The CIPS-SIGHAN CLP 2010 Chinese Word Segmentation Bakeoff

Hongmei Zhao and Qun Liu
Key Laboratory of Intelligent Information Processing,
Institute of Computing Technology, Chinese Academy of Sciences
{zhaohongmei,liuqun}@ict.ac.cn

Abstract

The CIPS-SIGHAN CLP 2010 Chinese Word Segmentation Bakeoff was held in the summer of 2010 to evaluate the current state of the art in word segmentation. It focused on the cross-domain performance of Chinese word segmentation algorithms. Eighteen groups submitted 128 results over two tracks (open training and closed training), four domains (literature, computer science, medicine and finance) and two subtasks (simplified Chinese and traditional Chinese). We found that compared with the previous Chinese word segmentation bakeoffs, the performance of cross-domain Chinese word segmentation is not much lower, and the out-of-vocabulary recall is improved.

1 Introduction

Chinese is written without inter-word spaces, so finding word-boundaries is an essential first step in many natural language processing tasks ranging from part of speech tagging to parsing, reference resolution and machine translation.

SIGHAN, the Special Interest Group for Chinese Language Processing of the Association for Computational Linguistics, successfully conducted four prior word segmentation bakeoffs, in 2003 (Sproat and Emerson, 2003), 2005 (Emerson, 2005), 2006 (Levow, 2006) and 2007 (Jin and Chen, 2007), and the bakeoff 2007 was jointly organized with the Chinese Information Processing Society of China (CIPS). These evaluations established benchmarks for word segmentation with which researchers evaluate their segmentation system.

After years of intensive researches, Chinese word segmentation has achieved a quite high precision, though the out-of-vocabulary problem is still a continuing challenge. However, the performance of segmentation is not so satisfying for out-of-domain text.

The CIPS-SIGHAN CLP 2010 Chinese Word Segmentation Bakeoff continues the ongoing series of the SIGHAN Chinese Word Segmentation Bakeoff. It was organized by Institute of Computing Technology, Chinese Academy of Sciences (abbreviated as ICT below). It focused on the cross-domain performance of Chinese word segmentation algorithms. And the bakeoff results will be reported in conjunction with the First CIPS-SIGHAN joint conference on Chinese Language Processing, Beijing, China.

2 Details of the Evaluation

2.1 Corpora

There are two kinds of corpora in the evaluation, with one using the simplified Chinese characters and another using the traditional Chinese characters. For the simplified Chinese corpora, the test corpora, reference corpora, and the unlabeled training corpora were provided by ICT, and the labeled training corpus (1 month data of The People's Daily in 1998) was provided by Peking University. For the traditional Chinese corpora, all the training, test and reference corpora were provided by the Hongkong City University.

There are four domains in this evaluation. Before the releasing of the test data, two of them (literature and computer science, we abbreviate “computer science” to “computer” below) are known to the participants (we provided the corresponding unlabeled training

corpora			Characters	Tokens	Word Types	TTR	OOV Rate
Simplified Chinese	Test	Literature	50,637	35,736	6,364	0.18	0.069
		Computer	53,382	35,319	4,150	0.12	0.152
		Medicine	50,969	31,490	5,076	0.16	0.11
		Finance	53,253	33,028	4,918	0.15	0.087
	Training	Labeled	1,820,456	1,109,947	55,303	0.05	
		Unlabeled-L	100,352				
		Unlabeled-C	103,764				
Traditional Chinese	Test	Literature	54,357	36,378	8,141	0.22	0.094
		Computer	67,321	43,499	6,197	0.14	0.094
		Medicine	68,090	43,458	6,510	0.15	0.075
		Finance	74,461	47,144	6,652	0.14	0.068
	Training	Labeled	1,863,298	1,146,988	63,588	0.06	
		Unlabeled-L	105,653				
		Unlabeled-C	109,303				

Table1. Overall corpus statistics

Site ID	Site Name	Contact	Simplified Chinese	Traditional Chinese
S1	College of Computer and Information Engineering, Anyang Normal University, Henan province, China	Jiangde Yu	◆◇	◆◇
S2	Institute of Intelligent Information Processing, Beijing Information Science & Technology University	Wenjie Su	◆	
S3	Beijing Institute of Technology	Huaping Zhang	◇	
S4	Center for Language Information Processing Institute，Beijing Language and Culture University	Zhiyong Luo	◇	
S5	Beijing University of Posts and Telecommunications	Caixia Yuan	◆◇	
S6	Dalian University of Technology	Huiwei Zhou	◆◇	
S7	Fudan University	Xipeng Qiu	◆	◆
S8	Shenzhen Graduate School Harbin Institute of Technology	Jianping Shen	◇	◇
S9	Language Technologies Institute, Carnegie Mellon University	Qin Gao	◆◇	
S10	National Central University, Taiwan	Yu-Chieh Wu	◆	◆
S11	Natural Language Processing Lab, Northeastern University, China	Huizhen Wang	◆	
S12	National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Science	Kun Wang	◆	
S13	Institute of Computer Science and Technology, Peking University	Liang Zong	◆	
S14	Institute of Computational Linguistics, Peking University	Mairgup	◆	
S15	Queensland University of Technology	Eric Tang	◆◇	◆◇
S16	Institute of Information Science, Academia Sinica, Taiwan	Cheng-Lung Sung	◆	◆
S17	Natural Language Processing Lab, Suzhou University	Junhui Li	◆	
S18	Anhui Speech and language Technology Engineering Research Center	Zhigang Chen	◆◇	

