Introduction to China's CWMT2008 Machine Translation Evaluation

Hongmei Zhao, Jun Xie, Qun Liu, Yajuan Lü
Key Laboratory of Intelligent Information Processing,
Institute of Computing Technology,
Chinese Academy of Sciences
No. 6 Kexueyuan South Road, Zhongguancun,
Haidian District,
Beijing, China 100190
{zhaohongmei,xiejun,liuqun,lvyajuan}@ict.ac.cn

Dongdong Zhang, Mu Li
Microsoft Research Asia
5F, Sigma Center, No. 49 Zhichun Road,
Haidian District,
Beijing, China 100190
{dozhang, muli}@microsoft.com

Abstract

This paper presents an overall introduction to the CWMT2008 evaluation and focuses on its two new metrics: BLEU-SBP (Chiang et al., 2008) and linguistic check-point method (Zhou et al., 2008). BLEU-SBP is a revised BLEU with strict brevity penalty. Our experiments validated BLEU-SBP's effectivity in resolving the nondecomposability problem of both NIST-BLEU and IBM-BLEU at sentence level. Linguistic check-point method (LCM) is a linguistic diagnostic evaluation method based on automatically constructed linguistic check-points, and our evaluation indicates that this method can be used to evaluate the capability of an MT system in translating various linguistic phenomena, and revealed good correlations between BLEU score and LCM scores in most tasks. With the aid of these metrics, we disclosed some detailed performance differences between statistical MT systems and rule-based MT systems. In addition, through the study on some practical cases we suggest that the high BLEU score doesn't necessarily mean high translation adequacy.

1 Introduction

China Workshop on Machine Translation 2008 (CWMT2008) Evaluation continues the ongoing series of evaluation of machine translation technology in China. Its predecessor is SSMT (Symposium on Statistical Machine Translation) Evaluation. CWMT evaluation series have been organized by the Institute of Computing Technology, Chinese Academy of Sciences.

There are four evaluation tracks in CWMT2008 evaluation, namely Chinese-to-English news machine translation, Chinese-to-English news system combination, English-to-Chinese news machine translation and English-to-Chinese scientific and technical literature machine translation.

We organized a system combination track in CWMT2008 evaluation, since system combination research has intensified over the past several years as significant performance gains have been achieved through various combination techniques (NIST, 2009). But can these techniques really result in a better translation quality rather than only a higher BLEU score? Callison-Burch et al. (2009) used the results of the manual evaluation to analyze the translation quality of the different systems that were submitted to the WMT09 and found that in general, system combinations performed as well as the best individual systems, but not statistically significantly better than them. Some practical cases in our evaluation suggest that the higher BLEU score doesn't always mean higher translation adequacy no matter in single system MT task or in system combination task.

For the evaluation measurement, we choose two novel metrics as our alternatives: BLEU-SBP (Chiang et al., 2008) and linguistic check-point method (Zhou et al., 2008).

We encountered two actual cases which happened to be very similar with those in (Chiang et al., 2008). These cases can be traced to the fact that BLEU (Papineni et al., 2002) is not decomposable at the sentence level, which means if a system generates a long translation for one sentence, it can generate a short translation for another sentence without facing a penalty. Our experiments validated BLEU-SBP's effectivity in resolving the nondecomposability problem of both NIST-BLEU and IBM-BLEU at sentence level.

On the other hand, we choose linguistic check-point method (LCM) as another alternative metric with an attempt to detect and report richer linguistic information on the system. Now most MT evaluation methods only generate a general similarity score. At the present time, there is no single metric that has been deemed to be completely indicative of all aspects of system performance (NIST, 2008). As a completely different metric, LCM (which is implemented in a platform called WoodPecker) can give scores of different linguistic categories to an MT system rather than a single general score, which helps us to dig into the multiple linguistic levels and find the concrete strength and flaws of the system and compare the systems with different architectures or systems with similar general scores (Zhou et al., 2008). With the aid of LCM, we succeeded in disclosing the latent linguistic differences of statistical MT systems (SMT) and rule-based MT (RBMT) systems.

