Fully Automatic Semantic MT Evaluation

Chi-kiu LO, Anand Karthik TUMULURU and **Dekai WU**
HKUST
Human Language Technology Center
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
{jackielo,aktumuluru,dekai}@cs.ust.hk

Abstract

We introduce the first fully automatic, fully semantic frame based MT evaluation metric, MEANT, that outperforms all other commonly used automatic metrics in correlating with human judgment on translation adequacy. Recent work on HMEANT, which is a human metric, indicates that machine translation can be better evaluated via semantic frames than other evaluation paradigms, requiring only minimal effort from monolingual humans to annotate and align semantic frames in the reference and machine translations. We propose a surprisingly effective Occam's razor automation of HMEANT that combines standard shallow semantic parsing with a simple maximum weighted bipartite matching algorithm for aligning semantic frames. The matching criterion is based on lexical similarity scoring of the semantic role fillers through a simple context vector model which can readily be trained using any publicly available large monolingual corpus. Sentence level correlation analysis, following standard NIST MetricsMATR protocol, shows that this fully automated version of HMEANT achieves significantly higher Kendall correlation with human adequacy judgments than BLEU, NIST, METEOR, PER, CDER, WER, or TER. Furthermore, we demonstrate that performing the semantic frame alignment automatically actually tends to be just as good as performing it manually. Despite its high performance, fully automated MEANT is still able to preserve HMEANT's virtues of simplicity, representational transparency, and inexpensiveness.

1 Introduction

We introduce the first fully automatic semantic-frame-based MT evaluation metric capable of outperforming all other commonly used automatic metrics like BLEU, NIST, METEOR, PER, CDER, WER, and TER for evaluating translation adequacy. This work, MEANT, can be seen as a fully automated version of HMEANT, which is a human metric, introduced by Lo and Wu (2011b). Despite its high performance, MEANT is still able to preserve HMEANT's virtues of Occam's razor simplicity, representational transparency, and inexpensiveness.

For the past decade, MT evaluation has relied heavily on inexpensive automatic metrics such as BLEU (Papineni *et al.*, 2002), NIST (Doddington, 2002), METEOR (Banerjee and Lavie, 2005), PER (Tillmann *et al.*, 1997), CDER (Leusch *et al.*, 2006), WER (Nießen *et al.*, 2000), and TER (Snover *et al.*, 2006). In large part, this is because automatic metrics significantly shorten the evaluation cycle by providing a fast, easy and cheap quantitative evaluation which can be effectively incorporated into modern SMT training methods.

Despite the fact that HMEANT, a human metric recently proposed by Lo and Wu (2011b,c,d), was shown to reflect translation adequacy more accurately than all of these automatic metrics, it is unfortunately infeasible to incorporate the HMEANT metrics directly into SMT training methods, due to the non-automatic processes of (1) semantic parsing and (2) aligning semantic frames. In this paper we introduce an automatic metric in which both the semantic parsing and the alignment of semantic frames are fully automated. Our aim is to show that even with full automation, this new metric still outperforms all the previous automatic metrics mentioned, thus providing a foundation for future incorporation into the training of SMT to drive system improvements in providing more adequate translation output.

N-gram oriented automatic MT evaluation metrics like BLEU perform well at capturing translation fluency, and ranking overall systems with respect to each other when their scores are averaged over entire documents or corpora. However, they do not fare so well in ranking translations of individual sentences. As MT systems improve, the n-gram based evaluation metrics have begun to show their limits. State-of-the-art MT systems are often able to output translations containing roughly the correct words, while failing to convey important aspects of the meaning of the input sentence. Cases where BLEU strongly disagrees with human judgment of translation quality were

reported in large scale MT evaluation tasks by Callison-Burch *et al.* (2006) and Koehn and Monz (2006).

Motivated by the goal of addressing the weaknesses of n-gram oriented automatic MT evaluation metrics at evaluating translation adequacy, the HMEANT metric assesses translation utility by matching the basic event structure—"who did what to whom, when, where and why" (Pradhan *et al.*, 2004)—representing the central meaning conveyed by sentences. As mentioned above, however, HMEANT requires humans to manually annotate semantic frames in the reference and machine translations, and then to align the semantic frames—making it difficult to incorporate HMEANT as an objective function in the MT system training, evaluating, and optimizing cycle.

We argue in this paper that both the human semantic parsing and the semantic frame alignment tasks performed within HMEANT can be successfully automated to produce a state-of-the-art automatic metric. Moreover, we show that the spirit of Occam's razor can be preserved even for the semantic frame alignment, by demonstrating the effectiveness of a simple maximum weighted bipartite matching algorithm based on the lexical similarity between semantic frames. In addition, we show empirically that performing this semantic frame alignment automatically tends to be just as good as performing it manually. Our results indicate that MEANT, the fully automatic version of HMEANT, achieves levels of correlation with human adequacy judgment (in our experiments, approximately 0.37) which significantly outperforms the commonly used automatic metrics BLEU, NIST, METEOR, PER, CDER, WER, and TER (in our experiments, ranging between 0.20 and 0.29).

