Improving machine translation into Chinese by tuning against Chinese MEANT

Chi-kiu Lo, Meriem Beloucif, Dekai Wu

HKUST
Human Language Technology Center
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
{jackielo|mbeloucif|dekai}@cs.ust.hk

Abstract

We present the first ever results showing that Chinese MT output is significantly improved by tuning a MT system against a semantic frame based objective function, MEANT, rather than an n-gram based objective function, BLEU, as measured across commonly used metrics and different test sets. Recent work showed that by preserving the meaning of the translations as captured by semantic frames in the training process, MT systems for translating into English on both formal and informal genres are constrained to produce more adequate translations by making more accurate choices on lexical output and reordering rules. In this paper we describe our experiments in IWSLT 2013 TED talk MT tasks on tuning MT systems against MEANT for translating into Chinese and English respectively. We show that the Chinese translation output benefits more from tuning a MT system against MEANT than the English translation output due to the ambiguous nature of word boundaries in Chinese. Our encouraging results show that using MEANT is a promising alternative to BLEU in both evaluating and tuning MT systems to drive the progress of MT research across different languages.

1. Introduction

We present the first ever results of tuning a MT system against a semantic frame based objective function in order to produce a more adequate Chinese translation output. We compare the performance of our systems in IWSLT 2013 TED talk MT tasks on Chinese-English and English-Chinese translation with that of the baseline SMT systems tuned against BLEU. We show that the improvement of tuning a MT system against MEANT on Chinese translation output is more significant because of the nature of ambiguous word boundaries in Chinese. Our encouraging results show that using MEANT is a promising alternative to BLEU in evaluating and tuning MT systems to drive the progress of MT research across different languages.

In the past decade, the progress of MT research is predominantly driven by the fast and cheap n-gram based MT evaluation metrics, such as BLEU [1], which assume that a good translation is one that shares the same lexical choices as the reference translation. Despite enforcing fluency, it has been established that these metrics do not enforce translation utility adequately and often fail to preserve meaning[2, 3]. Unlike BLEU, or other n-gram based MT evaluation metrics, the MEANT family of metrics [4, 5, 6] adopt at outset the principle that a good translation is one from which humans can successfully understand at least the central meaning of the input sentence as captured by the basic event structure— "*who did what to whom, when, where and why"*[7]. [6]MEANT measures similarity between the MT output and the reference translations by comparing the similarities between the semantic frame structures of output and reference translations. For evaluating English translations, we have shown that MEANT correlates better with human adequacy judgment than commonly used MT evaluation metrics, such as BLEU [1], NIST [8], METEOR [9], CDER [10], WER [11], and TER [12].

We recently showed that the translation adequacy across different genres (ranging from formal news to informal web forum) is improved by replacing surface oriented metrics like BLEU or TER with a semantic frame based objective function, MEANT, when tuning the parameters of MT systems [13, 14]. However, the question of whether the same approach of tuning MT systems against a semantic objective function might improve translation adequacy when translating into other

languages, such as Chinese, is left unanswered.

Although there exists no studies on correlation between human adequacy judgement and MEANT scores on Chinese output, we hypothesize that the benefits of tuning against MEANT that we see for English: better adequacy and fluency carries over into Chinese. It is because a high MEANT score is contingent on correct lexical choices as well as getting the syntactic and semantic structures right, which is language independent.

The proposed approach of incorporating semantic information into SMT by tuning the model against a semantic frame based evaluation metric is independent of assumptions about the underlying translation model architecture. Therefore, we show that MT systems from different SMT approaches, flat phrase-based and hierarchical phrase-based, both benefit from the semantic information incorporated through our approach.

2. Related work

2.1. MT evaluation metrics

N-gram or edit distance based metrics such as BLEU [1], NIST [8], METEOR [9], CDER [10], WER [11], and TER [12] do not correctly reflect the similarity of the basic event structure— "*who did what to whom, when, where and why*"— of the input sentence. In fact, a number of large scale meta-evaluations [2, 3] report cases where BLEU strongly disagrees with human judgments of translation adequacy.

