SMT Versus AI Redux: How Semantic Fames Evaluate MT More Accurately

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

Abstract

We argue for an alternative paradigm in evaluating machine translation quality that is strongly empirical but more accurately reflects the utility of translations, by returning to a representational foundation based on AI oriented lexical semantics, rather than the superficial flat n-gram and string representations recently dominating the field. Driven by such metrics as BLEU and WER, current SMT frequently produces unusable translations where the semantic event structure is mistranslated: *who did what to whom, when, where, why, and how?* We argue that it is time for a new generation of more intelligent" automatic and semi-automatic metrics, based clearly on getting the structure right at the lexical semantics level. We show empirically that it is possible to use simple PropBank style semantic frame representations to surpass all currently widespread metrics' correlation to human adequacy judgments, including even HTER. We also show that replacing human annotators with automatic semantic role labeling still yields much of the advantage of the approach. We combine the best of both worlds: from an SMT perspective, we provide superior yet low-cost quantitative objective functions for translation quality; and yet from an AI perspective, we regain the representational transparency and clear reflection of semantic utility of structural frame-based knowledge representations.

1 Introduction

For the past decade, progress in statistical machine translation has been largely measured via highly surface oriented "non-intelligent" 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), or word error rate (Nießen *et al.*, 2000). Despite their simplicity, such mechanical metrics have driven progress with great success, largely because subjective metrics are too expensive and insufficiently quantitative to incorporate into modern training methods.

However, such metrics have begun to show their limits as flat n-gram oriented statistical machine translation systems have plateaued. Surface oriented metrics have been quite successful in ranking overall systems with respect to each other when their scores are averaged over entire documents or corpora. But they do not fare so well at the level of ranking translations of individual sentences, in large part because they are poor at appropriately reflecting translation quality improvements stemming from better semantic predicate structure translations. Even when human judges clearly indicate that one sentence translation contains significantly more meaningful word sense or semantic frame choices, surface oriented metrics typically register little difference.

As a consequence, current statistical machine translation models that have been trained to optimize on surface oriented metrics frequently produce unusable translations where the semantic event structure is mistranslated: *who did what to whom, when, where, why, and how?* Such basic shallow semantic structures, which are the basis of lexical semantics representations from case grammar (Fillmore, 1968) to FrameNet (Baker *et al.*, 1998) to PropBank (Palmer *et al.*, 2005), have proven highly amenable to large-scale data-driven machine learning techniques (e.g., Pradhan *et al.* (2004) or Xue and Palmer (2005)). Nevertheless, they have remained overlooked in SMT until very recently.

Meanwhile, recent work on semantic SMT shows promise in employing lexical semantics models for both word sense disambiguation (WSD) and semantic role labeling (SRL). WSD models can be used to combine a wide range of context features into a single translation lexical choice prediction, as in the work of Carpuat and Wu (2007), Chan *et al.* (2007), and Giménez and Màrquez (2007). In particular, the Phrase Sense Disambiguation (PSD) model of Carpuat and Wu (2007) and Carpuat and Wu (2008) generalizes WSD in order to automatically provide fully phrasal translation lexicons with context-dependent probability distributions over the possible translation candidates.

Semantic SMT has also recently begun applying SRL models, facilitated by the increasing availability of annotated parallel corpora as in Palmer *et al.* (2005) and Xue and Palmer (2005), with automatic shallow parsers for English (Pradhan *et al.*, 2004), as well as Chinese (Fung *et al.*, 2006) achieving f-scores in the 82% range, plus cross-lingual semantic verb frame argument mappings in the 89% range (Fung *et al.*, 2007). Following on this, Wu and Fung (2009a) and Wu and Fung (2009b) apply SRL to SMT decoding, using

a SRL based reordering model that returns improved translations containing fewer semantic role confusion errors.

We argue that the time is ripe for a new generation of more "intelligent" automatic and semi-automatic metrics, based clearly on getting the structure right at the lexical semantics level. We detail studies using improved versions of one such metric, MEANT, and its human variant HMEANT (Lo and Wu, 2011), arising from methodologies proposed in (Lo and Wu, 2010a) and (Lo and Wu, 2010b).

Our work shows that, although "deep" AI representation may remain for the moment out of practical reach, it is empirically possible to use PropBank style semantic frame representations to do significantly better than all currently widespread metrics, including even the relatively expensive HTER metric used in the DARPA GALE program (Snover *et al.*, 2006), which is arguably the most developed highly of the semi-automatic string based metrics, in terms of correlation to human judgments.

This then leads us to ask *how and why* the lexical semantics approach outperforms the string based approaches. Specifically, which aspects of predicate-argument structure are more important to preserve in translation, so as to render the translation *useful* and *adequate* in the human readers' eyes?

