The First International Chinese Word Segmentation Bakeoff

Richard Sproat
AT&T Labs – Research
180 Park Avenue, Florham Park, NJ, 07932, USA
rws@research.att.com

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

Abstract

This paper presents the results from the ACL-SIGHAN-sponsored First International Chinese Word Segmentation Bakeoff held in 2003 and reported in conjunction with the Second SIGHAN Workshop on Chinese Language Processing, Sapporo, Japan. We give the motivation for having an international segmentation contest (given that there have been two within-China contests to date) and we report on the results of this first international contest, analyze these results, and make some recommendations for the future.

1 Introduction

Chinese word segmentation is a difficult problem that has received a lot of attention in the literature; reviews of some of the various approaches can be found in (Wang et al., 1990; Wu and Tseng, 1993; Sproat and Shih, 2001). The problem with this literature has always been that it is very hard to compare systems, due to the lack of any common standard test set. Thus, an approach that seems very promising based on its published report is nonetheless hard to compare fairly with other systems, since the systems are often tested on their own selected test corpora. Part of the problem is also that there is no single accepted segmentation standard: There are several, including the four standards used in this evaluation.

A number of segmentation contests have been held in recent years within Mainland China, in the context of more general evaluations for Chinese-English machine translation. See (Yao, 2001; Yao, 2002) for the first and second of these; the third evaluation will be held in August 2003. The test corpora were segmented according to the Chinese national standard GB 13715 (GB/T 13715–92, 1993), though some lenience was granted in the case of plausible alternative segmentations (Yao, 2001); so while GB 13715 specifies the segmentation 毛/泽东 for *Mao Zedong*, 毛泽东 was also allowed. Accuracies in the mid 80's to mid 90's were reported for the four systems that participated in the first evaluation, with higher scores (many in the high nineties) being reported for the second evaluation.

The motivations for holding the current contest are twofold. First of all, by making the contest international, we are encouraging participation from people and institutions who work on Chinese word segmentation anywhere in the world. The final set of participants in the bakeoff include two from Mainland China, three from Hong Kong, one from Japan, one from Singapore, one from Taiwan and four from the United States.

Secondly, as we have already noted, there are at least four distinct standards in active use in the sense that large corpora are being developed according to those standards; see Section 2.1. It has also been observed that different segmentation standards are appropriate for different purposes; that the segmentation standard that one might prefer for information retrieval applications is likely to be different from the one that one would prefer for text-to-speech synthesis; see (Wu, 2003) for useful discussion. Thus, while we do not subscribe to the view that any of the extant standards are, in fact, appropriate for any particular application, nevertheless, it seems desirable to have a contest where people are tested against more than one standard.

A third point is that we decided early on that we would not be lenient in our scoring, so that alternative segmentations as in the case of 毛泽东 *Mao Zedong*, cited above, would not be allowed. While it would be fairly straightforward (in many cases) to automatically score both alternatives, we felt we could provide a more objective measure if we went strictly by the particular segmentation standard being tested on, and simply did not get into the business of deciding upon allowable alternatives.

Comparing segmenters is difficult. This is not only because of differences in segmentation standards but also due to differences in the design of systems: Systems based exclusively (or even primarily) on lexical and grammatical analysis will often be at a disadvantage during the comparison compared to systems trained exclusively on the training data. Competitions also may fail to predict the performance of the segmenter on new texts outside the training and testing sets. The handling of out-of-vocabulary words becomes a much larger issue in these situations than is accounted for within the test environment: A system that performs admirably in the competition may perform poorly on texts from different registers.

Another issue that is not accounted for in the current collection of evaluations is the handling of short strings with minimal context, such as queries submitted to a search engine. This has been studied indirectly through the cross-language information retrieval work performed for the TREC 5 and TREC 6 competitions (Smeaton and Wilkinson, 1997; Wilkinson, 1998).

This report summarizes the results of this First International Chinese Word Segmentation Bakeoff, provides some analysis of the results, and makes specific recommendations for future bakeoffs. One thing we do not do here is get into the details of specific systems; each of the participants was required to provide a four page description of their system along with detailed discussion of their results, and these papers are published in this volume.

2 Details of the contest

2.1 Corpora

The corpora are detailed in Table 1. Links to descriptions of the corpora can be found at `http://www.sighan.org/bakeoff2003/bakeoff_instr.html`; publications on specific corpora are (Huang et al., 1997) (Academia Sinica), (Xia, 1999) (Chinese Treebank); the Beijing University standard is very similar to that outlined in (GB/T 13715–92, 1993). Table 1 lists the abbreviations for the four corpora that will be used throughout this paper. The suffixes "o" and "c" will be used to denote open and closed tracks, respectively: Thus "ASo,c" denotes the Academia Sinica corpus, both open and closed tracks; and "PKc" denotes the Beijing University corpus, closed track.