In the next section, we give an overall introduction to the CWMT2008 evaluation. In Section 3 and 4, we introduce BLEU-SBP, LCM and their results in our evaluation. In Section 5, we present the performance differences of statistical MT systems and rule-based MT systems under different tasks and metrics. In Section 6, we use some practical cases to state that higher BLEU score doesn't necessarily mean higher translation adequacy. Section 7 is the conclusion.

2 Overall Introduction to the CWMT2008 Evaluation

2.1 Evaluation Tracks

There are four tracks in CWMT2008 evaluation. Table 1 gives the evaluation tracks.

Language	Domain	Task
Chinese to English	News	Machine Translation
Chinese to English	News	System Combination
English to Chinese	News	Machine Translation
English to Chinese	S&T	Machine Translation

Table 1. Evaluation Tracks. S&T= Scientific and Technical Literature.

2.2 Participants and Primary Systems

There are 15 participants. Among these participants, some are from institutes and universities such as Chinese Academy of Sciences, Harbin Institute of Technology, some are from companies such as SYSTRAN Software, Inc. and Microsoft Research Asia. For each evaluation track, every participant should submit one primary result, and at most two contrast results. Table 2 gives the number of the primary systems of every track and the amount of primary systems of different architectures.

Track	# of P	# of SMT	# of RBMT
C2E News Translation	12	9	3
C2E System Combination	6	6	
E2C News Translation	11	6	5
E2C S&T Translation	9	6	3
Total	38	27	11

Table 2. Situation of the Primary Systems. P=primary systems.

2.3 Evaluation Data

2.3.1 Evaluation Data for MT Tracks

Training Data: We provided for the participants training data, which has 868,947 Chinese-English sentence pairs for the news domain and 620,985 Chinese-English sentence pairs for the scientific and technical literature domain, in addition the participants in the latter domain can also use the training data for the news domain. The participants were allowed to use the data not included in the training data list that we provided, however,

those using out-of-list data were marked in the evaluation result reports.

Test Data: The test data are collected from two domains: news and scientific & technical literature. For each MT source text, we prepared four references which were translated by different translators independently. Table 3 gives the size of the source text for each track.

Source language	domain	# of Chinese characters or English words
Chinese	News	41042
English	News	21767
English	S&T	13050

Table 3. CWMT2008 Evaluation Test Data. S&T= Scientific and Technical Literature.

2.3.2 Evaluation Data for System Combination Track

Test Data: When the MT task was finished, we collected all the N-best translations submitted by the participants of C2E news MT task. We used these translation results as the test data for system combination track, and sent them to the participants of system combination track, who performed combinations on these data and submit the combined results to us.

Development Data: Besides submitting the results on the CWMT2008 evaluation data, the participants of the C2E news MT task had also been asked to submit the results of the same participating system on the C2E news MT test data of SSMT2007 evaluation, which would be given to the participants of system combination task along with reference translations as the development data.

2.4 Performance Measurement

Besides the new BLEU-SBP and LCM, the automatic evaluation metrics of CWMT2008 evaluation also include: BLEU, NIST, GTM, mWER, mPER and ICT (a metric developed by the Institute of Computing Technology, CAS). All these metrics are case-sensitive. The evaluation of Chinese translation is based on Chinese characters instead of words.

2.5 Evaluation Results

Figures 1-4 show the evaluation results.

3 BLEU-SBP

3.1 BLEU's Deficiency

We encountered the following two practical cases in our evaluation.

In this paper, if not specified particularly, all the BLEU means NIST-BLEU.