To be able to explain these questions, the design of MEANT is also unlike that of some recent machine learning based approaches, in which flat classifier-based metrics aggregate large numbers of features of very mixed types. For instance, Giménez and Màrquez (2007, 2008) describe a trained model that aggregates hundreds of assorted lexical, n-gram, ordering, syntactic, semantic, and other features. While such models have the potential to be mechanically trained to fairly high accuracy, such models do not easily lend themselves to use as investigative tools for error analysis, because (a) the features they automatically extract tend to be more simplistic than in the present work (for example, Gimnez and Mrquez check if semantic roles are present, but do not check if the fillers of semantic roles match), and (b) too many different levels of analysis are aggregated at once to permit independent examination of the impact of individual factors.

In contrast, MEANT emphasizes *representational transparency* and *simplicity*so that the translation and representation of the structural knowledge is explicit. Our approach facilitates error analysis and credit/blame assignment. Just a dozen or so weights are used to directly represent the importance of the predicate and each type of argument: *agent*, *experiencer*, *locative*, etc., as well as modality and negation. We then analyze the utility of each type of structural element for translation, by measuring and optimizing correlation against human judgments of adequacy. This allows us to *quantitatively* address in this paper questions such as "just how important is it to translate locatives correctly?" Our results across multiple data samples indicate that the weights are robust and stable.

Finally, we demonstrate empirically that our approach to MT evaluation is *low cost*, in terms of both (a) the amounts of time required by human annotators, and (b) the easily found, non-expert, untrained humans that can be used as annotators.

2 Prelude: Do PRED and ARG roles correlate well with human translation assessments?

The studies reported in this paper adopt a recent incarnation of shallow lexical semantic frame representations popularized by PropBank (Palmer *et al.*, 2005). PropBank is widely used because its corpus-driven methodology guarantees high coverage by prioritizing high-frequency types, making it well suited to machine learning approaches. Aside from labeling predicates (PRED), it also labels numerous argument types including agents (ARG0), experiencers (ARG1), patients (ARG2), temporals (ARGM-TMP), locatives, purposes, extents, manners, modals, negations, and so forth.

However, right at the outset, a natural concern is whether this PRED and ARG*j* representation really is a good match to the problem of evaluating the usefulness of SMT output. Although preserving semantic frames across translations may intuitively sound like an obviously good idea, the history of SMT is littered with surprising failures in attempts to incorporate "obvious" candidates for syntactic models and, more recently, semantic models, so caution is warranted.

We therefore constructed a battery of experiments to test the correlation of each individual PRED and ARG*j* type against human judgments of translation adequacy. This provides an independent, neutral indication of the amount of information that might potentially available toward assembling a translation evaluation metric.

Experimental setup We had a series of three data samples annotated. Each sample was randomly drawn from a translation evaluation corpus containing Chinese input sentences, English reference translations, and the machine translation outputs from three different state-of-the-art systems. Two of the samples were separately drawn from the subset of the DARPA GALE program Phase 2.5 newswire evaluation dataset in which both the Chinese and English sentences have been annotated with PropBank semantic role labels: GALE-A with 40 sentences, and GALE-B with 35 sentences. A third sample was drawn from the NIST MetricsMaTr meta-evaluation dataset (Callison-Burch *et al.*, 2010), with 39 sentences of the broadcast news genre. Section 5 gives additional details on the annotation process.

For each ARG*j* type, as well as PRED, we manually compared each English MT output against its reference translation. Using the counts thus obtained, we computed the precision, recall, and f-score for PRED and each ARG*j* type.

The MetricsMatr dataset was annotated by NIST with 7-level adequacy judgments. For the GALE-A and GALE-B datasets, we had the adequacy of all English MT outputs independently ranked by human judges.

We then computed sentence-level correlations following the benchmark assessment procedure used by WMT and NIST MetricsMaTr (Callison-Burch *et al.*, 2008, 2010), which use ranking preference consistency, also known as Kendall's rank correlation coefficient, to evaluate the correlation of each f-score against human rankings of the three systems' translation adequacy on each individual sentence. A higher value for indicates more similarity to the ranking by the evaluation metric to the human judgment. The range of

Table 1: Correlations of individual roles against human adequacy judgments on GALE Phase 2.5 and MetricsMaTr data samples (see text).

	PRED	ARG0	ARG1	ARG2	ARGM TMP	ARGM LOC/ DIR	ARGM PRP/ CAU	ARGM EXT	ARGM MNR	ARGM MOD	ARGM NEG	ARGM *
	action	*agent*	*experiencer*	*patient*	*temporal*	*locative*	*purpose*	*extent*	*manner*	*modal*	*negation*	*other*
GALE-A	0.19	0.25	0.22	0.0	0.06	0.06	0.0	0.01	0.03	-	-	0.0
GALE-B (I)	0.09	0.14	0.13	0.0	0.02	0.05	0.0	0.04	0.0	0.0	0.0	0.0
GALE-B (II)	0.10	0.14	0.14	0.0	0.03	0.05	0.0	0.04	0.0	0.0	0.0	0.0
MetricsMaTr	0.17	0.23	0.22	0.01	0.17	0.17	0.0	0.01	0.0	0.0	0.0	0.0

possible values of correlation coefficient is $[-1, 1]$, where 1 means the systems are ranked in the same order as the human judgment, and -1 means the systems are ranked in the reverse order as the human judgment.