This has caused a recent surge of work on developing MT evaluation metrics that would outperforms BLEU in correlation with human judgment. AMBER [15] shows a high correlation with human adequacy judgment [16], however, it is very hard to interpret and indicate what errors the MT systems are making.

ULC [17, 18] is an automatic metric that incorporates several semantic similarity features and shows improved correlation with human judgement of translation quality [19, 17, 20, 18] but no work has been done towards tuning an SMT system using a pure form of ULC perhaps due to its expensive run time. Similarly, SPEDE [21] is an integrated probabilistic FSM and probabilistic PDA model that predicts the edit sequence needed for the MT output to match the reference. Sagan [22] is a semantic textual similarity metric based on a complex textual entailment pipeline. These aggregated metrics require sophisticated feature extraction steps; contain several dozens of parameters to tune and employ expensive linguistic resources, like WordNet and paraphrase tables. Like ULC, these metrices are not useful in the MT system development cycle for tuning due to expensive running time. The metrics themselves are also expensive in training and tuning due to the large number of parameters that need to be estimated.

ROSE [23] is a weighted linear model of shallow linguistic features which is cheaper in run time but still contains several dozens of weights that need to be tuned, which makes it hard to port the metric to different domains. TINE [24] is an automatic recall-oriented evaluation metric which aims to preserve the basic event structure. However, it performs comparably to BLEU and worse than METEOR on correlation with human adequacy judgment.

In contrast, there is very little work on designing MT evaluation metrics for evaluating Chinese or other languages with ambiguous word boundaries. For instance, studies show that simply adapting the commonly used MT evaluation metrics to evaluate Chinese on character-level showed a higher correlation with human judgment than the original word-level evaluation metrics [25]. Later, TESLA-CELAB is introduced as a hybrid character-level and word-level MT evaluation metric for evaluating Chinese [26]. Although TESLA-CELAB correlates significantly better with human judgment for evaluating Chinese than BLEU, no work has been done towards tuning an SMT system for translating into Chinese using it.

2.2. The MEANT family of metrics

MEANT [6], which is the weighted f-score over the matched semantic role labels of the automatically aligned semantic frames and role fillers, outperforms BLEU, NIST, METEOR, WER, CDER and TER in correlating with human adequacy judgment. MEANT is easily portable to other languages requiring only an automatic semantic parser and a large monolingual corpus in the output language for identifying the semantic structures and the lexical similarity between the semantic role fillers of the reference and translation.

Precisely, MEANT is computed as follows:

1. Apply an automatic shallow semantic parser to both the references and MT output. (Figure 1 shows examples of automatic shallow semantic parses on both reference and MT output.)