Table 2. Participating groups (◆=closed track, ◇=open track, there are four domains on every track)

corpora for each during the training phrase), and another two domains (medicine and finance) are unknown to the participants (without any in-domain training corpora). All corpora are UTF-8 encoded. Details on each corpus are provided in Table 1. We introduce a type-token ratio (TTR) to indicate the vocabulary diversity in each corpus.

During the process of building the reference corpora for the simplified Chinese word segmentation subtask, we manually check the automatically segmented results of the test data against the standard provided in "The Specification for the Basic Processing of Contemporary Chinese Corpus from Peking University". In this process, we refer to the labeled training data frequently with a view to keep the annotation consistency between these two kinds of corpora. Furthermore, we made a comparison test which compared the segmentation of the same character strings present in both corpora automatically, and corrected the inconsistent cases. However, in the labeled training corpus, there are minor incorrect segmentation cases against the standard from Peking University, such as "赢家" (yin2 jia1, with the meaning of "the winner", this word should be regarded as a word according to the above-mentioned standard), and there are also a few interior inconsistent cases in this corpus, such as "患有" and "患 有" (huan4 you3, with the meaning of "suffer from"). Whenever the segmentation of the reference corpora was different from the above-mentioned incorrect or inconsistent segmentation in the training corpus, we followed the standard from Peking University. All the evaluation corpora can be accessible from the Chinese Linguistic Data Consortium at: http://www.chineseldc.org.

2.2 Rules and Procedures

This bakeoff followed a strict set of guidelines and a rigid timetable. The detailed instructions for the bakeoff can be found at http://www.cipsc.org.cn/clp2010/cfpa.htm. The training material of simplified Chinese word segmentation was available starting April 1, the training material of traditional Chinese word segmentation was available April 23, testing material was available June 9, and the results had to be returned to the organizer by email by June 11 no later than 18:00 Beijing time.

The participating groups ("sites") of CIPS-SIGHAN CLP 2010 Bakeoff registered by email. There are two subtasks in this evaluation: word segmentation for simplified Chinese text and word segmentation for traditional Chinese text. The participating sites were required to declare which subtask they would participate in. The open and closed tracks were defined as follows:

- For the closed training evaluation, participants can only use data provided by the organizer to train their systems. Specifically, the following data resources and software tools are not permitted to be used in the training:
  1. Unspecified corpus;
  2. Unspecified dictionary, word list or character list: include the dictionaries of named entity, character lists for specific type of Chinese named entities, idiom dictionaries, semantic lexicons, etc.;
  3. Human-encoded rule bases;
  4. Unspecified software tools, include word segmenters, part-of-speech taggers, or parsers which are trained using unspecified data resources.

  The character type information to distinguish the following four character types can be used in training: Chinese characters, English letters, digits and punctuations.

- In the Open training evaluation, participants can use any language resources, including the training data provided by the organizer.

Participants were asked to submit their data using specific naming conventions, and from the result file name we can see in which track the result was run, as well as other necessary information. Of course, the results on both tracks are welcomed.

Scoring was done automatically using a combination of Perl and shell scripts. The scripts (Sproat and Emerson, 2003, 2005) used for scoring can be downloaded from http://www.sighan.org/bakeoff2005/. The bakeoff organizer provided an on-line scoring system to all the participants who had submitted their bakeoff results for their follow-up experiments.

2.3 Participating sites

Eighteen sites submitted results and a technical report. Mainland China had the greatest number with 14, followed by Taiwan (2), the United States (1) and Australia (1). A summary of participating groups and the tracks for which they submitted results can be found in Table 2 on the preceding page. There are more sites who had registered for the bakeoff. However, several of them withdrew due to technical difficulties or other problems. Altogether 128 runs were submitted for scoring.

3 Results

3.1 Baseline and topline experiments

Following previous bakeoffs, to provide a basis for comparison, we computed baseline and topline scores for each of the corpora. When computing a baseline, we compiled a dictionary of all the words in the labeled training corpus, and then we used this dictionary with a simple left-to-right maximal match algorithm to segment the test corpus. The results of this experiment are shown in Table 3. We expect systems to do at least as well as the baseline. The topline employed the same procedure, but instead used the dictionary of all the words in the test corpus. These results are presented in Table 4. We expect systems to generally underperform this topline, because no one could exactly know the set of words that occur in the test corpus.

