MEANT: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility via semantic frames

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

Abstract

We introduce a novel semi-automated metric, MEANT, that assesses translation utility by matching semantic role fillers, producing scores that correlate with human judgment as well as HTER but at much lower labor cost. As machine translation systems improve in lexical choice and fluency, the shortcomings of widespread n-gram based, fluency-oriented MT evaluation metrics such as BLEU, which fail to properly evaluate adequacy, become more apparent. But more accurate, non-automatic adequacy-oriented MT evaluation metrics like HTER are highly labor-intensive, which bottlenecks the evaluation cycle. We first show that when using untrained monolingual readers to annotate semantic roles in MT output, the non-automatic version of the metric HMEANT achieves a 0.43 correlation coefficient with human adequacy judgments at the sentence level, far superior to BLEU at only 0.20, and equal to the far more expensive HTER. We then replace the human semantic role annotators with automatic shallow semantic parsing to further automate the evaluation metric, and show that even the semi-automated evaluation metric achieves a 0.34 correlation coefficient with human adequacy judgment, which is still about 80% as closely correlated as HTER despite an even lower labor cost for the evaluation procedure. The results show that our proposed metric is significantly better correlated with human judgment on adequacy than current widespread automatic evaluation metrics, while being much more cost effective than HTER.

1 Introduction

In this paper we show that evaluating machine translation by assessing the translation accuracy of each argument in the semantic role framework correlates with human judgment on translation *adequacy* as well as HTER, at a significantly lower labor cost. The correlation of this new metric, MEANT, with human judgment is far superior to BLEU and other automatic n-gram based evaluation metrics.

We argue that BLEU (Papineni *et al.*, 2002) and other automatic n-gram based MT evaluation metrics do not adequately capture the similarity in meaning between the machine translation and the reference translation—which, ultimately, is essential for MT output to be useful. N-gram based metrics assume that "good" translations tend to share the same lexical choices as the reference translations. While BLEU score performs well in capturing the translation fluency, Callison-Burch *et al.* (2006) and Koehn and Monz (2006) report cases where BLEU strongly disagree with human judgment on translation quality. The underlying reason is that lexical similarity does not adequately reflect the similarity in meaning. As MT systems improve, the shortcomings of the n-gram based evaluation metrics are becoming more apparent. State-of-the-art MT systems are often able to output fluent translations that are nearly grammatical and contain roughly the correct words, but still fail to express meaning that is close to the input.

At the same time, although HTER (Snover *et al.*, 2006) is more adequacy-oriented, it is only employed in very large scale MT system evaluation instead of day-to-day research activities. The underlying reason is that it requires rigorously trained human experts to make difficult combinatorial decisions on the minimal number of edits so as to make the MT output convey the same meaning as the reference translation—a highly labor-intensive, costly process that bottlenecks the evaluation cycle.

Instead, with MEANT, we adopt at the outset the principle 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. It is true that limited tasks might exist for which inadequate translations are still useful. But for meaningful tasks, generally speaking, for a translation to be useful, at least the basic event structure must be correctly understood. Therefore, our objective is to evaluate *translation utility*: from a user's point of view, how well is

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 220–229,
Portland, Oregon, June 19-24, 2011. ©2011 Association for Computational Linguistics

the most essential semantic information being captured by machine translation systems?

In this paper, we detail the methodology that underlies MEANT, which extends and implements preliminary directions proposed in (Lo and Wu, 2010a) and (Lo and Wu, 2010b). We present the results of evaluating translation utility by measuring the accuracy within a semantic role labeling (SRL) framework. We show empirically that our proposed SRL based evaluation metric, which uses untrained monolingual humans to annotate semantic frames in MT output, correlates with human adequacy judgments as well as HTER, and far better than BLEU and other commonly used metrics. Finally, we show that replacing the human semantic role labelers with an automatic shallow semantic parser in our proposed metric yields an approximation that is about 80% as closely correlated with human judgment as HTER, at an even lower cost—and is still far better correlated than n-gram based evaluation metrics.

2 Related work

Lexical similarity based metrics BLEU (Papineni *et al.*, 2002) is 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 disagree with human judgment on 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) and WER (Nießen *et al.*, 2000), 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. State-of-the-art MT systems are often able to output translations containing roughly the correct words, yet expressing meaning that is not close to that of the input.

We argue that a translation metric that reflects meaning similarity is better based on similarity in semantic structure, rather than simply flat lexical similarity.

