# BiMEANT: Integrating Cross-Lingual and Monolingual Semantic Frame Similarities in the MEANT Semantic MT Evaluation Metric

Chi-kiu Lo and Dekai Wu(✉)

Human Language Technology Center,
Department of Computer Science and Engineering,
Hong Kong University of Science and Technology (HKUST),
Kowloon, Hong Kong
{jackielo,dekai}@cs.ust.hk

**Abstract.** We present experimental results showing that integrating cross-lingual semantic frame similarity into the semantic frame based automatic MT evaluation metric MEANT improves its correlation with human judgment on evaluating translation adequacy. Recent work shows that MEANT more accurately reflects translation adequacy than other automatic MT evaluation metrics such as BLEU or TER, and that moreover, optimizing SMT systems against MEANT robustly improves translation quality across different output languages. However, in some cases the human reference translation employs different scoping strategies from the input sentence and thus standard monolingual MEANT, which only assesses translation quality via the semantic frame similarity between the reference and machine translations, fails to fairly and accurately reward the adequacy of the machine translation. To address this issue we propose a new bilingual metric, BiMEANT, that correlates with human judgment more closely than MEANT by incorporating new cross-lingual semantic frame similarity assessments into MEANT.

## 1 Introduction

We show that a new bilingual version of MEANT (Lo *et al.* [19]) correlates with human judgments of translation adequacy even more closely than MEANT by integrating cross-lingual semantic frame similarity assessments. We assess cross-lingual semantic frame similarity by (1) incorporating BITG constraints for word alignment within the semantic role fillers, and (2) using simple lexical translation probabilities, instead of the monolingual context vector model used in MEANT for computing the semantic role fillers similarities. We then combine this cross-lingual semantic frame similarity into the MEANT score. Our results show that integrating cross-lingual semantic frame similarity into MEANT improves its correlation with human judgment on evaluating translation adequacy.

The MEANT family of metrics (Lo and Wu [20,22]; Lo *et al.* [19]) adopt the principle that a good translation is one where a human can successfully understand the central meaning of the foreign sentence as captured by the basic


L. Besacier et al. (Eds.): SLSP 2014, LNAI 8791, pp. 82–93, 2014.
DOI: 10.1007/978-3-319-11397-5_6

event structure: "*who did what to whom, for whom, when, where, how and why*" (Pradhan *et al.* [31]). 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. Previous work indicates that the MEANT family of metrics correlates better with human adequacy judgment than commonly used MT evaluation metrics (Lo and Wu [20,22]; Lo *et al.* [19]; Lo and Wu [24]; Macháček and Bojar [25]). In addition, MEANT has been shown to be tunable—translation adequacy across different genres (ranging from formal news to informal web forum and public speech) and different languages (English and Chinese) is improved by replacing BLEU or TER with MEANT during parameter tuning (Lo *et al.* [16]; Lo and Wu [23]; Lo *et al.* [18]).

Particularly for very different languages—Chinese and English, for instance—monolingual MT evaluation strategies that compare reference and machine translations, including MEANT, often fail to properly recognize cases where alternative strategies for scoping, topicalization, and the like are employed by the input sentence and the MT output, leading to artificial differences between the reference and machine translations. As pointed out in the empirical study of Addanki *et al.* [1], this can result in drastically different semantic frame annotations. To combat this, we propose a strategy where direct bilingual comparisons of the machine translation and the original input sentence are incorporated into MEANT.

## 2 Related Work

### 2.1 MT Evaluation Metrics

A number of large scale meta-evaluations (Callison-Burch *et al.* [6]; Koehn and Monz [13]) report cases where BLEU (Papineni *et al.* [30]) strongly disagrees with human judgments of translation adequacy. Other surface-form oriented metrics such as NIST (Doddington [8]), METEOR (Banerjee and Lavie [2]), CDER (Leusch *et al.* [14]), WER (Nießen *et al.* [27]), and TER (Snover *et al.* [35]) can also suffer from similar problems because the degree of n-gram match does not accurately reflect how well the "*who did what to whom, for whom, when, where, how and why*" is preserved across translation, particularly for very different language pairs where reference translations can be extremely non-deterministic.