3.1.1 The Sign Test

When we applied the sign test (Collins et al., 2005) for significance testing with BLEU, we encountered such problem (Table 4): when comparing system A and system B, if we select A as the baseline system, we found B is significantly better than A, but if we select B as the baseline system, we found A is significantly better than B. We tried two kinds of BLEU: NIST-BLUE and IBM-BLEU, the results are similar (Table 4). This is because the sign test requires a function (a_i,b_i) that indicates whether b_i is better, worse or same quality translation relative to a_i. Because BLEU is not defined on single sentences, Collins et al. (2005) use an approximation: for each i, form a composite set of outputs $a'=\{a_1,\ldots,a_{i-1}, b_i,a_{i+1},\ldots,a_n\}$, and compare the BLEU scores of a and a' . Because BLEU scores are highly context-dependent, for example, if the sentences in a are on average ε words longer than the reference sentences, then b_i can be as short as (N-1) ε words shorter than r_i without incurring the brevity penalty. Moreover, since the b_i is substituted in one at a time, we can do this for all of the b_i. Hence, b could have a disastrously low BLEU score (because of the brevity penalty) yet be found by the sign test to be significantly better than the baseline a. (Chiang et al., 2008).

System	A	B
NIST-BLEU	0.2611++	0.2532++
BP	1	0.9652
IBM-BLEU	0.2611++	0.2437++
BP	1	0.9289
BLEU-SBP	0.2579	0.2417
SBP	0.9877	0.9213

Table 4. Sign Test Experiment with NIST-BLEU, IBM-BLEU and BLEU-SBP. ++ represents significant improvement ($P<0.01$), significances are relative to other system. There are no significant differences in BLEU-SBP line. BP=brevity penalty, SBP=strict brevity penalty in BLEU-SBP.

3.1.2 Word and Sentence Deletion

We fabricated a C2E news translation result using one of the four references, deleted words to the degree that only 2-3 words remained in a sentence for some sentences ("word deletion") or deleted all the words of some sentences ("sentence deletion"), and found even deleting 10% words or sentences of this fabricated translation didn't result in any decrease of the NIST-BLEU score (Table 5). These results are also very similar to those of (Chiang et al., 2008). We tried the IBM-BLEU, and the results are more rational.

deletion	word	word	sentence	sentence
del%	2	10	2	10
NIST-BLEU4	1	1	1	1
IBM-BLEU4	0.9825	0.9014	0.9811	0.8933
BLEU4-SBP	0.9798	0.8862	0.9782	0.8769

Table 5. Word and Sentence Deletion. del%=percentage of words or sentences deleted.

3.2 Brief Introduction of BLEU-SBP

The cause of the above problems is that in the brevity penalty (which can be regarded as a stand-in for recall) of BLEU, the per-sentence score $\frac{|c_i|}{|r_i|}$ exceed unity, so Chiang made a simple fix of clipping the per-sentence recall scores in a similar fashion to the clipping of precision scores:

$$bp(\mathbf{c},\mathbf{r}) = \phi\left(\frac{\sum_i \min\{|c_i|,|r_i|\}}{\sum_i |r_i|}\right)$$

He used the above bp to replace NIST-BLEU's bp:

$$bp(\mathbf{c},\mathbf{r}) = \phi\left(\min\left\{1, \frac{\sum_i |c_i|}{\sum_i |r_i|}\right\}\right)$$

where

$$\phi(\chi) = \exp(1 - 1/x),$$

and he called this revised metric BLEU-SBP (for BLEU with strict brevity penalty). We tested BLEU-SBP in the problem cases described above (Tables 4, 5) and got rather rational results, which further proves the BLEU-SBP is effective in reducing the BLEU's nondecomposability at the sentence level.

3.3 Evaluation Results: a Comparison between BLEU and BLEU-SBP

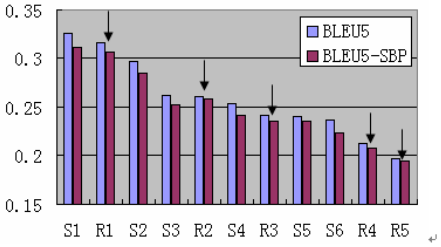

Figure 1. E2C News Translation Results.