In these and subsequent tables, we list the word count for the test corpus, 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 (Roov), and the recall on in-vocabulary words (Riv).

3.2 Raw scores

All the results are presented in Tables 5-20. Column headings are as above, except for “Cr” and “Cp” for which see Section 3.3. All tables are sorted by F score.

3.3 Statistical significance of the results

Following previous bakeoffs, let us assume that the recall rates represent the probability p that a word will be successfully identified, and let us further assume the binomial distribution is appropriate for this experiment. Given the Central Limit Theorem for Bernouli trials — e.g. (Grinstead and Snell, 1997), then the 95% confidence interval is given as $\pm 2\sqrt{p(1-p)/n}$, where n is the number of trials (words). The recall-based confidences ($\pm 2\sqrt{p(1-p)/n}$) are given as “Cr” in Tables 5-20. Similarly, we can assume the precision rates represent the probability that a character string that has been identified as a word is really a word. And the precision-based confidences are given as “Cp” in the tables. They can be interpreted as follows: To decide whether two systems are significantly different (at the 95% confidential level), one just has to compute whether their confidence intervals overlap. If at least one of the “Cr” and “Cp” are different, we can treat these two systems as significantly different (at the 95% confidential level). Using this criterion all systems in this bakeoff are significantly different from each other.

4 Discussion

4.1 Comparison between open and closed tracks

In this bakeoff, there are 8 systems that ran on both closed and open tracks, which result in 32 pairs of scores for F measure and OOV recall respectively. Table 21 shows the results of these systems. We can see that their scores of F measures on open track don’t have advantage over their counterparts on the closed track: only 14 scores (in 32 scores) on open track are higher than their counterparts on the closed track. This is different from the previous bakeoffs. But for OOV recall, the case is different. There are 23 scores (in 32 scores) on open track are higher than their counterparts on the closed track.

4.2 Improved OOV recall over the prior bakeoffs

From all the results, we can see that the widest variation among systems lies in the OOV recall rate. And dealing with unknown words is still the most difficult problem of Chinese word segmentation.

However, while comparing the top OOV recall rates of this bakeoff with those of the prior four bakeoffs, we found the OOV recall rates of this bakeoff achieved an obvious improvement. Table 22 shows the comparisons. We managed to find four pairs of test corpora with similar OOV rates for comparisons. In the comparisons, most top OOV recall rates of bakeoff 2010 are much higher than their counterparts of prior bakeoffs. An exception comes from the open track of medicine domain for traditional CWS subtask, and because only 3 systems submitted results, this comparison seems less meaningful.

4.3 Performance under different domains

We listed the top performance by F measure on every track, domain and subtask on Table 23.

Generally we think that cross-domain word segmentation will lead to a lower performance than in-domain word segmentation. In this bakeoff, it seems that the best performance of cross-domain word segmentation is at almost the same level of that of the prior bakeoffs. We know that the performance of different test set is incomparable. However, the performance in the out-of-domain text is somewhat surprising to us. We guess one reason may be the usage of domain adaptive technology, another reason may be the new technologies used by the participants. We hope to see the exact reasons in the technological reports of participants in the coming conference.

We provided unlabeled data to two domains. However, we did not see significant difference on the performance of closed test between these domains and other domains. Some participants pointed out that it is because the size of the unlabeled data is rather small.

And we found that among four domains, the performance (by the value of F measure and OOV recall, with scores in bold in the table) on finance is always the best or very close to the best. Perhaps this is because the OOV rate on finance test corpus is rather low.

Corpus		Word Count	R	P	F	OOV	R_{oov}	R_{iv}
Simplifed Chinese	L	35736	0.917	0.862	0.889	0.069	0.156	0.973
	C	35319	0.856	0.632	0.727	0.152	0.163	0.98
	M	31490	0.886	0.774	0.826	0.11	0.123	0.981
	F	33028	0.914	0.803	0.855	0.087	0.233	0.979
Traditional Chinese	L	36378	0.863	0.788	0.824	0.094	0.041	0.948
	C	43499	0.873	0.701	0.778	0.094	0.01	0.963
	M	43458	0.886	0.81	0.846	0.075	0.027	0.955
	F	47144	0.888	0.826	0.855	0.068	0.006	0.952

Table 3. Baseline scores: Results for maximum match with training vocabulary (L=literature, C=computer, M=medicine, F=finance)