2 Related Work

2.1 Automatic lexical similarity based metrics

BLEU (Papineni *et al.*, 2002) remains the most widely used MT evaluation metric despite the fact that a number of large scale meta-evaluations (Callison-Burch *et al.*, 2006; Koehn and Monz, 2006) report cases where it strongly disagrees with human judgments of translation accuracy. Other lexical similarity based automatic MT evaluation metrics, like NIST (Doddington, 2002), METEOR (Banerjee and Lavie, 2005), PER (Tillmann *et al.*, 1997), CDER (Leusch *et al.*, 2006), WER (Nießen *et al.*, 2000), and TER (Snover *et al.*, 2006), also perform well in capturing translation fluency, but share the same problem that although evaluation with these metrics can be done very quickly at low cost, their underlying assumption—that a "good" translation is one that shares the same lexical choices as the reference translation—is not justified semantically. Lexical similarity does not adequately reflect similarity in meaning.

Generating a translation that contains roughly the correct words may be necessary—but is far from sufficient—to preserve the essence of the meaning. We argue that a translation metric that reflects meaning similarity needs to be based on similarity of semantic structure, and not merely flat lexical similarity.

2.2 HMEANT (human SRL based metric)

As mentioned above, despite the fact that the semi-automatic HMEANT metric recently proposed by Lo and Wu (2011b,c,d) shows a higher correlation with human adequacy judgments than all commonly used automatic MT evaluation metrics, as with other human metrics like HTER (Snover *et al.*, 2006), it is unfortunately infeasible to incorporate the HMEANT metrics directly into SMT training methods. HMEANT requires non-automatic manual steps of (1) semantic parsing and (2) aligning semantic frames. Monolingual (or bilingual) annotators must label the semantic roles in both the reference and machine translations, and then to align the semantic predicates and role fillers in the MT output to the reference translations. These annotations allow HMEANT to then look at the aligned role fillers, and aggregate the translation accuracy for each role. In the spirit of Occam's razor and representational transparency, the HMEANT score is defined simply in terms of a weighted f-score over these aligned predicates and role fillers. More precisely, HMEANT is defined as follows:

1. Human annotators annotate the shallow semantic structures of both the references and MT output.
2. Human judges align the semantic frames between the references and MT output by judging the correctness of the predicates.
3. For each pair of aligned semantic frames,
   (a) Human judges determine the translation correctness of the semantic role fillers.
   (b) Human judges 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.

