The Third International Chinese Language Processing Bakeoff: Word Segmentation and Named Entity Recognition

Gina-Anne Levow
University of Chicago
1100 E. 58th St.
Chicago, IL 60637 USA
levow@cs.uchicago.edu

Abstract

The Third International Chinese Language Processing Bakeoff was held in Spring 2006 to assess the state of the art in two important tasks: word segmentation and named entity recognition. Twenty-nine groups submitted result sets in the two tasks across two tracks and a total of five corpora. We found strong results in both tasks as well as continuing challenges.

1 Introduction

Many important natural language processing tasks ranging from part of speech tagging to parsing to reference resolution and machine translation assume the ready availability of a tokenization into words. While such tokenization is relatively straight-forward in languages which use whitespace to delimit words, Chinese presents a significant challenge since it is typically written without such separation. Word segmentation has thus long been the focus of significant research because of its role as a necessary pre-processing phase for the tasks above. However, word segmentation remains a significant challenge both for the difficulty of the task itself and because standards for segmentation vary and human segmenters may often disagree.

SIGHAN, the Special Interest Group for Chinese Language Processing of the Association for Computational Linguistics, conducted two prior word segmentation bakeoffs, in 2003 and 2005(Emerson, 2005), which established benchmarks for word segmentation against which other systems are judged. The bakeoff presentations at SIGHAN workshops highlighted new approaches in the field as well as the crucial importance of handling out-of-vocabulary (OOV) words.

A significant class of OOV words is Named Entities, such as person, location, and organization names. These terms are frequently poorly covered in lexical resources and change over time as new individuals, institutions, or products appear. These terms also play a particularly crucial role in information retrieval, reference resolution, and question answering. As a result of this importance, and interest in expanding the scope of the bakeoff expressed at the Fourth SIGHAN Workshop, in the Winter of 2005 it was decided to hold a new bakeoff to evaluate both continued progress in Word Segmentation (WS) and the state of the art in Chinese Named Entity Recognition (NER).

2 Details of the Evaluation

2.1 Corpora

Five corpora were provided for the evaluation: three in Simplified characters and two in traditional characters. The Simplified character corpora were provided by Microsoft Research Asia (MSRA) for WS and NER, by University of Pennsylvania/University of Colorado (UPUC) for WS, and by the Linguistic Data Consortium (LDC) for NER. The Traditional character corpora were provided by City University of Hong Kong (CITYU) for WS and NER and by the Chinese Knowledge Information Processing Laboratory (CKIP) of the Academia Sinica, Taiwan for WS. Each data provider offered separate training and test corpora. General information for each corpus appears in Table 1.

All data providers were requested to supply the training and test corpora in both the standard local encoding and in Unicode (UTF-8) in a standard XML format with sentence and word tags, and named entity tags if appropriate. For

Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pages 108–117,
Sydney, July 2006.

Source	Encodings	Training (Wds/Types)	Test (Wds/Types)
CITYU	BIG5HKSCS/Unicode	1.6M/76K	220K/23K
CKIP	BIG5/Unicode	5.5M/146K	91K/15K
LDC	Unicode	632K (est. wds)	61K (est. wds)
MSRA	GB18030/Unicode	1.3M/63K	100K/13K
UPUC	GB/Unicode	509K/37K	155K/17K

Table 1: Overall corpus statistics

all providers except the LDC, missing encodings were transcoded by the organizers using the appropriate Python CJK codecs.

Primary training and truth data for word segmentation were generated by the organizers via a Python script by replacing sentence end tags with newlines and word end tags with a single whitespace character, deleting all other tags and associated newlines. For test data, end of sentence tags were replaced with newlines and all other tags removed. Since the UPUC truth corpus was only provided in white-space separated form, test data was created by automatically deleting line-internal whitespace.

Primary training and truth data for named entity recognition were converted from the provided XML format to a two-column format similar to that used in the CoNLL 2002 NER task(Sang, 2002) adapted for Chinese, where the first column is the current character and the second column the corresponding tag. Format details may be found at the bakeoff website (`http://www.sighan.org/bakeoff2006/`). For consistency, we tagged only "<NAMEX>" mentions, of either (PER)SON, (LOC)ATION, (ORG)ANIZATION, or (G)EO-(P)OLITICAL (E)NTITY as annotated in the corpora.[1] Test was generated as above.

