XMEANT: Better semantic MT evaluation without reference translations

Lo, Chi-kiu Beloucif, Meriem Saers, Markus Wu, Dekai
HKUST
Human Language Technology Center
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
{jackielo|mbeloucif|masaers|dekai}@cs.ust.hk

Abstract

We introduce XMEANT—a new *cross-lingual* version of the semantic frame based MT evaluation metric MEANT—which can correlate even more closely with human adequacy judgments than monolingual MEANT and eliminates the need for expensive human references. Previous work established that MEANT reflects translation adequacy with state-of-the-art accuracy, and optimizing MT systems against MEANT robustly improves translation quality. However, to go beyond tuning weights in the loglinear SMT model, a cross-lingual objective function that can deeply integrate semantic frame criteria into the MT training pipeline is needed. We show that cross-lingual XMEANT outperforms monolingual MEANT by (1) replacing the monolingual context vector model in MEANT with simple translation probabilities, and (2) incorporating bracketing ITG constraints.

1 Introduction

We show that XMEANT, a new cross-lingual version of MEANT (Lo et al., 2012), correlates with human judgment even more closely than MEANT for evaluating MT adequacy via semantic frames, despite discarding the need for expensive human reference translations. XMEANT is obtained by (1) using simple lexical translation probabilities, instead of the monolingual context vector model used in MEANT for computing the semantic role fillers similarities, and (2) incorporating bracketing ITG constrains for word alignment within the semantic role fillers. We conjecture that the reason that XMEANT correlates more closely with human adequacy judgement than MEANT is that on the one hand, the semantic structure of the MT output is closer to that of the input sentence than that of the reference translation, and on the other hand, the BITG constraints the word alignment more accurately than the heuristic bag-of-word aggregation used in MEANT. Our results suggest that MT translation adequacy is more accurately evaluated via the cross-lingual semantic frame similarities of the input and the MT output which may obviate the need for expensive human reference translations.

The MEANT family of metrics (Lo and Wu, 2011a, 2012; Lo et al., 2012) 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 event structure: "*who did what to whom, when, where and why*" (Pradhan et al., 2004). 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. It is well established that the MEANT family of metrics correlates better with human adequacy judgments than commonly used MT evaluation metrics (Lo and Wu, 2011a, 2012; Lo et al., 2012; Lo and Wu, 2013b; Macháček and Bojar, 2013). In addition, the 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., 2013a; Lo and Wu, 2013a; Lo et al., 2013b).

In order to continue driving MT towards better translation adequacy by deeply integrating semantic frame criteria into the MT training pipeline, it is necessary to have a *cross-lingual semantic objective function* that assesses the semantic frame similarities of input and output sentences. We therefore propose XMEANT, a cross-lingual MT evaluation metric, that modifies MEANT using (1) simple translation probabilities (in our experiments,

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 765–771,
Baltimore, Maryland, USA, June 23-25 2014.

from quick IBM-1 training), to replace the monolingual context vector model in MEANT, and (2) constraints from BITGs (bracketing ITGs). We show that XMEANT assesses MT adequacy more accurately than MEANT (as measured by correlation with human adequacy judgement) without the need for expensive human reference translations in the output language.

2 Related Work

2.1 MT evaluation metrics

Surface-form oriented metrics such as BLEU (Papineni et al., 2002), NIST (Doddington, 2002), METEOR (Banerjee and Lavie, 2005), CDER (Leusch et al., 2006), WER (Nießen et al., 2000), and TER (Snover et al., 2006) do not correctly reflect the meaning similarities of the input sentence. In fact, a number of large scale meta-evaluations (Callison-Burch et al., 2006; Koehn and Monz, 2006) report cases where BLEU strongly disagrees with human judgments of translation adequacy.