Corpus		Word Count	R	P	F	OOV	R_{oov}	R_{iv}
Simplifed Chinese	L	35736	0.986	0.99	0.988	0.069	0.996	0.985
	C	35319	0.991	0.993	0.992	0.152	0.99	0.991
	M	31490	0.989	0.991	0.99	0.11	0.98	0.99
	F	33028	0.994	0.995	0.994	0.087	0.995	0.994
Traditional Chinese	L	36378	0.981	0.988	0.985	0.094	0.998	0.979
	C	43499	0.988	0.991	0.99	0.094	0.996	0.987
	M	43458	0.984	0.989	0.986	0.075	0.992	0.983
	F	47144	0.981	0.986	0.984	0.068	0.997	0.98

Table 4. Topline scores: Results for maximum match with testing vocabulary (L=literature, C=computer, M=medicine, F=finance)

Site ID	Word Count	R	C_r	P	C_p	F	OOV	R_{oov}	R_{iv}
S5	35736	0.945	±0.00241	0.946	±0.00239	0.946	0.069	0.816	0.954
S6	35736	0.94	±0.00251	0.942	±0.00247	0.941	0.069	0.649	0.961
S12	35736	0.937	±0.00257	0.937	±0.00257	0.937	0.069	0.652	0.958
S10	35736	0.936	±0.00259	0.932	±0.00266	0.934	0.069	0.564	0.964
S11	35736	0.931	±0.00268	0.936	±0.00259	0.934	0.069	0.648	0.952
S18	35736	0.932	±0.00266	0.935	±0.00261	0.933	0.069	0.654	0.953
S14	35736	0.925	±0.00279	0.931	±0.00268	0.928	0.069	0.667	0.944
S9	35736	0.92	±0.00287	0.925	±0.00279	0.923	0.069	0.625	0.942
S7	35736	0.915	±0.00295	0.925	±0.00279	0.92	0.069	0.577	0.94
S13	35736	0.916	±0.00293	0.922	±0.00284	0.919	0.069	0.613	0.939
S16	35736	0.917	±0.00292	0.921	±0.00285	0.919	0.069	0.699	0.933
S1	35736	0.908	±0.00306	0.918	±0.00290	0.913	0.069	0.556	0.935
S17	35736	0.909	±0.00304	0.903	±0.00313	0.906	0.069	0.707	0.924
S15	35736	0.907	±0.00307	0.862	±0.00365	*0.884*	0.069	0.206	0.959
S2	35736	0.695	±0.00487	0.744	±0.00462	*0.719*	0.069	0.381	0.719

Table 5. Simplified Chinese: Literature -- Closed (italics indicate performance below baseline)

Site ID	Word Count	R	Cr	P	Cp	F	OOV	R_{oov}	R_{iv}
S6	35736	0.958	±0.00212	0.953	±0.00224	0.955	0.069	0.655	0.981
S3	35736	0.965	±0.00194	0.94	±0.00251	0.952	0.069	0.814	0.976
S18	35736	0.942	±0.00247	0.943	±0.00245	0.942	0.069	0.702	0.959
S9	35736	0.939	±0.00253	0.943	±0.00245	0.941	0.069	0.699	0.957
S1	35736	0.908	±0.00306	0.916	±0.00293	0.912	0.069	0.535	0.936
S5	35736	0.893	±0.00327	0.918	±0.00290	0.905	0.069	0.803	0.899
S4	35736	0.897	±0.00322	0.907	±0.00307	0.902	0.069	0.688	0.913
S15	35736	0.869	±0.00357	0.873	±0.00352	0.871	0.069	0.657	0.885
S8	35736	0.836	±0.00392	0.841	±0.00387	0.838	0.069	0.609	0.853

Table 6. Simplified Chinese: Literature --Open (italics indicate performance below baseline)

Site ID	Word Count	R	C_r	P	C_p	F	OOV	R_{oov}	R_{iv}
S6	35319	0.953	±0.00225	0.95	±0.00232	0.951	0.152	0.827	0.975
S11	35319	0.948	±0.00236	0.945	±0.00243	0.947	0.152	0.853	0.965
S12	35319	0.941	±0.00251	0.94	±0.00253	0.94	0.152	0.757	0.974
S9	35319	0.938	±0.00257	0.936	±0.00260	0.937	0.152	0.805	0.962
S13	35319	0.939	±0.00255	0.934	±0.00264	0.937	0.152	0.81	0.962
S18	35319	0.935	±0.00262	0.934	±0.00264	0.935	0.152	0.792	0.961
S5	35319	0.946	±0.00241	0.914	±0.00298	0.93	0.152	0.808	0.971
S14	35319	0.941	±0.00251	0.916	±0.00295	0.928	0.152	0.796	0.967
S7	35319	0.934	±0.00264	0.919	±0.00290	0.926	0.152	0.739	0.969
S10	35319	0.915	±0.00297	0.915	±0.00297	0.915	0.152	0.594	0.972
S17	35319	0.921	±0.00287	0.9	±0.00319	0.91	0.152	0.748	0.952
S1	35319	0.89	±0.00333	0.908	±0.00308	0.899	0.152	0.592	0.943
S15	35319	0.876	±0.00351	0.844	±0.00386	0.86	0.152	0.457	0.951
S16	35319	0.876	±0.00351	0.799	±0.00426	0.836	0.152	0.456	0.952
S2	35319	0.713	±0.00481	0.641	±0.00511	*0.675*	0.152	0.257	0.795

