Improved Statistical Machine Translation by Multiple Chinese Word Segmentation

Ruiqiang Zhang[1,2] and **Keiji Yasuda**[1,2] and **Eiichiro Sumita**[1,2]
[1]National Institute of Information and Communications Technology
[2]ATR Spoken Language Communication Research Laboratories
2-2-2 Hikaridai, Science City, Kyoto, 619-0288, Japan
{ruiqiang.zhang,keiji.yasuda,eiichiro.sumita}@atr.jp

Abstract

Chinese word segmentation (CWS) is a necessary step in Chinese-English statistical machine translation (SMT) and its performance has an impact on the results of SMT. However, there are many settings involved in creating a CWS system such as various *specifications* and *CWS methods*. This paper investigates the effect of these settings to SMT. We tested dictionary-based and CRF-based approaches and found there was no significant difference between the two in the qualty of the resulting translations. We also found the correlation between the CWS F-score and SMT BLEU score was very weak. This paper also proposes two methods of combining advantages of different specifications: a simple concatenation of training data and a feature interpolation approach in which the same types of features of translation models from various CWS schemes are linearly interpolated. We found these approaches were very effective in improving quality of translations.

1 Introduction

Chinese word segmentation (CWS) is a necessary step in Chinese-English statistical machine translation (SMT). The research on CWS independently from SMT has been conducted for decades. As an evidence, the CWS evaluation campaign, the Sighan Bakeoff (Emerson, 2005),[1], has been held four times since 2004. However, works on relations between CWS and SMT are scarce.

Generally, two factors need to be considered in constructing a CWS system. The first one is the *specifications* for CWS, i.e., the rules or guidelines for word segmentation, and the second one is the *CWS methods*. There are many CWS specifications used by different organizations. Unfortunately, these organizations do not seem to have any intention of reaching a unified specification. More than five or six specifications have been used in the four Sighan Bakeoffs. There is also significant disagreement on the specifications, although much of their contents is the same. One of the aims of this work was therefore to establish whether inconsistencies in specifications significantly affect the quality of SMT.

The second factor is CWS methods. We grouped all of the CWS methods into two classes: the class without out-of-vocabulary (OOV) recognition and the class with OOV recognition, represented by the dictionary-based CWS and the CRF-based CWS, respectively. Out-of-vocabulary recognition may have two-sided effects on SMT performance. The CRF-based CWS that supports OOV recognition produces word segmentations with a higher F-score, but a huge number of new words recognized correctly and incorrectly that can incur data sparseness in training the SMT models. On the other hand, the dictionary-based approach that does not support OOV recognition produced a lower F-score, but with a relatively weak data spareness problem. Which approach pro-

[1]A CWS competition organized by the ACL special interest group on Chinese.

Proceedings of the Third Workshop on Statistical Machine Translation, pages 216–223,
Columbus, June 2008.

Table 1: Examples of disagreement in segmentation guidelines

	ChineseName	EnglishName	Time
AS	DENGXIAOPING	*GEORGE BUSH*	1997YEAR 7MONTH 1DAY
CITYU	DENGXIAOPING	GEORGEBUSH	1997 YEAR 7 MONTH 1 DAY
MSR	DENGXIAOPING	GEORGEBUSH	1997YEAR7MONTH1DAY
PKU	*DENG XIAOPING*	GEORGEBUSH	1997YEAR 7MONTH 1DAY

Table 2: A second example of disagreement in segmentation guidelines

	Composite words	Composite words
AS	FUJITSUCOMPANY	EUROZONE
CITYU	*FUJITSU COMPANY*	EUROZONE
MSR	FUJITSUCOMPANY	*EURO ZONE*
PKU	*FUJITSU COMPANY*	EUROZONE

duces a better SMT result is our research interest in this work.

The performance of CWS is usually measured by the F-score, while that of SMT is measured using the BLEU score. Does a CWS with a higher F-score produce a better translation? In this paper we answer this question by comparing F-scores with BLEU scores.

In this work, we also propose approaches to make use of all the Sighan training data regardless of the specifications. Two methods are proposed: (1) a simple combination of all the training data, and (2) implementing linear interpolation of multiple translation models. Linear interpolation is widely used in language modeling for speech recognition. We interpolated multiple translation models generated by the CWS schemes and found our approaches were very effective in improving the translations.

