00150 | *0.930* | 0.026 | 0.617 | 0.933 |
| 8 | a | 106873 | 0.94 | ±0.00145 | 0.917 | ±0.00169 | *0.928* | 0.026 | 0.239 | 0.959 |
| 34 | | 106873 | 0.916 | ±0.00170 | 0.933 | ±0.00153 | *0.924* | 0.026 | 0.705 | 0.922 |
| 8 | c | 106873 | 0.928 | ±0.00158 | 0.913 | ±0.00172 | *0.920* | 0.026 | 0.355 | 0.944 |
| 8 | b | 106873 | 0.923 | ±0.00163 | 0.914 | ±0.00172 | *0.918* | 0.026 | 0.354 | 0.938 |
| 2 | | 106873 | 0.913 | ±0.00172 | 0.915 | ±0.00171 | *0.914* | 0.026 | 0.725 | 0.918 |
| 3 | | 106873 | 0.921 | ±0.00165 | 0.897 | ±0.00186 | *0.909* | 0.026 | 0.562 | 0.93 |
| 8 | d | 106873 | 0.92 | ±0.00166 | 0.889 | ±0.00192 | *0.904* | 0.026 | 0.332 | 0.936 |
| 8 | e | 106873 | 0.9 | ±0.00184 | 0.861 | ±0.00212 | *0.880* | 0.026 | 0.309 | 0.916 |
| 27 | c | 106873 | 0.865 | ±0.00209 | 0.844 | ±0.00222 | *0.855* | 0.026 | 0.391 | 0.878 |
| 23 | | 106873 | 0.788 | ±0.00250 | 0.818 | ±0.00236 | *0.803* | 0.026 | 0.37 | 0.8 |

Table 10: Microsoft Research — Open (italics indicate performance below baseline)

## 4 Discussion

Across all of the corpora the best performing system, in terms of F score, achieved a 0.972, with an average of 0.918 and median of 0.941. As one would expect the best F score on the open tests was higher than the best on the closed tests, 0.972 vs. 0.964, both on the MSR corpus. This result follows from the fact that systems taking part on the open test can utilize more information than those on the closed. Also interesting to compare are the OOV recall rates between the Open and Closed tracks. The best OOV recall in the open evaluation was 0.872 compared to just 0.813 on the closed track. These data indicate that OOV handling is still the Achilles heel of segmentation systems, even when the OOV rates are relatively small. These OOV recall scores are better than those observed in the first bakeoff in 2003, with similar OOV values, which suggests that advances in unknown word recognition have occurred. Nevertheless OOV is still the most significant problem in segmentation systems.

The best score on any track in the 2003 bakeoff was F=0.961, while the best for this evaluation was F=0.972, followed by 17 other scores above 0.961. This shows a general trend to a decrease in error rates, from 3.9% to 2.8%! These scores are still far below the theoretical 0.99 level reflected in the topline and the higher numbers often reflected in the literature. It is plain that one can construct a test set that any given system will achieve very high measures of precision and recall on, but these numbers must viewed with caution as they may not scale to other applications or other problem sets.

Three participants that used the scoring script in their system evaluation observed different behavior from that of the organizers in the