During the course of this bakeoff, a number of inconsistencies in segmentation were noted in the CTB corpus by one of the participants. This was done early enough so that it was possible for the CTB developers to correct some of the more common cases, both in the training and the test data. The revised training data was posted for participants, and the revised test data was used during the testing phase.

Inconsistencies were also noted by another participant for the AS corpus. Unfortunately this came too late in the process to correct the data. However, some informal tests on the revised testing data indicated that the differences were minor.

2.2 Rules and Procedures

The contest followed a strict set of guidelines and a rigid timetable. The detailed instructions for the bakeoff can be found at `http://www.sighan.org/bakeoff2003/bakeoff_instr.html` (with simplified and traditional Chinese versions also available). Training material was available starting March 15, testing material was available April 22, and the results had to be returned to the SIGHAN ftp site by April 25 no later than 17:00 EDT.

Upon initial registration sites were required to declare which corpora they would be training and testing on, and whether they would be participating in the open or closed tracks (or both) on each corpus,

Corpus	Abbrev.	Encoding	# Train. Words	# Test. Words
Academia Sinica	**AS**	Big Five (MS Codepage 950)	5.8M	12K
U. Penn Chinese Treebank	**CTB**	EUC-CN (GB 2312-80)	250K	40K
Hong Kong CityU	**HK**	Big Five (HKSCS)	240K	35K
Beijing University	**PK**	GBK (MS Codepage 936)	1.1M	17K

Table 1: Corpora used.

where these were defined as follows:

- For the **open** test sites were allowed to train on the training set for a particular corpus, and in addition they could use any other material including material from other training corpora, proprietary dictionaries, material from the WWW and so forth. However, if a site selected the open track the site was required to explain what percentage of the results came from which sources. For example, if the system did particularly well on out-of-vocabulary words then the participants were required to explain if, for example, those results could mostly be attributed to having a good dictionary.

- In the **closed** test, participants could only use training material from the training data for the particular corpus being testing on. No other material was allowed.

Other obvious restrictions applied: Participants were prohibited from testing on corpora from their own sites, and by signing up for a particular track, participants were declaring implicitly that they had not previously seen the test corpus for that track.

Scoring was completely automatic. Note that the scoring software does not correct for cases where a participant converted from one coding scheme into another, and any such cases were counted as errors. Results were returned to participants within a couple of days of submission of the segmented test data. The script used for scoring can be downloaded from `http://www.sighan.org/bakeoff2003/score`; it is a simple Perl script that depends upon a version of *diff* (e.g. GNU diffutils 2.7.2), that supports the *-y* flag for side-by-side output format.

2.3 Participating sites

Participating sites are shown in Table 2. These are a subset of the sites who had registered for the bakeoff, as some sites withdrew due to technical difficulties.

3 Further details of the corpora

An unfortunate, and sometimes unforseen, complexity in dealing with Chinese text on the computer is the plethora of character sets and character encodings used throughout Greater China. This is demonstrated in the Encoding column of Table 1:

1. Both AS and HK utilize complex-form (or "traditional") characters, using variants of the Big Five character set. The Academia Sinica corpus is composed almost entirely of characters in pure Big Five (four characters, 0xFB5B, 0xFA76, 0xFB7A, and 0xFAAF are outside the encoding range of Big Five), while the City University corpus utilizes 38 (34 unique) characters from the Hong Kong Supplementary Character Set (HKSCS) extension to Big Five.

2. The CTB and PK corpora each use simple-form (or "simplified") characters, using the EUC-CN encoding of the GB 2312-80 character set. However, The PKU corpus includes characters that are not part of GB 2312-80, but are encoded in GBK. GBK is an extension of GB 2312-80 that incorporates some 18,000 hanzi found in Unicode 2.1 within the GB-2312 code space. Only Microsoft's CP936 implements GBK.