The LDC required sites to download training data from their website directly in the ACE[2] evaluation format, restricted to "NAM" mentions. The organizers provided the sites with a Python script to convert the LDC data to the CoNLL format above, and the same script was used to create the truth data. Test data was created by splitting on newlines or Chinese period characters.

Comparable XML format data was also provided for all corpora and both tasks.

The segmentation and NER annotation standard, as appropriate, for each corpus was made available on the bakeoff website. As observed in previous evaluations, these documents varied widely in length, detail, and presentation language.

Except as noted above, no additional changes were made to the data furnished by the providers.

2.2 Rules and Procedures

The Third Bakeoff followed the structure of the first two word segmentation bakeoffs. Participating groups ("sites") registered by email form; only the primary contact was required to register, identifying the corpora and tasks of interest. Training data was released for download from the websites (both SIGHAN and LDC) on April 17, 2006. Test data was released on May 15, 2006 and results were due 14:00 GMT on May 17. Scores for all submitted runs were emailed to the individual groups by May 19, and were made available to all groups on a web page a few days later.

Groups could participate in either or both of two tracks for each task and corpus:

- In the *open* track, participants could use any external data they chose in addition to the provided training data. Such data could include external lexica, name lists, gazetteers, part-of-speech taggers, etc. Groups were required to specify this information in their system descriptions.

- In the *closed* track, participants could only use information found in the provided training data. Information such as externally obtained word counts, part of speech information, or name lists was excluded.

Groups were required to submit fully automatic runs and were prohibited from testing on corpora which they had previously used.

Scoring was performed automatically using a combination of Python and Perl scripts, facilitated by stringent file naming conventions. In cases

[1] Only the LDC provided GPE tags.

[2] http://www.ldc.upenn.edu/projects/ACE

where naming errors or minor divergences from required file formats arose, a mix of manual intervention and automatic conversion was employed to enable scoring. The primary scoring scripts were made available to participants for followup experiments.

3 Participating Sites

A total of 36 sites registered, and 29 submitted results for scoring. The greatest number of participants came from the People's Republic of China (11), followed by Taiwan (7), the United States (5), Japan (2), with one team each from Singapore, Korea, Hong Kong, and Canada. A summary of participating groups with task and track information appears in Table 2. A total of 144 official runs were scored: 101 for word segmentation and 43 for named entity recognition.

4 Results & Discussion

We report results below first for word segmentation and second for named entity recognition.

4.1 Word Segmentation Results

To provide a basis for comparison, we computed baseline and possible topline scores for each of the corpora. The baseline was constructed by left-to-right maximal match implemented by Python script, using the training corpus vocabulary. The topline employed the same procedure, but instead used the test vocabulary. These results are shown in Tables 3 and 4.

For the WS task, we computed the following measures using the *score*(Sproat and Emerson, 2003) program developed for the previous bakeoffs: recall (R), precision (P), equally weighted F-measure ($F = \frac{2PR}{(P+R)}$), the rate of out-of-vocabulary words (OOV rate) in the test corpus, the recall on OOV (R_{oov}), and recall on in-vocabulary words (R_{iv}). In and out of vocabulary status are defined relative to the training corpus. Following previous bakeoffs, we employ the Central Limit Theorem for Bernoulli trials (Grinstead and Snell, 1997) to compute 95% confidence interval as $\pm 2\sqrt{(\frac{p(1-p)}{n})}$, assuming the binomial distribution is appropriate. For recall, C_r, we assume that recall represents the probability of correct word identification. Symmetrically, for precision, we compute C_p, setting p to the precision value. One can determine if two systems may then be viewed as significantly different at a 95% confidence level by computing whether their confidence intervals overlap.

Word segmentation results for all runs grouped by corpus and track appear in Tables 5-12; all tables are sorted by F-score.