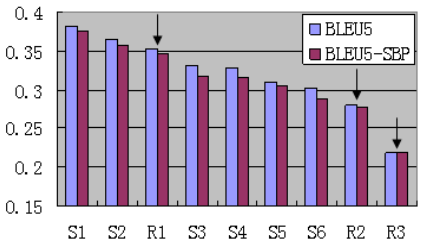

Figure 2. E2C S&T Translation Results.

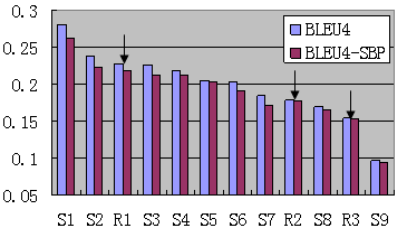

Figure 3. C2E News Translation Results.

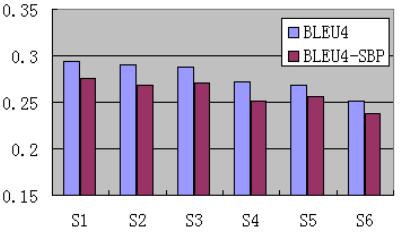

Figure 4. C2E System Combination Results.

Key: S=SMT system, R=RBMT system. The RBMT systems are distinguished from the SMT systems by arrows.

From above figures, we can see, in terms of BLEU-SBP, all the systems' scores have reduced because of the strict brevity penalty. A notable difference is that the RBMT systems (marked with arrows) are generally punished more slightly than the SMT systems, which is due to the fact that the SMT system is trained toward to the

maximum BLEU and the language model (LM) prefers shorter translation sentence. Table 6 presents a representative case.

Ref1	Ref2	Ref3	Ref4	SMT1	RBMT3
27181	27731	27304	25222	21524	29871

Table 6. Total Numbers of Words of Four References, One SMT System and One RBMT System in C2E News Translation Task.

In Table 6, SMT1 is a statistical machine translation system with the highest BLEU score and the biggest score decrease (nearly 2 percent point) from BLEU to BLEU-SBP, we'll discuss this system in Section 6, RBMT3 is a rule-based translation system with the least score decrease (only 0.1 percent point).

4 Linguistic Check-points Method

Inspired by (Yu, 1993), Zhou et al. (2008) proposed a linguistic check-point based automatic evaluation method (LCM), which currently can diagnose both Chinese-to-English and English-to-Chinese machine translation systems. Different from Yu's method, LCM constructs check-points automatically rather than manually. Meanwhile, LCM can distinguish machine translation systems more delicately as its evaluation score is computed based on n-gram partial matching rate instead of hard binary score. LCM regards a sentence as a collection of check-points with different types, defined by a linguistic taxonomy, hence it can reveal much linguistic information about the evaluated system by assigning scores to linguistic categories.

The process of this diagnostic evaluation consists of two main steps: check-point extraction and check-point evaluation. The purpose of check-point extraction is to extract the check-points automatically with the aid of word aligner and parser, the check-point evaluation aims to generate the diagnostic report by computing the matching rate between the candidate translation and the references of check-points. For more details, see Section 2 of (Zhou et al., 2008).

4.1 Applying Linguistic Check-points Method in CWMT2008 Evaluation

We use Woodpecker[1] as the linguistic check-points evaluation tool. To get the word alignment, dependency and constituent structures required by Woodpecker, we used the following tools: GIZA++ (Och et al., 2003) for word alignment, Stanford Parser for dependency and constituent structures, Berkeley Parser for constituent structure, and ICTCLAS (Zhang et al., 2003) for Chinese word segmentation. With the information acquired by above tools, linguistic check-points can be extracted automatically. Then the candidate translations are evaluated with Woodpecker by being matched to the references of linguistic check-points extracted.

4.2 Evaluation Results: Correlations between Scores of LCM and BLEU

The taxonomy includes typical check-points at word, phrase and sentence levels. The Figures 5-7 show BLEU score, the LCM's general score and these three levels' scores in three tasks. We calculated the correlations between BLEU and these LCM's scores of all primary systems with Spearman coefficient and Pearson coefficient (Table 7).