Results The detailed correlations are shown in Table 1. There are two rows for GALE-B because we had the MT output translations independently ranked for adequacy by two separate human judges, for the sake of verifying that results were not excessively sensitive to the idiosyncracies of individual human adequacy judges: (I) gives correlations against judge 1, and (II) against judge 2. In addition, for both GALE sample 2 and the MetricsMaTr sample, two additional roles were assessed (modality and negation).

At first blush, these preliminary correlations would seem to be a mixed bag. Clearly, the action (PRED), agent (ARG0), and experiencer (ARG1) all individually show significant correlation to human judgments of translation adequacy. To a lesser extent, temporal (ARGM-TMP) and locative (ARGM-LOC, ARGM-DIR) roles do as well. All these roles show encouraging levels of correlation, and so should certainly be helpful in assembling a semantic MT evaluation metric.

The remainder, though, show no signficant correlation including even the patient (ARG2) and purposive roles! However, bear in mind that the correlation of each role is being checked individually here. What this fails to capture is the likely dependencies between *combinations* of roles. In other words, such roles should still be useful if conditioned on other roles also being correctly translated.

3 Semantic frames evaluate MT more accurately

We now show that, as suggested by the results of the previous section, it is indeed possible to construct a semantic MT evaluation metric combining the effects of multiple roles, that is superior to all other widely used metrics including BLEU, NIST, METEOR, and even HTER. To do this, we define an improved version of the HMEANT metric originally described in Lo and Wu (2011), which essentially summarizes the degree of match between shallow semantic parses of the human versus machine translations of sentences. Specifically, we compute an f-score that balances weighted aggregate precision and recall.

We also facilitate a finer-grained measurement of utility by allowing the human annotators to mark each role filler translation not only as either "correct" or "incorrect", but alternatively as "partially correct", which is then incorporated into the precision and recall via a weight w_{partial} (estimated as described later), as follows.

$$
\begin{aligned}
C_{\text{pred}} &\equiv \text{\# correctly translated predicates} \\
M_{\text{pred}} &\equiv \text{total \# predicates in MT} \\
R_{\text{pred}} &\equiv \text{total \# predicates in REF} \\
C_{i,j} &\equiv \text{\# correct ARG}j \text{ fillers of PRED i in MT} \\
P_{i,j} &\equiv \text{\# partially correct ARG}j \text{ fillers of PRED i in MT} \\
M_{i,j} &\equiv \text{total \# ARG}j \text{ fillers of PRED i in MT} \\
R_{i,j} &\equiv \text{total \# ARG}j \text{ fillers of PRED i in REF}
\end{aligned}
$$

$$
\text{precision} \equiv \frac{w_{\text{pred}} C_{\text{pred}} + \sum_j \sum_{\text{correct}_i} w_j (C_{i,j} + w_{\text{partial}} P_{i,j})}{w_{\text{pred}} M_{\text{pred}} + \sum_j \sum_{\text{correct}_i} w_j M_{i,j}}
$$

$$
\text{recall} \equiv \frac{w_{\text{pred}} C_{\text{pred}} + \sum_j \sum_{\text{correct}_i} w_j (C_{i,j} + w_{\text{partial}} P_{i,j})}{w_{\text{pred}} R_{\text{pred}} + \sum_j \sum_{\text{correct}_i} w_j R_{i,j}}
$$

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

Note that "correct i" in the above definitions is short for "correctly translated predicate i".

The parameters w_{pred} and w_{j} determine the extent to which matching the predicate and matching arguments of type j contribute to the overall score. These weights can be viewed as the importance of the predicate and the different categories of semantic roles. The weight w_{partial} controls the degree to which "partially correct" translations are penalized. These weights can be automatically estimated, as discussed below.

If all the reconstructed semantic frames in the MT output are completely identical to the gold standard annotation in the reference translation, and all the arguments in the reconstructed frames are judged to express the same meaning as the corresponding arguments in the reference translations, then the f-score in the definition of HMEANT will be equal to 1.