4.2 Word Segmentation Discussion

Across all corpora, the best F-score was achieved in the MSRA Open Track at 0.979. Overall, as would be expected, the best results on Open track runs had higher F-scores than the best results for Closed Track runs on the same corpora. Likewise, the OOV recall rates for the best Open Track systems exceed those of the best Closed Track runs on comparable corpora by exploiting outside information. Unfortunately, few sites submitted runs in both conditions making strong direct comparisons difficult.

Many systems strongly outperformed the baseline runs, though none achieved the topline. The closest approach to the topline score was on the CITYU corpus, with the best performing runs achieving 99% of the topline F-score.

It is also informative to observe the rather wide variation in scores across the test corpora. The maximum scores were achieved on the MSRA corpus closely followed by the CITYU corpus. The best score achieved on the UPUC Open track condition, however, was lower than all scores but one on the MSRA Open track. However, a comparison of the baseline, topline, and especially the OOV rates may shed some light on this disparity. The UPUC training corpus was only about one-third the size of the MSRA corpus, and the OOV rate for UPUC was more than double that of any of the other corpora, yielding a challenging task, especially in the Closed track. This high OOV rate may also be attributed to a change in register, since the training data for UPUC had been drawn exclusively from the Chinese Treebank and the test data also included data from other newswire and broadcast news sources. In contrast, the MSRA corpus had both the highest baseline and highest topline scores, possibly indicating an easier corpus in some sense. The differences in topline also suggest a greater degree of variance in the UPUC, and in fact all other corpora, relative the MSRA corpus. These differences highlight the continuing challenges of handling out-of-vocabulary words and performing segmentation across different reg-

Site Name	Site ID	Country	CITYU WS	CKIP WS	MSRA WS	UPUC WS	CITYU NER	LDC NER	MSRA NER
Natural Language Processing Lab, Northeastern University of China	1	ZH	C	C	C	C			
Language Technologies Institute, Carnegie Mellon University	2	US	O	O	O	O			
National Institute of Information and Communications Technology, Japan	3	JP					C	C	C
Basis Technology Corp.	4	US	C	C	C	C			
Pattern Recognition and Intelligent System Laboratory,Beijing University of Posts and Telecommunications	5	ZH			C	C			
HKUST, Human Language Technology Center	6	HK					O	O	O
The University of Tokyo	7	JP			O	O		O	O
Institute of Software, Chinese Academy of Sciences	8	ZH	C	C	C	C	C	C	OC
Alias-i, Inc.	9	US	C	C	C	C	C		C
Beijing University of Posts and Telecommunications	10	ZH			O	O			O
France Telecom R&D Beijing	11	ZH	C		OC				O
NETEASE Information Technology (Beijing) Co., Ltd.	12	ZH				O			O
AI Lab., Dept of Information Management, Huafan University, Taiwan.	13	TW	OC	OC					
Nanjing University, China	14	ZH			O				OC
Intelligent Agent Systems Lab (IASL), Academia Sinica	15	TW	C	C	C				
Simon Fraser University	16	CA	C		C	C			
Tung Nan Institute of Technology	18	TW			C				
Institute of Information Science, Taiwan	19	TW					C		C
Microsoft Research Asia	20	ZH	OC	OC		OC			
Yahoo!	21	US					C		C
CKIP, Academia Sinica, Taiwan	22	TW	O						
Kookmin University	23	KO	C	C	C	C			
Shenyang Institute of Aeronautical Engineering	24	ZH			OC	OC			
Institute for Infocomm Research, Singapore	26	SG	C	C	C	C	C		C
National Taiwan University	29	TW					C		
ITNLP, Harbin Institute of Technology, China	30	ZH			OC				O
National Central University at Taiwan	31	TW				C	C		
National Laboratory on Machine Perception, Peking University, China	32	ZH	OC	OC	OC	OC			O
University of Texas at Austin	34	US	O	O	O	O			

Table 2: Participating Sites by Corpus, Task, and Track

Source	Recall	Precision	F-measure	OOV Rate	R_{oov}	R_{iv}
CITYU	0.930	0.882	0.906	0.040	0.009	0.969
CKIP	0.915	0.870	0.892	0.042	0.030	0.954
MSRA	0.949	0.900	0.924	0.034	0.022	0.981
UPUC	0.869	0.790	0.828	0.088	0.011	0.951