HTER (non-automatic) Despite the fact that Human-targeted Translation Edit Rate (HTER) as proposed by Snover *et al.* (2006) shows a high correlation with human judgment on translation adequacy, it is not widely used in day-to-day machine translation evaluation because of its high labor cost. HTER not only requires human experts to understand the meaning expressed in both the reference translation and the machine translation, but also requires them to propose the minimum number of edits to the MT output such that the post-edited MT output conveys the same meaning as the reference translation. Requiring such heavy manual decision making greatly increases the cost of evaluation, bottlenecking the evaluation cycle.

To reduce the cost of evaluation, we aim to reduce any human decisions in the evaluation cycle to be as simple as possible, such that even untrained humans can quickly complete the evaluation. The human decisions should also be defined in a way that can be closely approximated by automatic methods, so that similar objective functions might potentially be used for tuning in MT system development cycles.

Task based metrics (non-automatic) Voss and Tate (2006) proposed a task-based approach to MT evaluation that is in some ways similar in spirit to ours, but rather than evaluating how well people understand the meaning as a whole conveyed by a sentence translation, they measured the recall with which humans can extract *one* of the *who*, *when*, or *where* elements from MT output—and without attaching them to any predicate or frame. A large number of human subjects were instructed to extract only *one* particular type of *wh*-item from each sentence. They evaluated only whether the role fillers were correctly identified, without checking whether the roles were appropriately attached to the correct predicate. Also, the actor, experiencer, and patient were all conflated into the undistinguished *who* role, while other crucial elements, like the action, purpose, manner, were ignored.

Instead, we argue, evaluating meaning similarity should be done by evaluating the semantic structure as a whole: (a) *all* core semantic roles should be checked, and (b) not only should we evaluate the presence of semantic role fillers in isolation, but also their relations to the frames' predicates.

Syntax based metrics Unlike Voss and Tate, Liu and Gildea (2005) proposed a structural approach, but it was based on syntactic rather than semantic structure, and focused on checking the correctness of the role *structure* without checking the correctness of the role *fillers*. Their subtree metric (STM) and headword chain metric (HWC) address the failure of BLEU to evaluate translation *grammaticality*; however, the problem remains that a grammatical translation can achieve a high syntax-based score even if contains meaning errors arising from confusion of semantic roles.

STM was the first proposed metric to incorporate syntactic features in MT evaluation, and STM underlies most other recently proposed syntactic MT evaluation metrics, for example the evaluation metric based on lexical-functional grammar of Owczarzak *et al.* (2008). STM is a precision-based metric that measures what fraction of subtree *structures* are shared between the parse trees of

machine translations and reference translations (averaging over subtrees up to some depth threshold). Unlike Voss and Tate, however, STM does not check whether the role *fillers* are correctly translated.

HWC is similar, but is based on dependency trees containing lexical as well as syntactic information. HWC measures what fraction of headword chains (a sequence of words corresponding to a path in the dependency tree) also appear in the reference dependency tree. This can be seen as a similarity measure on n-grams of dependency chains. Note that the HWC's notion of lexical similarity still requires exact word match.

Although STM-like syntax-based metrics are an improvement over flat lexical similarity metrics like BLEU, they are still more fluency-oriented than adequacy-oriented. Similarity of syntactic rather than semantic structure still inadequately reflects meaning preservation. Moreover, properly measuring translation *utility* requires verifying whether role *fillers* have been correctly translated—verifying only the abstract structures fails to penalize when role fillers are confused.

Semantic roles 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: semantic role overlapping, semantic role matching, and semantic structure overlapping. Unlike Liu and Gildea (2007) who use discriminative training to tune the weight on each feature, ULC uses uniform weights. Although the 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 heterogenous types of features.

Specifically, note that the feature based representations of semantic roles used in these aggregate metrics do not actually capture the structural predicate-argument relations. "Semantic structure overlapping" can be seen as the shallow semantic version of STM: it only measures the similarity of the tree structure of the semantic roles, without considering the lexical realization. "Semantic role overlapping" calculates the degree of lexical overlap between semantic roles of the same type in the machine translation and its reference translation, using simple bag-of-words counting; this is then aggregated into an average over all semantic role types. "Semantic role matching" is just like "semantic role overlapping", except that bag-of-words degree of similarity is replaced (rather harshly) by a boolean indicating whether the role fillers are an exact string match. It is important to note that "semantic role overlapping" and "semantic role matching" both use flat feature based representations which do not capture the structural relations in semantic frames, i.e., the predicate-argument relations.

Like system combination approaches, ULC is a vastly more complex aggregate metric compared to widely used metrics like BLEU or STM. We believe it is important to retain a focus on developing *simpler* metrics which not only correlate well with human adequacy judgments, but nevertheless still directly 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. We also believe that to provide a foundation for better design of efficient automated metrics, making use of *humans* for annotating semantic roles and judging the role translation accuracy in MT output is an essential step that should not be bypassed, in order to adequately understand the upper bounds of such techniques.