This has caused a recent surge of work to develop better ways to automatically measure MT adequacy. Owczarzak et al. (2007a,b) improved correlation with human *fluency* judgments by using LFG to extend the approach of evaluating syntactic dependency structure similarity proposed by Liu and Gildea (2005), but did not achieve higher correlation with human *adequacy* judgments than metrics like METEOR. TINE (Rios et al., 2011) is a recall-oriented metric which aims to preserve the basic event structure but it performs comparably to BLEU and worse than METEOR on correlation with human adequacy judgments. ULC (Giménez and Màrquez, 2007, 2008) incorporates several semantic features and shows improved correlation with human judgement on translation quality (Callison-Burch et al., 2007, 2008) 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 (Wang and Manning, 2012) predicts the edit sequence for matching the MT output to the reference via an integrated probabilistic FSM and PDA model. Sagan (Castillo and Estrella, 2012) 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 or paraphrase tables; the expensive training, tuning, and/or running time makes them hard to incorporate into the MT development cycle.

1. Apply an automatic shallow semantic parser to both the references and MT output. (Figure 2 shows examples of automatic shallow semantic parses on both reference and MT.)

2. Apply the maximum weighted bipartite matching algorithm to align the semantic frames between the references and MT output according to the lexical similarities of the predicates.

3. For each pair of the aligned frames, apply the maximum weighted bipartite matching algorithm to align the arguments between the reference and MT output 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 according to the following definitions:

$$
\begin{aligned}
q^0_{i,j} &\equiv \text{ARG j of aligned frame i in MT} \\
q^1_{i,j} &\equiv \text{ARG j of aligned frame i in REF} \\
w^0_i &\equiv \frac{\text{\#tokens filled in aligned frame i of MT}}{\text{total \#tokens in MT}} \\
w^1_i &\equiv \frac{\text{\#tokens filled in aligned frame i of REF}}{\text{total \#tokens in REF}} \\
w_{\text{pred}} &\equiv \text{weight of similarity of predicates} \\
w_j &\equiv \text{weight of similarity of ARG j} \\
s_{i,\text{pred}} &\equiv \text{predicate similarity in aligned frame i} \\
s_{i,j} &\equiv \text{ARG j similarity in aligned frame i} \\
\text{precision} &= \frac{\sum_i w^0_i \frac{w_{\text{pred}} s_{i,\text{pred}} + \sum_j w_j s_{i,j}}{w_{\text{pred}} + \sum_j w_j |q^0_{i,j}|}}{\sum_i w^0_i} \\
\text{recall} &= \frac{\sum_i w^1_i \frac{w_{\text{pred}} s_{i,\text{pred}} + \sum_j w_j s_{i,j}}{w_{\text{pred}} + \sum_j w_j |q^1_{i,j}|}}{\sum_i w^1_i} \\
\text{MEANT} &= \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}
\end{aligned}
$$

Figure 1: Monolingual MEANT algorithm.

2.2 The MEANT family of metrics

MEANT (Lo et al., 2012), which is the weighted f-score over the matched semantic role labels of the automatically aligned semantic frames and role fillers, that outperforms BLEU, NIST, METEOR, WER, CDER and TER in correlation with human adequacy judgments. 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.

Figure 1 shows the algorithm and equations for computing MEANT. $q^0_{i,j}$ and $q^1_{i,j}$ are the argument of type j in frame i in MT and REF respectively. w^0_i and w^1_i 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 MT out-

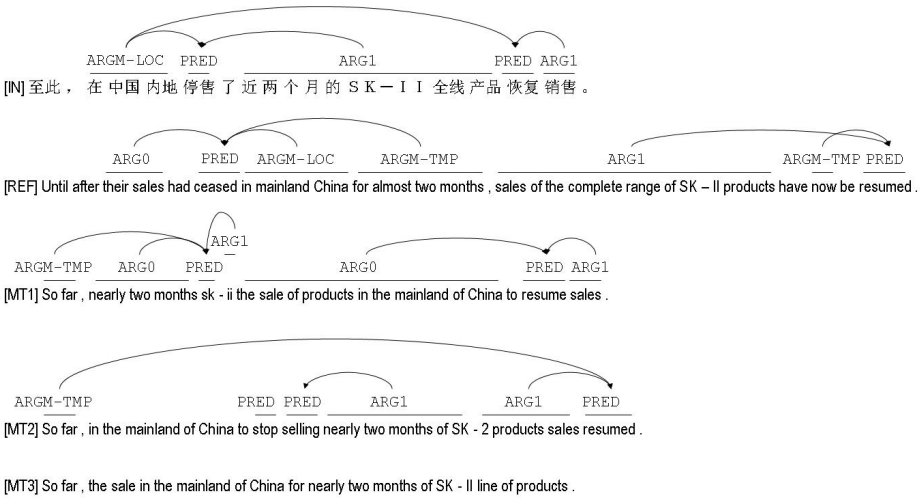

Figure 2: Examples of automatic shallow semantic parses. The input is parsed by a Chinese automatic shallow semantic parser. The reference and MT output are parsed by an English automatic shallow semantic parser. There are no semantic frames for MT3 since the system decided to drop the predicate.