2 CWS specifications and corpora from the second Sighan Bakeoff

A Chinese word is composed of one or more characters. There are no spaces between the words. Automatic word segmentation is required for machine translation. Usually a specification is needed to carry out word segmentation. Unfortunately, there are many different versions of specifications. Different tasks give rise to different requirements and the CWS specifications must be adjusted accordingly. For example, shorter segmentation has been shown to be better for speech recognition. A composite word (numbers, dates, times, etc.) is split into characters even if it is one word defined by linguists. In contrast, longer segmentation is preferred for named entity recognition consisting of longer character sequences, such as the name of people, places, and organizations.

This work investigated four well-known specifications created by four different organizations: Academia Sinica (AS), City University of Hong Kong (CITYU), Microsoft Research (Beijing) (MSR), and Beijing University (PKU). These specs were used in the second Sighan Bakeoff (Emerson, 2005). When we compared the four specifications and the manual segmentations in the Sighan Bakeoff training data, we found there were many inconsistencies among the four specifications. Some examples are shown in Table 1 and 2. For instance, the AS and PKU specifications are distinct in splitting both Chinese and English names. We also found the MSR specification generated more composite words and grouped longer character sequences into a word. Using this specification could generate tens of thousands of new words, which can cause data sparseness for SMT.

In addition to using the four specifications, we also downloaded the training and test corpora of the second Sighan Bakeoff. We used each of the training corpora provided to create a CWS scheme and evaluated the performance of the schemes on our test data. This enabled us to examine the effect of CWS specifications on SMT.

We used a Chinese word segmentation tool, Achilles, to implement word segmentation. Part of the work using this tool was described by (Zhang et al., 2006). The approach was reported to achieve the highest word segmentation accuracy using the data from the second Sighan Bakeoff. Moreover, this tool meets our need to test the effect of the two kinds of CWS approaches for SMT. We can easily train a dictionary-based and a CRF-based CWS by using this tool. By turning the program's option for the CRF model on and off, we can use the Achilles as a dictionary-based approach and as a CRF-based CWS. In fact, the dictionary-based approach is the default approach for Achilles.

3 Experiments

3.1 SMT resources

We followed the instructions for the 2005 NIST MT evaluation campaign. Training the translation models for our SMT system used the available LDC parallel data except the UN corpus. To train the language models for English, we used all the available English parallel data plus Xinhua News of the LDC Gigaword English corpus, LDC2005T12. In summary, we used 2.4 million parallel sentences for training the translation model. We used the test data defined in the NIST MT05 evaluation which is defined in the LDC corpus as LDC2006E38. We used the corpus, LDC2006E43, as the development data for loglinear model optimization.

We used a phrase-based SMT system that is based on a log-linear model incorporating multiple features. The training and decoding system of our SMT used the publicly available Pharaoh (Koehn et al., 2003)[2]. GIZA++ was used for word alignment.

The Pharaoh decoder was used exclusively in all the experiments. No additional features but the defaults defined by Pharaoh were used. The feature weights were optimized against the BLEU scores (Och, 2003).

We chose automatic metrics to evaluate CWS and SMT. We used the F-score for CWS and BLEU for SMT. The BLEU is BLEU4, computed using the NIST-provided "mt-eval" script.

[2]http://www.iccs.informatics.ed.ac.uk/~pkoehn

3.2 Implementation of CWS schemes

To determine the effect of CWS on SMT, we created 14 CWS schemes which are shown in Table 3. Schemes 1 to 12 were implemented using the in-house tool, Achilles, and schemes 13 and 14 using off-the-shelf tools. The CWS schemes are named according to the specifications (AS, CITYU, MSR, PKU), implementing methods (CRF-based or dictionary-based), and lexicon sources (Sighan or LDC corpus). The table also shows the results of segmentation on the SMT training and test data, i.e., number of total tokens, unique words, and OOV words.

We divided the schemes into two groups for simplicity. The first group includes schemes 1 to 12, which were trained using a specific Sighan corpus. For example, schemes 1 to 3 were trained using the AS corpus, schemes 4 to 6 using the CITYU corpus, and so on. The meaning of the name of the CWS scheme can be derived from the table – the name is defined by specifications, methods and lexicon sources. For example, the CRF-AS scheme performs CRF-based segmentation; and its lexicon is from the AS corpus provided by the Sighan. The CRF-AS segmenter can be easily trained, as described by Achilles.