We agree with Przybocki *et al.* (2010), who observe in the NIST MetricsMaTr 2008 report that "human [adequacy] assessments only pertain to the translations evaluated, and are of no use even to updated translations from the same systems". Instead, we aim for MT evaluation metrics that provide fine-grained scores in a way that also directly reflects interpretable insights on the strengths and weaknesses of MT systems rather than simply replicating human assessments.

3 MEANT: SRL for MT evaluation

A good translation is one from which human readers may successfully understand at least the basic event structure—"*who* did *what* to *whom*, *when*, *where* and *why*" (Pradhan *et al.*, 2004)—which represents the most essential meaning of the source utterances.

MEANT measures this as follows. First, semantic role labeling is performed (either manually or automatically) on both the reference translation and the machine translation. The semantic frame structures thus obtained for the MT output are compared to those in the reference translations, frame by frame, argument by argument. The frame translation accuracy is a weighted sum of the number of correctly translated arguments. Conceptually, MEANT is defined in terms of f-score, with respect to the precision/recall for sentence translation accuracy as calculated by averaging the translation accuracy for all frames in the MT output across the number of frames in the MT output/reference translations. Details are given below.

3.1 Annotating semantic frames

In designing a semantic MT evaluation metric, one important issue that should be addressed is how to evaluate the similarity of meaning objectively and systematically

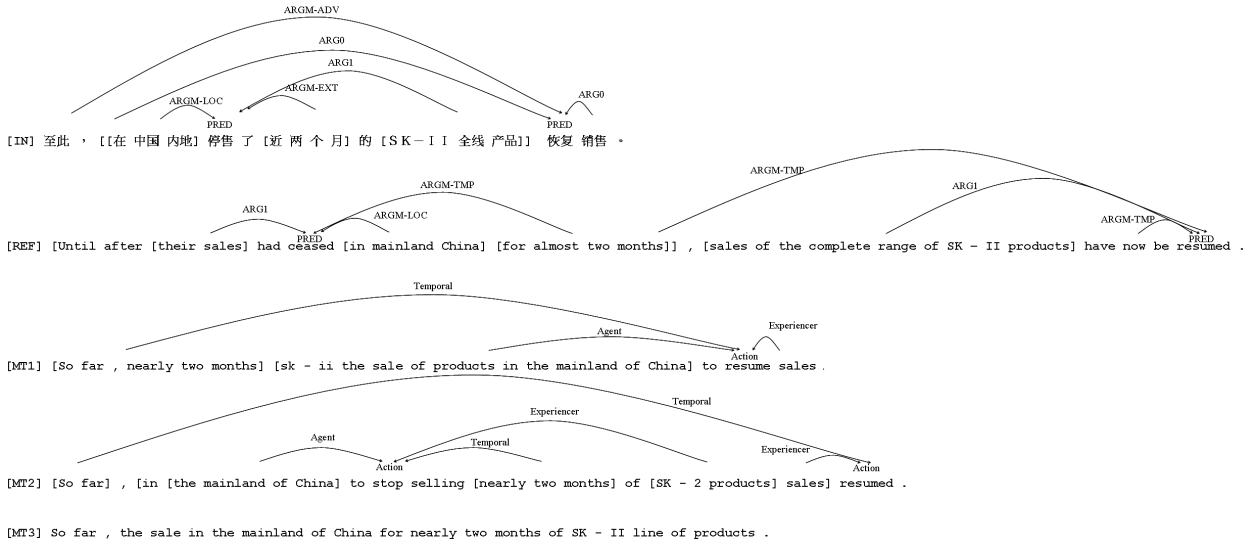

Figure 1: Example of source sentence and reference translation with reconstructed semantic frames in Propbank format and MT output with reconstructed semantic frames by minimal trained human annotators. Following Propbank, there are no semantic frames for MT3 because there is no predicate.

using fine-grained measures. We adopted the Propbank SRL style predicate-argument framework, which captures the basic event structure in a sentence in a way that clearly indicates many strengths and weaknesses of MT. Figure 1 shows the reference translation with reconstructed semantic frames in Propbank format and the corresponding MT output with reconstructed semantic frames by minimal trained human annotators.

3.2 Comparing semantic frames

After annotating the semantic frames, we must determine the translation accuracy for each semantic role filler in the reference and machine translations. Although ultimately it would be nice to do this automatically, it is essential to first understand extremely well the upper bound of accuracy for MT evaluation via semantic frame theory. Thus, instead of resorting to excessively permissive bag-of-words matching or excessively restrictive exact string matching, for the experiments reported here we employed a group of human judges to evaluate the correctness of each role filler translation between the reference and machine translations.