put. There is a total of 12 weights for the set of semantic role labels in MEANT as defined in Lo and Wu (2011b). For MEANT, they are determined using supervised estimation via a simple grid search to optimize the correlation with human adequacy judgments (Lo and Wu, 2011a). For UMEANT (Lo and Wu, 2012), 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.

$s_{i,\text{pred}}$ 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. (2012) and Tumuluru et al. (2012) described how the lexical and phrasal similarities of the semantic role fillers are computed. A subsequent variant of the aggregation function inspired by Mihalcea et al. (2006) that normalizes phrasal similarities according to the phrase length more accurately was used in more recent work (Lo et al., 2013a; Lo and Wu, 2013a; Lo et al., 2013b). In this paper, we employ a newer version of MEANT that uses f-score to aggregate individual token similarities into the composite phrasal similarities of semantic role fillers, as our experiments indicate this is more accurate than the previously used aggregation functions.

Recent studies (Lo et al., 2013a; Lo and Wu, 2013a; Lo et al., 2013b) 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 informal public speech.

2.3 MT quality estimation

Evaluating cross-lingual MT quality is similar to the work of MT quality estimation (QE). Broadly speaking, there are two different approaches to QE: surface-based and feature-based.

Token-based QE models, such as those in Gandrabur et al. (2006) and Ueffing and Ney (2005) fail to assess the overall MT quality because translation goodness is not a compositional property. In contrast, Blatz et al. (2004) introduced a sentence-level QE system where an arbitrary threshold is used to classify the MT output as *good* or *bad*. The fundamental problem of this approach is that it defines QE as a binary classification task rather than attempting to measure the degree of goodness of the MT output. To address this problem, Quirk (2004) related the sentence-level correctness of the QE model to human judgment and achieved a high correlation with human judgement for a small annotated corpus; however, the proposed model does not scale well to larger data sets.

Feature-based QE models (Xiong et al., 2010; He et al., 2011; Ma et al., 2011; Specia, 2011; Avramidis, 2012; Mehdad et al., 2012; Almaghout and Specia, 2013; Avramidis and Popović, 2013; Shah et al., 2013) throw a wide range of linguistic and non-linguistic features into machine learn-

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.*, 2004, 2007).)

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 section 2.2 except for replacing REF with IN in $q^1_{i,j}$ and w^1_i.

Figure 3: Cross-lingual XMEANT algorithm.

ing algorithms for predicting MT quality. Although the feature-based QE system of Avramidis and Popović (2013) slightly outperformed METEOR on correlation with human adequacy judgment, these "black box" approaches typically lack representational transparency, require expensive running time, and/or must be discriminatively retrained for each language and text type.

3 XMEANT: a cross-lingual MEANT

Like MEANT, XMEANT aims to evaluate how well MT preserves the core semantics, while maintaining full representational transparency. But whereas MEANT measures lexical similarity using a monolingual context vector model, XMEANT instead substitutes simple cross-lingual lexical translation probabilities.

XMEANT differs only minimally from MEANT, as underlined in figure 3. The same weights obtained by optimizing MEANT against human adequacy judgement were used for XMEANT. The weights can also be estimated in unsupervised fashion using the relative frequency of each semantic role label in the foreign input, as in UMEANT.

To aggregate individual lexical translation probabilities into phrasal similarities between cross-lingual semantic role fillers, we compared two natural approaches to generalizing MEANT's method of comparing semantic parses, as described below.