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

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

$M_{i,j} \equiv$ total # ARG j of aligned frame i in MT

$R_{i,j} \equiv$ total # ARG j of aligned frame i in REF

$C_{i,j} \equiv$ # correct ARG j of aligned frame i in MT

$P_{i,j} \equiv$ # partially correct ARG j of aligned frame i in MT

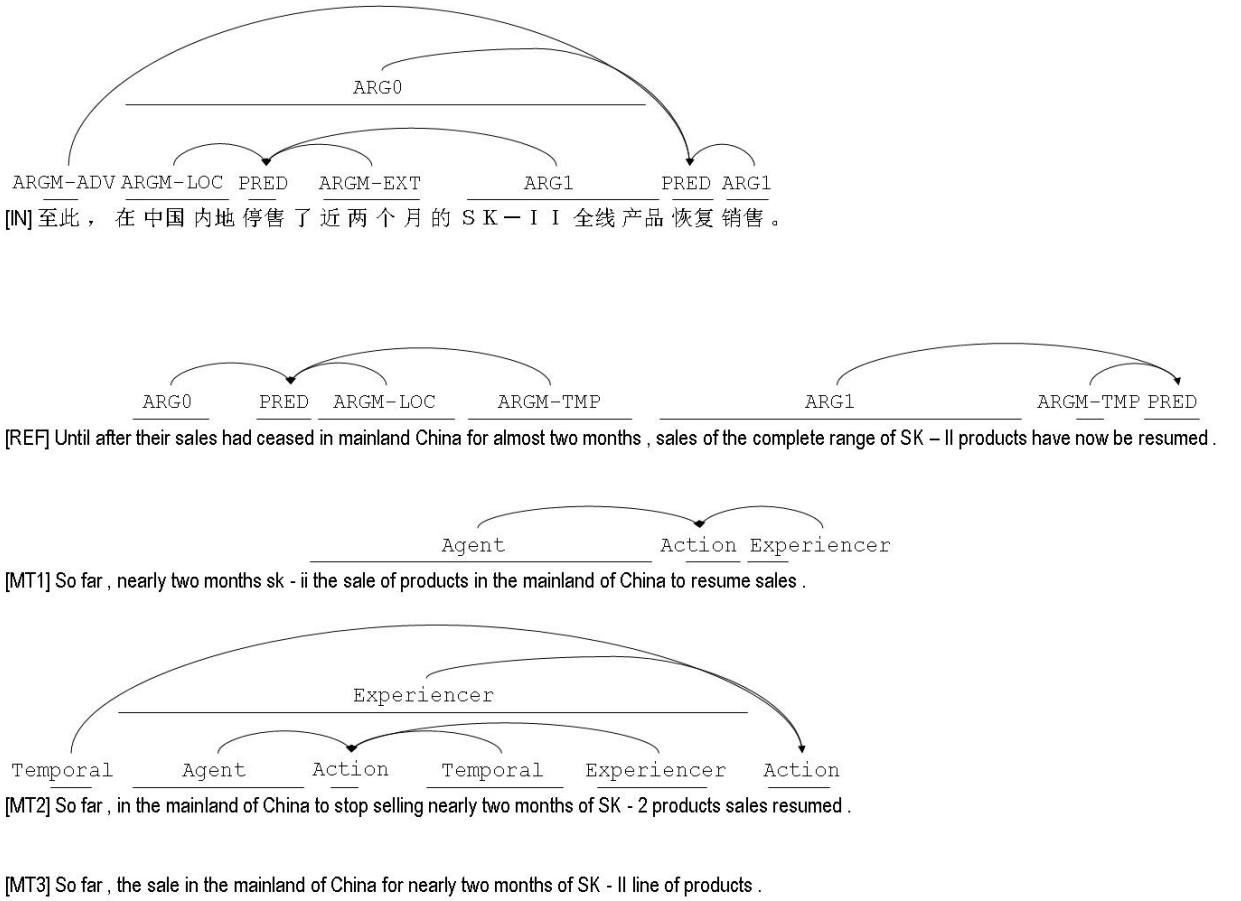

Figure 1: Examples of human semantic frame annotation. Semantic parses of the Chinese input and the English reference translation are from the Propbank gold standard. The MT output is semantically parsed by monolingual lay annotators according to the HMEANT guidelines. There are no semantic frames for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

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

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

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. Figure 1 shows examples of human semantic frame annotation on reference and machine translations as used in HMEANT. Table 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.

Unlike HMEANT, MEANT is fully automatic; but nevertheless, it adheres to HMEANT's principles of Occam's razor simplicity and representational transparency. These properties crucially facilitate error analysis and credit/blame assignment that are invaluable for MT system modeling.

Furthermore, being fully automatic, MEANT is even less expensive than HMEANT, which was already shown by Lo and Wu (2011b,c,d) to be significantly less expensive than HTER. This makes MEANT a much better candidate than HMEANT for future incorporation into the automatic training of SMT systems to drive improvements in translation adequacy.

2.3 Semantic role labels as features in aggregate metrics

Giménez and Màrquez (2007, 2008) introduced ULC, an automatic MT evaluation metric that aggregates many types of features, including several shallow semantic similarity features. However, unlike Lo and Wu (2011b),

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.

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

the ULC representation is based on flat semantic role label features that do not capture the structural *relations* in semantic frames, i.e., the predicate-argument relations. Also unlike HMEANT, which weights each semantic role type according to its empirically determined relative importance to the adequate preservation of meaning, ULC uses uniform weights. Although the automatic ULC metric shows an improved correlation with human judgment of translation quality (Callison-Burch *et al.*, 2007; Giménez and Màrquez, 2007; Callison-Burch *et al.*, 2008; Giménez and Màrquez, 2008), it is not commonly used in large-scale MT evaluation campaigns, perhaps due to its high time cost and/or the difficulty of interpreting its score because of its highly complex combination of many heterogeneous types of features.

Like system combination approaches, ULC is a vastly more complex aggregate metric compared to widely used metrics like BLEU. We believe it is important for automatic semantic MT evaluation metrics to provide representational transparency via simple, clear, and transparent scoring schemes that are (a) easily human readable to support error analysis, and (b) potentially directly usable for automatic credit/blame assignment in tuning tree-structured SMT systems.

3 MEANT: A fully automatic semantic MT evaluation metric

Like HMEANT, our guiding principle is that a good translation is one that is useful, in the sense that human readers may successfully understand at least the basic event structure—*who did what to whom, when, where and why* (Pradhan *et al.*, 2004)—representing the central meaning of the source utterances. Whereas HMEANT measures this using a f-score of correctly translated semantic roles in MT output that are annotated and compared by monolingual human annotators, MEANT *automates* HMEANT as follows (the differences from HMEANT are italicized):

1. *Apply an automatic shallow semantic parser* on both the references and MT output.

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

3. For each pair of aligned semantic frames,

   (a) *Lexical similarity scores* determine the *similarity* of the semantic role fillers.