Consider the example in Figure 1. There is one predicate in MT1, resumed. There are two predicates in REF, ceased and resume. Therefore, resume is considered a correctly translated predicate. In MT1, there is one agent (ARG0), one experiencer (ARG1), and one temporal (ARGM-TMP) associated with the frame of resumed; and in REF, there is one experiencer (ARG1) and two temporals (ARGM-TMP) associated with the corresponding frame of resume. The role filler

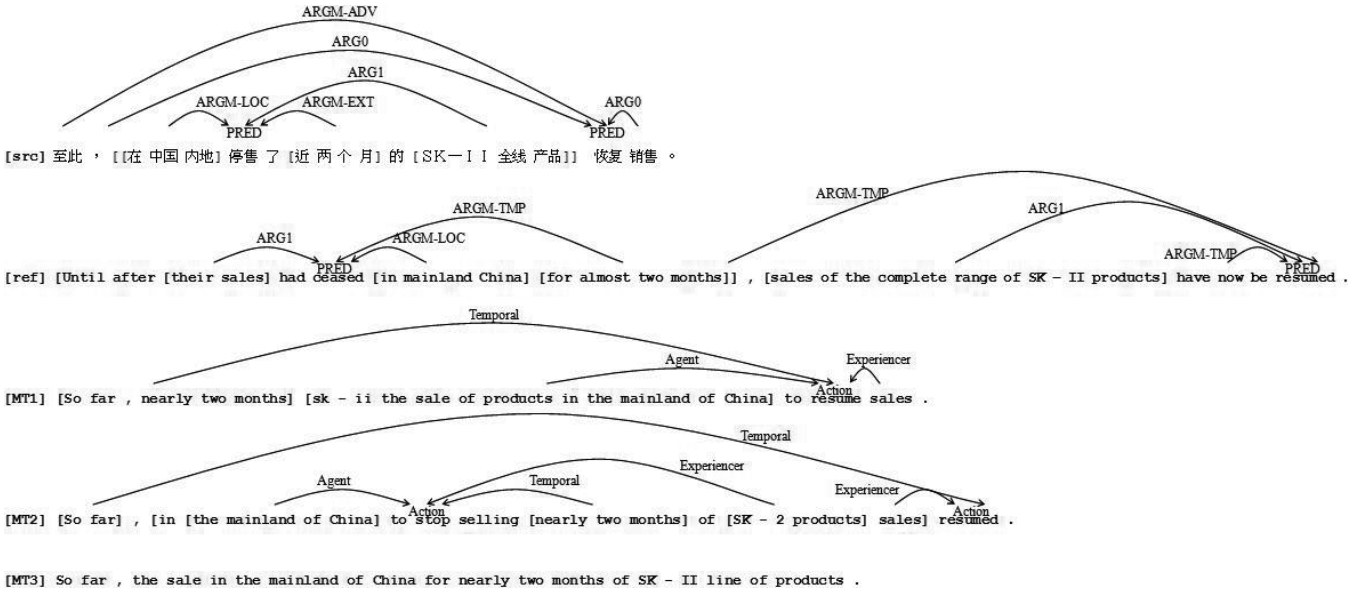

Figure 1: Example of source sentence and reference translation with reconstructed semantic frames in Propbank format and MT output with reconstructed semantic frames by minimal trained human annotators. Following PropBank, there are no semantic frames for MT3 because there is no predicate.

of experiencer (ARG1) in MT1, sales, is a partially correct translation of the role filler of experiencer (ARG1) in REF, sales of complete range of SK-II products. Similarly, the role filler of the temporal (ARGM-TMP) in MT1, So far, nearly two months, is a partially correct translation of the role filler of the corresponding temporal (ARGM-TMP) in REF, Until after, their sales had ceased in mainland China for almost two months. Assuming $w_{\text{partial}} = 0.5$, the HMEANT metric collects 10 types of semantic roles, with uniform weight for each role (optimization of weights will be discussed later), then $w_{\text{pred}} = w_j = 0.1$ giving a precision and recall of 0.5 and 0.4 respectively. Therefore, HMEANT's f-score for this example is 0.44.

Experimental setup The same GALE Phase 2.5 samples as described in Section 2 were employed. The weights w_{pred}, w_{j}, and w_{partial} were estimated by optimizing correlation against human adequacy judgments, using any of the many standard optimization search techniques (we found grid search to be quite adequate). To ensure no overfitting to the data, 4-fold cross-validation was performed. In practice, the optimal w_{partial} is always essentially 0.5, as would be expected.

For the baseline version of HMEANT, the f-score combines weighted precisions and recalls of the action (PRED) and its agent (ARG0), experiencer (ARG1), patient (ARG2), temporal (ARGM-TMP), locative (ARGM-LOC and ARGM-DIR), purpose (ARGM-PRP and ARGM-CAU), extent (ARGM-EXT), and manner (ARGM-MNR).

Since it has been often argued that a critical weakness of SMT systems is that they often fail to translate modality and negation correctly, for the GALE-B (I) and (II) samples we tested three additional variants that also include modality (ARGM-MOD), negation (ARGM-NEG), or both.