The second group contains two schemes 13 and 14. The ICTCLAS is a HHMM-based hierarchical HMM segmenter (Zhang et al., 2003) that uses the specifications of PKU. This segmenter incorporates parts-of-speech information in the probability models and generates multiple HMM models for solving segmentation ambiguities. The MSRSEG was developed by Gao et al. (Gao et al., 2004). This segmenter is based on the MSR specifications. It uses a log-linear model that integrates multiple features.

The segmenters of the first group, dict-AS and dict-LDC-AS, are two dictionary-based CWS schemes. They differ in lexicon size and lexicon extracting source. The former used a lexicon extracted directly from the Sighan AS training data while the latter used a lexicon from LDC parallel corpora. It took some efforts to get the lexicon. First, we used the CRF-AS to segment the LDC corpora. We extracted a unique word list from the segmented data and sorted it in decreasing order according to word frequency. Because OOV was recognized by

Table 3: Analysis of results of segmentation on LDC training and test data for all CWS schemes

No.	CWS schemes	Specifications	Methods	Lexicon	Tokens	Unique words	OOVs
1	CRF-AS	AS	CRF	Sighan	47,934,088	413,588	1,193
2	dict-AS	AS	Dict	Sighan	51,664,675	89,346	237
3	dict-LDC-AS	AS	Dict	LDC	48,665,364	102,919	273
4	CRF-CITYU	CITYU	CRF	Sighan	47,963,541	426,273	1,155
5	dict-CITYU	CITYU	Dict	Sighan	51,251,729	56,996	362
6	dict-LDC-CITYU	CITYU	Dict	LDC	48,787,154	102,754	217
7	CRF-MSR	MSR	CRF	Sighan	46,483,923	523,788	1,297
8	dict-MSR	MSR	Dict	Sighan	51,302,509	60,247	248
9	dict-LDC-MSR	MSR	Dict	LDC	47,469,271	102,390	217
10	CRF-PKU	PKU	CRF	Sighan	48,022,697	440,114	1,136
11	dict-PKU	PKU	Dict	Sighan	52,721,809	47,176	211
12	dict-LDC-PKU	PKU	Dict	LDC	48,721,795	102,213	256
13	ICTCLAS	PKU	HHMM	-	50,751,402	162,222	835
14	MSRSEG	MSR	-	-	48,734,113	274,411	1,443

the CRF-AS, a huge word list was generated(see Table 3). We chose the most frequent 100,000 words as the dictionary for the dict-LDC-AS [3]. The LM for the dict-AS was trained using the AS corpus while the LM for the dict-LDC-AS was trained using the segmented SMT training corpus.

Therefore, the dict-LDC-AS used a larger lexicon than the dict-AS. This lexicon contained the most frequent OOV words recognized by the CRF-AS. Our aim was to investigate whether the dict-LDC-AS, whose lexicon consisted of the lexicon of dict-AS and new words recognized by CRF-AS, could improve SMT.

As shown in Table 3, using CRF-AS generated a huge number of unique words for the training data and OOV words for the test data. We found that the CRF-AS generated three times more OOVs for the test data than the dictionary-based CWS,dict-AS (see OOVs in Table 3).

Other schemes in the first group were implemented similarly to the "AS".

Table 3 lists the segmentation statistics for the training and test data of all the tested CWS schemes, where "Tokens" indicates the total number of words in the training data. "Unique words" and "OOVs" mean the lexicon size of the segmented training data and the unknown words in the test data, respectively.

[3]Only those words that appeared at least five times in the lexicon were considered.

Table 4: BLEU scores for CWS schemes

CWS	AS	CITYU	MSR	PKU
CRF	23.70	23.55	22.50	23.61
dict	23.46	23.72	23.33	23.61
dict-LDC	23.52	23.36	23.16	23.74
ICTCLAS	-	-	-	24.12
MSRSEG	-	-	19.72	-
BEST	23.70	23.72	23.33	23.74 (24.12)

3.3 Effect of CWS specifications on SMT

Our first concern was the effect of CWS specifications on SMT. The results in Table 4 show the relationships that were found. The last row gives the best BLEU scores obtained for each of the CWS specifications. The scores for AS, CITYU, MSR and PKU were 23.70 (CRF-AS), 23.72 (dict-CITYU), 23.33 (dict-MSR) and 23.74 (dict-PKU-LDC), respectively. We found there were no observable differences between AS, CITYU, and PKU. However, the specification that produced the worst translations was the MSR. The MSR specification appears to have been designed for recognizing named entities (NE) (See the examples of segmentation in Table 1). Many NEs are regarded as words by MSR, while they are more appropriately split into separate words by other specifications. For example, the long word, "1997YEAR7MONTH1DAY" ("July 1, 1997"). As a result, the CRF-MSR generated 20% more words in the vocabulary than the other CWS schemes in segmenting the SMT training data. The larger vocabulary can trigger data sparseness problems and result in SMT degradation. The segmenter, MSRSEG, produced an even lower BLEU score (19.72) than the Achilles.