| Participant | Run ID | Word Count | R | Cr | P | Cp | F | OOV | Roov | Riv |
|---|---|---|---|---|---|---|---|---|---|---|
| 27 | | 104372 | 0.953 | ±0.00131 | 0.946 | ±0.00140 | 0.95 | 0.058 | 0.636 | 0.972 |
| 14 | | 104372 | 0.946 | ±0.00140 | 0.954 | ±0.00130 | 0.95 | 0.058 | 0.787 | 0.956 |
| 6 | a | 104372 | 0.952 | ±0.00132 | 0.945 | ±0.00141 | 0.949 | 0.058 | 0.673 | 0.969 |
| 6 | b | 104372 | 0.952 | ±0.00132 | 0.943 | ±0.00144 | 0.947 | 0.058 | 0.673 | 0.969 |
| 13 | | 104372 | 0.941 | ±0.00146 | 0.95 | ±0.00135 | 0.946 | 0.058 | 0.813 | 0.949 |
| 7 | | 104372 | 0.943 | ±0.00144 | 0.944 | ±0.00142 | 0.944 | 0.058 | 0.656 | 0.961 |
| 15 | b | 104372 | 0.93 | ±0.00158 | 0.951 | ±0.00134 | 0.941 | 0.058 | 0.76 | 0.941 |
| 4 | | 104372 | 0.954 | ±0.00130 | 0.927 | ±0.00161 | 0.941 | 0.058 | 0.518 | 0.981 |
| 34 | | 104372 | 0.938 | ±0.00149 | 0.942 | ±0.00145 | 0.94 | 0.058 | 0.767 | 0.948 |
| 15 | a | 104372 | 0.93 | ±0.00158 | 0.938 | ±0.00149 | 0.934 | 0.058 | 0.521 | 0.955 |
| 5 | | 104372 | 0.95 | ±0.00135 | 0.919 | ±0.00169 | 0.934 | 0.058 | 0.449 | 0.98 |
| 9 | | 104372 | 0.922 | ±0.00166 | 0.934 | ±0.00154 | 0.928 | 0.058 | 0.728 | 0.934 |
| 12 | | 104372 | 0.919 | ±0.00169 | 0.935 | ±0.00153 | 0.927 | 0.058 | 0.593 | 0.939 |
| 15 | c | 104372 | 0.904 | ±0.00182 | 0.93 | ±0.00158 | 0.917 | 0.058 | 0.325 | 0.94 |
| 29 | a | 104372 | 0.926 | ±0.00162 | 0.908 | ±0.00179 | 0.917 | 0.058 | 0.535 | 0.95 |
| 29 | c | 104372 | 0.918 | ±0.00170 | 0.915 | ±0.00173 | 0.917 | 0.058 | 0.621 | 0.936 |
| 33 | | 104372 | 0.929 | ±0.00159 | 0.904 | ±0.00182 | 0.916 | 0.058 | 0.252 | 0.971 |
| 21 | | 104372 | 0.9 | ±0.00186 | 0.925 | ±0.00163 | 0.912 | 0.058 | 0.389 | 0.931 |
| 29 | b | 104372 | 0.917 | ±0.00171 | 0.903 | ±0.00183 | 0.91 | 0.058 | 0.6 | 0.937 |
| 8 | a | 104372 | 0.906 | ±0.00181 | 0.886 | ±0.00197 | 0.896 | 0.058 | 0.29 | 0.943 |
| 8 | c | 104372 | 0.907 | ±0.00180 | 0.843 | ±0.00225 | 0.874 | 0.058 | 0.082 | 0.958 |
| 8 | b | 104372 | 0.906 | ±0.00181 | 0.842 | ±0.00226 | 0.873 | 0.058 | 0.081 | 0.956 |
| 3 | | 104372 | 0.843 | ±0.00225 | 0.737 | ±0.00273 | *0.786* | 0.058 | 0.153 | 0.885 |

Table 11: Peking University — Closed (italics indicate performance below baseline)

generation of the recall numbers, thereby affecting the F score. We were unable to replicate the behavior observed by the participant, nor could we determine a common set of software versions that might lead to the problem. We verified our computed scores on two different operating systems and two different hardware architectures. In each case the difference was in the participants favor (i.e., resulted in an increased F score) though the impact was minimal. If there is an error in the scripts then it affects all data sets identically, so we are confident in the scores as reported here. Nevertheless, we hope that further investigation will uncover the cause of the discrepancy so that it can be rectified in the future.

### 4.1 Future Directions

This second bakeoff was an unqualified success, both in the number of systems represented and in the demonstrable improvement in segmentation technology since 2003. However, there are still open questions that future evaluations can attempt to answer, including: how well a system trained on one genre performs when faced with text from a different register. This will stress OOV handling in the extreme. Consider a situation where a system trained on PRC newswire