To address these problems of n-gram based metrics, Owczarzak *et al.* [28,29] apply LFG to extend the approach of evaluating syntactic dependency structure similarity proposed by Liu and Gildea [15]. Although they showed improved correlation with human *fluency* judgments, they did not achieve higher correlation with human *adequacy* judgments than metrics like METEOR. TINE (Rios *et al.* [32]) is a recall-oriented evaluation metric which aims to preserve the basic event structure. However, its correlation with human adequacy judgments is similar to that of BLEU and worse than that of METEOR. For a semantic MT metric to work better at the current stage of technology, we believe that it is necessary to choose (1) a suitable abstraction level for the meaning representation, and (2) the right balance of precision and recall.

Instead of prioritizing simplicity and representational transparency as Rios *et al.* [32] and Owczarzak *et al.* [28,29] do, Giménez and Màrquez [11,12] incorporate several semantic similarity features into a huge collection of n-gram and syntactic features within ULC so as to improve correlation with human adequacy judgments (Callison-Burch *et al.* [4]; Giménez and Màrquez [11]; Callison-Burch *et al.* [5]; Giménez and Màrquez [12]). However, there is no work towards tuning an SMT system using a pure form of ULC perhaps due to its expensive run time. Both SPEDE (Wang and Manning [37]), an MT evaluation metric that predicts the edit sequence needed for the MT output to match the reference via an integrated probabilistic FSM and probabilistic PDA model, and Sagan (Castillo and Estrella [7]), a semantic textual similarity metric based on a complex textual entailment pipeline, may also be susceptible to similar problems. 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. Because of their expensive training, tuning and/or running times, such metrics become less useful in the MT system development cycle. We have taken the approach of keeping the representation of meaning simple and clear in MEANT, so that the resulting metric can not only be transparently understood when used in error analysis, but also be employed when scoring massive number of hypotheses for training and tuning MT systems.

### 2.2 The MEANT Family of Metrics

Addanki *et al.* [1] shows that for very different languages—Chinese and English, for example—there would be cases where alternative strategies for scoping or topicalization are employed by the input sentence and the MT output, leading to artificial differences between the reference and machine translations. This drives us to investigate avenues toward further improving MEANT by incorporating the cross-lingual semantic frame similarity into MEANT, so that translation output whose semantic structure is closer to the foreign input sentence than the human reference translation can be scored more fairly.

MEANT, which is a weighted f-score over the matched semantic role labels of the automatically aligned semantic frames and role fillers, has been shown to correlate with human adequacy judgments more highly than BLEU, NIST, TER, WER, CDER, and others (Lo *et al.* [19]). It is relatively easy to apply to other languages, requiring only an automatic semantic parser and a large monolingual corpus in the output language; these resources are used for identifying the semantic structures and the lexical similarity between the semantic role fillers of the reference and machine translations, respectively. 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 maximum weighted bipartite matching to align the semantic frames between the reference and machine translations, according to the lexical similarities of the predicates.