In order to facilitate a finer-grained measurement of utility, the human judges were not only allowed to mark each role filler translation as "correct" or "incorrect", but also "partial". Translations of role fillers are judged "correct" if they express the same meaning as that of the reference translations (or the original source input, in the bilinguals experiment discussed later). Translations may also be judged "partial" if only part of the meaning is correctly translated. Extra meaning in a role filler is not penalized unless it belongs in another role. We also assume that a wrongly translated predicate means that the entire semantic frame is incorrect; therefore, the "correct" and "partial" argument counts are collected only if their associated predicate is correctly translated in the first place.

Table 1 shows an example of SRL annotation of MT1 in Figure 1 by one of the annotators, along with the human judgment on translation accuracy of each argument. The predicate ceased in the reference translation did not match with any predicate annotated in MT1, while the predicate resumed matched with the predicate resume annotated in MT1. All arguments of the untranslated ceased are automatically considered incorrect (with no need to consider each argument individually), under our assumption that a wrongly translated predicate causes the entire event frame to be considered mistranslated. The ARGM-TMP argument, Until after their sales had ceased in mainland China for almost two months, in the reference translation is partially translated to ARGM-TMP argument, So far , nearly two months, in MT1. Similar decisions are made for the ARG1 argument and the other ARGM-TMP argument; now in the reference translation is missing in MT1.

3.3 Quantifying semantic frame match

To quantify the above in a summary metric, we define MEANT in terms of an f-score that balances the precision and recall analysis of the comparative matrices collected from the human judges, as follows.

$C_{i,j}$ = # correct fillers of ARG j for PRED i in MT

$P_{i,j}$ = # partial fillers of ARG j for PRED i in MT

$M_{i,j}$ = total # fillers of ~~ARG j for~~ PRED i in MT

$R_{i,j}$ = total # fillers of ~~ARG j of~~ PRED i in REF

Table 1: SRL annotation of MT1 in Figure 1 and the human judgment of translation accuracy for each argument (see text).

SRL	REF	MT1	Decision
PRED (Action)	ceased	–	no match
PRED (Action)	resumed	resume	match
ARG0 (Agent)	–	sk - ii the sale of products in the mainland of China	incorrect
ARG1 (Experiencer)	sales of complete range of SK - II products	sales	partial
ARGM-TMP (Temporal)	Until after , their sales had ceased in mainland China for almost two months	So far , nearly two months	partial
ARGM-TMP (Temporal)	now	–	incorrect

$$C_{\text{precision}} = \sum_{\text{matched i}} \frac{w_{\text{pred}} + \sum_j w_{\text{j}} C_{\text{i,j}}}{w_{\text{pred}} + \sum_j w_{\text{j}} M_{\text{i,j}}}$$

$$C_{\text{recall}} = \sum_{\text{matched i}} \frac{w_{\text{pred}} + \sum_j w_{\text{j}} C_{\text{i,j}}}{w_{\text{pred}} + \sum_j w_{\text{j}} R_{\text{i,j}}}$$

$$P_{\text{precision}} = \sum_{\text{matched i}} \frac{\sum_j w_{\text{j}} P_{\text{i,j}}}{w_{\text{pred}} + \sum_j w_{\text{j}} M_{\text{i,j}}}$$

$$P_{\text{recall}} = \sum_{\text{matched i}} \frac{\sum_j w_{\text{j}} P_{\text{i,j}}}{w_{\text{pred}} + \sum_j w_{\text{j}} R_{\text{i,j}}}$$

$$\text{precision} = \frac{C_{\text{precision}} + (w_{\text{partial}} \times P_{\text{precision}})}{\text{total \# predicates in MT}}$$

$$\text{recall} = \frac{C_{\text{recall}} + (w_{\text{partial}} \times P_{\text{recall}})}{\text{total \# predicates in REF}}$$

$$\text{f-score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

$C_{\text{precision}}$, $P_{\text{precision}}$, C_{recall} and P_{recall} are the sum of the fractional counts of correctly or partially translated semantic frames in the MT output and the reference, respectively, which can be viewed as the true positive for precision and recall of the whole semantic structure in one source utterence. Therefore, the SRL based MT evaluation metric is equivalent to the f-score, i.e., the translation accuracy for the whole predicate-argument structure.

Note that w_{pred}, w_{j} and w_{partial} are the weights for the matched predicate, arguments of type j, and partial translations. These weights can be viewed as the importance of meaning preservation for each different category of semantic roles, and the penalty for partial translations. We will describe below how these weights are estimated.