3.1 Applying MEANT's f-score within semantic role fillers

The first natural approach is to extend MEANT's f-score based method of aggregating semantic parse accuracy, so as to also apply to aggregating lexical translation probabilities *within* semantic role filler phrases. However, since we are missing structure information within the flat role filler phrases, we can no longer assume an injective mapping for aligning the tokens of the role fillers between the foreign input and the MT output. We therefore relax the assumption and thus for cross-lingual phrasal precision/recall, we align each token of the role fillers in the output/input string to the token of the role fillers in the input/output string that has the maximum lexical translation probability. The precise definition of the cross-lingual phrasal similarities is as follows:

$$
\begin{aligned}
\mathbf{e}_{i,\text{pred}} &\equiv \text{the output side of the pred of aligned frame } i \\
\mathbf{f}_{i,\text{pred}} &\equiv \text{the input side of the pred of aligned frame } i \\
\mathbf{e}_{i,j} &\equiv \text{the output side of the ARG } j \text{ of aligned frame } i \\
\mathbf{f}_{i,j} &\equiv \text{the input side of the ARG } j \text{ of aligned frame } i \\
p(e,f) &= \sqrt{t\,(e|f)\,t\,(f|e)} \\
\text{prec}_{\mathbf{e},\mathbf{f}} &= \frac{\sum_{e\in\mathbf{e}} \max_{f\in\mathbf{f}} p(e,f)}{|\mathbf{e}|} \\
\text{rec}_{\mathbf{e},\mathbf{f}} &= \frac{\sum_{f\in\mathbf{f}} \max_{e\in\mathbf{e}} p(e,f)}{|\mathbf{f}|} \\
s_{i,\text{pred}} &= \frac{2\cdot \text{prec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}} \cdot \text{rec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}}}{\text{prec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}} + \text{rec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}}} \\
s_{i,j} &= \frac{2\cdot \text{prec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}} \cdot \text{rec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}}}{\text{prec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}} + \text{rec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}}}
\end{aligned}
$$

where the joint probability p is defined as the harmonized the two directions of the translation table t trained using IBM model 1 (Brown et al., 1993). $\text{prec}_{\mathbf{e},\mathbf{f}}$ is the precision and $\text{rec}_{\mathbf{e},\mathbf{f}}$ is the recall of the phrasal similarities of the role fillers. $s_{i,\text{pred}}$ and $s_{i,j}$ are the f-scores of the phrasal similarities of the predicates and role fillers of the arguments of type j between the input and the MT output.

3.2 Applying MEANT's ITG bias within semantic role fillers

The second natural approach is to extend MEANT's ITG bias on compositional reordering, so as to also apply to aggregating lexical translation probabilities *within* semantic role filler phrases. Addanki et al. (2012) showed empirically that cross-lingual semantic role reordering of the kind that MEANT is based upon is fully covered within ITG constraints. In Wu et al. (2014), we extend ITG constraints into aligning the tokens within the semantic role fillers within monolingual MEANT, thus replacing its previous monolingual phrasal aggregation heuristic. Here we borrow the

idea for the cross-lingual case, using the length-normalized inside probability at the root of a BITG biparse (Wu, 1997; Zens and Ney, 2003; Saers and Wu, 2009) as follows:

$$
\begin{aligned}
G &\equiv \langle \{\mathrm{A}\}, \mathcal{W}^0, \mathcal{W}^1, \mathcal{R}, \mathrm{A} \rangle \\
\mathcal{R} &\equiv \{\mathrm{A} \rightarrow [\mathrm{AA}], \mathrm{A} \rightarrow \langle \mathrm{AA} \rangle, \mathrm{A} \rightarrow e/f\} \\
p([\mathrm{AA}]|\mathrm{A}) &= p(\langle \mathrm{AA} \rangle | \mathrm{A}) = 0.25 \\
p(e/f|\mathrm{A}) &= \frac{1}{2}\sqrt{t(e|f)\,t(f|e)} \\
s_{i,\mathrm{pred}} &= \frac{1}{1 - \frac{\ln\left(P\left(\mathrm{A} \stackrel{*}{\Rightarrow} \mathbf{e}_{i,\mathrm{pred}}/\mathbf{f}_{i,\mathrm{pred}}|G\right)\right)}{\max(|\mathbf{e}_{i,\mathrm{pred}}|,|\mathbf{f}_{i,\mathrm{pred}}|)}} \\
s_{i,j} &= \frac{1}{1 - \frac{\ln\left(P\left(\mathrm{A} \stackrel{*}{\Rightarrow} \mathbf{e}_{i,j}/\mathbf{f}_{i,j}|G\right)\right)}{\max(|\mathbf{e}_{i,j}|,|\mathbf{f}_{i,j}|)}}
\end{aligned}
$$

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 in both directions. To calculate the inside probability of a pair of segments, $P\left(\mathrm{A} \stackrel{*}{\Rightarrow} \mathbf{e}/\mathbf{f}|G\right)$, we use the algorithm described in Saers et al. (2009). $s_{i,\mathrm{pred}}$ and $s_{i,j}$ 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.