Table 3: Baselines: WS: Maximum match with training vocabulary

Source	Recall	Precision	F-measure	OOV Rate	R_{oov}	R_{iv}
CITYU	0.982	0.985	0.984	0.040	0.993	0.981
CKIP	0.980	0.987	0.983	0.042	0.997	0.979
MSRA	0.991	0.993	0.992	0.034	0.999	0.991
UPUC	0.961	0.976	0.968	0.088	0.989	0.958

Table 4: Toplines: WS: Maximum match with testing vocabulary

Site	RunID	R	C_r	P	C_p	F	R_{oov}	R_{iv}
15	d	0.973	±0.000691	0.972	±0.000703	0.972	0.787	0.981
15	b	0.973	±0.000691	0.972	±0.000703	0.972	0.787	0.981
20		0.972	±0.000703	0.971	±0.000715	0.971	0.792	0.979
32		0.969	±0.000739	0.970	±0.000727	0.970	0.773	0.978
1	a	0.971	±0.000715	0.965	±0.000783	0.968	0.719	0.981
15	c	0.965	±0.000783	0.967	±0.000761	0.966	0.792	0.972
15	a	0.966	±0.000772	0.967	±0.000761	0.966	0.786	0.973
26		0.968	±0.000750	0.961	±0.000825	0.965	0.633	0.983
11		0.962	±0.000815	0.962	±0.000815	0.962	0.722	0.972
16		0.963	±0.000805	0.958	±0.000855	0.961	0.689	0.974
9		0.966	±0.000772	0.957	±0.000865	0.961	0.555	0.983
1	b	0.958	±0.000855	0.963	±0.000805	0.960	0.714	0.968
8		0.952	±0.000911	0.954	±0.000893	0.953	0.747	0.960
23		0.950	±0.000929	0.949	±0.000938	0.949	0.638	0.963
4	b	0.845	±0.001543	0.844	±0.001547	0.844	0.632	0.854
4	a	0.841	±0.001559	0.844	±0.001547	0.843	0.506	0.855
13	1	0.589	±0.002097	0.589	±0.002097	0.589	0.022	0.613

Table 5: CITYU: Word Segmentation: Closed Track

Site	RunID	R	C_r	P	C_p	F	R_{oov}	R_{iv}
20		0.978	±0.000625	0.977	±0.000639	0.977	0.840	0.984
32		0.979	±0.000611	0.976	±0.000652	0.977	0.813	0.985
34		0.971	±0.000715	0.967	±0.000761	0.969	0.795	0.978
22		0.970	±0.000727	0.965	±0.000783	0.967	0.761	0.979
2		0.964	±0.000794	0.964	±0.000794	0.964	0.787	0.971
13	2	0.544	±0.002123	0.549	±0.002121	0.547	0.194	0.559
13	3	0.524	±0.002129	0.503	±0.002131	0.513	0.195	0.538
13	1	0.497	±0.002131	0.467	±0.002127	0.481	0.057	0.516

Table 6: CITYU: Word Segmentation: Open Track

Site	RunID	R	C_r	P	C_p	F	R_{oov}	R_{iv}
20		0.961	±0.001280	0.955	±0.001371	0.958	0.702	0.972
15	a	0.961	±0.001280	0.953	±0.001400	0.957	0.658	0.974
15	b	0.961	±0.001280	0.952	±0.001414	0.957	0.656	0.974
32		0.958	±0.001327	0.948	±0.001468	0.953	0.646	0.972
26		0.958	±0.001327	0.941	±0.001558	0.949	0.554	0.976
1	b	0.947	±0.001482	0.943	±0.001533	0.945	0.601	0.962
1	a	0.949	±0.001455	0.940	±0.001571	0.944	0.694	0.960
9		0.951	±0.001428	0.935	±0.001630	0.943	0.389	0.976
23		0.937	±0.001607	0.933	±0.001654	0.935	0.547	0.954
8		0.939	±0.001583	0.929	±0.001699	0.934	0.606	0.954
4	a	0.836	±0.002449	0.834	±0.002461	0.835	0.521	0.849
4	b	0.836	±0.002449	0.828	±0.002496	0.832	0.590	0.847
13	1	0.747	±0.002875	0.677	±0.003093	0.710	0.036	0.778

