Improving MEANT Based Semantically Tuned SMT

Meriem Beloucif, Chi-kiu Lo, Dekai Wu

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

Abstract

We discuss various improvements to our MEANT tuned system, previously presented at IWSLT 2013. In our 2014 system, we incorporate this year's improved version of MEANT, improved Chinese word segmentation, Chinese named entity recognition and dedicated proper name translation, and number expression handling. This results in a significant performance jump compared to last year's system. We also ran preliminary experiments on tuning to IMEANT, our new ITG based variant of MEANT. The performance of tuning to IMEANT is comparable to tuning on MEANT (differences are statistically insignificant). We are presently investigating if tuning on IMEANT can produce even better results, since IMEANT was actually shown to correlate with human adequacy judgment more closely than MEANT. Finally, we ran experiments applying our new architectural improvements to a contrastive system tuned to BLEU. We observed a slightly higher jump in comparison to last year, possibly due to mismatches of MEANT's similarity models to our new entity handling.

1. Introduction

In this paper we present an improved version of our MT system tuned against MEANT (Lo and Wu [1, 2]; Lo *et al*. [3]), a semantic MT evaluation metric which has been proven to highly correlate with human adequacy judgments. We employ an improved version of MEANT that correlates more closely with human adequacy judgments, resulting also in translation performance gains compared to the system tuned against our previous version of MEANT from the IWSLT 2013 evaluation campaign (Lo *et al.* [4]). This improved variant of MEANT uses f-score to aggregate lexical similarities within role filler phrases instead of linear average. We also introduced several changes to last year's baseline, including improved Chinese word segmentation, improved Chinese named entity recognition combined with dedicated proper name translation, and number expression handling.

We also experimented with tuning against IMEANT (Wu *et al.* [5]), a new inversion transduction grammar (ITG) version of MEANT, that was shown this year to correlate with human adequacy judgements more closely than MEANT. Despite this fact, we observed that tuning to IMEANT is statistically indistinguishable from tuning to MEANT.In the past few years, MT research has mainly focused on evaluation using fast and cheap n-gram based MT evaluation metrics such as BLEU [6] which assume that a good translation is one that has similar lexical n-grams as the reference translation. Although such metrics tend to enforce fluency, it has been shown that these metrics generally do not emphasize meaning preservation, and thus are weak at enforcing translation adequacy (Callison-Burch *et al.* [7]; Koehn and Monz [8]).

Unlike BLEU, or other n-gram based metrics, the MEANT family of metrics adopt the principle that a good translation is one in which humans can successfully understand the central meaning of the input sentence as captured by the basic event structure "*who did what to whom, when, where and why"* (Pradhan *et al.* [9]). MEANT measures similarity between an MT output and a reference translation by comparing the similarities between the semantic frame structures of the MT output and reference. We have shown that MEANT correlates better with human adequacy judgments than commonly used MT evaluation metrics such as BLEU [6], NIST [10], METEOR [11], CDER [12], WER [13], and TER [14].

2. Related work

Surface-form oriented metrics like BLEU [6], NIST [10], METEOR [11], CDER [12], WER [13], and TER [14] do not correctly reflect the meaning similarities of the basic event structure "*who did what to whom, when, where and why*" of the input sentence. In fact, many studies (Callison-Bursh *et al.* [7]; Koehn and Monz [8]) report cases where BLEU strongly disagrees with human adequacy judgment. This has caused a recent surge of work on developing MT evaluation metrics that outperforms BLEU in correlation with human judgment. AMBER [15] shows a high correlation with human adequacy judgment (Callison-Burch *et al.* [16]); however, it is very hard to indicate what errors the MT systems are making.