4 Results

Table 1 shows that for human adequacy judgments at the sentence level, the f-score based XMEANT (1) correlates significantly more closely than other commonly used monolingual automatic MT evaluation metrics, and (2) even correlates nearly as well as monolingual MEANT. This suggests that the semantic structure of the MT output is indeed closer to that of the input sentence than that of the reference translation.

Furthermore, the ITG-based XMEANT (1) significantly outperforms MEANT, and (2) is an automatic metric that is nearly as accurate as the HMEANT human subjective version. This indicates that BITG constraints indeed provide a more robust token alignment compared to the heuristics previously employed in MEANT. It is also consistent with results observed while estimating word alignment probabilities, where BITG constraints outperformed alignments from GIZA++ (Saers and Wu, 2009).

Table 1: Sentence-level correlation with HAJ (GALE phase 2.5 evaluation data)

Metric	*Kendall*
HMEANT	0.53
XMEANT (BITG)	**0.51**
MEANT (f-score)	0.48
XMEANT (f-score)	**0.46**
MEANT (2013)	0.46
NIST	0.29
BLEU/METEOR/TER/PER	0.20
CDER	0.12
WER	0.10

5 Conclusion

We have presented XMEANT, a new cross-lingual variant of MEANT, that correlates even more closely with human translation adequacy judgments than MEANT, without the expensive human references. This is (1) accomplished by replacing monolingual MEANT's context vector model with simple translation probabilities when computing similarities of semantic role fillers, and (2) further improved by incorporating BITG constraints for aligning the tokens in semantic role fillers. While monolingual MEANT alone accurately reflects adequacy via semantic frames and optimizing SMT against MEANT improves translation, the new cross-lingual XMEANT semantic objective function moves closer toward deep integration of semantics into the MT training pipeline.

The phrasal similarity scoring has only been minimally adapted to cross-lingual semantic role fillers in this first study of XMEANT. We expect further improvements to XMEANT, but these first results already demonstrate XMEANT's potential to drive research progress toward semantic SMT.

6 Acknowledgments

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.

References

Karteek Addanki, Chi-Kiu Lo, Markus Saers, and Dekai 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.

Hala Almaghout and Lucia Specia. A CCG-based quality estimation metric for statistical machine translation. In *Machine Translation Summit XIV (MT Summit 2013)*, Nice, France, 2013.

Eleftherios Avramidis and Maja Popović. Machine learning methods for comparative and time-oriented quality estimation of machine translation output. In *8th Workshop on Statistical Machine Translation (WMT 2013)*, 2013.

Eleftherios Avramidis. Quality estimation for machine translation output using linguistic analysis and decoding features. In *7th Workshop on Statistical Machine Translation (WMT 2012)*, 2012.

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.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. Confidence estimation for machine translation. In *20th international conference on Computational Linguistics*, 2004.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.

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.

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.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. Further meta-evaluation of machine translation. In *Third Workshop on Statistical Machine Translation (WMT-08)*, 2008.

Julio Castillo and Paula Estrella. Semantic textual similarity for MT evaluation. In *7th Workshop on Statistical Machine Translation (WMT 2012)*, 2012.

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.

Simona Gandrabur, George Foster, and Guy Lapalme. Confidence estimation for nlp applications. *ACM Transactions on Speech and Language Processing*, 2006.

Jesús Giménez and Lluís Màrquez. Linguistic features for automatic evaluation of heterogenous MT systems. In *Second Workshop on Statistical Machine Translation (WMT-07)*, pages 256–264, Prague, Czech Republic, June 2007.

Jesús Giménez and Lluís Màrquez. A smorgasbord of features for automatic MT evaluation. In *Third Workshop on Statistical Machine Translation (WMT-08)*, Columbus, Ohio, June 2008.

