On the reliability and inter-annotator agreement of human semantic MT evaluation via HMEANT

Chi-kiu Lo **Dekai Wu**
HKUST
Human Language Technology Center
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
{jackielo|dekai}@cs.ust.hk

Abstract

We present analyses showing that HMEANT is a reliable, accurate and fine-grained semantic frame based human MT evaluation metric with high inter-annotator agreement (IAA) and correlation with human adequacy judgments, despite only requiring a minimal training of about 15 minutes for lay annotators. Previous work shows that the IAA on the semantic role labeling (SRL) subtask within HMEANT is over 70%. In this paper we focus on (1) the IAA on the semantic role alignment task and (2) the overall IAA of HMEANT. Our results show that the IAA on the alignment task of HMEANT is over 90% when humans align SRL output from the same SRL annotator, which shows that the instructions on the alignment task are sufficiently precise, although the overall IAA where humans align SRL output from different SRL annotators falls to only 61% due to the pipeline effect on the disagreement in the two annotation task. We show that instead of manually aligning the semantic roles using an automatic algorithm not only helps maintaining the overall IAA of HMEANT at 70%, but also provides a finer-grained assessment on the phrasal similarity of the semantic role fillers. This suggests that HMEANT equipped with automatic alignment is reliable and accurate for humans to evaluate MT adequacy while achieving higher correlation with human adequacy judgments than HTER.

Keywords: semantic MT evaluation, HMEANT, inter-annotator agreement

1. Introduction

HMEANT is a human metric that fully realizes a semantic frame based approach to MT evaluation that we originally envisioned in Lo and Wu (2010) and then subsequently implemented and refined over a substantial series of development cycles (Lo and Wu, 2011a,d). In this paper we present new focused empirical analyses showing that HMEANT achieves high inter-annotation agreement (IAA) and correlation with human judgments of translation adequacy, despite requiring only minimal training for inexpensive lay annotators. Through extensive IAA analyses—particularly on the semantic frame alignment task, an interesting question raised for example by by Birch *et al.* (2013)—we show that annotators align semantic frames consistently when the SRL output comes from the same SRL annotator, although the pipeline effect of accumulating the disagreement in the two annotation tasks significantly degrades the overall IAA of HMEANT when the alignment annotators align the SRL output from different SRL annotators. However, our results further show that instead of manually aligning the semantic roles, using an automatic algorithm in the alignment task not only helps maintain a high overall IAA for HMEANT, and at the same time also provides a finer-grained assessment of the phrasal similarity of semantic role fillers, such that HMEANT achieves higher correlation with human adequacy judgments than HTER or any automatic metric. These results indicate that HMEANT equipped with automatic alignment is a reliable and accurate methodology for human subjective evaluation of MT adequacy.

The MEANT family also includes other fully automatic approximations of HMEANT, that are accurate, inexpensive, and tunable semantic frame based MT evaluation metrics quantifying the semantic similarity between reference and machine translations in terns of how well their semantic frames match. HMEANT (Lo and Wu, 2011a,d), the human variant in the family, correlates better with human adequacy judgments than HTER at a significantly lower labor cost. MEANT (Lo *et al.*, 2012), the fully automatic metric in the family, correlates better with human adequacy judgments than the other commonly used automatic MT evaluation metrics, such as, BLEU (Papineni *et al.*, 2002), NIST (Doddington, 2002), or TER (Snover *et al.*, 2006). Since a high MEANT score is contingent on correct lexical choices as well as syntactic and semantic structures, tuning MT systems against MEANT improves both adequacy and fluency and outperforms BLEU-tuned and TER-tuned systems across different languages and different genres, such as formal newswire, informal web forum and informal public speech (Lo *et al.*, 2013a; Lo and Wu, 2013a; Lo *et al.*, 2013b).

As we continue to investigate on how to leverage the MEANT family of metrics to improve actual MT utility, we revisit in this paper one of the important concerns about using HMEANT as a human MT evaluation metric: is HMEANT a reliable human MT evaluation metric? Given only minimal instructions on on the SRL and alignment annotation tasks, humans might label and align the semantic roles inconsistently, which would reduce the reliability of HMEANT. Lo and Wu (2011a) carried out an extensive IAA analysis on the SRL task showing that using monolingual annotators to label the semantic roles achieves 79% IAA on average and using bilingual annotators to label the semantic roles achieves 70% IAA on average. We avoid directly diving into aggregating the overall IAA, which might risk prematurely jumping to the conclusion that HMEANT is not reliable; instead, we take a cautious approach to first analyze the IAA solely on the semantic frame alignment task to ensure any inconsistency is not caused by the fun-