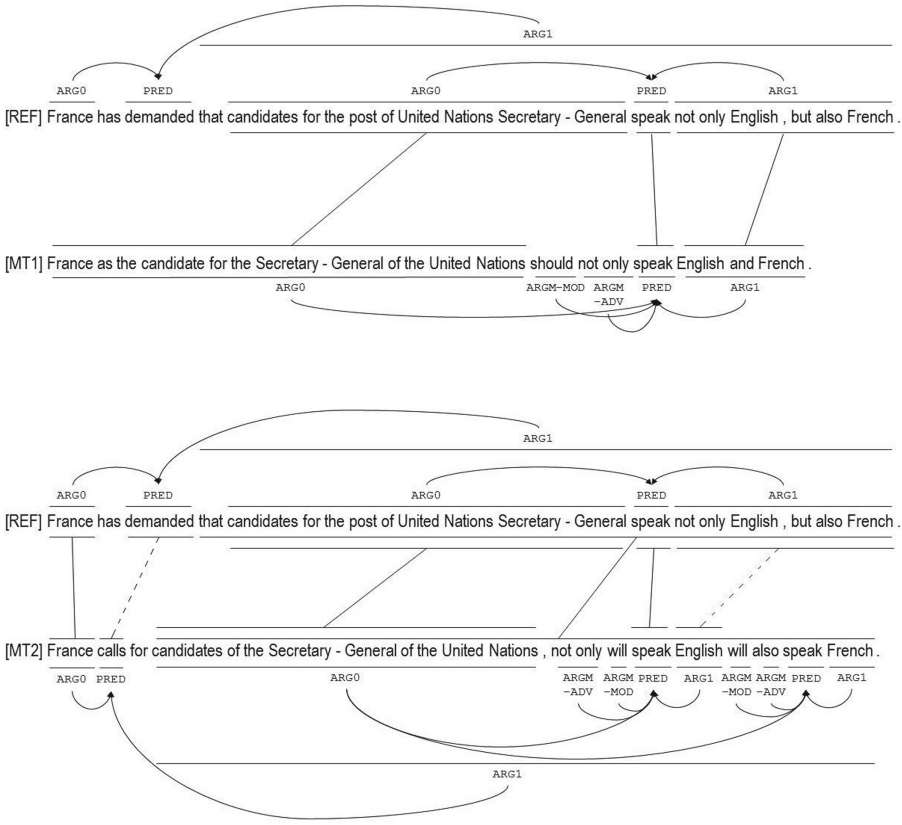


**Fig. 1.** Examples of monolingual semantic frame similarity captured by MEANT. MT2, the more adequate translation than MT1, is penalized by monolingual MEANT for producing translation output with more "inaccurate" semantic frames according to the reference translation. The dotted lines represent a low similarity (<0.5) semantic role alignments made by MEANT.

3. For each pair of the aligned frames, apply maximum weighted bipartite matching to align the arguments between the reference and machine translations, according to the lexical similarity of the semantic role fillers.
4. Compute the weighted f-score over the matching role labels of these aligned predicates and semantic role fillers according to the following definitions:

$$q^0_{i,j} \equiv \text{ARG j of aligned frame i in MT}$$
$$q^1_{i,j} \equiv \text{ARG j of aligned frame i in REF}$$
$$w^0_i \equiv \frac{\#\text{tokens filled in aligned frame i of MT}}{\text{total } \#\text{tokens in MT}}$$
$$w^1_i \equiv \frac{\#\text{tokens filled in aligned frame i of REF}}{\text{total } \#\text{tokens in REF}}$$

$$
\begin{aligned}
w_{\text{pred}} &\equiv \text{weight of similarity of predicates} \\
w_j &\equiv \text{weight of similarity of ARG j} \\
s_{i,\text{pred}} &\equiv \text{predicate similarity in aligned frame i} \\
s_{i,j} &\equiv \text{ARG j similarity in aligned frame i}
\end{aligned}
$$

$$
\text{precision} = \frac{\sum_i w_i^0 \frac{w_{\text{pred}} s_{i,\text{pred}} + \sum_j w_j s_{i,j}}{w_{\text{pred}} + \sum_j w_j \left| q_{i,j}^0 \right|}}{\sum_i w_i^0}
$$

$$
\text{recall} = \frac{\sum_i w_i^1 \frac{w_{\text{pred}} s_{i,\text{pred}} + \sum_j w_j s_{i,j}}{w_{\text{pred}} + \sum_j w_j \left| q_{i,j}^1 \right|}}{\sum_i w_i^1}
$$

$$
\text{MEANT} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} \cdot \text{recall}}
$$

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.