This variation of encoding is exacerbated by the usual lack of specific declaration in the files. Generally a file is said to be "Big Five" or "GB", when in actuality the file is encoded in a variation of these. This is problematic in systems that utilize Unicode

Site ID	Site Name	Domain	Contact	Tracks			
S01	Inst. of Comp. Tech.,CAS	CN	Huaping ZHANG	AS**o**	CTB**o**,c	HKc	PK**o**,c
S02	ICL, Beijing U	CN	Baobao CHANG		CTB**o**,c		
S03	HK Polytechnic University	HK	Qin LU	AS**o**	CTB**o**	HK**o**	PK**o**
S04	U of Hong Kong	HK	Guohong FU				PK**o**,c
S05	HK CityU	HK	Chunyu KIT	ASc	CTBc		PKc
S06	Nara IST	JP	Chooi Ling GOH	ASc	CTBc	HKc	PKc
S07	Inst. for Infocomm Research	SG	Guodong ZHOU				PKc
S08	CKIP Ac. Sinica Taiwan	TW	Wei Yun MA			HK**o**,c	PK**o**,c
S09	UC Berkeley	US	Aitao CHEN	ASc			PKc
S10	Microsoft Research	US	Andi WU		CTB**o**,c		PK**o**,c
S11	SYSTRAN Software, Inc.	US	Jin YANG	AS**o**	CTB**o**	HK**o**	PK**o**
S12	U Penn	US	Nianwen XUE	ASc		HKc	

Table 2: Participating sites and associated tracks.

internally, since transcoding back to the original encoding may lose information.

4 Results

4.1 Baseline and topline experiments

We computed a baseline for each of the corpora by compiling a dictionary of all and only the words in the training portion of the corpus. We then used this dictionary with a simple maximum matching algorithm to segment the test corpus. The results of this experiment are presented in Table 3. In this and subsequent tables, we list the *word count* for the test corpus, test *recall* (R), test *precision* (P), *F score*[1], the *out-of-vocabulary* (OOV) rate for the test corpus, the *recall on OOV* words (R_{oov}), and the *recall on in-vocabulary* (R_{iv}) words. Per normal usage, OOV is defined as the set of words in the test corpus not occurring in the training corpus.[2] We expect systems to do *at least* as well as this baseline.

As a nominal *topline* we ran the same maximum matching experiments, but this time populating the dictionary only with words from the test corpus; this is of course a "cheating" experiment since one could not reasonably know exactly the set of words that occur in the test corpus. Since this is better than one could hope for in practice, we would expect systems to generally underperform this topline. The results of this "cheating" experiment are given in Table 4.[3]

4.2 Raw scores

4.2.1 Closed Tests

Results for the closed tests are presented in Tables 5–8. Column headings are as above, except for "c_r", and "c_p" for which see Section 4.3.

4.2.2 Open Tests

Results for the open tests are presented in Tables 9–12; again, see Section 4.3 for the explanation of "c_r", and "c_p".

4.3 Statistical significance of the results

Let us assume that the recall rates for the various system represent the probability p that a word will be successfully identified, and let us further assume that a binomial distribution is appropriate for this experiment. Given the Central Limit Theorem for Bernouilli trials — e.g. (Grinstead and Snell, 1997, page 330), then the 95% confidence interval is given

[1] We use a balanced F score, so that $F = 2PR/(P + R)$.

[2] Note that the OOV recall in Table 3 should in theory be 0.0, but is not always zero because the maximum matching algorithm might get lucky. In particular, if the dictionary contains no word starting with some character c, then the maximum matching algorithm with move on to the next character, leaving c segmented as a word on its own. If it happens that c is in fact a single-character word, then the algorithm will have fortuitously done the right thing.

[3] If one did have the exact list of words occurring in the test corpus, one could still do better than the maximum matching algorithm, since the maximum matching algorithm cannot in general correctly resolve cases where more than one segmentation is possible given the dictionary. However as we can see from the scores in Table 4, such cases constitute at most about 1.5%.

Corpus	word count	R	P	F	OOV	R_{oov}	R_{iv}
AS	11,985	0.917	0.912	0.915	0.022	0.000	0.938
CTB	39,922	0.800	0.663	0.725	0.181	0.062	0.962
HK	34,955	0.908	0.830	0.867	0.071	0.037	0.974
PK	17,194	0.909	0.829	0.867	0.069	0.050	0.972

Table 3: **Baseline** scores: Results for maximum matching using only words from training data

Corpus	word count	R	P	F	OOV	R_{oov}	R_{iv}
AS	11,985	0.990	0.993	0.992	0.022	0.988	0.990
CTB	39,922	0.982	0.988	0.985	0.181	0.990	0.980
HK	34,955	0.986	0.991	0.989	0.071	0.996	0.985
PK	17,194	0.995	0.996	0.995	0.069	1.000	0.994

Table 4: **Topline** ("cheating") scores: Results for maximum matching using only words from testing data data

as $\pm 2\sqrt{p(1-p)/n}$, where n is the number of trials (words). The values for $\pm 2\sqrt{p(1-p)/n}$ are given in Tables 5–12, under the heading "c_r". They can be interpreted as follows: To decide whether two sites are significantly different (at the 95% confidence level) in their performance on a particular task, one just has to compute whether their confidence intervals overlap. Similarly one can treat the precision rates as the probability that a character string that has been identified as a word is really a word; these precision-based confidences are given as "c_p" in the tables.