Table 7. Simplified Chinese: Computer -- Closed (italics indicate performance below baseline)

Site ID	Word Count	R	Cr	P	Cp	F	OOV	R_{oov}	R_{iv}
S9	35319	0.95	±0.00232	0.95	±0.00232	0.95	0.152	0.82	0.973
S18	35319	0.948	±0.00236	0.946	±0.00241	0.947	0.152	0.812	0.973
S6	35319	0.948	±0.00236	0.929	±0.00273	0.939	0.152	0.735	0.986
S3	35319	0.951	±0.00230	0.926	±0.00279	0.938	0.152	0.775	0.982
S8	35319	0.951	±0.00230	0.915	±0.00297	0.932	0.152	0.77	0.983
S5	35319	0.918	±0.00292	0.896	±0.00325	0.907	0.152	0.771	0.945
S1	35319	0.893	±0.00329	0.908	±0.00308	0.9	0.152	0.607	0.944
S4	35319	0.892	±0.00330	0.88	±0.00346	0.886	0.152	0.791	0.91
S15	35319	0.859	±0.00370	0.878	±0.00348	0.868	0.152	0.668	0.893

Table 8. Simplified Chinese: Computer -- Open (italics indicate performance below baseline)

Site ID	Word Count	R	C_r	P	C_p	F	OOV	R_{oov}	R_{iv}
S6	31490	0.942	±0.00263	0.936	±0.00276	0.939	0.11	0.75	0.965
S18	31490	0.937	±0.00274	0.934	±0.00280	0.936	0.11	0.761	0.959
S5	31490	0.94	±0.00268	0.928	±0.00291	0.934	0.11	0.761	0.962
S7	31490	0.927	±0.00293	0.924	±0.00299	0.925	0.11	0.714	0.953
S10	31490	0.933	±0.00282	0.915	±0.00314	0.924	0.11	0.642	0.969
S11	31490	0.924	±0.00299	0.922	±0.00302	0.923	0.11	0.756	0.944
S12	31490	0.93	±0.00288	0.917	±0.00311	0.923	0.11	0.674	0.961
S14	31490	0.928	±0.00291	0.918	±0.00309	0.923	0.11	0.73	0.953
S9	31490	0.923	±0.00300	0.917	±0.00311	0.92	0.11	0.729	0.947
S13	31490	0.917	±0.00311	0.911	±0.00321	0.914	0.11	0.699	0.944
S1	31490	0.902	±0.00335	0.907	±0.00327	0.904	0.11	0.633	0.935
S16	31490	0.9	±0.00338	0.896	±0.00344	0.898	0.11	0.596	0.937
S17	31490	0.894	±0.00347	0.873	±0.00375	0.884	0.11	0.647	0.925
S15	31490	0.885	±0.00360	0.804	±0.00447	0.842	0.11	0.218	0.967
S2	31490	0.735	±0.00497	0.74	±0.00494	*0.738*	0.11	0.378	0.779

Table 9. Simplified Chinese: Medicine -- Closed (italics indicate performance below baseline)

Site ID	Word Count	R	Cr	P	Cp	F	OOV	R_{oov}	R_{iv}
S9	31490	0.94	±0.00268	0.936	±0.00276	0.938	0.11	0.768	0.962
S18	31490	0.941	±0.00266	0.935	±0.00278	0.938	0.11	0.787	0.96
S6	31490	0.951	±0.00243	0.92	±0.00306	0.935	0.11	0.67	0.986
S3	31490	0.953	±0.00239	0.913	±0.00318	0.933	0.11	0.704	0.984
S5	31490	0.917	±0.00311	0.907	±0.00327	0.912	0.11	0.704	0.943
S4	31490	0.91	±0.00323	0.901	±0.00337	0.906	0.11	0.725	0.933
S1	31490	0.904	±0.00332	0.906	±0.00329	0.905	0.11	0.635	0.937
S15	31490	0.865	±0.00385	0.846	±0.00407	0.855	0.11	0.559	0.903
S8	31490	0.839	±0.00414	0.832	±0.00421	*0.836*	0.11	0.618	0.866

Table 10. Simplified Chinese: Medicine -- Open (italics indicate performance below baseline)