The weights $w_{\text{pred}}$ 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 are a total of 12 weights for the set of semantic role labels in MEANT as defined in Lo and Wu [21]. For MEANT, $w_{\text{pred}}$ and $w_j$ are determined using supervised estimation via a simple grid search to optimize the correlation with human adequacy judgments (Lo and Wu [20]). For UMEANT (Lo and Wu [22]), $w_{\text{pred}}$ 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 human judgments on adequacy of the development set are unavailable.

$s_{i,\text{pred}}$ and $s_{i,j}$ are the lexical similarities 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. Lo *et al.* [19] and Tumuluru *et al.* [36] described how the lexical and phrasal similarities of the semantic role fillers are computed using geometric mean. A subsequent variant of the phrasal aggregation function that normalizes phrasal similarities according to the phrase length more accurately was proposed in Mihalcea *et al.* [26] and used in the work of Lo *et al.* [16]; Lo and Wu [23]; Lo *et al.* [18] and later further improved by a f-score aggregation in Lo *et al.* [17].

Recent studies (Lo *et al.* [16]; Lo and Wu [23]; Lo *et al.* [18]) show that tuning MT systems 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 public speech.

The promising results in evaluating and tuning with MEANT has led us to the present question: is it possible to further improve MEANT's correlation with human adequacy judgments by leveraging not only monolingual, but also cross-lingual, semantic frame similarities?

## 3 BiMEANT: Bilingual Semantic Frame Accuracy

Our new bilingual metric starts with monolingual MEANT's assessment of the degree of goodness of the translation, and then also integrates an assessment of roughly how well the translation captures the core semantics of the foreign input utterance. Whereas MEANT measures the lexical similarity using the monolingual context vector model and aggregates the lexical similarity into phrasal similarity using a variant of the aggregation function in Mihalcea *et al.* [26], for the new additional subtask of measuring semantic frame similarity cross-lingually, we propose to instead substitute simple lexical translation probabilities and a length-normalized inside probability at the root of the BITG biparse (Wu [38]; Zens and Ney [39]; Saers and Wu [34]; Adanki *et al.* [1]).

An example of the sorts of issues that the bilingual approach empirically helps to alleviate is shown in Fig. 1, which depicts examples of automatic shallow semantic parses on both reference and machine translations. In this case, the translation output MT2 is a more adequate translation than MT1, yet it is still too harshly penalized by monolingual MEANT for producing translation output with more "inaccurate" semantic frames as judged against the reference translation. This issue arises here because of a scoping choice in handling "not only": MT2 legitimately chooses to apply it to two separate semantic frames for the "speak" predicate, instead of the reference translation's choice of moving it inside to apply to the ARG1 of a single "speak" predicate. The dashed lines represent a low similarity (<0.5) semantic role alignments made by MEANT.

The bilingual approach, however, additionally incorporates a second way of assessing how well the semantic frames have been preserved in translation. Figure 2 shows how, by integrating a cross-lingual semantic frame similarity into MEANT, BiMEANT is able to reward the MT2 output that is closer to the semantic structure of the foreign input more fairly and accurately. To accomplish this, we compute cross-lingual semantic frame similarity in BiMEANT as follows (the differences from MEANT are underlined):

1. Apply an input language automatic shallow semantic parser to the foreign input and an output language automatic shallow semantic parser to the MT output. (Figure 2 shows examples of automatic shallow semantic parses on both foreign input and MT output. The Chinese semantic parser used in our experiments is C-ASSERT in Fung *et al.* [9,10].)
2. Apply the maximum weighted bipartite matching algorithm to align the semantic frames between the foreign input and MT output according to the lexical translation probabilities of the predicates.
3. For each pair of the aligned frames, apply the maximum weighted bipartite matching algorithm to align the arguments between the foreign input and MT output according to the aggregated phrasal translation probabilities of the role fillers.
4. Compute the weighted f-score over the matching role labels of these aligned predicates and role fillers according to the definitions similar to those in Sect. 2.2 except for replacing REF with IN in $q^1_{i,j}$ and $w^1_i$.