It seems reasonable to treat two systems as significantly different (at the 95% confidence level), if at least one of their recall-based or precision-based confidences are different. Using this criterion all systems are significantly different from each other except that on PK closed S10 is not significantly different from S09, and S07 is not significantly different from S04.

5 Discussion

5.1 Differences between "open" and "closed" performance

In Figure 1 we plot the F scores for all systems, all tracks. We include as "BASE", and "TOP" the baseline and topline scores discussed previously. In most cases people performed above the baseline, though well below the ideal topline; note though that the two participants in the Academia Sinica open track underperformed the baseline.

Performance on the Penn Chinese Treebank (CTB) corpus was generally lower than all the other corpora; omitting S02, which only ran on CTBo,c the scores for the other systems were uniformly higher on other corpora than they were on CTB, the single exception being S11 which did better on CTBo than on HKo. The baseline for CTB is also much lower than the baseline for other corpora, so one might be inclined to ascribe the generally lower performance to the smaller training data for this corpus. Also, the OOV rate for this corpus is much higher than all of the other corpora, and since error rates are generally higher on OOV, this is surely a contributing factor. However, this would only explain why CTB showed lower performance on the closed test; on the open test, one might expect the size of the training data to matter less, but there were still large differences between several systems' performance on CTB and their performance on other corpora. Note also that the topline for CTB is also lower than for the other corpora. What all of this suggests is that the CTB may simply be less consistent than the other corpora in its segmentation; indeed one of the participants (Andi Wu) noted a number of inconsistencies in both the training and the test data (though inconsistencies were also noted

Site	word count	R	c_r	P	c_p	**F**	OOV	R_{oov}	R_{iv}
S09	11,985	0.966	±0.0033	0.956	±0.0037	0.961	0.022	0.364	0.980
S12	11,985	0.961	±0.0035	0.958	±0.0037	0.959	0.022	0.729	0.966
S06	11,985	0.944	±0.0042	0.945	±0.0042	0.945	0.022	0.574	0.952
S05	11,985	0.952	±0.0039	0.931	±0.0046	0.942	0.022	0.043	0.972
S01	11,985	0.953	±0.0039	0.924	±0.0048	0.938	0.022	0.178	0.970

Table 5: Scores for AS closed, sorted by F.

Site	word count	R	c_r	P	c_p	**F**	OOV	R_{oov}	R_{iv}
S01	39,922	0.886	±0.0032	0.875	±0.0033	0.881	0.181	0.705	0.927
S02	39,922	0.892	±0.0031	0.856	±0.0035	0.874	0.181	0.644	0.947
S10	39,922	0.867	±0.0034	0.797	±0.0040	0.831	0.181	0.431	0.963
S06	39,922	0.852	±0.0036	0.807	±0.0040	0.829	0.181	0.412	0.949
S05	39,922	0.800	±0.0040	0.674	±0.0047	0.732	0.181	0.076	0.959

Table 6: Scores for CTB closed, sorted by F.

for the AS corpus).[4]

Systems that ran on both closed and open tracks for the same corpus generally did better on the open track, indicating (not surprisingly) that using additional data can help. However, the lower-than-baseline performance of S03 and S11 on ASo may reflect issues with tuning of these additional resources to the particular standard in question.

Finally note that the top performance of any system on any track was S09 on ASc (F=0.961). Since performances close to our ideal topline have occasionally been reported in the literature it is worth bearing the results of this bakeoff in mind when reading such reports.

5.2 Differences on OOV

Figure 2 plots the recall on out-of-vocabulary words (R_{oov}) for all systems and all tracks. For this measure, the performance of the baseline is only above 0.0 fortuitously, as we noted in Section 4.1. Similarly the topline performance is only less than 1.0 in cases where there are two or more possible decompositions of a string, and where the option with the longest prefix is not the correct one.

It is with OOV recall that we see the widest variation among systems, which in turn is consistent with the observation that dealing with unknown words is the major outstanding problem of Chinese word segmentation. While some systems performed little better than the baseline, others had a very respectable 0.80 recall on OOV. Again, there was clearly a benefit for many systems in using additional resources than what is in the training data: A number of systems that were run on both closed and open tracks showed significant improvements in the open track. For the closed-track entries that did well on OOV, one must conclude that they have effective unknown-word detection methods.