Many automatic metrics that aggregate semantic similarity have been introduced, but no tuning has been done using these metrics, because of their expensive run time. Gimenez and Marquez [17, 18] introduced ULC, an automatic metric that incorporates several semantic similarity features and shows improved correlation with human judgement of translation quality [19, 17, 20, 18]. SPEDE [21] is a metric that integrats probabilistic FSM and PDA models 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; furthermore, they typically rely on several dozens of parameters to tune and use expensive linguistic resources, like WordNet and paraphrase tables. These metrics themselves are expensive in training and tuning due to the large number of parameters that need to be estimated, thus to tune against these metrics can be extremely expensive.

3. The MEANT family of metrics

3.1. MEANT

MEANT (Lo *et al.* [3]) is a weighted f-score over the matched semantic role labels of automatically aligned semantic frames and role fillers. MEANT outperforms BLEU, NIST, METEOR, WER, CDER and TER in correlation 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 to establish the lexical similarity between the semantic role fillers of the reference and translation.

More precisely, MEANT is computed as follows:

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

2. Apply the maximum weighted bipartite matching algorithm to align the semantic frames between the reference and machine translations according to the lexical similarities of the predicates. ([23] proposed a backoff algorithm that evaluates the entire sentence of the MT output using the lexical similarity based on the context vector model, if the automatic shallow semantic parser fails to parse the reference or machine translations.)

3. For each pair of the aligned frames, apply the maximum weighted bipartite matching algorithm to align the arguments between the reference and machine translations according to the lexical similarity of role fillers.

4. Compute the weighted f-score over the matching role labels of these aligned predicates and role fillers as follow :

$$
\begin{aligned}
q^0_{i,j} &\equiv \text{ARG } j \text{ of aligned frame } i \text{ inMT} \\
q^1_{i,j} &\equiv \text{ARG } j \text{ of aligned frame } i \text{ in REF} \\
w^0_i &\equiv \frac{\text{\#tokens filled in aligned frame } i \text{ of MT}}{\text{total \#tokens in MT}} \\
w^1_i &\equiv \frac{\text{\#tokens filled in aligned frame } i \text{ of REF}}{\text{total \#tokens in REF}} \\
w_{\text{pred}} &\equiv \text{weight of similarity of predicates} \\
w_j &\equiv \text{weight of similarity of ARG } j \\
\mathbf{e}_{i,\text{pred}} &\equiv \text{the pred string of the aligned frame } i \text{ of MT} \\
\mathbf{f}_{i,\text{pred}} &\equiv \text{the pred string of the alignedframe } i \text{ of REF} \\
\mathbf{e}_{i,j} &\equiv \text{the role fillers of ARG } j \text{ of the aligned frame } i \\
\mathbf{f}_{i,j} &\equiv \text{the role fillers of ARG } j \text{ of the aligned frame } i \\
s(e,f) &= \text{lexical similarity of token } e \text{ and } f
\end{aligned}
$$

$$
\begin{aligned}
\text{prec}_{\mathbf{e},\mathbf{f}} &= \frac{\sum_{e\in\mathbf{e}} \max_{f\in\mathbf{f}} s(e,f)}{|\,\mathbf{e}\,|} \\
\text{rec}_{\mathbf{e},\mathbf{f}} &= \frac{\sum_{f\in\mathbf{f}} \max_{e\in\mathbf{e}} s(e,f)}{|\,\mathbf{f}\,|} \\
s_{i,\text{pred}} &= \frac{2\cdot \text{prec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}} \cdot \text{rec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}}}{\text{prec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}} + \text{rec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}}} \\
s_{i,j} &= \frac{2\cdot \text{prec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}} \cdot \text{rec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}}}{\text{prec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}} + \text{rec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}}}
\end{aligned}
$$