2. Apply the maximum weighted bipartite matching algorithm to align the semantic frames between

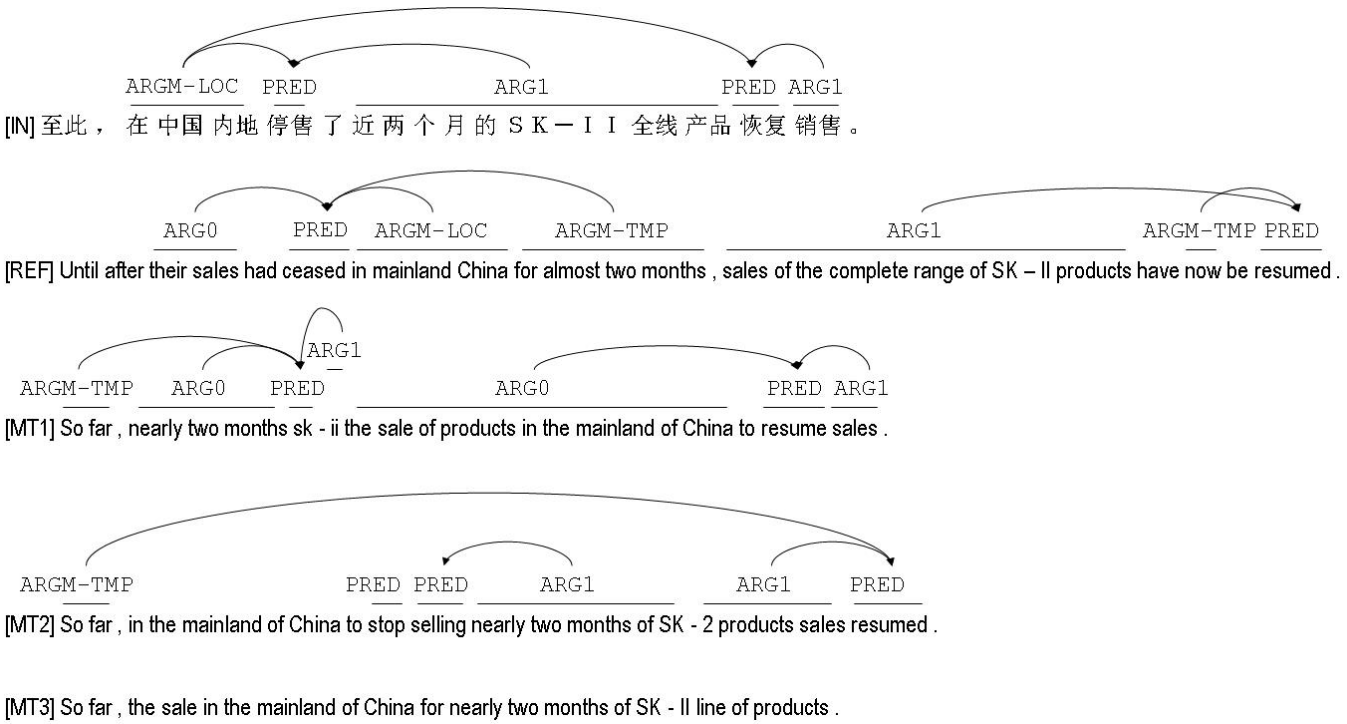

Figure 1: Examples of automatic shallow semantic parses. The input is parsed by a Chinese automatic shallow semantic parser. The reference and MT output are parsed by an English automatic shallow semantic parser. There are no semantic frames for MT3 since there is no predicate.

the references and MT output by the lexical similarities of the predicates.

3. For each pair of aligned semantic frames,
   (a) Determine the similarity of the semantic role fillers using Lexical similarity scores.
   (b) Apply the maximum weighted bipartite matching algorithm to align the semantic role fillers between the reference and MT output according to their lexical similarity.
4. Compute the weighted f-score over the matching role labels of these aligned predicates and role fillers acording to the mathematical definitions in the following.

$$
\begin{aligned}
M_{i,j} &\equiv \text{total \# ARG j of aligned frame i in MT} \\
R_{i,j} &\equiv \text{total \# ARG j of aligned frame i in REF} \\
S_{i,\text{pred}} &\equiv \text{similarity of predicate in aligned frame i} \\
S_{i,j} &\equiv \text{similarity of ARG j in aligned frame i} \\
w_{\text{pred}} &\equiv \text{weight of similarity of predicates} \\
w_j &\equiv \text{weight of similarity of ARG j}
\end{aligned}
$$

$$
\begin{aligned}
m_i &\equiv \frac{\text{\#tokens filled in aligned frame i of MT}}{\text{total \#tokens in MT}} \\
r_i &\equiv \frac{\text{\#tokens filled in aligned frame i of REF}}{\text{total \#tokens in REF}}
\end{aligned}
$$

$$
\text{precision} = \frac{\sum_i m_i \frac{w_{\text{pred}} S_{i,\text{pred}} + \sum_j w_j S_{i,j}}{w_{\text{pred}} + \sum_j w_j M_{i,j}}}{\sum_i m_i}
$$

$$
\text{recall} = \frac{\sum_i r_i \frac{w_{\text{pred}} S_{i,\text{pred}} + \sum_j w_j S_{i,j}}{w_{\text{pred}} + \sum_j w_j R_{i,j}}}{\sum_i r_i}
$$