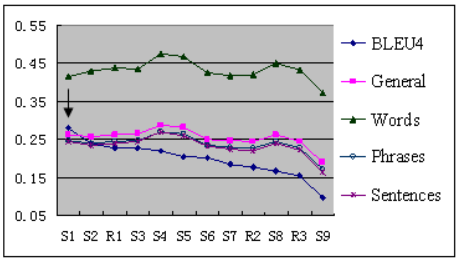

Figure 5. Scores of LCM and BLEU in C2E News Translation Task.

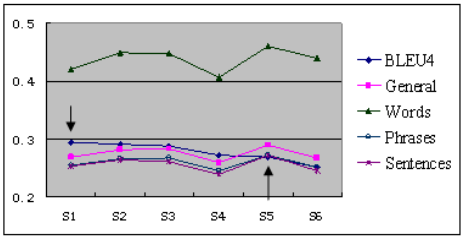

Figure 6. Scores of LCM and BLEU in C2E System Combination Task.

[1] http://research.microsoft.com/en-us/downloads/ad240799-a9a7-4a14-a556-d6a7c7919b4a/default.aspx

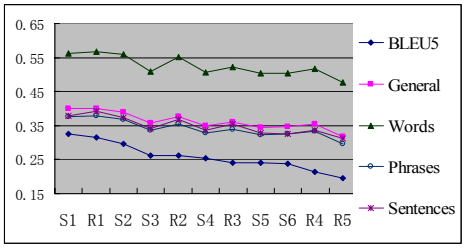

Figure 7. Scores of LCM and BLEU in E2C News Translation Task.

Task	N	Score	Spearman	Pearson
ce_ news	12	General	0.5594	0.7182
		Words	0.1678	0.4138
		Phrases	0.6224	0.7425
		Sentences	0.6923	0.7607
ce_ news_ comb	6	General	0.0857	0.2232
		Words	-0.1429	-0.0828
		Phrases	0.0857	0.2682
		Sentences	0.1429	0.3263
ec_ news	11	General	0.8727	0.9283
		Words	0.8273	0.8747
		Phrases	0.8727	0.9340
		Sentences	0.9182	0.8929
ec_ s&t	9	General	0.8500	0.9429
		Words	0.8333	0.9452
		Phrases	0.8500	0.94214
		Sentences	0.8167	0.9186

Table 7. Correlations between Scores of LCM and BLEU. N is the number of primary systems evaluated.

From Table 7, we can find good correlations between scores of LCM and BLEU in MT tasks. Considering that Woodpecker using n-gram matching strategy similar to that of BLEU, this result is rational.

We can also find the correlation for "ce-news" MT task is lower than those for "ec-news" and "ec-s&t" MT tasks. One possible reason is that the SMT1 (Table 6) with the highest BLEU score in "ce_news" MT task has rather low LCM scores (S1 in Figure 5, marked with arrow). We also notice that in the "ce_news_comb" task, the correlation between scores of LCM and BLEU is destroyed by S1 and S5 (Figure 6, marked with arrows). We'll discuss these phenomena in Section 6.

5 Comparison of Performances of SMT Systems and RBMT Systems

5.1 Performance Differences on Different Evaluation Tasks

From Figures 1-3, we can find, in terms of BLEU (LCM scores have the same tendency because of the high correlation with BLEU), the SMT systems have some advantage over the RBMT systems (marked with arrows) on all machine translation tasks.

5.2 Performance Differences under Different Metrics

BLEU-SBP: From Figure1-3, we can see that without the training course maximizing BLEU and the impact of linguistic models, the RBMT systems generally got less brevity penalty than the SMT systems, the RBMT systems' BLEU and BLEU-SBP scores are closer than those of the SMT systems.

LCM: From Figures 8-9, we can find that there are obvious differences in some categories of check-points according to their scores between the SMT systems and the RBMT systems (marked with arrows), such as in the C2E news translation task, the SMT systems perform better than the RBMT systems in "collocation" and "preposition" categories of the source language, and the RBMT systems perform obviously better than the SMT in "idiom" category of the source language; In E2C news translation task, the RBMT systems perform much better than the SMT systems in "preposition in dictionary source" and "quantity phrase" of the source language and "preposition" of the target language. These obvious differences prove the LCM can dig into the multiple linguistic levels and disclose the latent differences of the systems with different architectures.