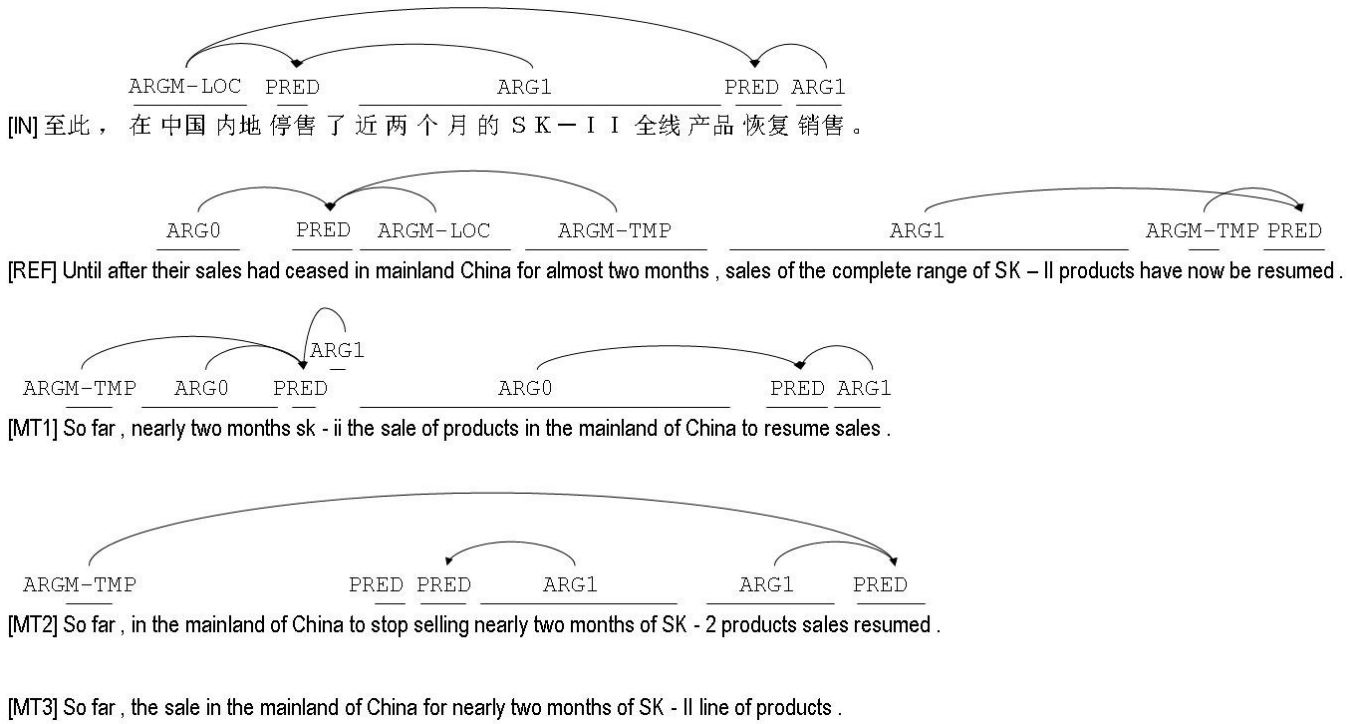

Figure 1: Examples of automatic shallow semantic parses. The input is parsed by a Chinese automatic shallow semantic parser. The reference and MT output are parsed by an English automatic shallow semantic parser. There are no semantic frames for MT3 since there is no predicate.

$$\begin{aligned}
\text{precision} &= \frac{\sum_i w_i^0 \frac{w_{\text{pred}} s_{i,\text{pred}} + \sum_j w_j s_{i,j}}{w_{\text{pred}} + \sum_j w_j |q_{i,j}^0|}}{\sum_i w_i^0} \\
\text{recall} &= \frac{\sum_i w_i^1 \frac{w_{\text{pred}} s_{i,\text{pred}} + \sum_j w_j s_{i,j}}{w_{\text{pred}} + \sum_j w_j |q_{i,j}^1|}}{\sum_i w_i^1} \\
\text{MEANT} &= \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}
\end{aligned}$$