Site ID	Word Count	R	C_r	P	C_p	F	OOV	R_{oov}	R_{iv}
S6	33028	0.959	±0.00218	0.96	±0.00216	0.959	0.087	0.827	0.972
S12	33028	0.957	±0.00223	0.956	±0.00226	0.957	0.087	0.813	0.971
S9	33028	0.956	±0.00226	0.955	±0.00228	0.956	0.087	0.857	0.965
S11	33028	0.953	±0.00233	0.956	±0.00226	0.955	0.087	0.871	0.961
S18	33028	0.955	±0.00228	0.956	±0.00226	0.955	0.087	0.848	0.965
S5	33028	0.956	±0.00226	0.952	±0.00235	0.954	0.087	0.849	0.966
S10	33028	0.945	±0.00251	0.941	±0.00259	0.943	0.087	0.666	0.972
S7	33028	0.94	±0.00261	0.942	±0.00257	0.941	0.087	0.719	0.961
S13	33028	0.943	±0.00255	0.94	±0.00261	0.941	0.087	0.773	0.959
S14	33028	0.948	±0.00244	0.928	±0.00284	0.937	0.087	0.761	0.965
S1	33028	0.925	±0.00290	0.938	±0.00265	0.931	0.087	0.664	0.95
S17	33028	0.935	±0.00271	0.915	±0.00307	0.925	0.087	0.736	0.954
S16	33028	0.91	±0.00315	0.906	±0.00321	0.908	0.087	0.562	0.943
S15	33028	0.904	±0.00324	0.865	±0.00376	0.884	0.087	0.321	0.96
S2	33028	0.736	±0.00485	0.752	±0.00475	*0.744*	0.087	0.23	0.784

Table 11. Simplified Chinese: Finance -- Closed (italics indicate performance below baseline)

Site ID	Word Count	R	Cr	P	Cp	F	OOV	R_{oov}	R_{iv}
S9	33028	0.96	±0.00216	0.96	±0.00216	0.96	0.087	0.847	0.971
S6	33028	0.964	±0.00205	0.95	±0.00240	0.957	0.087	0.763	0.983
S18	33028	0.948	±0.00244	0.955	±0.00228	0.951	0.087	0.853	0.957
S3	33028	0.963	±0.00208	0.938	±0.00265	0.95	0.087	0.758	0.982
S1	33028	0.925	±0.00290	0.937	±0.00267	0.931	0.087	0.669	0.95
S5	33028	0.928	±0.00284	0.934	±0.00273	0.931	0.087	0.808	0.939
S8	33028	0.893	±0.00340	0.896	±0.00336	0.894	0.087	0.796	0.902
S4	33028	0.885	±0.00351	0.893	±0.00340	0.889	0.087	0.757	0.897
S15	33028	0.853	±0.00390	0.85	±0.00393	*0.851*	0.087	0.438	0.893

Table 12. Simplified Chinese: Finance -- Open (italics indicate performance below baseline)

Site ID	Word Count	R	C_r	P	C_p	F	OOV	R_{oov}	R_{iv}
S10	36378	0.942	±0.00245	0.942	±0.00245	0.942	0.094	0.788	0.958
S1	36378	0.888	±0.00331	0.905	±0.00307	0.896	0.094	0.728	0.904
S7	36378	0.869	±0.00354	0.91	±0.00300	0.889	0.094	0.698	0.887
S16	36378	0.871	±0.00351	0.891	±0.00327	0.881	0.094	0.67	0.891
S15	36378	0.864	±0.00359	0.789	±0.00428	0.825	0.094	0.105	0.943

Table 13. Traditional Chinese: Literature -- Closed (italics indicate performance below baseline)

Site ID	Word Count	R	C_r	P	C_p	F	OOV	R_{oov}	R_{iv}
S1	36378	0.905	±0.00307	0.9	±0.00315	0.902	0.094	0.775	0.918
S8	36378	0.868	±0.00355	0.802	±0.00418	0.834	0.094	0.503	0.905
S15	36378	0.804	±0.00416	0.722	±0.00470	*0.761*	0.094	0.234	0.863

Table 14. Traditional Chinese: Literature -- Open (italics indicate performance below baseline)

Site ID	Word Count	R	C_r	P	C_p	F	OOV	R_{oov}	R_{iv}
S10	43499	0.948	±0.00213	0.957	±0.00195	0.952	0.094	0.666	0.977
S7	43499	0.933	±0.00240	0.949	±0.00211	0.941	0.094	0.791	0.948
S1	43499	0.908	±0.00277	0.931	±0.00243	0.919	0.094	0.684	0.931
S16	43499	0.913	±0.00270	0.917	±0.00265	0.915	0.094	0.663	0.939
S15	43499	0.868	±0.00325	0.85	±0.00342	0.859	0.094	0.316	0.926

Table 15. Traditional Chinese: Computer -- Closed (italics indicate performance below baseline)