   (b) *Apply maximum weighted bipartite matching algorithm to* align the semantic role fillers between the reference and MT output according to their *lexical similarity*.

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

3.1 Automatic semantic parsing

To automate the process of human semantic role labeling, we apply an automatic shallow semantic parser on both the reference and MT output that takes the raw translation as input and outputs the corresponding predicate-argument structure. We choose to semantically parse the translation independently, instead of inducing the parses

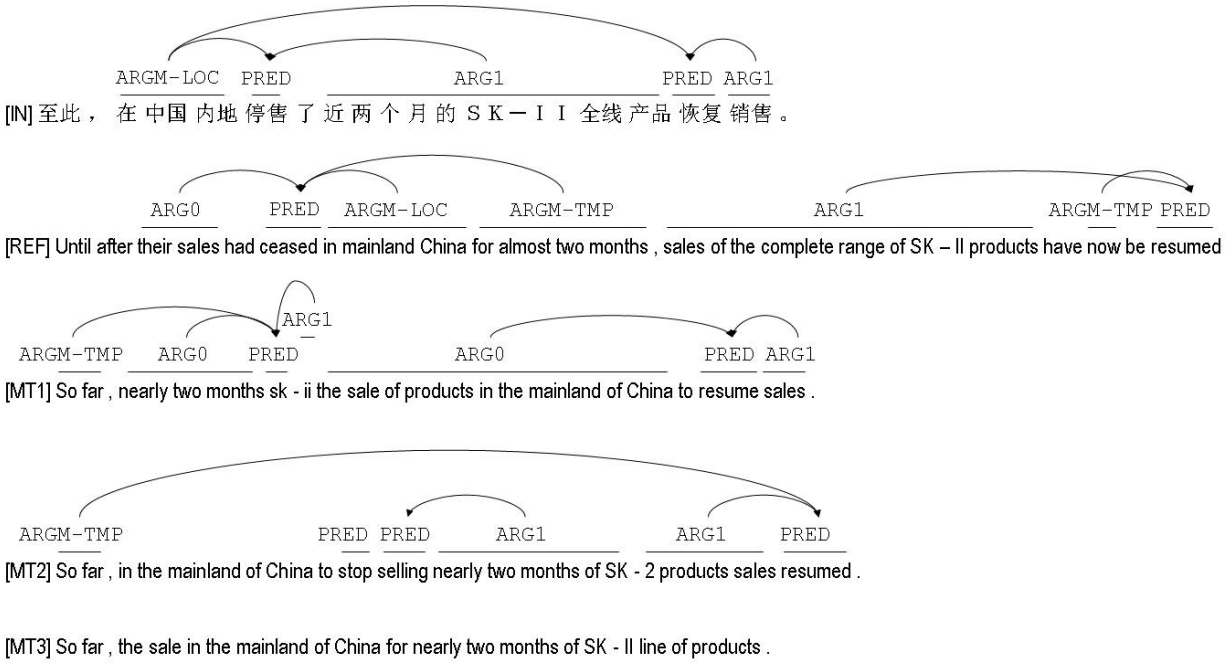

Figure 2: Examples of automatic shallow semantic parses. The Chinese input is parsed by a Chinese automatic shallow semantic parser. The English reference and machine translations are parsed by an English automatic shallow semantic parser. There are no semantic frames for mt3 since there is no predicate.

from the input, because it captures the raw meaning conveyed in the translation rather than predicting the meaning conveyed in the translation from the input. Figure 2 shows examples of automatic shallow semantic parses on both reference and machine translations.

3.2 Automatic semantic frame alignment

After reconstructing the shallow semantic parse, the manual semantic frame alignment process is automated by applying the maximum weighted bipartite matching algorithm where the weights of the edges represent the lexical similarity of the predicates. A wide range of lexical similarity measures are available to us, including for example BLEU, METEOR, cosine similarity based on context vector models (Dagan, 2000), and so forth. In Section 4, we will show the performance of the fully automatic semantic MT evaluation metric, MEANT ,couple with different lexical similarity metrics and other commonly used automatic MT evaluation metrics. In Section 6, we will discuss aligning the semantic frames according to all semantic role fillers, instead of the predicates only.

Then, for each pair of aligned semantic frames, we estimate the similarity of the semantic role fillers by summing all the lexical similarity of all the pairwise combination of tokens between the references and MT output. After obtaining the similarity of the semantic role fillers, we again apply the maximum weighted bipartite matching algorithm to align the semantic role fillers between the references and MT output. Table 2 shows examples of the human judges' decisions on semantic frame alignment and translation correctness for each semantic role in the "MT2" output from Figure 2.