6 Summary and recommendations

We feel that this First International Chinese Word Segmentation Bakeoff has been useful in that it has provided us with a good sense of the range of performance of various systems, both from academic and industrial institutions. There is clearly no single *best* system, insofar as there is no system that con-

[4]For example, Wu notes that 二十世纪 (*20th Century*) is consistently segmented as two words in the training data, but as one word in the test data. Similarly 副总经理 (*(corporate) vice president*) is segmented as one word in training data but as two words (副/总经理) in the testing data. As a final example, superlatives such as 最佳 (*best*) should be segmented as a single word if the adjective is monosyllabic, and it is not being used predicatively; however this principle is not consistently applied.

Wu also notes that the test data is different from the training data in several respects. Most of the training data comprise texts about Mainland China, whereas most of the testing data is about Taiwan. The test data contains classes of items, such as URL's and English page designations ("p. 64"), that never appeared in the test data.

Site	word count	R	c_r	P	c_p	**F**	OOV	R_{oov}	R_{iv}
S08	34,955	0.947	±0.0024	0.934	±0.0027	0.940	0.071	0.625	0.972
S06	34,955	0.940	±0.0025	0.908	±0.0031	0.924	0.071	0.415	0.980
S12	34955	0.917	±0.0030	0.915	±0.0030	0.916	0.071	0.670	0.936
S01	34,955	0.931	±0.0027	0.873	±0.0036	0.901	0.071	0.243	0.984

Table 7: Scores for HK closed, sorted by F.

Site	word count	R	c_r	P	c_p	**F**	OOV	R_{oov}	R_{iv}
S01	17,194	0.962	±0.0029	0.940	±0.0036	0.951	0.069	0.724	0.979
S10	17,194	0.955	±0.0032	0.938	±0.0037	0.947	0.069	0.680	0.976
S09	17,194	0.955	±0.0032	0.938	±0.0037	0.946	0.069	0.647	0.977
S07	17,194	0.936	±0.0037	0.945	±0.0035	0.940	0.069	0.763	0.949
S04	17,194	0.936	±0.0037	0.942	±0.0036	0.939	0.069	0.675	0.955
S08	17,194	0.939	±0.0037	0.934	±0.0038	0.936	0.069	0.642	0.961
S06	17,194	0.933	±0.0038	0.916	±0.0042	0.924	0.069	0.357	0.975
S05	17,194	0.923	±0.0041	0.867	±0.0052	0.894	0.069	0.159	0.980

Table 8: Scores for PK closed, sorted by F.

sistently outperformed all the others on all tracks. Even if there were, the most one could say is that for the four different segmentation standards and associated corpora, this particular system outperformed the others: But there could be no implication that said system would be the most appropriate for all applications.

One thing that we have not explicitly discussed in this paper is which type of *approach* shows the most promise, given the different submissions. While we are familiar with the approaches taken in several of the tested systems, we leave it up to the individual participants to describe their approaches and hopefully elucidate which aspects of their approaches are most responsible for their successes and failures; the participants' papers all appear in this volume. We leave it up to the research community as a whole to decide whether one approach or another shows most promise.

We believe that there should be future competitions of this kind, possibly not every year, but certainly every couple of years and we have some specific recommendations on how things might be improved in such future competitions:

1. It may be a good idea to insist that all participants participate in all tracks, subject of course to the restriction that participants may not be evaluated on data from their own institution. The decision this time to let people pick and choose was motivated in part by the concern that if we insisted that people participate in all tracks, some participants might be less inclined to participate. It was also motivated in part by the different Chinese coding schemes used by the various corpora, and the possibility that someone's system might work on one coding scheme, but not the other.

   However with sufficient planning, perhaps giving people a longer period of time for training their systems than was possible with this contest, it should be possible to impose this restriction without scaring away potential participants.

2. We would like to see more testing data developed for the next bakeoff. While the test sets turned out to be large enough to measure significant differences between systems in most cases, a larger test set would allow even better statistics. In some cases, more training data will also be needed.

   Given the problems noted by some of the participants with some of the data, we would also

Site	word count	R	c_r	P	c_p	**F**	OOV	R_{oov}	R_{iv}
S11	11,985	0.915	±0.0051	0.894	±0.0056	0.904	0.022	0.426	0.926
S03	11,985	0.892	±0.0057	0.853	±0.0065	0.872	0.022	0.236	0.906

Table 9: Scores for AS open, sorted by F.