The results were verified by significance test (Zhang et al., 2004). We found the systems with the BLEU scores higher than 23.70 were significantly better than those lower than 23.70.

3.4 Correlation between BLEU score and F-score

The values of the F-scores and BLEU scores are listed in parallel in Table 5. We tied the F-scores and specifications together because comparing the value of the F-score across specs is meaningless. We separated the F-score and BLEU score for different corpus. The F-score was calculated using the Sighan test data. The CRF-based approach usually gives a higher F-score, but its corresponding BLEU scores were not always higher. The F-score and BLEU score correlated well for ICTCLAS and CRF-AS but less well for CRF-CITYU, CRF-PKU and CRF-MSR. Obviously, there is no strong correlation between the F-score and BLEU score.

4 Effect of combining multiple CWS schemes

We used the Sighan Bakeoff corpora of different CWS specifications separately in the previous experiments. Here, we propose two approaches to using all the resources combined. The first approach is to concatenate all the training data of the Sighan Bakeoff, regardless of the specifications and training a new CWS for segmenting SMT training data. The second approach involves linear integration of translation models. We found that both approaches produced an improvement in translation quality.

4.1 Effect of combining training data from multiple CWS specifications

The CWS specifications are very different and the corresponding Sighan training data are segmented in different ways. We used these data separately in the previous work as if they were incompatible. However, creating data manually is laborious and costly. It would therefore be a significant advantage if all the data could be used, regardless of the different specifications. We therefore created a new CWS scheme, called "dict-hybrid". This CWS was trained by concatenating all the Sighan Bakeoff corpora regardless of the different specifications. The "dict-hybrid" was trained using Achilles. It uses a dictionary-based approach, and its lexicon and language model were obtained as follows.

First, we created a hybrid corpus by combining all the Sighan training corpora: AS, CITYU, MSR, PKU. The hybrid corpus was used to train a CRF-based CWS. This CWS was then used to segment the SMT training corpus and then we extracted a lexicon of 100,000 from the top frequent words of the segmented SMT corpus. This lexicon was used as the lexicon of the "dict-hybrid." The LM of "dict-hybrid" was also trained on the segmented corpus. Note a lexicon and a LM are the only needed resources for building a dictionary-based CWS, like the "dict-hybrid." (Zhang et al., 2006)

We used the "dict-hybrid" to segment the SMT training corpus and test data. This segmentation generated 49,546,231 tokens, 112,072 unique words for the training data and 693 OOVs for the test data.

The segmentation data were used for training a new SMT model. We tested the model using the same approach and found the BLEU score obtained by this CWS scheme was 23.91. This score was better than those in Table 4 obtained by any of the Achilles CWS schemes except ICTCLAS. Therefore, the CWS scheme "dict-hybrid" produced better translations than other schemes implemented using Achilles, indicating that using multiple CWS corpora can improve SMT even if their specifications are different.

Significance testing also showed that the results for ICTCLAS and "dict-hybrid" were not significantly different. The results of "dict-hybrid" are significantly better than those in the Table 4 which have

Table 5: Correlation between F-score and BLEU

PKU		
	F-score	BLEU
CRF	0.939	23.61
dict	0.930	23.61
dict-LDC	0.931	23.74
ICTCLAS	0.948	24.12

MSR		
	F-score	BLEU
CRF	0.954	22.50
dict	0.947	23.22
dict-LDC	0.928	23.16
MSRSEG	0.969	19.72

CITYU		
	F-score	BLEU
CRF	0.920	23.55
dict	0.873	23.72
dict-LDC	0.886	23.36

AS		
	F-score	BLEU
CRF	0.922	23.70
dict	0.896	23.46
dict-LDC	0.878	23.52

a BLEU score lower than 23.70.