3.3 Scoring the semantic similarity

After aligning the semantic frames automatically, the computation of the MEANT score is largely the same as stated in Lo and Wu (2011d), except that we now replace the counts of correctly and partially correctly translated semantic role fillers by the similarity scores of the predicates and arguments between the references and MT output.

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

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

$M_{i,j}$ ≡ total # ARG j of aligned frame i in MT

$R_{i,j}$ ≡ total # ARG j of aligned frame i in REF

$S_{i,\text{pred}}$ ≡ sim. of pred of REF and MT in aligned frame i

$S_{i,j}$ ≡ sim. of ARG j of REF and MT in aligned frame i

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

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

Table 2: Automatic semantic frame alignment of the MT2 output from figure 2, along with the automatic lexical similarity scoring on translation correctness for each argument.

REF roles	**REF**	**MT2 roles**	**MT2**	**similarity**
PRED	ceased	PRED	stop	0.0377
ARG0	their sales	—	—	—
ARGM-LOC	in mainland China	—	—	—
ARGM-TMP	for almost two months	—	—	—
—	—	PRED	selling	—
—	—	ARG1	nearly two months of SK	—
PRED	resumed	PRED	resumed	1.0
ARG1	sales of complete range of SK - II products	ARG1	2 products sales	0.0836
ARGM-TMP	now	ARGM-TMP	So far	0.0459

where m_{i}, r_{i}, $M_{\text{i,j}}$, $R_{\text{i,j}}$ are defined the same as in HMEANT, and $S_{i,\text{pred}}$ and $S_{i,j}$ are the lexical similarities (BLEU, METEOR, cosine similarity based on a context vector model, and so on, as discussed in the following section) of the predicates and arguments of type j between the reference translations and the MT output.

4 MEANT outperforms all automatic metrics

We will first show that the fully automatic semantic MT evaluation metric, MEANT, outperforms all the other commonly used automatic metrics.

4.1 Experimental setup

For assessing lexical similarity, a wide range of lexical similarity scoring models are available. We describe a representative subset of a wide range of experiments we have performed using all the most typical and commonly used measures. On one hand, we report experiments with integrating two commonly used MT evaluation metrics, BLEU and METEOR, as the lexical similarity. On the other hand, we also report experiments on integrating two common similarity measures—cosine similarity measure and min/max with mutual information (Dagan, 2000)—that are based on context vector models, and trained from the Gigaword corpus with window sizes of 3 and 5.

The cosine similarity between two sequences of word tokens, $\overrightarrow{u}$ and $\overrightarrow{v}$, is defined as follows:

$$\begin{aligned} \overrightarrow{w_x} &= \text{context vector of word token x} \\ w_{xi} &= \text{attribute i of context vector } \overrightarrow{w_x} \end{aligned}$$

$$f(x, w_{xi}) = \frac{\text{count}(x, w_{xi})}{\text{count}\,(w_{xi})}$$

$$\text{cosine}(x, y) = \frac{\sum_i f(x, w_{xi}) \times f(y, w_{y_i})}{\sqrt{\sum_i f(x, w_{xi})^2}\sqrt{\sum_i f(y, w_{y_i})^2}}$$

$$\text{cosine}(\overrightarrow{u}, \overrightarrow{v}) = \sum_i \sum_j \text{cosine}(u_i, v_j)$$

Using the same definition of w_{xi}, the min/max with mutual information similarity between two sequences of word tokens, $\overrightarrow{u}$ and $\overrightarrow{v}$, is defined as follows:

$$P(w_{xi} \mid x) = \frac{\text{count}(x, w_{xi})}{\sum_i \text{count}(x, w_{xi})}$$

$$P\,(w_{xi}) = \frac{\sum_y \text{count}(y, w_{xi})}{\sum_y \sum_j \text{count}(y, w_{xj})}$$

$$\text{MI}(x, w_{xi}) = \log\left(\frac{P(w_{xi} \mid x)}{P\,(w_{xi})}\right)$$

$$\text{MinMax-MI}(x, y) = \frac{\sum_i \min\left(\text{MI}(\,x, w_{xi}), \text{MI}(y, w_{y_i}\,)\right)}{\sum_i \max\left(\text{MI}(\,x, w_{xi}), \text{MI}(y, w_{y_i}\,)\right)}$$

$$\text{MinMax-MI}(\overrightarrow{u}, \overrightarrow{v}) = \sum_i \sum_j \text{MinMax-MI}(u_i, v_j)$$

For our benchmark comparison, the evaluation data for our experiments is the same two sets of sentences, GALE-A and GALE-B that were used in Lo and Wu (2011d), where GALE-A is used for estimating the weight parameters of the metric by optimizing the correlation with human adequacy judgment, and then the learned weights are applied to testing on GALE-B.

For the automatic semantic role labeling, we used the publicly available off-the-shelf shallow semantic parser, ASSERT (Pradhan *et al.*, 2004).