Site ID	Word Count	R	C_r	P	C_p	F	OOV	R_{oov}	R_{iv}
S1	43499	0.911	±0.00273	0.924	±0.00254	0.918	0.094	0.698	0.933
S8	43499	0.875	±0.00317	0.829	±0.00361	0.851	0.094	0.594	0.904
S15	43499	0.789	±0.00391	0.736	±0.00423	*0.761*	0.094	0.35	0.834

Table 16. Traditional Chinese: Computer -- Open (italics indicate performance below baseline)

Site ID	Word Count	R	C_r	P	C_p	F	OOV	R_{oov}	R_{iv}
S10	43458	0.953	±0.00203	0.957	±0.00195	0.955	0.075	0.798	0.966
S7	43458	0.908	±0.00277	0.932	±0.00242	0.92	0.075	0.771	0.919
S1	43458	0.905	±0.00281	0.924	±0.00254	0.914	0.075	0.725	0.919
S16	43458	0.9	±0.00288	0.915	±0.00268	0.908	0.075	0.668	0.919
S15	43458	0.871	±0.00322	0.815	±0.00373	0.842	0.075	0.115	0.932

Table 17. Traditional Chinese: Medicine -- Closed (italics indicate performance below baseline)

Site ID	Word Count	R	C_r	P	C_p	F	OOV	R_{oov}	R_{iv}
S1	43458	0.903	±0.00284	0.903	±0.00284	0.903	0.075	0.729	0.917
S8	43458	0.879	±0.00313	0.814	±0.00373	0.846	0.075	0.48	0.912
S15	43458	0.811	±0.00376	0.74	±0.00421	*0.774*	0.075	0.254	0.856

Table 18. Traditional Chinese: Medicine -- Open (italics indicate performance below baseline)

Site ID	Word Count	R	C_r	P	C_p	F	OOV	R_{oov}	R_{iv}
S10	47144	0.964	±0.00172	0.962	±0.00176	0.963	0.068	0.812	0.975
S7	47144	0.925	±0.00243	0.939	±0.00220	0.932	0.068	0.793	0.935
S16	47144	0.922	±0.00247	0.929	±0.00237	0.925	0.068	0.732	0.935
S1	47144	0.891	±0.00287	0.912	±0.00261	0.901	0.068	0.676	0.907
S15	47144	0.875	±0.00305	0.834	±0.00343	*0.854*	0.068	0.169	0.926

Table 19. Traditional Chinese: Finance -- Closed (italics indicate performance below baseline)

Site ID	Word Count	R	C_r	P	C_p	F	OOV	R_{oov}	R_{iv}
S1	47144	0.903	±0.00273	0.916	±0.00256	0.91	0.068	0.721	0.916
S8	47144	0.832	±0.00344	0.76	±0.00393	*0.794*	0.068	0.356	0.866
S15	47144	0.811	±0.00361	0.753	±0.00397	*0.781*	0.068	0.235	0.853

Table 20. Traditional Chinese: Finance -- Open (italics indicate performance below baseline)

Subtask	Site ID	Track	Literature		Computer		Medicine		Finance	
			F	R_{oov}	F	R_{oov}	F	R_{oov}	F	R_{oov}
Simplified Chinese	S1	◆	0.913	0.556	0.899	0.592	0.904	0.633	0.931	0.664
		◇	0.912	0.535	0.9	0.607	0.905	0.635	0.931	0.669
	S5	◆	0.946	0.816	0.93	0.808	0.934	0.761	0.954	0.849
		◇	0.905	0.803	0.907	0.771	0.912	0.704	0.931	0.808
	S6	◆	0.941	0.649	0.951	0.827	0.939	0.75	0.959	0.827
		◇	0.955	0.655	0.939	0.735	0.935	0.67	0.957	0.763
	S9	◆	0.923	0.625	0.937	0.805	0.92	0.729	0.956	0.857
		◇	0.941	0.699	0.95	0.82	0.938	0.768	0.96	0.847
	S15	◆	0.884	0.206	0.86	0.457	0.842	0.218	0.884	0.321
		◇	0.871	0.657	0.868	0.668	0.855	0.559	0.851	0.438
	S18	◆	0.933	0.654	0.935	0.792	0.936	0.761	0.955	0.848
		◇	0.942	0.702	0.947	0.812	0.938	0.787	0.951	0.853
Traditional Chinese	S1	◆	0.896	0.728	0.919	0.684	0.914	0.725	0.901	0.676
		◇	0.902	0.775	0.918	0.698	0.903	0.729	0.91	0.721
	S15	◆	0.825	0.105	0.859	0.316	0.842	0.115	0.854	0.169
		◇	0.761	0.234	0.761	0.35	0.774	0.254	0.781	0.235