Table 7: CKIP: Word Segmentation: Closed Track

Site	RunID	R	C_r	P	C_p	F	R_{oov}	R_{iv}
20		0.964	±0.001232	0.955	±0.001371	0.959	0.704	0.975
34		0.959	±0.001311	0.949	±0.001455	0.954	0.672	0.972
32		0.958	±0.001327	0.948	±0.001468	0.953	0.647	0.972
2	a	0.953	±0.001400	0.946	±0.001495	0.949	0.679	0.965
2	b	0.951	±0.001428	0.944	±0.001521	0.948	0.676	0.964
13	2	0.724	±0.002956	0.668	±0.003115	0.695	0.161	0.749
13	3	0.736	±0.002915	0.653	±0.003148	0.692	0.160	0.761
13	1	0.654	±0.003146	0.573	±0.003271	0.611	0.057	0.680

Table 8: CKIP: Word Segmentation: Open Track

Site	RunID	R	C_r	P	C_p	F	R_{oov}	R_{iv}
32		0.964	±0.001176	0.961	±0.001222	0.963	0.612	0.976
26		0.961	±0.001222	0.953	±0.001336	0.957	0.499	0.977
9		0.959	±0.001252	0.955	±0.001309	0.957	0.494	0.975
1	a	0.955	±0.001309	0.956	±0.001295	0.956	0.650	0.966
15	d	0.953	±0.001336	0.956	±0.001295	0.955	0.574	0.966
11	a	0.955	±0.001309	0.953	±0.001336	0.954	0.575	0.969
15	b	0.952	±0.001350	0.956	±0.001295	0.954	0.575	0.966
15	c	0.949	±0.001389	0.957	±0.001281	0.953	0.673	0.959
15	a	0.949	±0.001389	0.958	±0.001266	0.953	0.672	0.959
16		0.952	±0.001350	0.954	±0.001323	0.953	0.604	0.964
11	b	0.950	±0.001376	0.954	±0.001323	0.952	0.602	0.962
5		0.956	±0.001295	0.947	±0.001414	0.951	0.493	0.972
1	b	0.946	±0.001427	0.952	±0.001350	0.949	0.568	0.959
18	c	0.950	±0.001376	0.930	±0.001611	0.940	0.272	0.974
30	a	0.963	±0.001192	0.918	±0.001732	0.940	0.175	0.991
18	b	0.954	±0.001323	0.921	±0.001703	0.937	0.163	0.981
8		0.933	±0.001578	0.942	±0.001476	0.937	0.640	0.943
23		0.933	±0.001578	0.939	±0.001511	0.936	0.526	0.948
24		0.923	±0.001683	0.929	±0.001621	0.926	0.554	0.936
18	a	0.949	±0.001389	0.897	±0.001919	0.922	0.022	0.982
4	a	0.830	±0.002371	0.832	±0.002360	0.831	0.473	0.842
4	b	0.817	±0.002441	0.821	±0.002420	0.819	0.491	0.829

Table 9: MSRA: Word Segmentation: Closed Track

Site	RunID	R	C_r	P	C_p	F	R_{oov}	R_{iv}
11	a	0.980	±0.000884	0.978	±0.000926	0.979	0.839	0.985
11	b	0.977	±0.000946	0.976	±0.000966	0.977	0.840	0.982
14		0.975	±0.000986	0.976	±0.000966	0.975	0.811	0.981
32		0.977	±0.000946	0.971	±0.001059	0.974	0.675	0.988
10		0.970	±0.001077	0.970	±0.001077	0.970	0.804	0.976
30	a	0.977	±0.000946	0.960	±0.001237	0.968	0.624	0.989
34		0.959	±0.001252	0.961	±0.001222	0.960	0.711	0.968
2		0.949	±0.001389	0.954	±0.001323	0.952	0.692	0.958
7		0.953	±0.001336	0.940	±0.001499	0.947	0.503	0.969
24		0.938	±0.001522	0.946	±0.001427	0.942	0.706	0.946