where $q_{i,j}^0$ and $q_{i,j}^1$ are the argument of type j in frame i in MT and REF respectively. w_i^0 and w_i^1 are the weights for frame i in MT/REF respectively. These weights estimate the degree of contribution of each frame to the overall meaning of the sentence. w_{pred} and w_j are the weights of the lexical similarities of the predicates and role fillers of the arguments of type j of all frame between the reference translations and the machine translations. There is a total of 12 weights for the set of semantic role labels in MEANT as defined in Lo and Wu [24]. For MEANT, they are determined using supervised estimation via a simple grid search to optimize the correlation with human adequacy judgments (Lo and Wu [1]). For UMEANT (Lo and Wu [2]), they are estimated in an unsupervised manner using relative frequency of each semantic role label in the references and thus UMEANT is useful when human judgments on adequacy of the development set are unavailable.

3.2. IMEANT

IMEANT (Wu *et al.* [5]) is an inversion transduction grammar based variant of MEANT. IMEANT uses a a length-normalized weighted BITG [25, 26, 27, 28] to constrain permissible token alignment patterns between aligned role filler phrases. More precisely, IMEANT differs from MEANT in the definition of $s_{i,\text{pred}}$ and $s_{i,j}$, as follows:

$$\begin{aligned}
G &\equiv \langle \{\mathrm{A}\}, \mathcal{W}^0, \mathcal{W}^1, \mathcal{R}, \mathrm{A} \rangle \\
\mathcal{R} &\equiv \{\mathrm{A} \rightarrow [\mathrm{AA}], \mathrm{A} \rightarrow \langle \mathrm{AA} \rangle, \mathrm{A} \rightarrow e/f\}
\end{aligned}$$

$$\begin{aligned}
p([\mathrm{AA}]|\mathrm{A}) &= p(\langle \mathrm{AA} \rangle | \mathrm{A}) = 1 \\
p(e/f|\mathrm{A}) &= s(e,f) \\
s_{i,\text{pred}} &= \lg^{-1}\left(\frac{\lg\left(P\left(\mathrm{A} \overset{*}{\Rightarrow} \mathbf{e}_{i,\text{pred}}/\mathbf{f}_{i,\text{pred}}|G\right)\right)}{\max(|\mathbf{e}_{i,\text{pred}}|, |\mathbf{f}_{i,\text{pred}}|)}\right) \\
s_{i,j} &= \lg^{-1}\left(\frac{\lg\left(P\left(\mathrm{A} \overset{*}{\Rightarrow} \mathbf{e}_{i,j}/\mathbf{f}_{i,j}|G\right)\right)}{\max(|\mathbf{e}_{i,j}|, |\mathbf{f}_{i,j}|)}\right)
\end{aligned}$$

where G is a bracketing ITG whose only non terminal is A, and $\mathcal{R}$ is a set of transduction rules with $e \in \mathcal{W}^0 \cup \{\epsilon\}$ denoting a token in the MT output (or the *null* token) and $f \in \mathcal{W}^1 \cup \{\epsilon\}$ denoting a token in the reference translation (or the *null* token).

The rule weight function p is set to be 1 for structural transduction rules, and for lexical transduction rules it is defined using MEANT's context vector model based lexical similarity measure. The Saers *et al.* [29] algorithm is used to compute the inside probability of a pair of segments, $P\left(\mathrm{A} \stackrel{*}{\Rightarrow} \mathbf{e}/\mathbf{f}|G\right)$.

Given this, $s_{i,\text{pred}}$ and $s_{i,j}$ now represent the length normalized BITG parse scores of the predicates and role fillers of the arguments of type j between the reference and machine translations.

4. Baseline

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

4.1. Data

Our main goal for 2014 was to improve our MEANT tuned system and compare the results to our 2013 system. For this purpose, we deliberately constrained our training data to 2013 in-domain data only. Thus we use the English-Chinese parallel data from the IWSLT 2013 training set and used the output side to train the language model.

Similarly, our development set was restricted to the IWSLT 2013 development set. Since our main focus was to test our performance in comparison to 2013, we purposely targeted the IWSLT 2013 set more than the IWSLT 2014 set. However, we do present IWSLT 2014 results for our BLEU tuned system for both English-Chinese and Chinese-English.