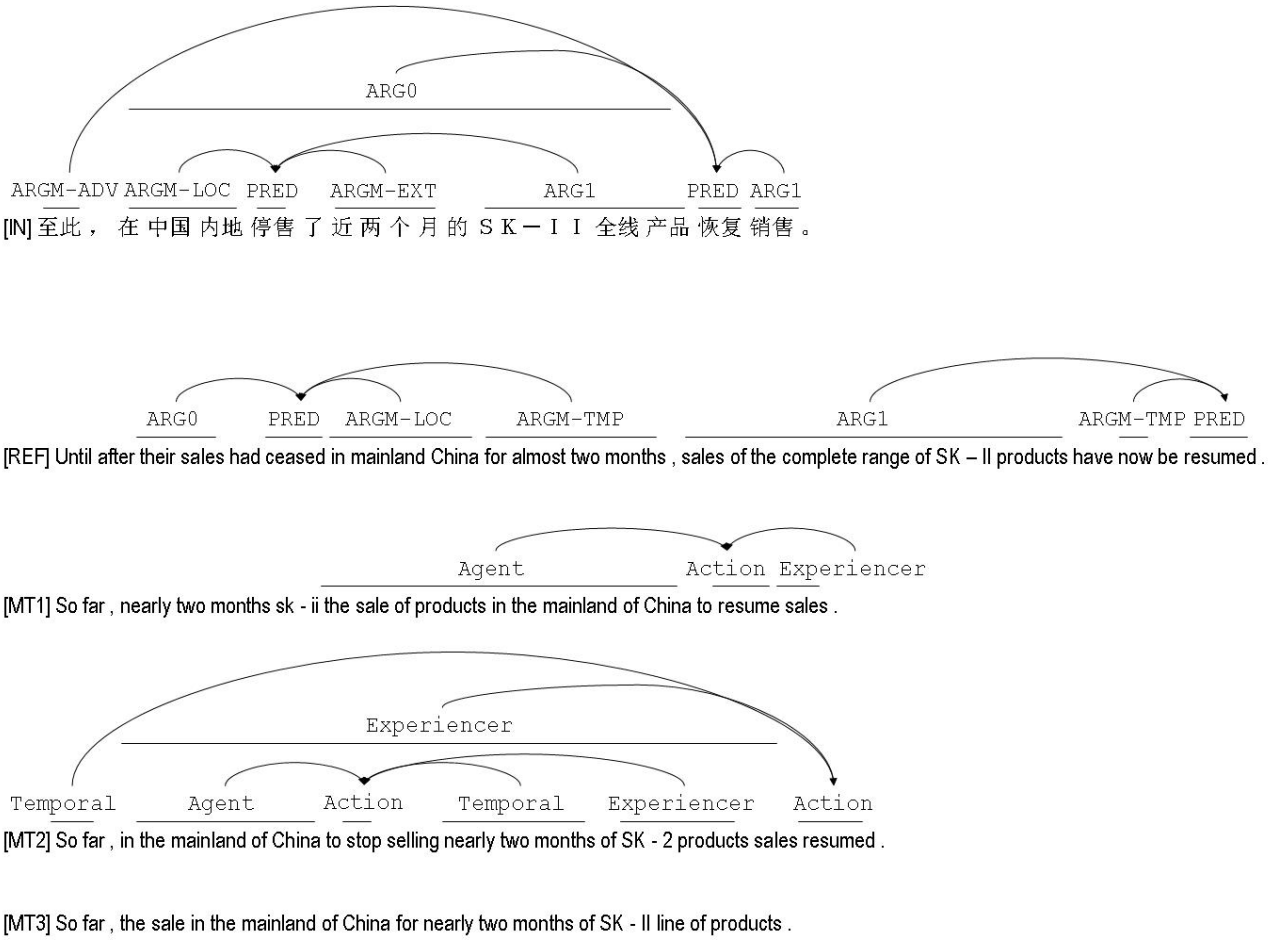

Figure 1: Examples of human semantic frame annotation. Semantic parses of the Chinese input and the English reference translation are from the Propbank gold standard. The MT output is semantically parsed by monolingual lay annotators according to the HMEANT guidelines. There are no semantic frames for MT3 because there is no predicate.

damental design of HMEANT.

Based on our findings in Lo *et al.* (2012) that MEANT equipped with automatic SRL and automatic semantic role alignment outperforms HMEANT equipped with automatic SRL and manual semantic role alignment, we evaluate the feasibility and reliability of replacing human semantic frame alignment with an automatic alignment algorithm in HMEANT. Since automatic alignment aligns semantic frames more consistently and measures phrasal similarity of the role fillers in a finer-grained manner, we believe HMEANT using human SRL and automatic alignment will be more reliable in terms of IAA and more accurate in correlation with human adequacy judgment.