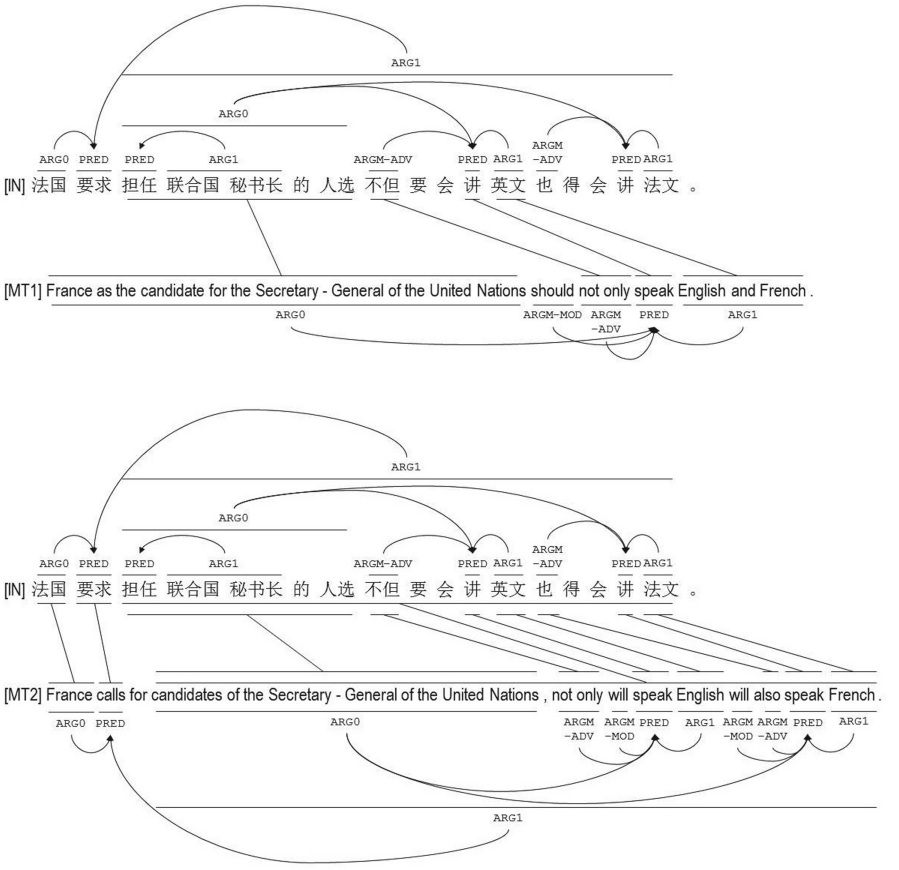


**Fig. 2.** Examples of bilingual semantic frame similarity captured by BiMEANT. MT2, the more adequate translation than MT1, is now fairly rewarded by the bilingual BiMEANT for producing translation with more accurate semantic frames according to the foreign input.

$$
\begin{aligned}
\mathbf{e}_{i,\text{pred}} &\equiv \text{the output side of the pred of aligned frame } i \\
\mathbf{f}_{i,\text{pred}} &\equiv \text{the input side of the pred of aligned frame } i \\
\mathbf{e}_{i,j} &\equiv \text{the output side of the ARG } j \text{ of aligned frame } i \\
\mathbf{f}_{i,j} &\equiv \text{the input side of the ARG } j \text{ of aligned frame } i \\
G &\equiv \langle \{\mathrm{A}\}, \mathcal{W}^0, \mathcal{W}^1, \mathcal{R}, \mathrm{A} \rangle \\
\mathcal{R} &\equiv \{\mathrm{A} \rightarrow [\mathrm{AA}], \mathrm{A} \rightarrow \langle \mathrm{AA} \rangle, \mathrm{A} \rightarrow e/f\} \\
p([\mathrm{AA}]|\mathrm{A}) &= p(\langle \mathrm{AA} \rangle | \mathrm{A}) = 0.25 \\
p(e/f|\mathrm{A}) &= \frac{1}{2}\sqrt{t(e|f)\, t(f|e)}
\end{aligned}
$$