The English sentences were normalized for punctuation, tokenization, and truecasing.

Obviously, higher scores could have been obtained by training on the IWSLT 2014 data set instead of 2013.

4.2. SMT pipeline

With the goal of improving MT utility by using our new improved version of MEANT as an objective function to drive minimum error rate training (MERT) [30] of state-of-the-art MT systems, we set up our baseline using the translation toolkit Moses [31]. In our experiments, we are using the flat phrase-based MT. The language models are trained using the SRI language model toolkit [32]. For both translation tasks, we used a 6-gram language model. We use ZMERT [33] to tune the baseline since it is a reliable implementation of MERT and is fully configurable and extensible allowing us to easily incorporate our new evaluation metrics.

5. Experiments

5.1. MEANT improvements

This year's system incorporated new improvements to the MEANT metric, consisting of using f-score in order to aggregate lexical similarities *within* semantic role filler phrases instead of Mihalcea's [34] method used in our last year system. We also tried to extend the window-size from 3 to 5 for the context vector model trained on the word segmented monolingual English gigaword corpus.

Since UMEANT (Lo and Wu [35]) has been shown to be more stable when evaluating translations across different language pairs (Machacek and Bojar [36]), we use UMEANT for evaluating our output.

5.2. Tuning to IMEANT

In this paper, we also ran preliminary experiments on tuning to IMEANT [5], the new inversion transduction grammar based variant of MEANT, that achieves higher correlation with human adequacy judgments of MT output quality than MEANT and its variants. Addanki *et al.* [28] showed empirically that the semantic role reordering that MEANT uses is covered by ITG constraints.

5.3. Word segmentation improvements

For Chinese sentences, we improved the segmentation of Chinese words. We performed extensive comparisons between four word segmentation approaches. The results reported this year were obtained using the ICTCLAS word segmenter [37].

5.4. Named entity translation improvements

We also used our own new implementation of Chinese named entity recognition and a dedicated proper name translation, where we use our own library translator based on Wikipedia data. We implemented an adequate library generator for our new named entity recognizer.

Table 1: Translation quality of the participated Chinese-English MT systems on the IWSLT 2013 test set: (a) 2013 MEANT-tuned system, (b) 2014 improved MEANT-tuned system.

	uncased (internal)							
System	BLEU	NIST	METEOR	TER	WER	PER	CDER	MEANT
2013 MEANT-tuned system	10.49	4.54	4.24	73.97	75.77	59.17	70.94	31.42
2014 MEANT-tuned system	**13.56**	**4.97**	**4.69**	**70.48**	**73.98**	**56.19**	**69.18**	**39.79**

Table 2: Translation quality of the participated Chinese-English MT systems on the IWSLT 2013 test set tuned against MEANT and IMEANT respectively.

	uncased (internal)							
System	BLEU	NIST	METEOR	TER	WER	PER	CDER	MEANT
MEANT-tuned	**13.56**	4.97	**4.69**	70.48	73.98	56.19	69.18	**39.79**
IMEANT-tuned	13.55	**4.99**	4.68	70.48	**73.60**	**55.78**	**68.85**	34.21

5.5. Number expression translation improvements

We incorporated our HKUST number expression recognition and translation module this year.

6. Results

For IWSLT 2014 we submitted our new architecturally changed baseline for the BLEU tuned system for both, English-Chinese TED talks and Chinese-English TED talks as a primitive task. We also include our latest results on the MEANT-tuned Moses flat phrase-based system MT system, as well as our IMEANT-tuned system for Chinese-English TED talks MT task.

Table 1 shows that our new MEANT tuning using f-score as an aggregation function outperforms 2013 system. We see a high jump in terms of BLEU scores between all our MEANT tuned systems for last year and this year.