Site	word count	R	c_r	P	c_r	**F**	OOV	R_{oov}	R_{iv}
S02	39,922	0.916	±0.0028	0.907	±0.0029	0.912	0.181	0.766	0.949
S10	39,922	0.911	±0.0029	0.891	±0.0031	0.901	0.181	0.738	0.949
S11	39,922	0.891	±0.0031	0.877	±0.0033	0.884	0.181	0.733	0.925
S01	39,922	0.887	±0.0032	0.876	±0.0033	0.881	0.181	0.707	0.927
S03	39,922	0.853	±0.0035	0.806	±0.0040	0.829	0.181	0.578	0.914

Table 10: Scores for CTB open, sorted by F.

like to see more consistently annotated training and test data, and test data that is more representative of what was seen in the training data.

3. We would like to expand the testing data to include texts of various lengths, particularly short strings, in order to emulate query strings seen in commercial search engines.

4. Finally, one question that we did not ask that should have been asked was whether the tested system is used as part of a commercial product or not. It is often believed of natural language and speech applications that deployed commercial systems are about a generation behind the systems being developed in research laboratories. It would be interesting to know if this is true in the domain of Chinese word segmentation, which should be possible to find out if we get a good balance of both.

For the present, we will make the training and test data for the bakeoff available via `http://www.sighan.org/bakeoff2003` (subject to the restrictions of the content providers), so that others can better study the results of this contest.

Acknowledgements

First and foremost we wish to thank the following institutions for providing the training and testing data for this bakeoff:

- Institute of Linguistics, Academia Sinica.
- Institute of Computational Linguistics, Beijing University.
- Language Information Sciences Research Centre, City University of Hong Kong.
- The Chinese Treebank Project, University of Pennsylvania, and the Linguistic Data Consortium.

Without the generous contribution of these resources, this competition would not have been possible.

We would also like to thank Martha Palmer for making funds available to pay for translations of the detailed bakeoff instructions, and to Fu-Dong Chiou, Susan Converse and Nianwen Xue for their work on the translations. Andi Wu and Aitao Chen provided useful feedback on errors in some of the corpora. The first author wishes to thank Bill DuMouchel of AT&T Labs for advice on the statistics. We also wish to thank Professor Tianshun Yao of Northeast (Dongbei) University for sending us the reports of the Chinese national competitions. Finally we thank Fei Xia and Qing Ma for their work on the Second meeting of SIGHAN of which this bakeoff is a part.

References

GB/T 13715–92. 1993. Contemporary Chinese language word-segmentation specification for information processing. Technical report, , Beijing.

Site	word count	R	c_r	P	c_r	**F**	OOV	R_{oov}	R_{iv}
S08	34,955	0.958	±0.0021	0.954	±0.0022	0.956	0.071	0.788	0.971
S03	34,955	0.909	±0.0031	0.863	±0.0037	0.886	0.071	0.579	0.935
S11	34,955	0.898	±0.0032	0.860	±0.0037	0.879	0.071	0.616	0.920

Table 11: Scores for HK open, sorted by F.

Site	word count	R	c_r	P	c_r	**F**	OOV	R_{oov}	R_{iv}
S10	17,194	0.963	±0.0029	0.956	±0.0031	0.959	0.069	0.799	0.975
S01	17,194	0.963	±0.0029	0.943	±0.0035	0.953	0.069	0.743	0.980
S08	17,194	0.939	±0.0037	0.938	±0.0037	0.938	0.069	0.675	0.959
S04	17,194	0.933	±0.0038	0.942	±0.0036	0.937	0.069	0.712	0.949
S03	17,194	0.940	±0.0036	0.911	±0.0043	0.925	0.069	0.647	0.962
S11	17,194	0.905	±0.0045	0.869	±0.0051	0.886	0.069	0.503	0.934

Table 12: Scores for PK open, sorted by F.

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

Chu-Ren Huang, Keh-Jiann Chen, Chang Lili, and Feng-yi Chen. 1997. Segmentation standard for Chinese natural language processing. *International Journal of Computational Linguistics and Chinese Language Processing*, 2(2):47–62.

Alan Smeaton and Ross Wilkinson. 1997. Spanish and chinese document retrieval at TREC-5. In E.M. Vorhees and D.K. Harman, editors, *Proceedings of the Fifth Text REtrieval Conference*.

Richard Sproat and Chilin Shih. 2001. Corpus-based methods in Chinese morphology and phonology. Technical report, Linguistic Society of America Summer Institute, Santa Barbara, CA. `http://www.research.att.com/~rws/newindex/notes.pdf`.

Yongheng Wang, Haiju Su, and Yan Mo. 1990. Automatic processing of Chinese words. *Journal of Chinese Information Processing*, 4(4):1–11.