If all the reconstructed semantic frames in the MT output are completely identical to those annotated in the reference translation, and all the arguments in the reconstructed frames express the same meaning as the corresponding arguments in the reference translations, then the f-score will be equal to 1.

For instance, consider MT1 in Figure 1. The number of frames in MT1 and the reference translation are 1 and 2, respectively. The total number of participants (including both predicates and arguments) of the resume frame in both MT1 and the reference translation is 4 (one predicate and three arguments), with 2 of the arguments (one ARG1/experiencer and one ARGM-TMP/temporal) only partially translated. Assuming for now that the metric aggregates ten types of semantic roles with uniform weight for each role (optimization of weights will be discussed later), then $w_{\text{pred}} = w_j = 0.1$, and so $C_{\text{precision}}$ and C_{recall} are both zero while $P_{\text{precision}}$ and $P_{\text{recall}}$are both 0.5. If we further assume that $w_{\text{partial}} = 0.5$, then precison and recall are 0.25 and 0.125 respectively. Thus the f-score for this example is 0.17.

Both human and semi-automatic variants of the MEANT translation evaluation metric were meta-evaluated, as described next.

4 Meta-evaluation methodology

4.1 Evaluation Corpus

We leverage work from Phase 2.5 of the DARPA GALE program in which both a subset of the Chinese source sentences, as well as their English reference, are being annotated with semantic role labels in Propbank style. The corpus also includes three participating state-of-the-art MT systems' output. For present purposes, we randomly drew 40 sentences from the newswire genre of the corpus to form a meta-evaluation corpus. To maintain a controlled environment for experiments and consistent comparison, the evaluation corpus is fixed throughout this work.

4.2 Correlation with human judgements on adequacy

We followed the benchmark assessment procedure in WMT and NIST MetricsMaTr (Callison-Burch *et al.*, 2008, 2010), assessing the performance of the proposed evaluation metric at the sentence level using ranking preference consistency, which also known as Kendall's τ rank correlation coefficient, to evaluate the correlation of the proposed metric with human judgments on translation adequacy ranking. A higher value for τ indicates more similarity to the ranking by the evaluation metric to the human judgment. The range of possible values of correlation coefficient is [-1,1], where 1 means the systems are ranked

Table 2: List of semantic roles that human judges are requested to label.

Label	Event	Label	Event
Agent	who	Location	where
Action	did	Purpose	why
Experiencer	what	Manner	how
Patient	whom	Degree or Extent	how
Temporal	when	Other adverbial arg.	how

in the same order as the human judgment and -1 means the systems are ranked in the reverse order as the human judgment.

5 Experiment: Using human SRL

The first experiment aims to provide a more concrete understanding of one of the key questions as to the upper bounds of the proposed evaluation metric: how well can human annotators perform in reconstructing the semantic frames in MT output? This is important since MT output is still not close to perfectly grammatical for a good syntactic parsing—applying automatic shallow semantic parsers, which are trained on grammatical input and valid syntactic parse trees, on MT output may significantly underestimate translation utility.

5.1 Experimental setup

We thus introduce HMEANT, a variant of MEANT based on the idea that semantic role labeling can be simplified into a task that is easy and fast even for untrained humans. The human annotators are given only very simple instructions of less than half a page, along with two examples. Table 2 shows the list of labels annotators are requested to annotate, where the semantic role labeling instructions are given in the intuitive terms of "who did what to whom, when, where, why and how". To facilitate the inter-annotator agreement experiments discussed later, each sentence is independently assigned to at least two annotators.

After calculating the SRL scores based on the confusion matrix collected from the annotation and evaluation, we estimate the weights using grid search to optimize correlation with human adequacy judgments.

5.2 Results: Correlation with human judgement

Table 3 shows results indicating that HMEANT correlates with human judgment on adequacy as well as HTER does (0.432), and is far superior to BLEU (0.198) or other surface-oriented metrics.

Inspection of the cross validation results shown in Table 4 indicates that the estimated weights are not overfitting. Recall that the weights used in HMEANT are globally estimated (by grid search) using the evaluation corpus. To analyze stability, the corpus is also partitioned randomly into four folds of equal size. For each fold, another grid search is also run. R_{HMEANT} is the rank at which the Kendall's correlation for HMEANT is found, if the Kendall's correlations for all points in the grid search space are sorted. Many similar weight-vectors produce the same Kendall's correlation score, so "distinct R" shows how many distinct Kendall's correlation scores exist in each case—between 16 and 29. HMEANT's weight settings always produce Kendall's correlation scores among the top 5, regardless of which fold is chosen, indicating good stability of HMEANT's weight-vector.