The correlation with human adequacy judgments on sentence-level system ranking is assessed by the standard NIST MetricsMaTr procedure (Callison-Burch *et al.*, 2010) using Kendall correlation coefficients.

Table 3: Sentence-level correlation with human adequacy judgment on GALE-A (training) and GALE-B (testing) comparing all commonly used MT evaluation metrics against our proposed new fully automatic semantic frame based MT evaluation metric integrated with various lexical similarity scores between semantic role fillers: (a) BLEU, (b) METEOR, (c) cosine similarity and (d) MinMax with mutual information.

	GALE-A (training)	**GALE-B (testing)**
Human metrics		
HMEANT	0.49	0.27
HTER	0.43	0.20
Automatic metrics		
MEANT	—	—
- with MinMax-MI on context vector model of window size 3	**0.37**	0.19
- with MinMax-MI on context vector model of window size 5	0.37	0.17
- with Cosine on context vector model of window size 3	0.32	0.13
- with Cosine on context vector model of window size 5	0.30	0.08
- with METEOR	0.17	—
- with BLEU	0.00	—
METEOR	0.20	**0.21**
NIST	0.29	0.09
TER	0.20	0.10
BLEU	0.20	0.12
PER	0.20	0.07
WER	0.10	0.11
CDER	0.12	0.10

4.2 Results

Table 3 shows that MEANT significantly outperforms all the other automatic MT evaluation metrics when integrated with a simple similarity measure based on word context vectors trained from a large monolingual corpus. We can also see that using min/max with mutual information is significantly better than using cosine similarity. Furthermore, context vector models using a window size of 3 appear to be as good or better than those using a window size of 5.

Although the human metrics, HMEANT and HTER, obviously remain superior, MEANT performs far better than almost all other automatic metrics. The only exception is the GALE-B dataset, where METEOR performs marginally better than MEANT and even HTER. Data analysis shows that the marginally higher correlation of METEOR on the GALE-B dataset is a statistical outlier; it is quite rare for a lexically based automatic metric to outperform even the *human-driven* HTER metric.

Interestingly and somewhat surprisingly, using the n-gram based MT evaluation metrics BLEU and METEOR as lexical similarity scores does not work well at all for this purpose, even on the training data (thus obviating the need to obtain results on the testing data). Analysis indicates that the reason for this is that variation between alternative paraphrasing of the role fillers makes the number of matching n-grams quite small, since there are many synonyms and few exact consecutive n-gram matches.

Table 4: Sentence-level correlation with human adequacy judgment on GALE-A (training) and GALE-B (testing) for aligning sematnic frame automatically and manually.

Semantic frame alignment	**GALE-A**	**GALE-B**
Automatic	0.37	0.19
Manual	0.35	0.17

In the following sections, we turn to considering several questions that naturally arise following these strong results.

5 Don't align semantic frames manually

One obvious question is whether the automatic alignment of semantic frames degrades MEANT's accuracy, and if so, the extent to which it hurts.

5.1 Experimental setup

To test this question, we compare the best fully automatic results of the previous section against a semi-automatic variant of our proposed metric. In the semi-automatic variant, the semantic parsing is still performed automatically. However, the semantic frame alignment is instead done manually by human annotators.

The rest of the experimental setup is the same as that used in Section 4.

5.2 Results

Table 4 shows that performing the alignment of semantic frames automatically is as good—or even better than—doing the alignment manually. We believe the success of automatic semantic frame alignment reflects the high degree of reliability of our chosen lexical similarity metric, when the candidates for role fillers are restricted to the fairly small set defined by the sentence pairs.

6 Look only at predicates when aligning semantic frames

Given the positive results of the previous sections, it is worth asking a deeper question: would it further improve the correlation with human adequacy judgment of the metric if the semantic frames were aligned not only by matching predicates (as HMEANT did), but in addition by trying to also maximize the match of the semantic role fillers?

The reason to revisit this question is that even though Lo and Wu (2011a) showed that in the case of HMEANT it is effective for *human* annotators to align semantic frames according to the predicates only, this could easily be due to the mental challenge for lay annotators to compare and keep in mind all the semantic role fillers at the same time. But in the case of a fully automatic metric, on the other hand, it is easy for an algorithm to compute the individual similarities between all the semantic role fillers and consider the aggregate similarity when optimizing the alignment of semantic frames.

Surprisingly, however, the results will show that even in the automated case, this still does *not* help improve the correlation with human adequacy judgments.

6.1 Experimental setup

To align semantic frames using all semantic roles, we aggregate the lexical similarity of all the semantic role fillers into a semantic frame similarity score. We experiment on two variations of the aggregation function (1) simple linear average of the lexical similarity over the number of aligned semantic roles in the frames; or (2) the inverse of the sum of the negative log of the role fillers similarity.

The rest of the experimental setup is the same as that used in Section 4.