Ross Wilkinson. 1998. Chinese document retrieval at TREC-6. In E.M. Vorhees and D.K. Harman, editors, *Proceedings of the Sixth Text REtrieval Conference*.

Zimin Wu and Gwyneth Tseng. 1993. Chinese text segmentation for text retrieval: Achievements and problems. *Journal of the American Society for Information Science*, 44(9):532–542.

Andi Wu. 2003. Customizable segmentation of morphologically derived words in Chinese. *International Journal of Computational Linguistics and Chinese Language Processing*, 8(1): forthcoming.

Fei Xia. 1999. Segmentation guideline, Chinese Treebank Project. Technical report, University of Pennsylvania. `http://morph.ldc.upenn.edu/ctb/`.

Tianshun Yao (姚天顺). 2001. 汉英机器翻译评测系统(第一次评测总结与汇报). Technical report, Northeast University (东北大学), China, January.

Tianshun Yao (姚天顺). 2002. 汉英机器翻译评测系统(第二阶段评测方案). Technical report, Northeast University (东北大学), China, August.

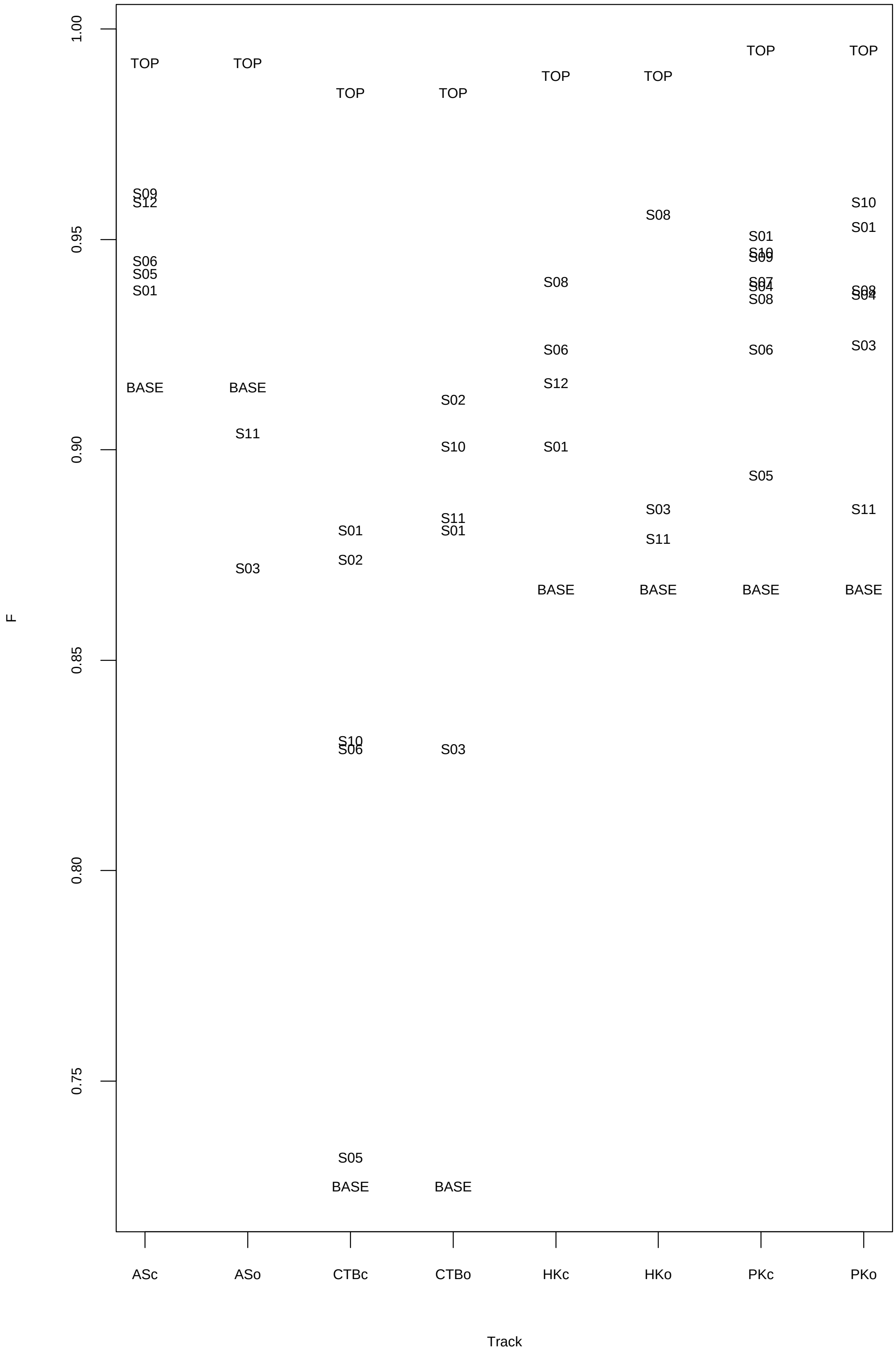

Figure 1: F scores for all systems, all tracks.