Table 10: MSRA: Word Segmentation: Open Track

Site	RunID	R	C_r	P	C_p	F	R_{oov}	R_{iv}
20		0.940	±0.001207	0.926	±0.001330	0.933	0.707	0.963
32		0.936	±0.001244	0.923	±0.001355	0.930	0.683	0.961
1	a	0.940	±0.001207	0.914	±0.001425	0.927	0.634	0.969
26	a	0.936	±0.001244	0.917	±0.001402	0.926	0.617	0.966
26	b	0.932	±0.001279	0.910	±0.001454	0.921	0.577	0.966
16		0.929	±0.001305	0.909	±0.001462	0.919	0.628	0.958
5		0.932	±0.001279	0.904	±0.001497	0.918	0.546	0.969
1	b	0.922	±0.001363	0.914	±0.001425	0.918	0.637	0.949
8		0.922	±0.001363	0.912	±0.001440	0.917	0.680	0.945
31	1	0.917	±0.001402	0.904	±0.001497	0.910	0.676	0.940
9		0.919	±0.001387	0.895	±0.001558	0.907	0.459	0.964
23		0.915	±0.001417	0.896	±0.001551	0.905	0.565	0.949
24		0.902	±0.001511	0.887	±0.001609	0.895	0.568	0.934
4	a	0.831	±0.001905	0.819	±0.001957	0.825	0.487	0.864
4	b	0.809	±0.001998	0.827	±0.001922	0.818	0.637	0.825

Table 11: UPUC: Word Segmentation: Closed Track

Site	RunID	R	C_r	P	C_p	F	R_{oov}	R_{iv}
34		0.949	±0.001118	0.939	±0.001216	0.944	0.768	0.966
2		0.942	±0.001188	0.928	±0.001314	0.935	0.711	0.964
20		0.940	±0.001207	0.927	±0.001322	0.933	0.741	0.959
7		0.944	±0.001169	0.922	±0.001363	0.933	0.680	0.970
12		0.933	±0.001271	0.916	±0.001410	0.924	0.656	0.959
32		0.940	±0.001207	0.907	±0.001476	0.923	0.561	0.976
24		0.928	±0.001314	0.906	±0.001483	0.917	0.660	0.954
10		0.925	±0.001339	0.897	±0.001545	0.911	0.593	0.957

Table 12: UPUC: Word Segmentation: Open Track

isters and writing styles.

4.3 Named Entity Results

We employed a slightly modified version of the CoNLL 2002 scoring script to evaluate NER task submissions. For each submission, we compute overall phrase precision (P), recall(R), and balanced F-measure (F), as well as F-measure for each entity type (PER-F,ORG-F,LOC-F,GPE-F).

For each corpus, we compute a baseline performance level as follows. Based on the training data, using a left-to-right pass over the test data, we assign a named entity tag to a span of characters if it was tagged with a single unique NE tag (PER/LOC/ORG/GPE) in the training data.[3] All In the case of overlapping spans, we tag the maximal span. These scores for all NER corpora are found in Table 13.

4.4 Named Entity Discussion

Though fewer sites participated in the NER task, performances overall were very strong, with only two runs performing below baseline. The best F-score overall on the MSRA Open Track reached 0.912, with ten other scores for MSRA and CITYU Open Track above 0.85. Only two sites submitted runs in both Open and Closed Track conditions, and few Open Track runs were submitted at all, again limiting comparability. For the only corpus with substantial numbers of both Open and Closed Track runs, MSRA, the top three runs outperformed all Closed Track runs.

System scores and baselines were much higher for the CITYU and MSRA corpora than for the LDC corpus. This disparity can, in part, also be attributed to a substantially smaller training corpus for the LDC than the other two collections. The presence of an additional category, Geo-political entity, which is potentially confused for either location or organization also enhances the difficulty of this corpus. Training requirements, variation across corpora, and most extensive tag sets will continue to raise challenges for named entity recognition.