where m_i and r_i are the weights for frame, i, in the MT/REF respectively. These weights estimate the degree of contribution of each frame to the overall meaning of the sentence. $M_{\text{i,j}}$ and $R_{\text{i,j}}$ are the total counts of argument of type j in frame i in the MT and REF respectively. $S_{i,\text{pred}}$ and $S_{i,j}$ are the lexical similarities (as computed based on a context vector model) of the predicates and role fillers of the arguments of type j between the reference translations and the MT output. The weights w_{pred} and w_j are the weights of the lexical similarities of the predicates and role fillers of the arguments of type j between the reference translations and the MT output. There is a total of 12 weights for the set

of semantic role labels in MEANT as defined in [27]. For MEANT, w_{pred} and w_j are determined using supervised estimation via a simple grid search to optimize the correlation with human adequacy judgments [4]. For UMEANT, w_{pred} and w_j are estimated in an unsupervised manner using relative frequency of each semantic role label in the reference translations. UMEANT can thus be used when when human judgments on adequacy of the development set are unavailable [5].

2.3. Tuning against better evaluation metrics

Previous works show that tuning MT system against better evaluation metrics improve the translation quality [28, 29]. Recent studies [13, 14] also shows that tuning MT system against MEANT produces more robustly adequate translations than the common practice of tuning against BLEU or TER across different data genres, such as formal newswire text, informal web forum text and informal public speech. Therefore, we believe that tuning MT systems against MEANT would improve the adequacy on Chinese MT output.

3. Experimental setup

In this section, we describe the details of our systems for the English-Chinese and Chinese-English TED talk MT tasks in terms of data, preprocesing, SMT pipeline and MEANT settings.

3.1. Data and preprocessing

Since our focus in this evaluation campaign is running contrastive experiments on tuning different MT systems against different MT evaluation metrics, we have deliberately constrained our training data to in-domain data only. For the translation model we have only used the officially released parallel training data, while for the language model we have only used the output side of the released training data. Similarly, no additional data was used as a part of development set other than the officially released development set. In order to test the consistency of the experimental results the test sets of IWSLT 2011 and 2012 were used in addition to the IWSLT 2013 test set. We perform minimal preprocessing on the training data running a maximum entropy Chinese segmenter [30] along with numex/timex segmenter on the Chinese data and punctuation tokenization and true casing on the English data.

3.2. SMT pipeline

With the goal of improving MT utility by using MEANT as an objective function to drive minimum error rate training (MERT) [31] of state-of-the-art MT systems, we setup our baseline using Moses [32], an off-the-shelf translation toolkit. In this paper we have two baselines: a flat phrase-based MT and a hierarchical phrase-based MT [33]. This allows us to use Moses to compare the performance of MEANT-tuned systems in these two different MT paradigms.

The language models are trained using the SRI language model toolkit [34]. For both translation tasks, we used a 6-gram language model. We use ZMERT [35] to tune the baseline because it is a widely used, highly competitive, robust, and reliable implementation of MERT that is also fully configurable and extensible with regard to incorporating new evaluation metrics.

3.3. MEANT for evaluating Chinese

Since UMEANT is shown to be more stable when evaluating translations across different language pairs [36], we use a UMEANT framework along the lines described in [37] for evaluating both English and Chinese.

However, for evaluating Chinese, MEANT has to be equipped with a Chinese shallow semantic parser in order to capture the semantic frames in the Chinese translation output. For this purpose, we used C-ASSERT [38] because of its high accuracy.

Since the primary objective in this experiment is studying the feasibility of tuning MT systems against Chinese MEANT, we limited ourselves to using a window-size-3 context vector model trained on the word segmented monolingual Chinese gigaword corpus, for estimating the phrasal similarity of the semantic role fillers, rather than investigating which combination of window-size, similarity function and phrasal aggregation function that would perform the best in evaluating Chinese.

3.4. Submitted systems

For the English-Chinese TED talks MT task, we submitted translation output from three systems. The primary system is our MEANT-tuned Moses flat phrase-based MT system and the two contrastive systems are the BLEU-tuned Moses flat phrase-based and BLEU-tuned Moses hierarchical phrase-based systems. In this paper, we have also include our latest results on the MEANT-tuned Moses hierarchical phrase-based system.