$$\mathrm{xs}_{i,\mathrm{pred}} = \frac{1}{1 - \frac{\ln\left(P\left(\mathrm{A} \overset{*}{\Rightarrow} \mathbf{e}_{i,\mathrm{pred}}/\mathbf{f}_{i,\mathrm{pred}}|G\right)\right)}{\max(|\mathbf{e}_{i,\mathrm{pred}}|,|\mathbf{f}_{i,\mathrm{pred}}|)}}$$

$$\mathrm{xs}_{i,j} = \frac{1}{1 - \frac{\ln\left(P\left(\mathrm{A} \overset{*}{\Rightarrow} \mathbf{e}_{i,j}/\mathbf{f}_{i,j}|G\right)\right)}{\max(|\mathbf{e}_{i,j}|,|\mathbf{f}_{i,j}|)}}$$

$$\mathrm{xp} = \frac{\sum_i w_i^0 \frac{w_{\mathrm{pred}}\mathrm{xs}_{i,\mathrm{pred}} + \sum_j w_j \mathrm{xs}_{i,j}}{w_{\mathrm{pred}} + \sum_j w_j |q_{i,j}^0|}}{\sum_i w_i^0}$$

$$\mathrm{xr} = \frac{\sum_i w_i^1 \frac{w_{\mathrm{pred}}\mathrm{xs}_{i,\mathrm{pred}} + \sum_j w_j \mathrm{xs}_{i,j}}{w_{\mathrm{pred}} + \sum_j w_j |q_{i,j}^1|}}{\sum_i w_i^1}$$

$$\mathrm{xf} = \frac{2 \cdot \mathrm{xp} \cdot \mathrm{xr}}{\mathrm{xp} \cdot \mathrm{xr}}$$

$$\mathrm{BiMEANT} = [\alpha * \mathrm{MEANT} + (1 - \alpha) * \mathrm{xf}]$$

where $G$ is a bracketing ITG, whose only nonterminal is $A$, and where $\mathcal{R}$ is a set of transduction rules where $e \in \mathcal{W}^0 \cup \{\epsilon\}$ is an output token (or the *null* token), and $f \in \mathcal{W}^1 \cup \{\epsilon\}$ is an input token (or the *null* token). The rule probability function $p$ is defined using fixed probabilities for the structural rules, and a translation table $t$ trained using IBM model 1 (Brown *et al.* [3]) in both directions. A small constant ($10^{-5}$) is used when one of the $t$s is undefined. To calculate the inside probability of a pair of segments, $P\left(\mathrm{A} \overset{*}{\Rightarrow} e/f|G\right)$, we use the algorithm described in Saers *et al.* [33]. $\mathrm{xs}_{i,\mathrm{pred}}$ and $\mathrm{xs}_{i,j}$ are the length normalized BITG parsing probabilities of the predicates and role fillers of the arguments of type $j$ between the input and the MT output. xp, xr and xf are the precision, recall and f-score of the cross-lingual semantic frame similarity computed by aggregating the BITG parsing probabilities of the predicates and role fillers in the same way as MEANT.

## 4 Results

Table 1 shows that BiMEANT significantly outperforms MEANT on sentence-level correlation with human adequacy judgment. This occurs despite the fact that only minimal adaptation has been done on the phrasal similarities for the cross-lingual semantic role fillers, suggesting that the performance of BiMEANT may be even better when settings are optimized.

Preliminary analysis indicates two reasons that BiMEANT improves correlation with human adequacy judgement. First, the semantic structure of the MT output often tends to be closer to that of the input sentence than that of the reference translation, due to somewhat arbitrary choices in scoping, topicalization, and similar phenomena. Secondly, the BITG constraints used in the cross-lingual assessment provide a more robust phrasal similarity aggregation function compared to the naive bag-of-words based heuristics previously employed in MEANT.