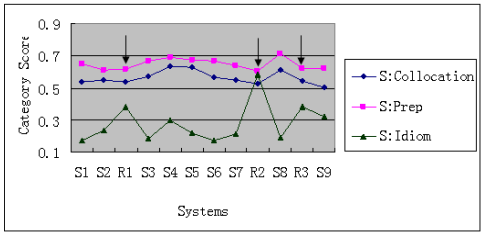

Figure 8. Comparison of Performances of SMT and RBMT Systems in C2E News Translation Task.

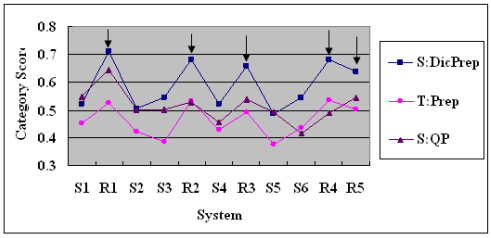

Figure 9. Comparison of Performances of SMT and RBMT Systems in E2C News Translation Task.

6 Case Study: High BLEU but Low Adequacy

We've mentioned that the SMT1 (Sections 3.3 and 4.2) with the highest BLEU score has suffered the biggest strict brevity penalty and has rather low LCM scores. We've also found the S1 and S5 (SC1 and SC5 in Table 8) in figure 6 destroyed the correlation between scores of LCM and BLEU. In order to study these cases, we selected 20 sentences uniformly distributed in the test data of "ce_news" task as our new test data, and then conducted a small-sized human evaluation on the results of these three systems and other two systems: a RBMT system and a SMT system (RBMT1 and SMT2 in Table 8). We assigned each sentence a subjective 1-5 score along two axes: adequacy and fluency (LDC, 2005). The result is showed in Table 8.

		SMT1	RBMT1	SMT2	SC1	SC5
NIST-BLEU		0.2809	0.2275	0.2264	0.2944	0.2679
IBM-BLEU		0.2661	0.2215	0.2137	0.2792	0.2588
BLEU-SBP		0.2631	0.2193	0.2122	0.2758	0.2560
LCM scores	G	0.2629	0.2618	0.2649	0.2686	0.2887
	W	0.4146	0.4377	0.4354	0.4209	0.4608
	P	0.2480	0.2446	0.2482	0.2536	0.2719
	S	0.2459	0.2401	0.2457	0.2519	0.2709
Human evaluation	F	3.4	3.45	3.2	3.2	3.45
	A	3	3.7	3.25	3	3.6

Table 8. Comparison of Five Systems in C2E News Translation and System Combination Tasks. G=General, W=Words, P=Phrases, S=Sentences, F=Fluency, A=Adequacy.

We can see that SMT1 has the lowest word level score (which can represent 1-gram matching degree to some extent) and adequacy score among three MT systems. As to the system combination, we found, except the three kinds of BLEU scores, all the LCM scores and human evaluation scores of SC1 (with the highest BLEU) are lower than those of SC5, especially the adequacy score (which is 0.6 point lower than that of SC5). We examined the output of SC1 and found its translation sentences are very similar to those of the single system SMT1 (Top 1 in Figure 10). The sentence in Figure 10 is an example. This is because SC1 used the technique of sentence-level system combination and assigned the top 1 hypothesis (SMT1) the highest score. We also examined the system description of SC5 and found that this system used the word-level system combination technique, although the BLEU scores are lower, its translation adequacy is much better. We can also see this from Figure 10.