Table 1: Translation quality of the participated English-Chinese MT systems on the IWSLT 2013 test set where (p) indicates our primary submission; (c1) and (c2) indicate the two contrastive submissions and (n) indicates our not-submitted system.

	char-level		word-level								
	official		official		internal						
System	BLEU	TER	BLEU	TER	BLEU	TER	NIST	WER	PER	CDER	MEANT
(p) MEANT-tuned flat	**18.66**	**70.36**	**10.85**	**78.12**	11.44	79.24	4.25	83.07	64.80	77.04	25.65
(c1) BLEU-tuned flat	18.08	72.00	10.38	82.02	10.93	83.58	4.06	87.19	69.07	81.03	24.88
(c2) BLEU-tuned hier	18.02	72.12	10.37	81.80	10.88	83.63	4.05	87.16	69.44	81.07	23.98
(n) MEANT-tuned hier	—	—	—	—	**11.83**	**72.31**	**4.59**	**76.09**	**58.86**	**70.78**	**25.72**

Table 2: Translation quality of the participated English-Chinese MT systems on the IWSLT 2012 test set where (p) indicates our primary submission; (c1) and (c2) indicate the two contrastive submissions and (n) indicates our not-submitted system.

	word-level (internal)						
System	BLEU	TER	NIST	WER	PER	CDER	MEANT
(p) MEANT-tuned flat	**10.89**	81.20	**4.18**	84.61	67.54	79.38	**24.24**
(c1) BLEU-tuned flat	10.65	86.23	3.98	89.33	72.81	84.17	23.76
(c2) BLEU-tuned hier	10.47	86.53	3.95	89.56	73.34	84.34	22.37
(n) MEANT-tuned hier	9.04	**78.33**	3.73	**81.42**	**66.15**	**76.60**	22.93

For the Chinese-English TED talks MT task, we submitted translation output from four systems. In addition to the primary MEANT-tuned Moses flat phrase-based MT system and the two contrastive BLEU-tuned Moses flat phrase-based and BLEU-tuned Moses hierarchical phrase-based systems, we have also submitted translation output from the contrastive MEANT-tuned Moses hierarchical phrase-based system.

4. Results

Table 1, 2, and 3 show that the MEANT-tuned systems in the English-Chinese TED talks MT task achieves significantly better scores than the two contrastive BLEU-tuned systems across all evaluation metrics on all three test sets. The results is surprising because MEANT-tuned system beats BLEU-tuned systems even on BLEU, the metric which the BLEU-tuned systems are highly optimized on. This encouraging results confirm that MEANT is a better metric for evaluating and tuning MT system on Chinese.

Table 4, 5 and 6 show that the BLEU-tuned systems in the Chinese-English TED talks MT task only performs well on BLEU, the metric which they are highly optimized on. However, MEANT-tuned systems beats the BLEU-tuned systems on other evaluation metrics across all three test sets. More precisely, MEANT-tuned Moses flat phrase-based system achieves the best error metric scores (TER, WER, CDER) while the MEANT-tuned Moses hierarchical phrase-based system achieves better scores in NIST, PER and METEOR. This results show that tuning MT system against BLEU would easily result in overfitting instead of producing good translation in practice. On the other hand, since a high MEANT score rely on correct lexical choices as well as syntactic and semantic structures, tuning MT systems against MEANT would hardly result in overfitting while producing translations that more robustly express the meaning in the original input accurately.

5. Conclusion

In this paper, we have presented the first ever results that tuning a MT system for translating into Chinese against MEANT significantly improves translation quality, instead of tuning against BLEU. MEANT-tuned English-Chinese MT system successfully achieves the best scores across commonly used metrics on all test sets. Since a high MEANT score rely on correct lexical choices as well as syntactic and semantic structures, tuning MT systems against MEANT would hardly result in overfitting while producing translations that more robustly accurately express the meaning in the original input. This effect is more obvious when we are translating into a non-English language.