Similar results have been observed while trying to estimate word alignment probabilities where BITG constraints outperformed alignments from GIZA++ (Saers and Wu [34]).

**Table 1.** Sentence-level correlation with human adequacy judgement (GALE phase 2.5 evaluation data)

| | *Kendall* |
|---|---|
| BiMEANT | **0.50** |
| MEANT | 0.46 |
| NIST | 0.29 |
| BLEU/METEOR/TER/PER | 0.20 |
| CDER | 0.12 |
| WER | 0.10 |

## 5 Conclusion

We have presented a new bilingual automatic MT evaluation metric, BiMEANT, that correlates even more closely with human judgments of translation adequacy than standard monolingual MEANT. While previous work has established that MEANT accurately reflects translation adequacy via semantic frames and that optimizing SMT against MEANT improves translation quality, for very different languages the performance of purely monolingual metrics such as MEANT can be degraded by surface differences in choices such as scoping or topicalization, that lead to artificial differences between the reference and machine translations. The bilingual strategy employed by BiMEANT combats this by incorporating cross-lingual similarity assessments directly between the semantic frames of the input and output sentences. This is accomplished by (1) incorporating bracketing ITG constraints for aligning the lexicons in semantic role fillers, and (2) replacing the monolingual context vector model in MEANT with simple translation probabilities for computing the similarities of the semantic role fillers.

We would like to note that in this first study on a bilingual semantic frame based MT evaluation metric, we have performed minimal adaptation on the phrasal similarity assessments for the cross-lingual semantic role fillers. It is reasonable to expect that the performance of BiMEANT may further improve when the settings are optimized. The encouraging results suggest interesting potential for BiMEANT, especially across very different languages.

**Acknowledgment.** This material is based upon work supported in part 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;

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. Thanks to Markus Saers, Meriem Beloucif, and Karteek Addanki for supporting work, and to Pascale Fung, Yongsheng Yang and Zhaojun Wu for sharing the maximum entropy Chinese segmenter and C-ASSERT, the Chinese semantic parser.

## References

1. Addanki, K., Lo, C., Saers, M., Wu, D.: 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
2. Banerjee, S., Lavie, A.: 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
3. Brown, P.F., Della, P., Stephen, A., Della, P., Vincent, J., Mercer, R.L.: The mathematics of machine translation: parameter estimation. Comput. Linguist. **19**(2), 263–311 (1993)
4. Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., Schroeder, J.: (meta-) evaluation of machine translation. In: Second Workshop on Statistical Machine Translation (WMT-07) (2007)
5. Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., Schroeder, J.: Further meta-evaluation of machine translation. In: Third Workshop on Statistical Machine Translation (WMT-08) (2008)
6. Callison-Burch, C., Osborne, M., Koehn, P.: 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)
7. Castillo, J., Estrella, P.: Semantic textual similarity for MT evaluation. In: 7th Workshop on Statistical Machine Translation (WMT 2012) (2012)
8. Doddington, G.: 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)
9. Fung, P., Ngai, G., Yang, Y., Chen, B.: A maximum-entropy chinese parser augmented by transformation-based learning. ACM Trans. Asian Lang. Inf. Process. (TALIP) **3**(2), 159–168 (2004)
10. Fung, P., Wu, Z., Yang, Y., Wu, D.: Learning bilingual semantic frames: shallow semantic parsing vs. semantic role projection. In: The 11th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-07), Skovde, Sweden, pp. 75–84 (2007)
11. Giménez, J., Màrquez, L.: 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 (2007)
12. Giménez, J., Màrquez, L.: A smorgasbord of features for automatic MT evaluation. In: Third Workshop on Statistical Machine Translation (WMT-08), Columbus, Ohio, June 2008
13. Koehn, P., Monz, C.: Manual and automatic evaluation of machine translation between european languages. In: Workshop on Statistical Machine Translation (WMT-06) (2006)