Table 3: Sentence-level correlation with human adequacy judgments, across the evaluation metrics.

Metrics	Kendall τ
HMEANT	**0.4324**
HTER	**0.4324**
NIST	0.2883
BLEU	0.1982
METEOR	0.1982
TER	0.1982
PER	0.1982
CDER	0.1171
WER	0.0991

Table 4: Analysis of stability for HMEANT's weight settings, with R_{HMEANT} rank and Kendall's τ correlation scores (see text).

	Fold 0	Fold 1	Fold 2	Fold 3
R_{HMEANT}	3	1	3	5
distinct R	16	29	19	17
τ_{HMEANT}	0.33	0.48	0.48	0.40
τ_{HTER}	0.59	0.41	0.44	0.30
τ_{CV} train	0.45	0.42	0.40	0.43
τ_{CV} test	0.33	0.37	0.48	0.40

Next, Kendall's τ correlation scores are shown for HMEANT on each fold. They vary from 0.33 to 0.48, and are at least as stable as those shown for HTER, where τ varies from 0.30 to 0.59.

Finally, τ_{CV} shows Kendall's correlations if the weight-vector is instead subjected to full cross-validation training and testing, again demonstrating good stability. In fact, the correlations for the training set in three of the folds (0, 2, and 3) are identical to those for HMEANT.

5.3 Results: Cost of evaluating

The time needed for training non-expert humans to carry out our annotation protocol is significantly less than HTER and gold standard Propbank annotation. The half-page instructions given to annotators required only between 5 to 15 minutes for all annotators, including time

for asking questions if necessary. Aside from providing two annotated examples, no further training was given.

Similarly, the time needed for running the evaluation metric is also significantly less than HTER—under at most 5 minutes per sentence, even for non-expert humans using no computer-assisted UI tools. The average time used for annotating each sentence was lower bounded by 2 minutes and upper bounded by 3 minutes, and the time used for determing the translation accuracy of role fillers averaged under 2 minutes.

Note that these figures are for unskilled non-experts. These times tend to diminish significantly after annotators acquire experience.

6 Experiment: Monolinguals vs. bilinguals

We now show that using monolingual annotators is essentially just as effective as using more expensive bilingual annotators. We study the cost/benefit trade-off of using human annotators from different language backgrounds for the proposed evaluation metric, and compare whether providing the original source text helps. Note that this experiment focuses on the SRL annotation step, rather than the judgments of role filler paraphrasing accuracy, because the latter is only a simple three-way decision between "correct", "partial", and "incorrect" that is far less sensitive to the annotators' language backgrounds.

MT output is typically poor. Therefore, readers of MT output often guess the original meaning in the source input using their own language background knowledge. Readers' language background thus affects their understanding of the translation, which could affect the accuracy of capturing the key semantic roles in the translation.

6.1 Experimental Setup

Both English monolinguals and Chinese-English bilinguals (Chinese as first language and English as second language) were employed to annotate the semantic roles. For bilinguals, we also experimented with the difference in guessing constraints by optionally providing the original source input together with the translation. Therefore, there are three variations in the experiment setup: monolinguals seeing translation output only; bilinguals seeing translation output only; and bilinguals seeing both input and output.

The aim here is to do a rough sanity check on the effect of the variation of language background of the annotators; thus for these experiments we have not run the weight estimation step after SRL based f-score calculation. Instead, we simply assigned a uniform weight to all the semantic elements, and evaluated the variation under the same weight settings. (The correlation scores reported in this section are thus expected to be lower than that reported in the last section.)

Table 5: Sentence-level correlation with human adequacy judgments, for monolinguals vs. bilinguals. Uniform rather than optimized weights are used.

Metrics	**Kendall τ**
HMEANT - bilinguals	0.3514
HMEANT - monolinguals	0.3153
HMEANT - bilinguals with input	0.3153

6.2 Results

Table 5 of our results shows that using more expensive bilinguals for SRL annotation instead of monolinguals improves the correlation only slightly. The correlation coefficient of the SRL based evaluation metric driven by bilingual human annotators (0.351) is slightly better than that driven by monolingual human annotators (0.315); however, using bilinguals in the evaluation process is more costly than using monolinguals.

The results show that even allowing the bilinguals to see the input as well as the translation output for SRL annotation does not help the correlation. The correlation coefficient of the SRL based evaluation metric driven by bilingual human annotators who see also the source input sentences is 0.315 which is the same as that driven by monolingual human annotators. We find that the correlation coefficient of the proposed with human judgment on adequacy drops when bilinguals are shown to the source input sentence during annotation. Error analyses lead us to believe that annotators will drop some parts of the meaning in the translations when trying to align them to the source input.