We have to point out that in this feasibility study we have done minimal adaptation on the settings of MEANT for evaluating Chinese. We expect the performance of

Table 3: Translation quality of the participated English-Chinese MT systems on the IWSLT 2011 test set where (p) indicates our primary submission; (c1) and (c2) indicate the two contrastive submissions and (n) indicates our not-submitted system.

	word-level (internal)						
System	BLEU	TER	NIST	WER	PER	CDER	MEANT
(p) MEANT-tuned flat	**12.24**	79.56	**4.39**	82.42	65.66	77.28	**25.87**
(c1) BLEU-tuned flat	11.12	85.12	4.12	87.94	71.44	82.57	25.33
(c2) BLEU-tuned hier	10.89	83.63	4.05	87.16	69.44	81.07	23.16
(n) MEANT-tuned hier	10.14	**76.66**	3.96	**79.21**	**64.20**	**74.41**	23.51

Table 4: Translation quality of the participated Chinese-English MT systems on the IWSLT 2013 test set where cased and uncased BLEU and TER are the official results. (p) indicates our primary submission; (c1), (c2) and (c3) indicate the three contrastive submissions. MET stands for METEOR.

	cased		uncased									
	official		official		internal							
System	BLEU	TER	BLEU	TER	BLEU	TER	NIST	MET	WER	PER	CDER	MEANT
(p) MEANT-tuned flat	9.58	**74.82**	10.17	**73.75**	10.61	**73.82**	4.57	42.49	**75.66**	58.97	**70.81**	31.42
(c1) MEANT-tuned hier	10.20	75.92	10.79	74.83	11.29	74.59	**4.65**	43.05	77.32	58.96	71.73	**32.50**
(c2) BLEU-tuned flat	10.16	76.05	10.84	74.88	11.32	74.54	**4.65**	43.24	77.05	**58.94**	71.70	31.46
(c3) BLEU-tuned hier	**10.24**	76.95	**10.90**	75.76	**11.41**	75.17	4.62	**43.30**	78.07	59.72	72.39	31.86

MEANT-tuned systems to be even better when the optimal settings are used. This encouraging results show that using MEANT is a promising alternative to BLEU in both evaluating and tuning MT systems to drive the progress of MT research across different languages.

6. Acknowledgment

This material is based upon work supported in part by the Defense Advanced Research Projects Agency (DARPA) under BOLT contract no. HR0011-12-C-0016, and GALE contract nos. HR0011-06-C-0022 and HR0011-06-C-0023; by the European Union under the FP7 grant agreement no. 287658; and by the Hong Kong Research Grants Council (RGC) research grants GRF620811, GRF621008, and GRF612806. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA, the EU, or RGC. We are grateful to Pascale Fung, Yongsheng Yang and Zhaojun Wu for sharing the maximum entropy Chinese segmenter and C-ASSERT, the Chinese semantic parser, with us.

7. References

[1] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "BLEU: a method for automatic evaluation of machine translation," in *40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, Philadelphia, Pennsylvania, July 2002, pp. 311–318.

[2] Chris Callison-Burch, Miles Osborne, and Philipp Koehn, "Re-evaluating the role of BLEU in machine translation research," in *11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, 2006, pp. 249–256.

[3] Philipp Koehn and Christof Monz, "Manual and automatic evaluation of machine translation between european languages," in *Workshop on Statistical Machine Translation (WMT-06)*, 2006, pp. 102–121.

[4] Chi-kiu Lo and Dekai Wu, "MEANT: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles," in *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*, 2011.

[5] ——, "Unsupervised vs. supervised weight estimation for semantic MT evaluation metrics," in *Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-6)*, 2012.

[6] Chi-kiu Lo, Anand Karthik Tumuluru, and Dekai Wu, "Fully automatic semantic MT evaluation," in *7th Workshop on Statistical Machine Translation (WMT 2012)*, 2012.

[7] Sameer Pradhan, Wayne Ward, Kadri Hacioglu, James H. Martin, and Dan Jurafsky, "Shallow semantic parsing using support vector machines," in *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2004)*, 2004.