Table 2: Kendall's coefficients for different MT evaluation metrics correlated against human judgment of translation adequacy (see text for description of samples). HMEANT is our metric with 10 semantic role features; +modal is HMEANT with the "modal" label in addition; +negation is HMEANT with the "negation" label in addition; +mod+neg includes both.

Metric	GALE-A	GALE-B (I)	GALE-B (II)
HTER	0.43	0.16	0.20
METEOR	0.20	0.19	0.21
NIST	0.29	0.14	0.09
PER	0.20	-0.03	0.07
TER	0.20	0.08	0.10
WER	0.10	0.06	0.11
BLEU	0.20	0.14	0.12
CDER	0.12	0.04	0.10
HMEANT	**0.47**	0.26	**0.28**
+modal	-	**0.27**	**0.28**
+negation	-	**0.27**	**0.28**
+mod+neg	-	**0.27**	**0.28**

Results As shown in Table 2, our semantic frame based HMEANT metric shows significantly higher correlation than all other metrics in common use, outperforming even the much more expensive HTER metric used in the GALE evaluations.

Perhaps surprisingly given the frequent complaints about SMT handling of modality and negation, the results indicate that neither modality (ARGM-MOD) nor negation (ARGM-

Table 3: Weights for semantic roles as learned from combined optimization, extracted from models optimized on (a) GALE-A, (b) GALE-B (I), (c) GALE-B (II), and (d) NIST MetricsMaTr.

	PRED	ARG0	ARG1	ARG2	ARGM TMP	ARGM LOC/ DIR	ARGM PRP/ CAU	ARGM EXT	ARGM MNR	ARGM MOD	ARGM NEG	ARGM *
	action	*agent*	*experiencer*	*patient*	*temporal*	*locative*	*purpose*	*extent*	*manner*	*modal*	*negation*	*other*
HMEANT (a)	0.09	0.36	0.09	0.01	0.01	0.09	0.02	0.03	0.03	-	-	0.27
HMEANT (b)	0.09	0.36	0.18	0.01	0.03	0.09	0.02	0.02	0.02	-	-	0.18
+modal (b)	0.09	0.36	0.18	0.01	0.03	0.09	0.02	0.02	0.09	0.02	-	0.09
+negation (b)	0.09	0.36	0.18	0.01	0.03	0.09	0.01	0.02	0.02	-	0.01	0.18
+mod+neg (b)	0.09	0.36	0.18	0.01	0.03	0.09	0.01	0.02	0.09	0.03	0.01	0.09
HMEANT (c)	0.09	0.36	0.18	0.01	0.03	0.09	0.02	0.02	0.02	-	-	0.18
+modal (c)	0.09	0.36	0.18	0.01	0.03	0.09	0.02	0.02	0.09	0.02	-	0.09
+negation (c)	0.09	0.36	0.18	0.01	0.03	0.09	0.01	0.02	0.02	-	0.01	0.18
+mod+neg (c)	0.09	0.36	0.18	0.01	0.03	0.09	0.01	0.02	0.09	0.03	0.01	0.09
HMEANT (d)	0.09	0.45	0.27	0.01	0.03	0.09	0.01	0.02	0.03	-	-	0.09
+modal (d)	0.09	0.36	0.27	0.01	0.03	0.09	0.01	0.02	0.09	0.02	-	0.09
+negation (d)	0.09	0.45	0.27	0.01	0.03	0.09	0.01	0.01	0.03	-	0.01	0.09
+mod+neg (d)	0.09	0.36	0.27	0.01	0.03	0.09	0.01	0.01	0.09	0.03	0.01	0.09

NEG) appear to improve correlation with human adequacy judgments. Closer error analysis explains why. In the case of modality, which occurs in 46% of the sentences, over half were translated correctly. In the case of negation, which occurs in 17% of the sentences, nearly all were translated correctly.

The large difference between the magnitudes of the correlations for GALE-A versus GALE-B are due to significant differences in the character of the sentences and translations. Specifically, for a number of sentences in GALE-B, all three MT systems produced almost identical translations. This makes both the humans' and metrics' sentence translation rankings highly arbitrary, and the inconsistency depresses the correlation scores.

4 Which roles need to be translated correctly?

The representational transparency of the HMEANT metric allows us now to come back to the earlier question raised in Section 2: on average, which semantic roles need to be translated correctly, when in the context of a full sentence translation that is likely to have many dependencies between semantic roles?

