Structured vs. Flat Semantic Role Representations for Machine Translation Evaluation

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

Abstract

We argue that failing to capture the degree of contribution of each semantic frame in a sentence explains puzzling results in recent work on the MEANT family of semantic MT evaluation metrics, which have disturbingly indicated that dissociating semantic roles and fillers from their predicates actually improves correlation with human adequacy judgments even though, intuitively, properly segregating event frames should more accurately reflect the preservation of meaning. Our analysis finds that both properly structured and flattened representations fail to adequately account for the contribution of each semantic frame to the overall sentence. We then show that the correlation of HMEANT, the human variant of MEANT, can be greatly improved by introducing a simple length-based weighting scheme that approximates the degree of contribution of each semantic frame to the overall sentence. The new results also show that, without flattening the structure of semantic frames, weighting the degree of each frame's contribution gives HMEANT higher correlations than the previously best-performing flattened model, as well as HTER.

1 Introduction

In this paper we provide a more concrete answer to the question: what would be a better representation, structured or flat, of the roles in semantic frames to be used in a semantic machine translation (MT) evaluation metric? We compare recent studies on the MEANT family of semantic role labeling (SRL) based MT evaluation metrics (Lo and Wu, 2010a,b, 2011a,b) by (1) contrasting their variations in semantic role representation and observing disturbing comparative results indicating that segregating the event frames in structured role representation actually *damages* correlation against human adequacy judgments and (2) showing how SRL based MT evaluation can be improved beyond the current state-of-the-art compared to previous MEANT variants as well as HTER, through the introduction of a simple weighting scheme that reflects the degree of contribution of each semantic frame to the overall sentence. The weighting scheme we propose uses a simple length-based heuristic that reflects the assumption that a semantic frame that covers more tokens contributes more to the overall sentence translation. We demonstrate empirically that when the degree of each frame's contribution to its sentence is taken into account, the properly structured role representation is more accurate and intuitive than the flattened role representation for SRL MT evaluation metrics.

For years, the task of measuring the performance of MT systems has been dominated by lexical n-gram based machine translation evaluation 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) and WER (Nießen *et al.*, 2000). These metrics are excellent at ranking overall systems by averaging their scores over entire documents. However, as MT systems improve, the shortcomings of such metrics are becoming more apparent. Though containing roughly the correct words, MT output at the sentence remains often quite incomprehensible, and fails to preserve the meaning of the input. This results from the fact that n-gram based metrics are not as reliable at ranking the adequacy of translations of individual sentences, and are particularly

Proceedings of SSST-5, Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 10–20,
ACL HLT 2011, Portland, Oregon, USA, June 2011.

poor at reflecting translation quality improvements involving more meaningful word sense or semantic frame decisions—which human judges have no trouble distinguishing. Callison-Burch *et al.* (2006) and Koehn and Monz (2006), for example, study situations where BLEU strongly disagrees with human judgment of translation quality.

Newer avenues of research seek substitutes for n-gram based MT evaluation metrics that are better at evaluating translation adequacy, particularly at the sentence level. One line of research emphasizes more the structural correctness of translation. Liu and Gildea (2005) propose STM, a metric based on syntactic structure, that addresses the failure of lexical similarity based metrics to evaluate translation grammaticality. However, the problem remains that a grammatical translation can achieve a high syntax-based score yet still make significant errors arising from confusion of semantic roles. On the other hand, despite the fact that non-automatic, manually evaluated metrics, such as HTER (Snover *et al.*, 2006), are more adequacy oriented exhibit much higher correlation with human adequacy judgment, their high labor cost prohibits widespread use. There has also been work on explicitly evaluating MT adequacy by aggregating over a very large set of linguistic features (Giménez and Màrquez, 2007, 2008) and textual entailment (Pado *et al.*, 2009).