Named entity recognition results for all runs grouped by corpus and track appear in Tables 14-19; all tables are sorted by F-score.

[3] If the span was a single character and appeared UNtagged in the corpus, we exclude it. Longer spans are retained for tagging even if they might appear both tagged and untagged in the training corpus.

[4] This result indicates a rescoring of the run below with all

Source	P	R	F	PER-F	ORG-F	LOC-F	GPE-F
CITY	0.611	0.467	0.529	0.587	0.516	0.503	N/A
LDC	0.493	0.378	0.428	0.395	0.29	0.259	0.539
MSRA	0.59	0.488	0.534	0.614	0.469	0.531	N/A

Table 13: Baselines: NER: Maximal match with unique tag in training corpus

Site	RunID	P	R	F	ORG-F	LOC-F	PER-F
3		0.9143	0.8676	0.8903	0.8046	0.9211	0.9087
19	ccrf	0.9201	0.8545	0.8861	0.8054	0.9251	0.8872
21	a	0.9266	0.8475	0.8853	0.7973	0.9232	0.8937
21	b	0.9242	0.8491	0.8850	0.7976	0.9236	0.8920
19	avdic	0.9079	0.8626	0.8847	0.7984	0.9233	0.8914
8		0.9276	0.8181	0.8694	0.7707	0.9114	0.8769
21	f	0.9188	0.8231	0.8683	0.7852	0.9105	0.8652
21	g	0.9164	0.8246	0.8681	0.7842	0.9114	0.8636
9		0.8690	0.8417	0.8551	0.7541	0.8861	0.8845
19	bme	0.8742	0.8307	0.8519	0.7667	0.9015	0.8395
26		0.8466	0.8061	0.8259	0.7467	0.8863	0.7927
31	1	0.9035	0.6973	0.7871	0.7703	0.8905	0.5974
29		0.7703	0.6447	0.7019	0.4948	0.7613	0.7531

Table 14: CITYU: Named Entity Recognition: Closed Track

Site	RunID	P	R	F	ORG-F	LOC-F	PER-F
6		0.8692	0.7498	0.8051	0.6801	0.8604	0.8098

Table 15: CITYU: Named Entity Recognition: Open Track

Site	RunID	P	R	F	ORG-F	LOC-F	PER-F	GPE-F
3		0.8026	0.7265	0.7627	0.6585	0.3046	0.7884	0.8204
8		0.8143	0.5953	0.6878	0.5855	0.1705	0.6571	0.7727

Table 16: LDC: Named Entity Recognition: Closed Track

Site	RunID	P	R	F	ORG-F	LOC-F	PER-F	GPE-F
7		0.7616	0.6621	0.7084	0.5209	0.2857	0.7422	0.7930
6	GPE-LOC[4]	0.6720	0.6554	0.6636	0.4553	0.7078	0.7420	
6		0.3058	0.2982	0.3019	0.4553	0.0370	0.7420	0.0

Table 17: LDC: Named Entity Recognition: Open Track

Site	RunID	P	R	F	ORG-F	LOC-F	PER-F
14		0.8894	0.8420	0.8651	0.8310	0.8545	0.9009
21	a	0.9122	0.8171	0.8620	0.8196	0.9053	0.8257
21	b	0.8843	0.8288	0.8556	0.7698	0.9013	0.8495
3		0.8814	0.8234	0.8514	0.8150	0.9062	0.7938
21	f	0.8845	0.7931	0.8363	0.8071	0.9003	0.7568
21	g	0.8661	0.8032	0.8335	0.7742	0.8991	0.7753
19	avdic	0.8637	0.7767	0.8179	0.8138	0.8233	0.8126
19	dcrf	0.8623	0.7740	0.8158	0.8141	0.8207	0.8093
9		0.8188	0.8097	0.8142	0.7295	0.8529	0.8196
19	cnoword	0.8670	0.7554	0.8074	0.8100	0.8257	0.7764
19	bvoting	0.8583	0.7606	0.8065	0.8145	0.8133	0.7899
26		0.8105	0.7882	0.7992	0.7491	0.8385	0.7699
21	r	0.8748	0.7168	0.7880	0.7288	0.8604	0.7107
8		0.8164	0.3124	0.4519	0.4591	0.5084	0.3521