6.2 Results

Table 5 shows that to align semantic frames, using only the lexical similarity of the predicates between the frames in the reference translations and the MT output (0.37 Kendall in GALE-A and 0.19 Kendall in GALE-B) is more robust than either of the two natural ways of aggregating the lexical similarity of the aligned semantic role fillers. Aggregating by linear average yields a lower 0.35 Kendall in GALE-A and 0.10 Kendall in GALE-B. Aggregating by the inverse of the sum of negative logs yields a lower 0.30 Kendall in GALE-A and 0.17 Kendall in GALE-B.

Table 5: Sentence-level correlation with human adequacy judgments on GALE-A (training set) and GALE-B (testing set) for aligning semantic frames using predicate only vs. using all semantic role fillers aggregated by (1) the linear average of the lexical similarity vs. (2) the inverse of the sum of negative log of the lexical similarity.

Frame alignment	**GALE-A**	**GALE-B**
Predicate only	0.37	0.19
Linear average	0.35	0.10
Inverse of sum of neg. log	0.30	0.17

What might explain this perhaps surprising result? Our conjecture is that aggregating the lexical similarities of the semantic role fillers fails to help find better semantic frame alignments because the lexical similarities are aggregated with uniform weight across different types of role fillers. Therefore, the aggregation ignores the fact that different types of role types contribute to a widely varying degree to the meaning of an entire semantic frame—in reality, some role types are much more important than others. However, the complexity of the metric would be greatly increased if we added weights for each semantic roles type for semantic frame alignment process, and this would not be likely to be worthwhile given that automatic alignment is already performing as well as human alignment of semantic frames.

7 Don't word align semantic role fillers

Another question that naturally arises from the positive results above is: when aligning the semantic frames, would word-aligning the tokens within role fillers help? Specifically, if we had word alignments for every candidate pair of role filler strings, we could sum the lexical similarities only between the aligned tokens—instead of what we did above, which was to sum the lexical similarities of *all* pairwise combinations of tokens.

However, experimental results will show that, surprisingly, to judge the similarity of semantic role fillers, summing the lexical similarities over only word-aligned tokens—instead of all pairwise combinations of tokens—does *not* help to improve the correlation of the semantic MT evaluation with human adequacy judgment.

7.1 Experimental setup

To avoid the danger of aligning a token in one segment to excessive numbers of tokens in the other segment, we adopt a variant of competitive linking by Melamed (1996). Competitive linking is a greedy best-first word alignment algorithm.

Table 6: Sentence-level correlation with human adequacy judgments on GALE-A (training set) and GALE-B (testing set) for judging semantic role fillers similarity using pairwise tokens vs. only aligned tokens.

Semantic role filler similarity	GALE-A	GALE-B
All pairwise tokens	0.37	0.19
Only aligned tokens	0.36	0.17

The rest of the experimental setup is the same as that used in Section 4.

7.2 Results

Table 6 shows that, surprisingly, judging semantic role filler similarity using only the aligned tokens (selected by competitive linking word alignment algorithm) does *not* help the correlation with human adequacy judgment. This is surprising as, intuitively, using only the aligned tokens should avoid the introduction of noise in judging the similarity between semantic role fillers because it avoids adding in similarities for words within semantic role fillers whose meanings are not close to each other.

How might this outcome be explained? We conjecture that the word alignments over-constrain the calculation of segment similarities. The individual lexical similarities are already weighted fairly accurately, so the lexical similarities between words that do not correspond do not hurt since they are already close to zero. On the other hand, in cases where the word alignment is ambiguous, it is better to aggregate over different possible pairwise alignments—strictly obeying a hard word alignment undesirably forces dropping of some individual lexical similarity scores that are actually relevant.

8 Conclusion

We have introduced a new fully automatic semantic MT evaluation metric, MEANT, that is fundamentally based on semantic frames, that is the first such metric to outperform all other commonly used automatic MT evaluation metrics. Experimental results following the standard NIST MetricsMATR protocol indicate that our proposed metric achieves levels of correlation with human adequacy judgment (in our experiments, approximately 0.37) that significantly outperform BLEU, NIST, METEOR, PER, CDER, WER, and TER (in our experiments, ranging between 0.20 and 0.29).

We have also shown in this paper that the spirit of Occam's razor of HMEANT can be preserved even under full automation by (1) replacing human semantic role annotation with automatic shallow semantic parsing and (2) replacing human semantic frame alignment with a simple maximum weighted bipartite matching algorithm based on the lexical similarity between semantic frames. Under analysis, we have further shown empirically that performing this semantic frame alignment automatically tends to be just as good as performing it manually. Furthermore, we have shown surprisingly that (1) for aligning semantic frames, using *only* the similarity of predicates is more accurate than also taking into account the similarity of semantic role fillers, and (2) to judge similarity between semantic role fillers, aggregating similarity of *all* pairwise combination of word tokens is more accurate than considering only the similarity of the tokens that obey word alignments.