The optimized weights w_{pred} and w_{j} provide the analytical tool to get quantitative answers. Inspecting these weights, shown in Table 3, indicates that when used in combination:

- The agent and experiencer (ARG0, and ARG1) remain the most important roles (*who, what*).
- The action (PRED) and locative (ARGM-LOC and ARGM-DIR) also remain important (*did, where*).
- Temporals (ARGM-TMP) are weighted less highly than Section 2 might suggest (*when*).
- The patient (ARG2), suprisingly, remains relatively unimportant, though of more utility than indicated by the zero correlation found in Section 2 (*to whom*).
- Similarly, the purpose (ARGM-PRP and ARGM-CAU), extent (ARGM-EXT), and manner (ARGM-MNR) remain relatively unimportant but still of nonzero utility (*why, how*).
- The weights on modality (ARGM-MOD) and negation (ARGM-NEG) remain low.

Note that the weights are quite robust and stable. Regardless of which data sample the weights are optimized upon (notated as (a), (b) (c), and (d) in Table 3), almost the same weights are chosen (which thus produce nearly identical f-scores). The biggest variation occurs in the relative weight given to ARG1 in the GALE newswire versus MetricsMaTr broadcast news datasets, which may be due to the spoken language genre.

On one hand, from an AI standpoint of semantic representations, it might seem disturbing that so many of the standard semantic role types play such a proportionally small part in correlating against human adequacy judgments. However, they do nevertheless remain of nonzero utility, and they contribute in combination toward surpassing all the existing surface oriented MT evaluation metrics, as well as the expensive HTER metric used in the DARPA GALE program.

On the other hand, from an SMT standpoint of evaluation metrics, this is the first time that a pure semantic role based assessment paradigm has been empirically shown to be quantitatively superior to all other surface oriented metrics, as well as HTER. Moreover, as this is early in the development of semantic frame oriented MT evaluation metrics, we believe there is a great deal of headroom for future improvement.

Table 4: Timing statistics for human semantic role annotation ("annot") and role filler comparison ("comp") tasks, for both the MetricsMaTr and GALE-B samples. t/s, t/f, t/r, and t/w indicate time per sentence, frame, role, word, respectively.

	#frames	#roles	#words	min t/s	max t/s	avg t/s	SD t/s	avg t/f	avg t/r	avg t/w
MetricsMaTr REF annot	1.85	6.86	12.69	15.00	485.00	127.12	106.62	68.59	18.53	5.01
MetricsMaTr MT1 annot	1.57	5.65	11.10	10.00	261.00	76.06	55.00	43.69	12.12	3.08
MetricsMaTr MT2 annot	1.47	5.75	10.97	2.00	428.00	73.79	59.09	50.31	12.83	3.36
MetricsMaTr MT3 annot	1.13	4.17	9.71	4.00	353.00	77.97	63.92	69.21	18.68	4.02
MetricsMaTr MT1 comp	——	——	——	8.00	150.00	32.50	33.14	5.53	1.53	0.39
MetricsMaTr MT2 comp	——	——	——	8.00	183.00	30.38	31.57	5.52	1.41	0.37
MetricsMaTr MT3 comp	——	——	——	5.00	98.00	17.38	19.30	4.11	1.11	0.24
GALE-B REF annot	2.79	11.07	21.44	18.00	416.00	131.30	83.99	47.13	11.71	3.06
GALE-B MT1 annot	2.76	9.89	18.14	31.00	325.00	105.80	58.60	38.37	10.70	2.92
GALE-B MT2 annot	2.56	8.93	10.97	4.00	329.00	85.09	60.12	33.27	9.53	2.31
GALE-B MT3 annot	2.16	3.55	17.49	14.00	376.00	97.77	61.96	45.32	12.86	2.80
GALE-B MT1 comp	——	——	——	76.00	401.00	165.50	74.63	41.38	13.35	5.71
GALE-B MT2 comp	——	——	——	62.00	174.00	131.10	37.41	37.44	10.8	4.36
GALE-B MT3 comp	——	——	——	59.00	165.00	127.40	32.68	46.00	15.16	4.60

5 Semantic MT evaluation is low cost

The other primary goal of this work, aside from representational transparency and simplicity, is low cost of evaluation. To assess this, we conducted new in-depth experiments measuring the time required by human judges to perform either the semantic frame annotation and comparison task, on two different data sets. The collected timing data is detailed in Table 4 in terms of sentences, frames, roles, as well as words.

The results bear out the low cost of our approach, in spite of the fact that annotation was performed solely by inexpensive, non-expert, untrained computer science undergraduate students. The annotation protocol was on average carried out in about 11.5 minutes per sentence, depending on the complexity of the sentencesmuch less time than required for either HTER and gold standard Propbank annotation. Annotators were given only a half page of instructions with a lay person's guide based on *who did what to whom, when, where, why, and how?* as shown in Table 8. The instructions required only at most 5 to 15 minutes of preparation, including any necessary time for asking questions. Moreover, in a separate study, monolingual English speakers were shown to perform essentially just as well as more expensive bilinguals (Lo and Wu, 2011); that study also confirmed high inter-annotator agreement for both monolinguals and bilinguals, whether or not they were allowed to see the Chinese input sentences. No additional training was given, aside from providing two annotated examples.