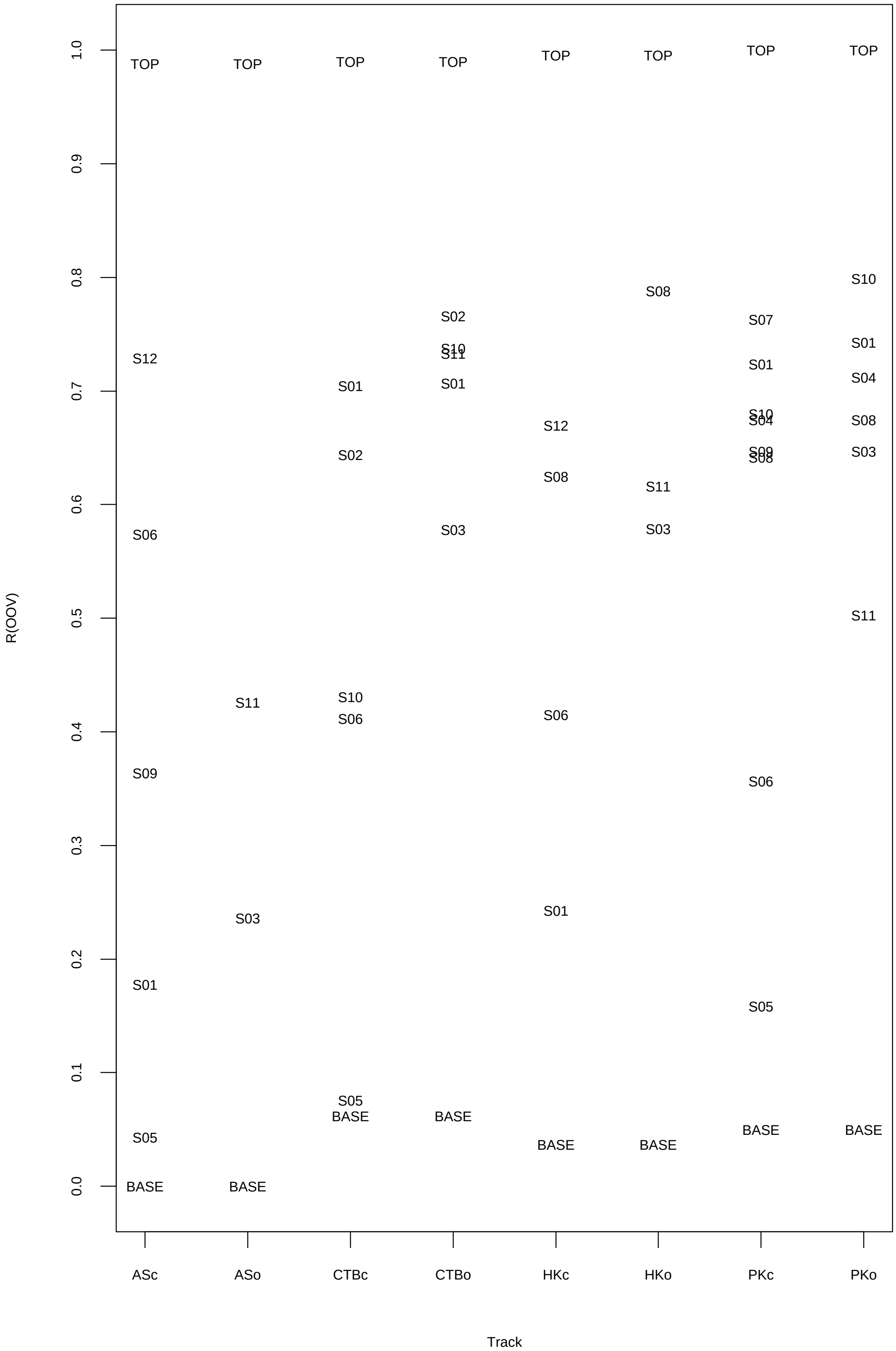

Figure 2: R_{oov} scores for all systems, all tracks.