This suggests that HMEANT requires only monolingual English annotators, who can be employed at low cost.

7 Inter-annotator agreement

One of the concerns of the proposed metric is that, given only minimal training on the task, humans would annotate the semantic roles so inconsistently as to reduce the reliability of the evaluation metric. Inter-annotator agreement (IAA) measures the consistency of human in performing the annotation task. A high IAA suggests that the annotation is consistent and the evaluation results are reliable and reproducible.

To obtain a clear analysis on where any inconsistency might lie, we measured IAA in two steps: role identification and role classification.

7.1 Experimental setup

Role identification Since annotators are not consistent in handling articles or punctuation at the beginning or the end of the annotated arguments, the agreement of semantic role identification is counted over the matching of

Table 6: Inter-annotator agreement rate on role identification (matching of word span)

Experiments	REF	MT
bilinguals working on output only	76%	72%
monolinguals working on output only	93%	75%
bilinguals working on input-output	75%	73%

Table 7: Inter-annotator agreement rate on role classification (matching of role label associated with matched word span)

Experiments	Ref	MT
bilinguals working on output only	69%	65%
monolinguals working on output only	88%	70%
bilinguals working on input-output	70%	69%

word span in the annotated role fillers with a tolerance of ±1 word in mismatch. The inter-annotator agreement rate (IAA) on the role identification task is calculated as follows. A_1 and A_2 denote the number of annotated predicates and arguments by annotator 1 and annotator 2 respectively. M_{span} denotes the number of annotated predicates and arguments with matching word span between annotators.

$$P_{\text{identification}} = \frac{M_{\text{span}}}{A_1}$$
$$R_{\text{identification}} = \frac{M_{\text{span}}}{A_2}$$
$$\text{IAA}_{\text{identification}} = \frac{2 * P_{\text{identification}} * R_{\text{identification}}}{P_{\text{identification}} + R_{\text{identification}}}$$

Role classification The agreement of classified roles is counted over the matching of the semantic role labels within two aligned word spans. The IAA on the role classification task is calculated as follows. M_{label} denotes the number of annotated predicates and arguments with matching role label between annotators.

$$P_{\text{classification}} = \frac{M_{\text{label}}}{A_1}$$
$$R_{\text{classification}} = \frac{M_{\text{label}}}{A_2}$$
$$\text{IAA}_{\text{classification}} = \frac{2 * P_{\text{classification}} * R_{\text{classification}}}{P_{\text{classification}} + R_{\text{classification}}}$$

7.2 Results

The high inter-annotator agreement suggests that the annotation instructions provided to the annotators are in general sufficient and the evaluation is repeatable and could be automated in the future. Table 6 and 7 show the annotators reconstructed the semantic frames quite consistently, even they were given only simple and minimal training.

We have noticed that the agreement on role identification is higher than that on role classification. This suggests that there are role confusion errors among the annotators. We expect a slightly more detailed instructions and explanations on different roles will further improve the IAA on role classification.

The results also show that monolinguals seeing output only have the highest IAA in semantic frame reconstruction. Data analyses lead us to believe the monolinguals are the most constrained group in the experiments. The monolingual annotators can only guess the meaning in the MT output using their English language knowledge. Therefore, they all understand the translation almost the same way, even if the translation is incorrect.

On the other hand, bilinguals seeing both the input and output discover the mistranslated portions, and often unconsciously try to compensate by re-interpreting the MT output with information not necessarily appearing in the translation, in order to better annotate what they think it should have conveyed. Since there are many degrees of freedom in this sort of compensatory re-interpretation, this group achieved a lower IAA than the monolinguals.

Bilinguals seeing only output appear to take this even a step further: confronted with a poor translation, they often unconsciously try to guess what the original input might have been. Consequently, they agree the least, because they have the most freedom in applying their own knowledge of the unseen input language, when compensating for poor translations.

8 Experiment: Using automatic SRL

In the previous experiment, we showed that the proposed evaluation metric driven by human semantic role annotators performed as well as HTER. It is now worth asking a deeper question: can we further reduce the labor cost of MEANT by using automatic shallow semantic parsing instead of humans for semantic role labeling?

Note that this experiment focuses on understanding the cost/benefit trade-off for the semantic frame reconstruction step. For SRL annotation, we replace humans with automatic shallow semantic parsing. We decouple this from the ternary judgments of role filler accuracy, which are still made by humans. However, we believe the evaluation of role filler accuracy will also be automatable.