| Participant | Run ID | Word Count | R | Cr | P | Cp | F | OOV | Roov | Riv |
|---|---|---|---|---|---|---|---|---|---|---|
| 19 | | 104372 | 0.968 | ±0.00109 | 0.969 | ±0.00107 | 0.969 | 0.058 | 0.838 | 0.976 |
| 4 | | 104372 | 0.968 | ±0.00109 | 0.966 | ±0.00112 | 0.967 | 0.058 | 0.826 | 0.977 |
| 13 | | 104372 | 0.964 | ±0.00115 | 0.97 | ±0.00106 | 0.967 | 0.058 | 0.864 | 0.97 |
| 27 | a | 104372 | 0.964 | ±0.00115 | 0.966 | ±0.00112 | 0.965 | 0.058 | 0.841 | 0.971 |
| 6 | a | 104372 | 0.961 | ±0.00120 | 0.969 | ±0.00107 | 0.965 | 0.058 | 0.872 | 0.966 |
| 6 | b | 104372 | 0.961 | ±0.00120 | 0.966 | ±0.00112 | 0.963 | 0.058 | 0.869 | 0.966 |
| 7 | | 104372 | 0.959 | ±0.00123 | 0.965 | ±0.00114 | 0.962 | 0.058 | 0.853 | 0.966 |
| 5 | | 104372 | 0.964 | ±0.00115 | 0.96 | ±0.00121 | 0.962 | 0.058 | 0.788 | 0.974 |
| 3 | | 104372 | 0.959 | ±0.00123 | 0.954 | ±0.00130 | 0.957 | 0.058 | 0.776 | 0.97 |
| 34 | | 104372 | 0.944 | ±0.00142 | 0.961 | ±0.00120 | 0.952 | 0.058 | 0.869 | 0.948 |
| 16 | | 104372 | 0.945 | ±0.00141 | 0.956 | ±0.00127 | 0.951 | 0.058 | 0.79 | 0.955 |
| 31 | | 104372 | 0.952 | ±0.00132 | 0.951 | ±0.00134 | 0.951 | 0.058 | 0.784 | 0.962 |
| 8 | a | 104372 | 0.943 | ±0.00144 | 0.944 | ±0.00142 | 0.943 | 0.058 | 0.737 | 0.955 |
| 12 | | 104372 | 0.932 | ±0.00156 | 0.944 | ±0.00142 | 0.938 | 0.058 | 0.755 | 0.943 |
| 8 | b | 104372 | 0.886 | ±0.00197 | 0.919 | ±0.00169 | 0.902 | 0.058 | 0.561 | 0.905 |
| 27 | b | 104372 | 0.877 | ±0.00203 | 0.904 | ±0.00182 | 0.89 | 0.058 | 0.72 | 0.886 |
| 23 | | 104372 | 0.781 | ±0.00256 | 0.846 | ±0.00223 | *0.813* | 0.058 | 0.628 | 0.791 |

Table 12: Peking University — Open (italics indicate performance below baseline)

text is given the Chinese translation of the Arabic *al Jazeera* newspaper. A more detailed evaluation of different techniques for dealing with certain constructs is also in order, finding the right balance of learned and heuristic knowledge is paramount. Tied to the accuracy performance of such hybrid systems is the runtime speed: the trade-off between accuracy and throughput is vitally important as more and more data becomes computerized. The overall effects of the various segmentation standards on the comparison of disparate systems has yet to be studied. In particular, a categorization of the differences in standards and the prevalence of the features reflected would be a worth while study. Xia (2000) compares the Penn Chinese Treebank's standard with those used in Taiwan and China, and concludes that, "most disagreements among these three guidelines do not make much difference in bracketing or sentence interpretation." This is probably not so transparent when evaluating segmentation accuracy, however.

No segmentation study has yet to examine the handling of short strings where there is little surrounding context, as in search engine queries. Future evaluations should be designed to focus on these and other specific areas of interest.

## Acknowledgments

This bakeoff could not have taken place without the following institutions who provided training and testing data:

- Institute of Linguistics, Academia Sinica, Taipei, Taiwan
- Institute for Computational Linguistics, Beijing University, Beijing, China
- Language Information Sciences Research Centre, City University of Hong Kong, Hong Kong SAR
- Microsoft Research Asia, Beijing, China

I would like to thank Gina Lavow and Chu-Ren Huang for their organization of the fourth SIGHAN workshop of which this bakeoff is part, and John O'Neil for his comments on an earlier draft of this paper. Finally I would also like to thank the participants for their interest and hard work in making this bakeoff a success.

## References

Charles M. Grinstead and J. Laurie Snell. 1997. *Introduction to Probability.* American Mathematical Society, Providence, RI, 2nd Edition.

Richard Sproat and 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.

Fei Xia. 2000. *The Segmentation Guidelines for the Penn Chinese Treebank (3.0).*