Table 21. Comparison: closed track vs. open track (◆=closed track, ◇=open track)

bakeoff	corpus	characters	OOV rate	word count	closed track		open track	
					R_{oov}	F	R_{oov}	F
2007	CKIP	traditional Chinese	0.074	90678	0.740	0.947	0.780	0.956
2010	medicine		0.075	43458	0.798	0.955	0.729	0.903
2006	UPUC	simplified Chinese	0.088	155K	0.707	0.933	0.768	0.944
2010	finance		0.087	33028	0.871	0.955	0.853	0.951
2005	CityU	traditional Chinese	0.074	41K	0.736	0.941	0.806	0.962
2010	medicine		0.075	43458	0.798	0.955	0.729	0.903
2003	PK	simplified Chinese	0.069	17K	0.763	0.940	0.799	0.959
2010	literature		0.069	35736	0.816	0.946	0.814	0.952

Table 22. Comparisons of top OOV recall rates of different bakeoffs on the test corpora with similar OOV rates (2003, 2005, 2006 and 2007 represent the SIGHAN bakeoff 2003, 2005, 2006 and 2007 respectively, and 2010 represents the CIPS-SIGHAN CLP 2010 bakeoff)

			Closed Track						Open Track					
		OOV	ID	R	P	F	R_{oov}	R_{iv}	ID	R	P	F	R_{oov}	R_{iv}
S	L	0.069	S5	0.945	0.946	0.946	0.816	0.954	S6	0.958	0.953	0.955	0.655	0.981
	C	0.152	S6	0.953	0.95	0.951	0.827	0.975	S9	0.95	0.95	0.95	0.82	0.973
	M	0.11	S6	0.942	0.936	0.939	0.75	0.965	S9	0.94	0.936	0.938	0.768	0.962
									S18	0.941	0.935	0.938	0.787	0.96
	F	0.087	S6	0.959	0.96	**0.959**	**0.827**	0.972	S9	0.96	0.96	**0.96**	**0.847**	0.971
T	L	0.094	S10	0.942	0.942	0.942	0.788	0.958	S1	0.905	0.9	0.902	**0.775**	0.918
	C	0.094	S10	0.948	0.957	0.952	0.666	0.977	S1	0.911	0.924	0.918	0.698	0.933
	M	0.075	S10	0.953	0.957	0.955	0.798	0.966	S1	0.903	0.903	0.903	0.729	0.917
	F	0.068	S10	0.964	0.962	**0.963**	**0.812**	0.975	S1	0.903	0.916	**0.91**	**0.721**	0.916

Table 23. Top performance on every subtask, domain, and track (S=simplified Chinese test, T=traditional Chinese test, L=literature, C=computer, M=medicine, F=finance)

5 Conclusions & Future Directions

The CIPS-SIGHAN CLP 2010 Chinese Word Segmentation Bakeoff successfully brought together a collection of 18 strong research groups to assess the progress of this fundamental research in Chinese language processing.

There is clearly no single best system. And the participating sites S1, S10, S9, S6, S5 and S18 have all achieved respectable scores on different track runs of this bakeoff. An improvement on the OOV recall over the prior bakeoffs has been observed.

It is the first time to apply word segmentation bakeoff on four domains. It's also the first time to use unlabeled training corpora in the bakeoff to test the unsupervised or semi-supervised learning ability of the segmentation system. Unsupervised or Semi-supervised learning needs to incorporate large amounts of unlabeled data. We design the evaluation with two unknown domains without any in-domain training corpora, compared with two known domains each with an in-domain unlabeled training corpus. Although no significant difference has been found, it's still worth it. The size of our unlabeled training corpora was too small in this bakeoff, and we hope to improve this in next evaluation.

The word segmentation is a necessary pre-processing phase for the downstream processing tasks. In future evaluations, we hope to see the integration of word segmentation task with a higher level task such as machine translation, with a view to exactly evaluate the impact of improvements in word segmentation on broader downstream applications.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (Grant No. 90920004). We gratefully acknowledge the generous assistance of the organizations listed below who provided the data and the Chinese word segmentation standard for this bakeoff; without their support, it could not have taken place:

- City University of Hong Kong
- Institute for Computational Linguistics, Beijing University, Beijing, China

We thank Le Sun for his organization of the First CIPS-SIGHAN Joint Conference on Chinese Language Processing of which this bakeoff is part. We thank Lili Zhao for her segmentation-inconsistency-checking program and other preparation works for this bakeoff, and thank Guanghui Luo and Siyang Cao for the online scoring system they set up and maintained. We thank Yajuan Lü for her helpful suggestions on this paper. Finally we thank all the participants for their interest and hard work in making this bakeoff a success.

References

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

Gina-Anne Levow. 2006. *The third international Chinese language processing bakeoff: Word segmentation and named entity recognition.* In Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pages 108–117, Sydney, Australia, July. Association for Computational Linguistics.

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, India.

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.

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.