Table 18: MSRA: Named Entity Recognition: Closed Track

Site	RunID	P	R	F	ORG-F	LOC-F	PER-F
10		0.9220	0.9018	0.9118	0.8590	0.9034	0.9604
14		0.9076	0.8922	0.8999	0.8397	0.9099	0.9261
11	b	0.8767	0.8753	0.8760	0.7611	0.8976	0.9225
11	a	0.8645	0.8399	0.8520	0.6945	0.8745	0.9199
32		0.8397	0.8184	0.8289	0.7261	0.8804	0.8207
7		0.8468	0.7822	0.8132	0.6958	0.8552	0.8280
6		0.8195	0.6926	0.7507	0.6443	0.8291	0.6955
30	a	0.8697	0.6556	0.7476	0.5841	0.7029	0.8987
8		0.8320	0.6703	0.7424	0.5651	0.8000	0.7565
12	b	0.7083	0.5464	0.6169	0.4168	0.6154	0.7171
12	a	0.7395	0.5186	0.6096	0.4168	0.6154	0.7074

Table 19: MSRA: Named Entity Recognition: Open Track

5 Conclusions & Future Directions

The Third SIGHAN Chinese Language Processing Bakeoff successfully brought together a collection of 29 strong research groups to assess the progress of research in two important tasks, word segmentation and named entity recognition, that in turn enable other important language processing technologies. The individual group presentations at the SIGHAN workshop detail the approaches that yielded strong performance for both tasks. Issues of out-of-vocabulary word handling, annotation consistency, character encoding and code mixing of Chinese and non-Chinese text all continue to challenge system designers and bakeoff organizers alike.

In future analyses, we hope to develop additional analysis tools to better assess progress in these fundamental tasks, in a more corpus independent fashion. Microsoft Research Asia has been pursuing work along these lines focusing on improvements in F-score and OOV F-score relative to more intrinsic corpus measures, such as baselines and toplines.[5] Such developments will guide the planning of future evaluations.

Finally, while word segmentation and named entity recognition are important in themselves, it is also important to assess the impact of improvements in these enabling technologies on broader downstream applications. More tightly coupled experiments that involve joint word segmentation and named entity recognition could provide insight. Integration of WS and NER with a higher level task such as parsing, reference resolution, or machine translation could allow the development of more refined, task-oriented metrics to evaluate WS and NER and focus attention on improvements to the fundamental techniques which enhance performance on higher level tasks.

GPE tags in the truth data mapped to LOC, since no GPE tags were present in the results.

[5]Personal communication, Mu Li, Microsoft Research Asia.

Acknowledgements

We gratefully acknowledge the generous assistance of the organizations and individuals listed below who provided the data for this bakeoff; without their support, it could not have taken place:

- Chinese Knowledge Information Processing Group, Academia Sinica, Taiwan: Keh-Jiann Chen, Henning Chiu
- City University of Hong Kong: Benjamin K. Tsou, Olivia Oi Yee Kwong
- Linguistic Data Consortium: Stephanie Strassel
- Microsoft Research Asia: Mu Li
- University of Pennsylvania/University of Colorado, USA: Martha Palmer, Nianwen Xue

We also thank Hwee Tou Ng and Olivia Oi Yee Kwong, the co-organizers of the fifth SIGHAN workshop, in conjunction with which this bakeoff takes place. Olivia Kwong merits special thanks both for her help in co-organizing this bakeoff and in coordinating publications. Finally, we thank all the participating sites who enabled the success of this bakeoff.

References

Thomas Emerson. 2005. The Second International Chinese Word Segmentation Bakeoff. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, Jeju Island, Republic of Korea.

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

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Richard Sproat and Thomas Emerson. 2003. The First International Chinese Word Segmentation Bakeoff. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, Sapporo, Japan.