In this paper we focus on the problem of inter-annotator agreement for the semantic frame alignment task in HMEANT, and on evaluating the feasibility and reliability of replacing HMEANT's human semantic role filler alignment step with a cheaper yet more accurate automatic alignment algorithm instead.

2. Related work

2.1. The MEANT family of metrics

2.1.1. HMEANT

HMEANT, proposed in Lo and Wu (2011a,c,d), has been found to correlate significantly better with human adequacy judgments than other commonly used automatic MT evaluation metrics, as well as other human metrics like HTER (Snover *et al.*, 2006). HMEANT consists of two manual steps: (1) human semantic role labeling, which labels aspects of the meaning of the reference and machine translations in terms of semantic predicate-argument structure; and (2) human semantic frame alignment, which aligns the annotated semantic predicates and role fillers. Monolingual (or bilingual) human annotators label the semantic roles and fillers in both the reference and machine translations, so that human semantic frame aligners can align the predicates and semantic role fillers in the MT output to the reference translations. These human annotations (semantic role labeling and semantic frame alignment) allow HMEANT to then aggregate the translation accuracy for each role into semantic frame accuracy, which is then aggregated into the overall sentence accuracy in meaning. The HMEANT score is simply defined in terms of a modified weighted f-score over these aligned predicates and role fillers. More precisely, HMEANT is computed as follows:

1. Human labelers annotate the shallow semantic structures of both the references and MT output.

2. Human aligners align the semantic frames between the references and MT output by judging the correctness of

REF roles	REF	MT2 roles	MT2	decision
PRED	ceased	Action	stop	match
ARG0	their sale	—	—	incorrect
ARGM-LOC	in mainland China	Agent	the mainland of China	correct*
ARGM-TMP	for almost two months	Temporal	nearly two months	correct
—	—	Experiencer	SK - 2 products	incorrect
PRED	resumed	Action	resume	match
ARG0	sales of complete range of SK - II products	Experiencer	in the mainland of China to stop selling nearly two months of SK - 2 products sales	incorrect
ARGM-TMP	Until after , their sales had ceased in mainland China for almost two months	Temporal	So far	partial
ARGM-TMP	now	—	—	incorrect

Table 1: Example of SRL annotation for the MT2 output from figure 1 along with the human judgements of translation correctness for each argument. *Notice that although the decision made by the human judge for "in mainland China" in the reference translation and "the mainland of China" in MT2 is "correct", nevertheless the HMEANT computation will not count this as a match since their role labels do not match.

the predicates.

3. For each pair of aligned semantic frames,
   (a) Human aligners determine the translation correctness of the semantic role fillers.
   (b) Human aligners align the semantic role fillers between the reference and MT output according to the correctness of the semantic role fillers.
4. Compute the weighted f-score over the matching role labels of these aligned predicates and role fillers.

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

$M_{i,j}$ ≡ total # ARG j of aligned frame i in MT
$R_{i,j}$ ≡ total # ARG j of aligned frame i in REF
$C_{i,j}$ ≡ # correct ARG j of aligned frame i in MT
$P_{i,j}$ ≡ # partially correct ARG j of aligned frame i in MT
w_{pred} ≡ weight of similarity of predicates
w_j ≡ weight of similarity of ARG j

$$\text{precision} = \frac{\sum_i m_i \frac{w_{\text{pred}} + \sum_j w_j (C_{i,j} + w_{\text{partial}} P_{i,j})}{w_{\text{pred}} + \sum_j w_j M_{i,j}}}{\sum_i m_i}$$
$$\text{recall} = \frac{\sum_i r_i \frac{w_{\text{pred}} + \sum_j w_j (C_{i,j} + w_{\text{partial}} P_{i,j})}{w_{\text{pred}} + \sum_j w_j R_{i,j}}}{\sum_i r_i}$$

where m_i and r_i are the weights for frame, i, in the MT/REF respectively. These weights estimate the degree of contribution of each frame to the overall meaning of the sentence. $M_{i,j}$ and $R_{i,j}$ are the total counts of argument of type j in frame i in the MT and REF respectively. $C_{i,j}$ and $P_{i,j}$ are the count of the correctly and partial correctly translated argument of type j in frame i in the MT output. w_{pred} and w_j are the aligned predicates and the aligned 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 HMEANT as defined in Lo and Wu (2011c) and they are determined in a supervised learning manner by optimizing the correlation with human adequacy judgments through simple grid searching (Lo and Wu, 2011a). Figure 1 shows examples of human semantic frame annotation on reference and machine translations as used in HMEANT. Table 2.1.1. shows examples of human judges' decisions for semantic frame alignment and translation correctness for each semantic roles, for the "MT2" output from Figure 1. Birch *et al.* (2013) reported that the final IAA of HMEANT drops below 50% due to the pipelining effect, where annotation disagreements in the SRL task and the semantic role alignment task accumulate.