Source: 张秀华家挂的胡、温画像是经过电脑处理，原来画面的其他人员已经被掩盖，只有两个人握手的画面。
SC1: Zhang Xiuhua, a computer processing, and other personnel have been only two people.
SC5: Zhang Xiuhua home hanging on Hu, warm portrait is through the computer processing, so that the other personnel have been covered, there are only two shake hands.
Top 1: Zhang Xiuhua, after computer processing, and other personnel have been only two people.
Top 2:, moustache after the computer , the picture of the other has been masked , only two shook hands .
Top 3: Zhang Xiuhua family hung Hu and Wen Hua like after the computer , the picture of the other has been covered up , only two shook hands .
Reference 1: The portrait of Hu and Wen hung in Zhang Xiuhua's home has been processed by the computer; the other officials present were edited out to show only the two shaking hands.

Figure 10. Example of System Combination. Top 1~3 are three single systems with the highest BLEU in training data.

Though the n-gram matching with the reference translation of the candidate system translations increases, which leads to higher BLEU score, plenty of linguistic information in the test data is lost. This might be the major reason of the high BLEU but low adequacy.

7 Conclusion and Future Work

Our experiments validated BLEU-SBP's effectivity in resolving the nondecomposability problem of both NIST-BLEU and IBM-BLEU at sentence level. The results of our evaluation also indicate that the LCM is a valid metric to evaluate the capability of an MT system in translating various linguistic phenomena. By means of these metrics,

we disclose some latent performance differences of the SMT systems and RBMT systems. Through case study, we suggest that the higher BLEU score doesn't always mean higher translation adequacy.

More information will be fed back to the LCM's developer in our future evaluation to improve this promising metric.

Acknowledgments

This research is supported by the High Technology Research and Development Program of China (Grant No. 2006AA010108), and the National Science Foundations of China (Grant Nos. 60736014 and 60603095). We would like to thank and acknowledge all supporters and participants of CWMT2008 evaluation for providing us with the evaluation tools and data. We thank the anonymous reviewers for their insightful comments. We are also grateful to Wenbin Jiang for his helpful feedback.

References

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In Proc. EACL 2009, pages 1-28.

David Chiang, Steve DeNeefe, Yee Seng Chan, and Hwee Tou Ng. 2008. *Decomposability of translation metrics for improved evaluation and efficient algorithms*. In Proc. EMNLP 2008, pages 610-619.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. *Clause restructuring for statistical machine translation*. In Proc. ACL 2005, pages 531–540.

LDC. 2005. Linguistic data annotation specification: Assessment of fluency and adequacy in translation. Revision 1.5.

NIST. 2008. *NIST 2008 Machine Translation Evaluation - (Open MT-08) Official Evaluation Results*. http://www.itl.nist.gov/iad/mig//tests/mt/2008/doc/mt08_official_results_v0.html

NIST. 2009. *The 2009 NIST Open Machine Translation Evaluation Plan (MT09)*. http://www.itl.nist.gov/iad/mig/tests/mt/2009/MT09_EvalPlan.pdf

Franz Josef Och, Hermann Ney. 2003. *A Systematic Comparison of Various Statistical Alignment Model*, Computational Linguistics, volume 29, number 1, pp. 19-51 March 2003.

Kishore Papineni, Salim Roukos, Todd Ward, WeiJing Zhu. 2002. *BLEU: a method for automatic evaluation of machine translation,* In Proceedings of the ACL 2002.

Shiwen YU, *Automatic Evaluation of Output Quality for Machine Translation Systems*, Machine Translation, 1993, 8:117-126, Kluwer Academic publisher, printed in the Netherlands

Huaping Zhang, Qun Liu, Xueqi Cheng, Hao Zhang and Hongkui Yu. 2003. *Chinese Lexical Analysis Using Hierarchical Hidden Markov Model.* In Second SIGHAN Workshop Affiliated with 41th ACL, Sapporo Japan, 2003, pp. 63-70.

Ming Zhou, Bo Wang, Shujie Liu, Mu Li, Dongdong Zhang and Tiejun Zhao. 2008. *Diagnostic Evaluation of Machine Translation Systems Using Automatically Constructed Linguistic Check-Points*. In Proceedings Coling 2008, pages 1121–1128.