The time used for comparing the role fillers between the semantic frames in the reference and machine translations, similarly, averaged under 2 minutes per sentence.

Thus, the total time per sentence, around 3.5 minutes per sentence combining both annotation and comparison phases, is only half of that required to evaluate HTER for the same types of text.

Furthermore, note that these timing figures are for completely unskilled non-experts. In fact, the time required tends to decrease even further as annotators gain experience.

6 Untrained humans still label semantic roles consistently

One of the concerns in employing untrained humans for semantic frame reconstruction is whether inconsistencies in their manual efforts might 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, making evaluation results more reliable and reproducible.

6.1 Experimental setup

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

Role identification Since the handling of articles or punctuation at the beginning or the end of annotated arguments is somewhat arbitrary, agreement for semantic role identification is assessed on the matching of word spans 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 assessed on the matching of the semantic role labels within two

Table 5: Inter-annotator agreement rate on role identification (matching of word span)

Experiments	Ref	MT
GALE-A	93%	75%
GALE-B	75%	73%
MetricsMaTr	76%	70%

Table 6: Inter-annotator agreement rate on role classification (matching of role labels associated with matched word span)

Experiments	Ref	MT
GALE-A	88%	70%
GALE-B	70%	69%
MetricsMaTr	70%	67%

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.

$$P_{\text{classification}} = \frac{M_{\text{label}}}{A_1}$$

$$R_{\text{classification}} = \frac{M_{\text{label}}}{A_2}$$

$$\text{IAA}_{\text{classification}} = \frac{2 * P_{\text{classification}} * R_{\text{classification}}}{P_{\text{classification}} + R_{\text{classification}}}$$

6.2 Results

The high inter-annotator agreement results across multiple datasets, as shown in Tables 5 and 6, suggest that the lay person's guide we designed succeeds in simplifying the semantic role annotation task to be very intuitive. The minimal training instructions provided to the annotators are in general sufficient and the evaluation is repeatable and could be automated in the future. The annotators reconstructed the semantic frames quite consistently, despite being given only simple and minimal training.

We have observed that the agreement on role identification is higher than that on role classification, suggesting that role confusion errors among the annotators. Thus, giving slightly more detailed instructions and explanations of the different roles could further improve the IAA on role classification.

7 Automating the SRL still outperforms automatic MT metrics

Having established how to regain representational transparency in human MT evaluation metrics, we now turn to the next obvious question: is it possible to take advantage of modern shallow semantic parsing methods to further automate our semantic MT metrics?

7.1 Experimental setup

We applied automatic SRL to both the reference and machine translations, in order to assess the extent to which performance would be degraded compared to HMEANT, where human SRL produced the semantic frame reconstructions. We used ASSERT (Pradhan *et al.*, 2004) for the automatic shallow semantic parsing.

Table 7: Kendall's coefficients for different MT evaluation metrics correlated against human judgment of translation adequacy (see text for description of samples). HMEANT is our human SRL based metric with 10 semantic role features; +modal is HMEANT with the "modal" label in addition; +negation is HMEANT with the "negation" label in addition; +mod+neg includes both. MEANT is our automatic SRL based metric.

Metric	GALE-A	GALE-B (I)	GALE-B (II)
HTER	0.43	0.16	0.20
METEOR	0.20	0.19	0.21
NIST	0.29	0.14	0.09
PER	0.20	-0.03	0.07
TER	0.20	0.08	0.10
WER	0.10	0.06	0.11
BLEU	0.20	0.14	0.12
CDER	0.12	0.04	0.10
HMEANT	**0.47**	0.26	**0.28**
+modal	-	**0.27**	**0.28**
+negation	-	**0.27**	**0.28**
+mod+neg	-	**0.27**	**0.28**
MEANT	*0.33*	*0.15*	*0.18*

7.2 Results

Table 7 compares results for MEANT against the earlier HMEANT results from Table 2. Using automatic SRL for different datasets, MEANT still achieves in the range of 76–93% of the correlation levels of HTER, in spite of its much lower labor cost. Moreover, MEANT greatly outperforms all common automated MT metrics, including BLEU, NIST, METEOR, WER, PER, CDER, and TER.

8 Conclusion

We have presented an alternative approach to evaluating MT utility that combines the advantages of both worlds: (a) from a pattern recognition perspective, the HMEANT metrics provide *simple* quantitative *low-cost* objective functions for translation quality that surpass the correlations against human judgments achieved by all currently widespread surface-oriented automatic and semi-automatic metrics, including even the expensive HTER metric; and yet (b) from an AI perspective, we regain the *representational transparency* and clear reflection of semantic utility of structural frame-based knowledge representations.