4.2 Effect of feature interpolation of translation models

We investigated the effect of linearly integrating multiple features of the same type. We generated multiple translation models by using different word segmenters. Each translation model corresponded to a word segmenter. The same type of features as in the log-linear model were added linearly. For example, the phrase translation model $p(e|f)$ can be linearly interpolated as, $p(e|f) = \sum_{i=1}^{S} \alpha_i p_i(e|f)$ where $p_i(e|f)$ is the phrase translation model corresponding to the i-th CWSs. α_i is the weight, and S is the total number of models. $\sum_{i=1}^{S} \alpha_i = 1$.

αs can be obtained by maximizing the likelihood or BLEU scores of the development data. Optimizing the α has been described elsewhere (Foster and Kuhn, 2007). $p(e|f)$ is the phrase translation model generated.

In addition to the phrase translation model, we used the same approach to integrate three other features: phrase inverse probability $p(f|e)$, lexical probability $lex(e|f, a)$, and lexical inverse probability $lex(f|e, a)$.

We integrated the CWS schemes ranked in the top five in Table 4: ICTCLAS, dict-hybrid, dict-LDC-PKU, dict-CITYU, and CRF-AS. We labeled the five schemes A, B, C, D, and E, respectively, as shown in Table 6. The first line of Table 6 represents the test data segmented by the five CWS schemes. "tst-A" means the test data was segmented by ICTCLAS. "tst-B" means the test data segmented by "dict-hybrid", and so on. The second line gives baseline results showing the original results without the use of feature integration. For different test data, the baseline is different. The baseline of ICTCLAS was tested on "tst-A" only. The baseline of "dict-hybrid" was tested on "tst-B" only. From the third line we gradually added a translation model to the models used in the baseline. For example, "A+B" integrates models made using ICTCLAS and "dict-hybrid." Each integration models were tested only on the test data participated in the integration. Hence, some slots in Table 6 are blank.

We did not carry out parameter optimization with regards to the αs. Instead, we used equal αs for all the features. For example, all αs equal 0.5 for A+B, and 0.25 for A+B+C+D. Each cell in Table 6 indicates the BLEU score of the integration in relation to the test data. We found our approach improved the baseline results significantly. The more models integrated, the better the results. The improvement was positive for all of the test data. With regards to the integration, if a phrase pair exists in one model only, we suppose the values of probabilities are zero in other models.

To better understand the effects of feature interpolation, we blended the features of the translation models, as shown in Table 7, by simply combining the phrase pairs without probability interpolation. When we merged two models, we defined one model as the master model and the other as the supplementary model. Only phrase pairs that were in the supplementary models but not in the master model were appended to the master model. Their feature probabilities were not changed. Hence, the combined model was a blend of phrase pairs from the master model and supplementary model. There was no probability integration, that was significantly different from the feature interpolation approach. For each set of test data in Table 7, the master model was the model using the same CWS as the test data. While there was one row for each type of combination, the cells in the row contained different models. For example, "A+B" for test data "A" uses "A" as the master model and "B" as the supplementary model, while the opposite holds for test data "B".

Comparing Table 6 and 7 showed that feature interpolation outperformed feature blending. Feature interpolation yielded surprisingly good results. The performance consistently improved when more models were integrated, but this was not the case for feature blending. This shows that probability integration is very effective. Increasing the size of phrase pairs, as feature blending does, is not as effective.

We used equal values for the αs. Optimal values may be obtained using the optimization approach of maximizing BLEU or the likelihood of development data as has been reported previously (Foster and Kuhn, 2007). However, optimization is computationally expensive and the effect was not satisfactory. Therefore, we decided not optimizing the αs in this work.

5 Related work and Discussions

CWS has been the subject of intensive research in recent years, as is evident from the last four international evaluations, the Sighan Bakeoffs, and many approaches have been proposed over the past decade. Segmentation performance has been improved significantly, from the earliest maximal match (dictionary-based) approaches to CRF (Peng and McCallum, 2004) approach. We used dictionary-based and CRF-based CWS approaches to demonstrate the effect of CWS on SMT, both without and with OOV recognition.