Table 5: Translation quality of the participated Chinese-English MT systems on the IWSLT 2012 test set. (p) indicates our primary submission; (c1), (c2) and (c3) indicate the three contrastive submissions.

	uncased (internal)							
System	BLEU	TER	NIST	METEOR	WER	PER	CDER	MEANT
(p) MEANT-tuned flat	10.22	**72.73**	4.47	41.63	**74.42**	59.60	**70.12**	31.94
(c1) MEANT-tuned hier	10.73	73.66	**4.60**	42.00	75.93	**59.32**	70.89	32.29
(c2) BLEU-tuned flat	**10.90**	73.84	4.59	**42.28**	75.80	59.74	71.00	31.45
(c3) BLEU-tuned hier	10.71	73.94	4.56	41.59	76.39	59.76	71.18	**32.60**

Table 6: Translation quality of the participated Chinese-English MT systems on the IWSLT 2011 test set. (p) indicates our primary submission; (c1), (c2) and (c3) indicate the three contrastive submissions.

	uncased (internal)							
System	BLEU	TER	NIST	METEOR	WER	PER	CDER	MEANT
(p) MEANT-tuned flat	11.17	**71.48**	4.66	43.87	**73.05**	58.13	**68.98**	34.12
(c1) MEANT-tuned hierarchical	11.97	71.90	**4.82**	**44.60**	74.16	**57.58**	69.30	**35.47**
(c2) BLEU-tuned flat	11.89	72.30	4.77	44.09	74.43	58.40	69.87	34.39
(c3) BLEU-tuned hierarchical	**12.06**	72.61	4.77	44.16	74.94	58.27	69.89	34.76

[8] George Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics," in *The second international conference on Human Language Technology Research (HLT '02)*, San Diego, California, 2002, pp. 138–145.

[9] Satanjeev Banerjee and Alon Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, Michigan, June 2005, pp. 65–72. [Online]. Available: http://www.aclweb.org/anthology/W/W05/W05-0909

[10] Gregor Leusch, Nicola Ueffing, and Hermann Ney, "CDer: Efficient MT evaluation using block movements," in *11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, 2006.

[11] Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney, "A evaluation tool for machine translation: Fast evaluation for MT research," in *The Second International Conference on Language Resources and Evaluation (LREC 2000)*, 2000.

[12] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul, "A study of translation edit rate with targeted human annotation," in *7th Biennial Conference Association for Machine Translation in the Americas (AMTA 2006)*, Cambridge, Massachusetts, August 2006, pp. 223–231.

[13] Chi-kiu Lo, Karteek Addanki, Markus Saers, and Dekai Wu, "Improving machine translation by training against an automatic semantic frame based evaluation metric," in *51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, 2013.

[14] Chi-kiu Lo and Dekai Wu, "Can informal genres be better translated by tuning on automatic semantic metrics?" in *14th Machine Translation Summit (MT Summit XIV)*, 2013.

[15] Boxing Chen, Roland Kuhn, and George Foster, "Improving AMBER, an MT evaluation metric," in *7th Workshop on Statistical Machine Translation (WMT 2012)*, 2012, pp. 59–63.

[16] Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia, "Findings of the 2012 workshop on statistical machine translation," in *7th Workshop on Statistical Machine Translation (WMT 2012)*, 2012, pp. 10–51.

[17] Jesús Giménez and Lluís Màrquez, "Linguistic features for automatic evaluation of heterogenous MT systems," in *Second Workshop on Statistical Machine Translation (WMT-07)*, Prague, Czech Republic, June 2007, pp. 256–264. [Online]. Available: http://www.aclweb.org/anthology/W/W07/W07-0738

[18] ——, "A smorgasbord of features for automatic MT evaluation," in *Third Workshop on Statistical Machine Translation (WMT-08)*, Columbus, Ohio, June 2008, pp. 195–198. [Online]. Available: http://www.aclweb.org/anthology/W/W08/W08-0332

[19] Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder, "(meta-) evaluation of machine translation," in *Second Workshop on Statistical Machine Translation (WMT-07)*, 2007, pp. 136–158.

[20] ——, "Further meta-evaluation of machine translation," in *Third Workshop on Statistical Machine Translation (WMT-08)*, 2008, pp. 70–106.