Yifan He, Yanjun Ma, Andy Way, and Josef van Genabith. Rich linguistic features for translation memory-inspired consistent translation. In *13th Machine Translation Summit (MT Summit XIII)*, 2011.

Philipp Koehn and Christof Monz. Manual and automatic evaluation of machine translation between european languages. In *Workshop on Statistical Machine Translation (WMT-06)*, 2006.

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.

Ding Liu and Daniel Gildea. 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.

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.

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.

Chi-kiu Lo and Dekai Wu. Unsupervised vs. supervised weight estimation for semantic MT evaluation metrics. In *Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-6)*, 2012.

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.

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.

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

Chi-kiu Lo, Karteek Addanki, Markus Saers, and Dekai Wu. Improving machine translation by training against an automatic semantic frame based eval-

uation metric. In *51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, 2013.

Chi-kiu Lo, Meriem Beloucif, and Dekai Wu. Improving machine translation into Chinese by tuning against Chinese MEANT. In *International Workshop on Spoken Language Translation (IWSLT 2013)*, 2013.

Yanjun Ma, Yifan He, Andy Way, and Josef van Genabith. Consistent translation using discriminative learning: a translation memory-inspired approach. In *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*. Association for Computational Linguistics, 2011.

Matouš Macháček and Ondřej Bojar. Results of the WMT13 metrics shared task. In *Eighth Workshop on Statistical Machine Translation (WMT 2013)*, Sofia, Bulgaria, August 2013.

Yashar Mehdad, Matteo Negri, and Marcello Federico. Match without a referee: evaluating mt adequacy without reference translations. In *7th Workshop on Statistical Machine Translation (WMT 2012)*, 2012.

Rada Mihalcea, Courtney Corley, and Carlo Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. In *The Twenty-first National Conference on Artificial Intelligence (AAAI-06)*, volume 21. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

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.

Karolina Owczarzak, Josef van Genabith, and Andy Way. Dependency-based automatic evaluation for machine translation. In *Syntax and Structure in Statistical Translation (SSST)*, 2007.

Karolina Owczarzak, Josef van Genabith, and Andy Way. Evaluating machine translation with LFG dependencies. *Machine Translation*, 21:95–119, 2007.

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)*, pages 311–318, Philadelphia, Pennsylvania, July 2002.

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.

Christopher Quirk. Training a sentence-level machine translation confidence measure. In *Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal, May 2004.

Miguel Rios, Wilker Aziz, and Lucia Specia. Tine: A metric to assess MT adequacy. In *Sixth Workshop on Statistical Machine Translation (WMT 2011)*, 2011.

Markus Saers and Dekai 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.

Markus Saers, Joakim Nivre, and Dekai 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.

Kashif Shah, Trevor Cohn, and Lucia Specia. An investigation on the effectiveness of features for translation quality estimation. In *Machine Translation Summit XIV (MT Summit 2013)*, Nice, France, 2013.

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)*, pages 223–231, Cambridge, Massachusetts, August 2006.

Lucia Specia. Exploiting objective annotations for measuring translation post-editing effort. In *15th Conference of the European Association for Machine Translation*, pages 73–80, 2011.

Anand Karthik Tumuluru, Chi-kiu Lo, and Dekai Wu. 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.

Nicola Ueffing and Hermann Ney. Word-level confidence estimation for machine translation using phrase-based translation models. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 763–770, 2005.

Mengqiu Wang and Christopher D. Manning. SPEDE: Probabilistic edit distance metrics for MT evaluation. In *7th Workshop on Statistical Machine Translation (WMT 2012)*, 2012.

Dekai Wu, Chi-kiu Lo, Meriem Beloucif, and Markus Saers. IMEANT: Improving semantic frame based MT evaluation via inversion transduction grammars. Forthcoming, 2014.

Dekai Wu. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403, 1997.

Deyi Xiong, Min Zhang, and Haizhou Li. Error detection for statistical machine translation using linguistic features. In *48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, 2010.

Richard Zens and Hermann Ney. A comparative study on reordering constraints in statistical machine translation. In *41st Annual Meeting of the Association for Computational Linguistics (ACL-2003)*, pages 144–151, Stroudsburg, Pennsylvania, 2003.