SMT is a very complicated system to study. Its response to CWS schemes is intractable and it is very hard to use one or two measures to describe the relationship between CWS and SMT, in a similar way to describing the relationship between the alignment error rate (AER) and SMT (Fraser and Marcu, 2007). The CWS and SMT are related by a series of factors such as the specifications, OOVs, lexicons, and F-scores. None of these factors can be directly related to the SMT. While we have completed many experiments, based on changing the CWS specifications and methods used, to determine the relationship between CWS and SMT, we have not established any overwhelming rules. However, we believe the following guidelines are appropriate in considering a CWS system for SMT. Firstly, the F-score is not a reliable guide to SMT quality. A very high F-score may produce the lowest quality translations, as was found for the MSRSEG. Secondly, it is better to design a specification with smaller word units to reduce data sparseness. Specifications like those for MSR will produce an inferior translation. Thirdly, do not use a huge lexicon for word segmentation. A huge lexicon will result in data sparseness and segmentation complexity. And lastly, using multiple word segmentation results and approaches does work. We used two approaches that combined multiple word segmentation - dict-hybrid and feature integration - and both improved the translations significantly.

The BLEU scores in our experiments were relatively low in comparison with current state-of-the art results. However, our system was very similar to the system (Koehn et al., 2005) that gave a BLEU score of 24.3, comparable to ours. The BLEU score can be raised if we do post-editing, use more data for language modeling and other methods.

6 Conclusions

We investigated the effect of CWS on SMT from two points of view. Firstly, we analyzed multiple CWS specifications and built a CWS for each one to examine how they affected translations. Secondly, we investigated the advantages and disadvantages of various CWS approaches, both dictionary-based and CRF-based, and built CWSs using these approaches to examine their effect on translations.

We proposed a new approach to linear interpolation of translation features. This approach produced a significant improvement in translation and

Table 6: Feature interpolation of translation models: A=ICTCLAS, B=dict-hybrid, C=dict-PKU-LDC, D=dict-CITYU, E=CRF-AS

Model	tst-A	tst-B	tst-C	tst-D	tst-E
Baseline	24.12	23.91	23.74	23.72	23.70
A+B	24.25	24.20			
A+B+C	24.49	24.31	23.84		
A+B+C+D	24.60	24.43	24.05	24.27	
A+B+C+D+E	24.61	24.55	24.16	24.39	24.17

Table 7: Feature blending of translation models

Model	tst-A	tst-B	tst-C	tst-D	tst-E
Baseline	24.12	23.91	23.74	23.72	23.70
A+B	24.20	24.24			
A+B+C	24.27	24.14	23.69		
A+B+C+D	23.92	24.29	23.61	24.00	
A+B+C+D+E	23.86	24.31	23.69	24.05	23.76

achieved the best BLEU score of all the CWS schemes.

We have published a much more detailed paper (Zhang et al., 2008) to describe the relations between CWS and SMT.

References

Thomas Emerson. 2005. The second international chinese word segmentation bakeoff. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, Jeju, Korea.

George Foster and Roland Kuhn. 2007. Mixture-model adaptation for SMT. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 128–135, Prague, Czech Republic, June. Association for Computational Linguistics.

Alexander Fraser and Daniel Marcu. 2007. Measuring word alignment quality for statistical machine translation. In *Computational linguistics, Squibs Discussion*, volume 33 of *3*, pages 293–303, September.

Jianfeng Gao, Andi Wu, Mu Li, Chang-Ning Huang, Hongqiao Li, Xinsong Xia, and Haowei Qin. 2004. Adaptive chinese word segmentation. In *ACL-2004*, pages 462–469, Barcelona, July.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *HLT-NAACL 2003: Main Proceedings*, pages 127–133.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Miles Osborne Chris Callison-Burch, David Talbot, and Michael White. 2005. Edinburgh system description for the 2005 nist mt evaluation. In *Proceedings of Machine Translation Evaluation Workshop*.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167.

Fuchun Peng and Andrew McCallum. 2004. Chinese segmentation and new word detection using conditional random fields. In *Proc. of Coling-2004*, pages 562–568, Geneva, Switzerland.

Huaping Zhang, HongKui Yu, Deyi xiong, and Qun Liu. 2003. HHMM-based Chinese lexical analyzer ICTCLAS. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, pages 184–187.

Ying Zhang, Stephan Vogel, and Alex Waibel. 2004. Interpreting BLEU/NIST scores: How much improvement do we need to have a better system? In *Proceedings of the LREC*.

Ruiqiang Zhang, Genichiro Kikui, and Eiichiro Sumita. 2006. Subword-based tagging by conditional random fields for chinese word segmentation. In *Proceedings of the HLT-NAACL, Companion Volume: Short Papers*, pages 193–196, New York City, USA, June.

Ruiqiang Zhang, Keiji Yasuda, and Eiichiro Sumita. 2008. Chinese word segmentation and statistical machine translation. *ACM Trans. Speech Lang. Process.*, 5(2), May.