2 SRL based MT evaluation metrics

A blueprint for more direct assessment of meaning preservation across translation was outlined by Lo and Wu (2010a), in which translation utility is manually evaluated with respect to the accuracy of semantic role labels. 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. Adopting this principle, the MEANT family of metrics compare the semantic frames in reference translations against those that can be reconstructed from machine translation output.

Preliminary results reported in (Lo and Wu, 2010b) confirm that the blueprint model outperforms BLEU and similar n-gram oriented evaluation metrics in correlation against human adequacy judgments, but does not fare as well as HTER. The more complete study of Lo and Wu (2011a) introduces MEANT and its human variants HMEANT, which implement an extended version of blueprint methodology. Experimental results show that HMEANT correlates against human adequacy judgments as well as the more expensive HTER, even though HMEANT can be evaluated using low-cost untrained monolingual semantic role annotators while still maintaining high inter-annotator agreement (both are far superior to BLEU or other surface oriented evaluation metrics). The study also shows that replacing the human semantic role labelers with an automatic shallow semantic parser yields an approximation that is still vastly superior to BLEU while remaining about 80% as closely correlated with human adequacy judgments as HTER. Along with additional improvements to the accuracy of the MEANT family of metrics, Lo and Wu (2011b) study the impact of each individual semantic role to the metric's correlation against human adequacy judgments, as well as the time cost for humans to reconstruct the semantic frames and compare the translation accuracy of the role fillers.

In general, the MEANT family of SRL MT evaluation metrics (Lo and Wu, 2011a,b) evaluate the translation utility as follows. First, semantic role labeling is performed (either manually or automatically) on both the reference translation (**REF**) and the machine translation output (**MT**) to obtain the semantic frame structure. Then, the semantic predicates, roles and fillers reconstructed from the MT output are compared to those in the reference translations. The number of correctly and partially correctly annotated arguments of each type in each frame of the MT output are collected in this step:

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

In the following three subsections, we describe how the translation utility is calculated using these counts in (a) the original blueprint model, (b) the first version of HMEANT and MEANT using structured role representations, and (c) the more accu-

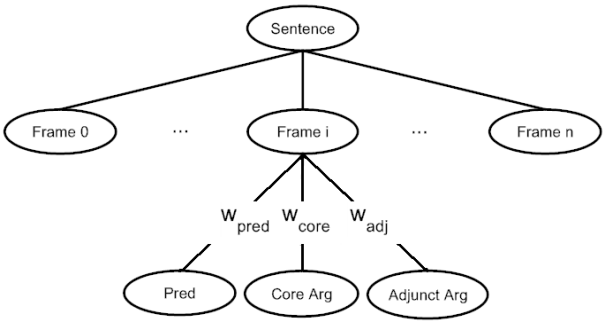

Figure 1: The structured role representation for the blueprint SRL-based MT evaluation metric as proposed in Lo and Wu (2010a,b), with arguments aggregated into core and adjunct classes.

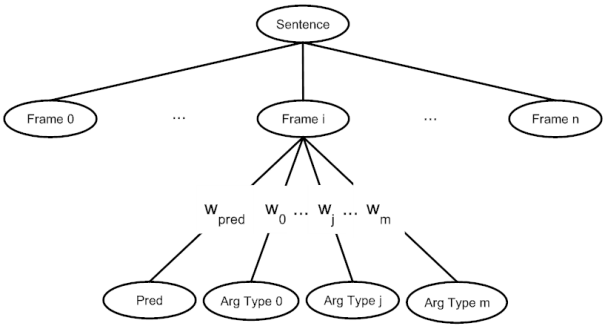

Figure 2: The structured role representation for the MEANT family of metrics as proposed in Lo and Wu (2011a).

rate flattened-role implementation of HMEANT and MEANT.