Table 2 shows also that IMEANT, the ITG variant of MEANT, produces almost identical results in comparison to our MEANT-tuned system. The differences are statistically insignificant. We are presently investigating whether tuning to IMEANT can produce even better results, since IMEANT was actually shown to correlate more closely with human adequacy judgment than MEANT.

Tables 3 and 4 show that our new word segmentation, named entity translation modules, and number expression translation modules incorporated in this year's system improved the performance of our BLEU and TER tuned systems respectively in comparison to our 2013 BLEU and TER tuned systems.

Tables 5 and 6 represent our official submitted systems for IWSLT 2014 evaluation campaign for Chinese-English and English-Chinese. We evaluate on both the 2013 and 2014 test sets. For English-Chinese translations, only the character level BLEU and TER were given.

7. Conclusion

In this paper we have presented an improved version of our MEANT tuned system which shows significant improvements over last year's model. The major changes to the system include improved Chinese word segmentation, improved Chinese named entity recognition, a new dedicated proper name translation and new number expression handling. We also experimented with tuning against IMEANT, our ITG based variant of MEANT. IMEANT performance was surprisingly similar to that of MEANT despite the fact that IMEANT has been shown to correlate better with human adequacy judgment than MEANT. We are currently looking at the possible reasons behind such a result.

8. Acknowledgment

This material is based upon work supported in part by the Hong Kong Research Grants Council (RGC) research grants GRF620811, GRF621008; and by the Defense Advanced Research Projects Agency (DARPA) under BOLT contract nos. HR0011-12-C-0014 and 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; 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 the EU, DARPA or RGC.

Table 3: Translation quality of the participated Chinese-English MT systems on the IWSLT 2013 test set: (a) 2013 BLEU-tuned system, (b) 2014 improved BLEU-tuned system.

	uncased (internal)							
System	BLEU	NIST	METEOR	TER	WER	PER	CDER	MEANT
2013 BLEU-tuned system	11.16	4.61	4.32	74.69	77.17	59.15	71.84	31.46
2014 BLEU-tuned system	**13.85**	**5.01**	**4.55**	**68.70**	**72.27**	**54.91**	**67.45**	**32.93**

Table 4: Translation quality of the participated Chinese-English MT systems on the IWSLT 2013 test set: (a) 2013 TER-tuned system, (b) 2014 improved TER-tuned system.

	uncased (internal)							
System	BLEU	NIST	METEOR	TER	WER	PER	CDER	MEANT
2013 TER-tuned system	10.65	2.96	3.33	71.09	71.51	60.72	69.10	38.38
2014 TER-tuned system	**11.16**	**4.01**	**3.97**	**66.49**	**68.43**	**56.93**	**65.78**	**39.18**

9. References

[1] C.-k. Lo and D. 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.

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

[3] C. Lo, A. K. Tumuluru, and D. Wu, "Fully automatic semantic MT evaluation," in *7th Workshop on Statistical Machine Translation (WMT 2012)*, 2012.

[4] C.-k. Lo, M. Beloucif, and D. Wu, "Improving machine translation into Chinese by tuning against Chinese MEANT," in *International Workshop on Spoken Language Translation (IWSLT 2013)*, 2013.

[5] D. Wu, C.-k. Lo, M. Beloucif, and M. Saers, "Better semantic frame based mt evaluation via inversion transduction grammars," 2014, sSST.

[6] K. Papineni, S. Roukos, T. Ward, and W.-J. 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.

[7] C. Callison-Burch, M. Osborne, and P. 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.

[8] P. Koehn and C. Monz, "Manual and automatic evaluation of machine translation between european languages," in *Workshop on Statistical Machine Translation (WMT-06)*, 2006.

[9] S. Pradhan, W. Ward, K. Hacioglu, J. H. Martin, and D. 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.

[10] G. 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.

[11] S. Banerjee and A. 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. [Online]. Available: http://www.aclweb.org/anthology/W/W05/W05-0909

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

[13] S. Nießen, F. J. Och, G. Leusch, and H. 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.