8.1 Experimental setup

We performed three variations of the experiments to assess the performance degradation from the automatic approximation of semantic frame reconstruction in each translation (reference translation and MT output): we applied automatic shallow semantic parsing on the MT output only; on the reference translation only; and on both reference translation and MT output. For the semantic

Table 8: Sentence-level correlation with human adequacy judgments. *The weights for individual roles in the metric are tuned by optimizing the correlation.

Metrics	Kendall τ
HTER	0.4324
HMEANT gold - monolinguals *	0.4324
HMEANT auto - monolinguals *	0.3964
MEANT gold - auto *	0.3694
MEANT auto - auto *	0.3423
NIST	0.2883
BLEU / METEOR / TER / PER	0.1982
CDER	0.1171
WER	0.0991

parser, we used ASSERT (Pradhan *et al.*, 2004) which achieves roughly 87% semantic role labeling accuracy.

8.2 Results

Table 8 shows that the proposed SRL based evaluation metric correlates slightly worse than HTER with a much lower labor cost. The correlation with human judgment on adequacy of the fully automated SRL annotation version, i.e., applying ASSERT on both the reference translation and the MT output, of the SRL based evaluation metric is about 80% of that of HTER. The results also show that the correlation with human judgment on adequacy of either one side of translation using automatic SRL is in the 85% to 95% range of that HTER.

9 Conclusion

We have presented MEANT, a novel semantic MT evaluation metric that assesses the translation accuracy via Propbank-style semantic predicates, roles, and fillers. MEANT provides an intuitive picture on how much information is correctly translated in the MT output.

MEANT can be run using inexpensive untrained monolinguals and yet correlates with human judgments on adequacy as well as HTER with a lower labor cost. In contrast to HTER, which requires rigorous training of human experts to find a minimum edit of the translation (an exponentially large search space), MEANT requires untrained humans to make well-defined, bounded decisions on annotating semantic roles and judging translation correctness. The process by which MEANT reconstructs the semantic frames in a translation and then judges translation correctness of the role fillers conceptually models how humans read and understand translation output.

We also showed that using automatic shallow semantic parser to further reduce the labor cost of the proposed metric successfully approximates roughly 80% of the correlation with human judgment on adequacy. The results suggest future potential for a fully automatic variant of MEANT that could out-perform current automatic MT evaluation metrics and still perform near the level of HTER.

Numerous intriguing questions arise from this work. A further investigation into the correlation of each of the individual roles to human adequacy judgments is detailed elsewhere, along with additional improvements to the MEANT family of metrics (Lo and Wu, 2011). Another interesting investigation would then be to similarly replicate this analysis of the impact of each individual role, but using automatically rather than manually labeled semantic roles, in order to ascertain whether the more difficult semantic roles for automatic semantic parsers might also correspond to the less important aspects of end-to-end MT utility.

Acknowledgments

This material is based upon work supported in part by the Defense Advanced Research Projects Agency (DARPA) under GALE Contract Nos. HR0011-06-C-0022 and HR0011-06-C-0023 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 Defense Advanced Research Projects Agency.

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.

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.

Ding Liu and Daniel Gildea. Syntactic Features for Evaluation of Machine Translation. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, page 25, 2005.

Ding Liu and Daniel Gildea. Source-Language Features and Maximum Correlation Training for Machine Translation Evaluation. In *Proceedings of the 2007 Conference of the North American Chapter of the Association of Computational Linguistics (NAACL-07)*, 2007.

Chi-kiu Lo and Dekai Wu. Evaluating machine translation utility via semantic role labels. In *Seventh International Conference on Language Resources and Evaluation (LREC-2010)*, pages 2873–2877, Malta, May 2010.

Chi-kiu Lo and Dekai Wu. Semantic vs. syntactic vs. n-gram structure for machine translation evaluation. In Dekai Wu, editor, *Proceedings of SSST-4, Fourth Workshop on Syntax and Structure in Statistical Translation (at COLING 2010)*, pages 52–60, Beijing, Aug 2010.

Chi-kiu Lo and Dekai Wu. SMT vs. AI redux: How semantic frames evaluate MT more accurately. In *22nd International Joint Conference on Artificial Intelligence (IJCAI-11)*, Barcelona, Jul 2011. To appear.

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.

Karolina Owczarzak, Josef van Genabith, and Andy Way. Evaluating machine translation with LFG dependencies. *Machine Translation*, 21:95–119, 2008.

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.

Mark Przybocki, Kay Peterson, Sébastien Bronsart, and Gregory Sanders. The NIST 2008 Metrics for Machine Translation Challenge - Overview, Methodology, Metrics, and Results. *Machine Tr*, 23:71–103, 2010.

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.

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 *Proceedings of the 11th Annual Conference of the European Association for Machine Translation (EAMT-2006)*, pages 203–212, Oslo, Norway, June 2006.