[21] Mengqiu Wang and Christopher D. Manning, "SPEDE: Probabilistic edit distance metrics for MT evaluation," in *7th Workshop on Statistical Machine Translation (WMT 2012)*, 2012, pp. 76–83.

[22] Julio Castillo and Paula Estrella, "Semantic textual similarity for MT evaluation," in *7th Workshop on Statistical Machine Translation (WMT 2012)*, 2012, pp. 52–58.

[23] Xingyi Song and Trevor Cohn, "Regression and ranking based optimisation for sentence level machine translation evaluation," in *Sixth Workshop on Statistical Machine Translation (WMT 2011)*, 2011, pp. 123–129.

[24] Miguel Rios, Wilker Aziz, and Lucia Specia, "Tine: A metric to assess MT adequacy," in *Sixth Workshop on Statistical Machine Translation (WMT 2011)*, 2011, pp. 116–122.

[25] Maoxi Li, Chengqing Zpng, and Hwee Tou Ng, "Automatic evaluation of Chinese translation output: Word-level or character-level?" in *49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*. Association for Computational Linguistics, 2011, pp. 159–164.

[26] Chang Liu and Hwee Tou Ng, "Character-level machine translation evaluation for language with ambiguous word boundaries," in *50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*. Association for Computational Linguistics, 2012, pp. 921–929.

[27] Chi-kiu Lo and Dekai Wu, "SMT vs. AI redux: How semantic frames evaluate MT more accurately," in *Twenty-second International Joint Conference on Artificial Intelligence (IJCAI-11)*, 2011.

[28] Patrik Lambert, Jesús Giménez, Marta R Costa-jussá, Enrique Amigó, Rafael E Banchs, Lluıs Márquez, and JAR Fonollosa, "Machine translation system development based on human likeness," in *IEEE/ACL 2006 Workshop on Spoken Language Technology (SLT 2006)*. IEEE, 2006, pp. 246–249.

[29] Chang Liu, Daniel Dahlmeier, and Hwee Tou Ng, "Better evaluation metrics lead to better machine translation," in *2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, 2011, pp. 375–384.

[30] Pascale Fung, Grace Ngai, Yongsheng Yang, and Benfeng Chen, "A maximum-entropy Chinese parser augmented by transformation-based learning," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 3, no. 2, pp. 159–168, 2004.

[31] Franz Josef Och, "Minimum error rate training in statistical machine translation," in *41st Annual Meeting of the Association for Computational Linguistics (ACL-2003)*, Sapporo, Japan, July 2003, pp. 160–167. [Online]. Available: http://www.aclweb.org/anthology/P03-1021

[32] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst, "Moses: Open source toolkit for statistical machine translation," in *Interactive Poster and Demonstration Sessions of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, Prague, Czech Republic, June 2007, pp. 177–180.

[33] David Chiang, "Hierarchical phrase-based translation," *Computational Linguistics*, vol. 33, no. 2, pp. 201–228, 2007. [Online]. Available: http://aclweb.org/anthology-new/J/J07/J07-2003.pdf

[34] Andreas Stolcke, "SRILM – an extensible language modeling toolkit," in *7th International Conference on Spoken Language Processing (ICSLP2002 - INTERSPEECH 2002)*, Denver, Colorado, September 2002, pp. 901–904.

[35] Omar F. Zaidan, "Z-MERT: A fully configurable open source tool for minimum error rate training of machine translation systems," *The Prague Bulletin of Mathematical Linguistics*, vol. 91, pp. 79–88, 2009.

[36] Matouš Macháček and Ondřej Bojar, in *8th Workshop on Statistical Machine Translation (WMT 2013)*. Association for Computational Linguistics, 2012, pp. 921–929.

[37] Chi-kiu Lo and Dekai Wu, "MEANT at WMT 2013: A tunable, accurate yet inexpensive semantic frame based mt evaluation metric," in *8th Workshop on Statistical Machine Translation (WMT 2013)*, 2013.

[38] Zhaojun Wu, Yongsheng Yang, and Pascale Fung. (2006) C-ASSERT: Chinese shallow semantic parser. Website. HKUST. [Online]. Available: http://hlt030.cse.ust.hk/research/c-assert/