[14] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. 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.

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

Table 5: Translation quality of the submitted Chinese-English MT systems on : (a) IWSLT 14 test set, (b) IWSLT 13 test set.

System	BLEU-cased	TER-cased	BLEU-uncased	TER-uncased
2014 ZH-EN BLEU-tuned	09.64	76.66	10.83	74.15
2014 ZH-EN BLEU-tuned	11.89	72.32	13.08	70.09

Table 6: Translation quality of the submitted English-Chinese MT systems on : (a) IWSLT 14 test set, (b) IWSLT 13 test set.

System	Character BLEU	Character TER
EN-ZH BLEU-tuned	18.81	70.94
EN-ZH BLEU-tuned	16.41	74.34

[16] C. Callison-Burch, P. Koehn, C. Monz, M. Post, R. Soricut, and L. Specia, "Findings of the 2012 workshop on statistical machine translation," in *7th Workshop on Statistical Machine Translation (WMT 2012)*, 2012, pp. 10–51.

[17] J. Giménez and L. 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. [Online]. Available: http://www.aclweb.org/anthology/W/W08/W08-0332

[19] C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder, "(meta-) evaluation of machine translation," in *Second Workshop on Statistical Machine Translation (WMT-07)*, 2007.

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

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

[22] J. Castillo and P. Estrella, "Semantic textual similarity for MT evaluation," in *7th Workshop on Statistical Machine Translation (WMT 2012)*, 2012.

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

[24] ——, "SMT vs. AI redux: How semantic frames evaluate MT more accurately," in *Twenty-second International Joint Conference on Artificial Intelligence (IJCAI-11)*, 2011.

[25] D. Wu, "Stochastic inversion transduction grammars and bilingual parsing of parallel corpora," *Computational Linguistics*, vol. 23, no. 3, pp. 377–403, 1997.

[26] R. Zens and H. Ney, "A comparative study on reordering constraints in statistical machine translation," in *41st Annual Meeting of the Association for Computational Linguistics (ACL-2003)*, Stroudsburg, Pennsylvania, 2003, pp. 144–151. [Online]. Available: http://dx.doi.org/10.3115/1075096.1075115

[27] M. Saers and D. Wu, "Improving phrase-based translation via word alignments from stochastic inversion transduction grammars," in *Third Workshop on Syntax and Structure in Statistical Translation (SSST-3)*, Boulder, Colorado, June 2009, pp. 28–36. [Online]. Available: http://www.aclweb.org/anthology/W/W09/W09-2304

[28] K. Addanki, C.-k. Lo, M. Saers, and D. Wu, "LTG vs. ITG coverage of cross-lingual verb frame alternations," in *16th Annual Conference of the European Association for Machine Translation (EAMT-2012)*, Trento, Italy, May 2012.

[29] M. Saers, J. Nivre, and D. Wu, "Learning stochastic bracketing inversion transduction grammars with a cubic time biparsing algorithm," in *11th International Conference on Parsing Technologies (IWPT'09)*, Paris, France, October 2009, pp. 29–32.

[30] F. J. 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

[31] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. 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.

[32] A. 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.

[33] O. 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.

[34] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," in *The Twenty-first National Conference on Artificial Intelligence (AAAI-06)*, vol. 21, 2006.

[35] C.-k. Lo and D. 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.

[36] M. Macháček and O. Bojar, "Results of the WMT13 metrics shared task," in *Eighth Workshop on Statistical Machine Translation (WMT 2013)*, Sofia, Bulgaria, August 2013.

[37] H.-P. Zhang, H.-K. Yu, D.-Y. Xiong, and Q. Liu, "Hhmm-based chinese lexical analyzer ictclas," in *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing - Volume 17*, ser. SIGHAN '03. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003, pp. 184–187. [Online]. Available: http://dx.doi.org/10.3115/1119250.1119280