Papineni et al. (2002) stated in their conclusion that "We believe that BLEU will accelerate the MT R&D cycle by allowing researchers to rapidly home in on effective modeling ideas." since fully automatic metrics allow inexpensive training and tuning of SMT systems. Developments in the past decade have more than borne witness to this statement. However, SMT has progressed to the stage where simple metrics like BLEU are no longer capable of driving progress toward preservation of meaning with respect to proper event structure. We believe that MEANT that rapidly and accurately reflects the translation adequacy of MT output by directly assessing *who did what to whom, when, where and why* is needed to bring MT R&D to a new level of improvement in generating more meaningful MT output.

Acknowledgments

This material is based upon work supported in part by the Defense Advanced Research Projects Agency (DARPA) under BOLT contract no. HR0011-12-C-0016, and GALE contract nos. HR0011-06-C-0022 and HR0011-06-C-0023; by the European Union under the FP7 grant agreement no. 287658; and by the Hong Kong Research Grants Council (RGC) research grants GRF621008, GRF612806, DAG03/04.EG09, RGC6256/00E, and RGC6083/99E. 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 RGC, EU, or DARPA.

References

Satanjeev Banerjee and Alon Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the 43th Annual Meeting of the Association of Computational Linguistics (ACL-05)*, pages 65–72, 2005.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluating the role of BLEU in Machine Translation Research. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*, pages 249–256, 2006.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. (Meta-) evaluation of

Machine Translation. In *Proceedings of the 2nd Workshop on Statistical Machine Translation*, pages 136–158, 2007.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. Further Meta-evaluation of Machine Translation. In *Proceedings of the 3rd Workshop on Statistical Machine Translation*, pages 70–106, 2008.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Pryzbocki, and Omar Zaidan. Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53, Uppsala, Sweden, 15-16 July 2010.

Ido Dagan. Contextual word similarity. In Robert Dale, Herman Moisl, and Harold Somers, editors, *Handbook of Natural Language Processing*, pages 459–476. Marcel Dekker, New York, 2000.

G. Doddington. Automatic Evaluation of Machine Translation Quality using N-gram Co-occurrence Statistics. In *Proceedings of the 2nd International Conference on Human Language Technology Research (HLT-02)*, pages 138–145, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

Jesús Giménez and Lluís Màrquez. Linguistic Features for Automatic Evaluation of Heterogenous MT Systems. In *Proceedings of the 2nd Workshop on Statistical Machine Translation*, pages 256–264, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

Jesús Giménez and Lluís Màrquez. A Smorgasbord of Features for Automatic MT Evaluation. In *Proceedings of the 3rd Workshop on Statistical Machine Translation*, pages 195–198, Columbus, OH, June 2008. Association for Computational Linguistics.

Philipp Koehn and Christof Monz. Manual and Automatic Evaluation of Machine Translation between European Languages. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 102–121, 2006.

Gregor Leusch, Nicola Ueffing, and Hermann Ney. CDer: Efficient MT Evaluation Using Block Movements. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*, 2006.

Chi-kiu Lo and Dekai Wu. A Radically Simple, Effective Annotation and Alignment Methodology for Semantic Frame Based SMT and MT Evaluation. In *Proceedings of International Workshop on Using Linguistic Information for Hybrid Machine Translation (LiHMT 2011), organized by OpenMT-2.*, 2011.

Chi-kiu Lo and Dekai Wu. MEANT: An Inexpensive, High-Accuracy, Semi-Automatic Metric for Evaluating Translation Utility based on Semantic Roles. In *Proceedings of the Joint conference of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies (ACL-HLT-11)*, 2011.

Chi-kiu Lo and Dekai Wu. SMT vs. AI redux: How semantic frames evaluate MT more accurately. In *Proceedings of the 22nd 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 *Proceedings of the 5th Workshop on Syntax and Structure in Statistical Translation (SSST-5)*, 2011.

I. Dan Melamed. Automatic construction of clean broad-coverage translation lexicons. In *Proceedings of the 2nd Conference of the Association for Machine Translation in the Americas (AMTA-1996)*, 1996.

Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney. A Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC-2000)*, 2000.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, pages 311–318, 2002.

Sameer Pradhan, Wayne Ward, Kadri Hacioglu, James H. Martin, and Dan Jurafsky. Shallow Semantic Parsing Using Support Vector Machines. In *Proceedings of the 2004 Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL-04)*, 2004.

Matthew Snover, Bonnie J. Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA-06)*, pages 223–231, 2006.

Christoph Tillmann, Stephan Vogel, Hermann Ney, Arkaitz Zubiaga, and Hassan Sawaf. Accelerated DP Based Search For Statistical Translation. In *Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH-97)*, 1997.