2.1.2. MEANT and UMEANT

Unlike HMEANT, MEANT (Lo *et al.*, 2012) is fully automatic; but nevertheless, it adheres to HMEANT's principles of Occam's razor simplicity and representational transparency and outperforms BLEU, NIST, METEOR, WER, CDER and TER in correlation with human adequacy judgments. MEANT automates HMEANT by replacing the human semantic role labelers with shallow semantic parsers and replacing the human semantic frame aligners with the maximum weighted bipartite matching algorithm based on a context vector model that computes the lexical similarity of the semantic role fillers. The minimal changes in the mathematics formula from HMEANT to MEANT are illustrated as follow:

$S_{i,\text{pred}}$ ≡ predicate similarity in aligned frame i
$S_{i,j}$ ≡ ARG j similarity in aligned frame i

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

where $S_{i,\text{pred}}$ and $S_{i,j}$ are the lexical and phrasal 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. The lexical similarities of the semantic role fillers can be computed using different statistical similarity measures while the phrasal similarities can be aggregated from lexical similarities using different heuristics, like the geometric mean used in Lo *et al.* (2012) and Tumuluru *et al.* (2012) or the normalized phrasal aggregation (Mihalcea *et al.*, 2006) used subsequently in Lo *et al.* (2013a); Lo and Wu (2013a); Lo *et al.* (2013b). In MEANT, the weights w_{pred} and w_j are estimated in the same way as HMEANT, i.e. by optimizing the correlation with human adequacy judgments through simple grid searching. As for UMEANT (Lo and Wu, 2013b), these weights 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.

Lo *et al.* (2012) show that fully automated MEANT outperforms semi-automated HMEANT (automatic SRL and human semantic frame alignment) in correlating with human adequacy judgments. 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, including formal newswire text, informal web forum text, and informal public speech. The work shows that the automatic alignment algorithm aligns semantic frames more consistently and measures phrasal similarity of the role fillers in a finer-grained manner, and thus suggests that the reliability of HMEANT would be improved by automatically aligning the manually labeled semantic frames.

2.2. Other human MT evaluation

HTER (Snover *et al.*, 2006) is only used in large-scale MT evaluation campaign because of its high labor cost. It not only requires well-trained professional human translators reading and understanding both the reference translation and the MT output, but also relies on the minimum edits made by those translator on the MT output so as to match the meaning expressed in the edited MT output with that in the reference translation. This requirement of heavy manual decision making in HTER greatly increases the cost of evaluation.

In contrast, task based human MT evaluation in Voss and Tate (2006) reduces labor cost by requiring human evaluators to finish some simple question answering tasks after reading the MT output. However, task based human MT evaluation do not generalize across different test sets.

3. IAA on the alignment task in HMEANT

To address the interesting questions raised by Birch *et al.* (2013), we systematically analyze the IAA for semantic frame alignment task by asking the alignment annotators to align SRL output from the same SRL annotators. This avoids directly diving into rough aggregation of the overall IAA for the entire evaluation pipeline, which might misleadingly jump to the conclusion that HMEANT is not reliable.

Annotator pairs	IAA
S1-A1 vs. S1-A2	90%
S2-A1 vs. S2-A2	91%

Table 2: IAA on the alignment task

Annotator pairs	IAA
S1-A1 vs. S2-A2	63%
S2-A1 vs. S1-A2	61%

Table 3: IAA on the overall annotation pipeline

3.1. Setup

For our benchmark comparison, the evaluation data for our experiments is the same set of sentences, GALE-A, that were used in Lo and Wu (2012). The reference and each of the MT system outputs are labeled by two SRL annotators for IAA analysis. For the purpose of cross-validation, we setup two rounds of alignment tasks. In the first round, two alignment annotators align the SRL output from the first SRL annotator. In the second round, the two alignment annotators align the SRL output from the second SRL annotator. As described in Lo and Wu (2011b), in all the human SRL task and the alignment task, we supplement annotators with one double sided sheet with three examples. As a result, we have alignment output from four combinations of the two SRL annotators and the two alignment annotators. For inter-annotator agreement, we follow the definition as in Lo and Wu (2011a).