We have also shown that taking advantage of modern shallow semantic parsing methods allows MEANT to further automate semantic MT evaluation, while retaining 76–93% of the correlation levels achieved by the heavily labor-intensive HTER, and still significantly outperforming currently used automated surface-oriented metrics.

The only procedure that remains non-automatic is comparing the accuracy of role filler. We believe this step will also be automatable. Another alternative complementary approach

Table 8: Lay person's simplification of semantic roles, as given to untrained human annotators.

Label	Role	Label	Role
agent	who	*locative*	where
action	did	*purpose*	why
experiencer	what	*manner*	how
patient	whom	*degree or extent*	how
temporal	when	*other*	how

would also be to incorporate the human-in-the-loop strategy of (Zaidan and Callison-Burch, 2009).

9 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

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The Berkeley FrameNet project. In *COLING-ACL'98*, Montreal,Canada, August 1998.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgement. In *Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics (ACL-2005)*, Ann Arbor, Michigan, June 2005.

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 *Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53, Uppsala, Sweden, Jul 2010.

Marine Carpuat and Dekai Wu. Improving statistical machine translation using word sense disambiguation. In *2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, pages 61–72, Prague, Jun 2007.

Marine Carpuat and Dekai Wu. Evaluation of context-dependent phrasal translation lexicons for statistical machine translation. In *Sixth International Conference on Language Resources and Evaluation (LREC-2008)*, Marrakech, May 2008. European Language Resources Association (ELRA).

Yee Seng Chan, Hwee Tou Ng, and David Chiang. Word sense disambiguation improves statistical machine translation. In *45th Annual Meeting of the Association for Computational Linguistics (ACL-07)*, Prague, Jun 2007.

George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Human Language Technology conference (HLT-2002)*, San Diego, CA, 2002.

Charles J. Fillmore. The case for case. In Bach and Harms, editors, *Universals in Linguistic Theory*, pages 1–88. Holt, Rinehart, and Winston, 1968.

Pascale Fung, Zhaojun Wu, Yongsheng Yang, and Dekai Wu. Automatic learning of Chinese-English semantic structure mapping. In *IEEE/ACL 2006 Workshop on Spoken Language Technology (SLT 2006)*, Aruba, Dec 2006.

Pascale Fung, Zhaojun Wu, Yongsheng Yang, and Dekai Wu. Learning bilingual semantic frames: Shallow semantic parsing vs. semantic role projection. In *11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI-07)*, pages 75–84, Skövde, Sweden, Sep 2007.

Jesús Giménez and Lluís Màrquez. Context-aware discriminative phrase selection for statistical machine translations. In *Workshop on Statistical Machine Translation*, Prague, Jun 2007.

Jesús Giménez and Lluís Màrquez. A smorgasbord of features for automatic MT evaluation. In *3rd ACL Workshop on Statistical Machine Translation*, pages 195–198, Columbus, OH, Jun 2008.

Gregor Leusch, Nicola Ueffing, and Hermann Ney. Cder: Efficient mt evaluation using block movements. In *EACL-2006 (11th Conference of the European Chapter of the Association for Computational Linguistics)*, pages 241–248, Trento, Italy, April 2006.

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. MEANT: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility via semantic frames. In *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*, Portland, Oregon, Jun 2011.

Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney. An evaluation tool for machine translation: Fast evaluation for MT research. In *Second International Conference on Language Resources and Evaluation (LREC-2000)*, 2000.

Martha Palmer, Paul Kingsbury, and Daniel Gildea. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–105, Mar 2005.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translations. In *40th Annual Meeting of the Association for Computational Linguistics (ACL-2002)*, pages 311–318, Philadelphia, Jul 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 Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *AMTA*, pages 223–231, Boston, MA, 2006. Association for Machine Translation in the Americas.

Christoph Tillmann, Stefan Vogel, Hermann Ney, A. Zubiaga, and H. Sawaf. Accelerated DP-based search for statistical translation. In *Fifth European Conference on Speech Communication and Technology (Eurospeech'97)*, pages 2667–2670, Rhodes, Greece, 1997.

Dekai Wu and Pascale Fung. Can semantic role labeling improve SMT? In *13th Annual Conference of the European Association for Machine Translation (EAMT 2009)*, pages 218–225, Barcelona, May 2009.

Dekai Wu and Pascale Fung. Semantic roles for SMT: A hybrid two-pass model. In *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2009)*, Boulder, CO, Jun 2009.

Nianwen Xue and Martha Palmer. Automatic semantic role labeling for chinese verbs. In *19th International Joint Conference on Artificial Intelligence (IJCAI-05)*, Edinburgh, 2005.

Omar F. Zaidan and Chris Callison-Burch. Feasibility of human-in-the-loop minimum error rate training. In *2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, Singapore, Aug 2009.