14. Leusch, G., Ueffing, N., Ney, H.: CDer: Efficient MT evaluation using block movements. In: 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006) (2006)
15. Liu, D., Gildea, D.: Syntactic features for evaluation of machine translation. In: Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, Michigan, June 2005
16. Lo, C., Addanki, K., Saers, M., Wu, D.: 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)
17. Lo, C., Beloucif, M., Saers, M., Wu, D.: XMEANT: better semantic MT evaluation without reference translations. In: 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014) (2014)
18. Lo, C., Beloucif, M., Wu, D.: Improving machine translation into Chinese by tuning against Chinese MEANT. In: International Workshop on Spoken Language Translation (IWSLT 2013) (2013)
19. Lo, C., Tumuluru, A.K., Wu, D.: Fully automatic semantic MT evaluation. In: 7th Workshop on Statistical Machine Translation (WMT 2012) (2012)
20. Lo, C., Wu, D.: 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)
21. Lo, C., Wu, D.: SMT vs. AI redux: how semantic frames evaluate MT more accurately. In: 22nd International Joint Conference on Artificial Intelligence (IJCAI-11) (2011)
22. Lo, C., Wu, D.: Unsupervised vs. supervised weight estimation for semantic MT evaluation metrics. In: Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-6) (2012)
23. Lo, C., Wu, D.: Can informal genres be better translated by tuning on automatic semantic metrics? In: 14th Machine Translation Summit (MT Summit XIV) (2013)
24. Lo, C., Wu, D.: 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)
25. Macháček, M., Bojar, O.: Results of the WMT13 metrics shared task. In: 8th Workshop on Statistical Machine Translation (WMT 2013), Sofia, Bulgaria, August 2013
26. Mihalcea, R., Corley, C., Strapparava, C.: Corpus-based and knowledge-based measures of text semantic similarity. In: The 21st National Conference on Artificial Intelligence (AAAI-06), vol. 21 (2006)
27. Nießen, S., Och, F. J., Leusch, G., Ney, H.: A evaluation tool for machine translation: fast evaluation for MT research. In: The 2nd International Conference on Language Resources and Evaluation (LREC 2000) (2000)
28. Owczarzak, K., van Genabith, J., Way, A.: Dependency-based automatic evaluation for machine translation. In: Syntax and Structure in Statistical Translation (SSST) (2007)
29. Owczarzak, K., van Genabith, J., Way, A.: Evaluating machine translation with LFG dependencies. Mach. Transl. **21**, 95–119 (2007)
30. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: 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 (2002)

31. Pradhan, S., Ward, W., Hacioglu, K., Martin, J. H., Jurafsky, D.: 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)
32. Rios, M., Aziz, W., Specia, L.: TINE: a metric to assess MT adequacy. In: 6th Workshop on Statistical Machine Translation (WMT 2011) (2011)
33. Saers, M., Nivre, J., Wu, D.: 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 (2009)
34. Saers, M., Wu, D.: 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 (2009)
35. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: 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 (2006)
36. Tumuluru, A. K., Lo, C., Wu, D.: Accuracy and robustness in measuring the lexical similarity of semantic role fillers for automatic semantic MT evaluation. In: 26th Pacific Asia Conference on Language, Information, and Computation (PACLIC 26) (2012)
37. Wang, M., Manning, C.D.: SPEDE: probabilistic edit distance metrics for MT evaluation. In: 7th Workshop on Statistical Machine Translation (WMT 2012) (2012)
38. Wu, D.: Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. Comput. Linguist. **23**(3), 377–403 (1997)
39. Zens, R., Ney, H.: A comparative study on reordering constraints in statistical machine translation. In: 41st Annual Meeting of the Association for Computational Linguistics (ACL-2003), Stroudsburg, Pennsylvania, pp. 144–151 (2003)