3.2. Results

Table 3.1. shows the IAA on the alignment task of HMEANT is over 90% consistently when the alignment annotators align the SRL output from the same SRL annotator. This shows that the instructions on the alignment task is sufficient and effective.

However, table 3.1. shows the final IAA where the alignment annotators align the SRL output from different SRL annotators falls to only 61% due to the pipelining effect on the disagreement in the two annotation tasks.

4. Don't align semantic frames manually

Lo *et al.* (2012) reported that MEANT equipped with automatic SRL and automatic semantic frame alignment outperforms HMEANT equipped with automatic SRL and manual semantic role alignment. The natural question following such findings is whether the reliability of HMEANT improves by replacing human semantic roles alignment with automatic alignment algorithm, and if so, the extent to which it helps.

4.1. Setup

We run MEANT's automatic alignment algorithm on the SRL output from the two SRL annotators in previous experiment. We use a HMEANT implementation along the lines described in Lo and Wu (2012), except the set of weights

Annotator pairs	IAA
S1-auto vs. S2-auto	70%

Table 4: IAA on the overall annotation pipeline where the human alignment annotators are replaced by an automatic alignment algorithm

	Kendall
Human metrics	
HMEANT(S2-auto)	0.53
HMEANT(S1-auto)	0.53
HMEANT(S2-A2)	0.49
HMEANT(S2-A1)	0.49
HMEANT(S1-A1)	0.49
HMEANT(S1-A2)	0.47
HTER	0.43
Automatic metrics	
MEANT	0.39
NIST	0.29
METEOR	0.20
BLEU	0.20
TER	0.20
PER	0.20
CDER	0.12
WER	0.10

Table 5: Sentence-level correlation with human adequacy judgment on GALE-A

is estimated in an unsupervised manner like UMEANT (Lo and Wu, 2013b).

4.2. Results

Table 4.1. shows the IAA on HMEANT using automatic semantic role alignment algorithm rises to 70%. These results are expected because the automatic alignment algorithm handles the partial alignment more consistently, especially on cases where the role fillers of the semantic roles in the reference is split into role fillers in more than one role in the MT output.

Table 4.1. shows that performing the semantic frame alignment automatically is better than aligning manually. The results are in line with the findings in Lo *et al.* (2012). Since automatic alignment algorithm aligns semantic roles more consistently and measures phrasal similarity of the role fillers in a finer-grained manner, we believe HMEANT using human SRL and automatic alignment is more reliable in term of IAA and accurate in correlating with human adequacy judgments.

5. Conclusion

We have shown that HMEANT is a reliable, accurate and fine-grained semantic frame based human MT evaluation metric with high IAA and correlation with human adequacy judgment, despite requiring only minimal training for lay annotators. Our results show that the IAA on the semantic frame alignment task of HMEANT is over 90% when the human annotators align SRL output from the same SRL annotator, although the final IAA of HMEANT based on alignment results from annotators aligning SRL output from different SRL annotators falls to only 61% due to the pipelining effect of the disagreement in the two annotation tasks.

More importantly, we have shown that to improve the reliability of HMEANT, completely replacing the manual semantic frame alignment with fully automatic alignment not only helps to maintain the overall IAA of HMEANT at a 70% level, but also provides a finer-grained assessment on the phrasal similarity of the semantic role fillers, so that HMEANT achieves higher correlation with human judgments of translation adequacy than HTER. This has the additional important benefit of making HMEANT even more cost effective. The results show that HMEANT equipped with automatic alignment algorithm is a highly reliable and accurate methodology for MT evaluation.

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.

7. References

Alexandra Birch, Barry Haddow, Ulrich Germann, Maria Nadejde, Christian Buck, and Philipp Koehn. The feasibility of HMEANT as a human MT evaluation metric. In *8th Workshop on Statistical Machine Translation (WMT 2013)*, 2013.

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)*, pages 138–145, San Diego, California, 2002.

Chi-Kiu Lo and Dekai Wu. Evaluating machine translation utility via semantic role labels. In *The Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, 2010.

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. A radically simple, effective annotation and alignment methodology for semantic frame based SMT and MT evaluation. In *International Workshop on Using Linguistic Information for Hybrid Machine Translation (LIHMT-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. Structured vs. flat semantic role representations for machine translation evaluation. In *Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-5)*, 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 evaluation 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.

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, page 775. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

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.

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.

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.

Clare R. Voss and Calandra R. Tate. Task-based evaluation of machine translation (MT) engines: Measuring how well people extract who, when, where-type elements in MT output. In *11th Annual conference of the European Association for Machine Translation (EAMT 2006)*, pages 203–212, Oslo, Norway, June 2006.