2.1 Structured core vs. adjunct role representation

Figure 1 depicts the semantic role representation in the blueprint model of SRL MT evaluation metric proposed by Lo and Wu (2010a,b). Each sentence consists of a number of frames, and each frame consists of a predicate and two classes of arguments, either core or adjunct. The frame precision/recall is the weighted sum of the number of correctly translated roles (where arguments are grouped into the core and adjunct classes) in a frame normalized by the weighted sum of the total number of all roles in that frame in the MT/REF respectively. The sentence precision/recall is the sum of the frame precision/recall for all frames averaged by the total number of frames in the MT/REF respectively. The SRL evaluation metric is then defined in terms of f-score in order to balance the sentence precision and recall. More precisely, assuming the above definitions of $C_{i,j}$, $P_{i,j}$, $M_{i,j}$ and $R_{i,j}$, the sentence precision and recall are defined as follows.

$$\text{precision} = \frac{\sum_i \frac{w_{\text{pred}} + \sum_t w_t \left(\sum_{j \in t} (C_{i,j} + w_{\text{partial}} P_{i,j}) \right)}{w_{\text{pred}} + \sum_t w_t \left(\sum_{j \in t} M_{i,j} \right)}}{\text{\# frames in MT}}$$

$$\text{recall} = \frac{\sum_i \frac{w_{\text{pred}} + \sum_t w_t \left(\sum_{j \in t} (C_{i,j} + w_{\text{partial}} P_{i,j}) \right)}{w_{\text{pred}} + \sum_t w_t \left(\sum_{j \in t} R_{i,j} \right)}}{\text{\# frames in REF}}$$

where w_{pred} is the weight for predicates, and w_{t} where $t \in \{\text{core}, \text{adj}\}$ is the weight for core arguments and adjunct arguments. These weights represent the degree of contribution of the predicate and different classes of arguments (either core or adjunct) to the overall meaning of the semantic frame they attach to. In addition, w_{partial} is a weight controlling the degree to which "partially correct" translations are penalized. All the weights can be automatically estimated by optimizing the correlation with human adequacy judgments.

We conjecture that the reason for the low correlation with human adequacy judgments of this model as reported in Lo and Wu (2010b) is that the abstraction of arguments actually reduces the representational power of the original predicate-argument structure in SRL. Under this representation, all the arguments in the same class, e.g. all adjunct arguments, are weighted uniformly. The assumption that all types of arguments in the same class have the same degree of contribution to their frame is obviously wrong, and the empirical results confirm that the assumption is too coarse.

2.2 Structured role representation

Figure 2 shows the structured role representation used in the MEANT family of metrics as proposed in Lo and Wu (2011a), which avoids aggregating arguments into core and adjunct classes. The design of the MEANT family of metrics addresses the incorrect assumption in the blueprint model by assuming each type of argument has a unique weight representing its degree of contribution to the overall sentence translation. Thus, the number of dimensions of

the weight vector is increased to allow an independent weight to be assigned to each type of argument. Unlike the previous representation in the blueprint model, there is no aggregation of arguments into core and adjunct classes. Each sentence consists of a number of frames, and each frame consists of a predicate and a number of arguments of type j.

Under the new approach, the frame precision/recall is the weighted sum of the number of correctly translated roles in a frame normalized by the weighted sum of the total number of all roles in that frame in the MT/REF respectively. Similar to the previous blueprint representation, the sentence precision/recall is the sum of the frame precision/recall for all frames averaged by the total number of frames in the MT/REF respectively. More precisely, following the previous definitions of $C_{i,j}$, $P_{i,j}$, $M_{i,j}$, $R_{i,j}$, w_{pred} and w_{partial}, the sentence precision and recall are redefined as follows.

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

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

where w_{j} is the weight for the arguments of type j. These weights represent the degree of contribution of different types of arguments to the overall meaning of their semantic frame.

2.3 Flat role representation

Figure 3 depicts the flat role representation used in the more accurate variants of MEANT as proposed in Lo and Wu (2011b). This representation is motivated by the studies of the impact of individual semantic role. The highly significant difference between this flat representation and both of the previous two structured role representations is that the semantic frames in the sentence are no longer segregated.

The flat role representation desegregates the frame structure, resulting in a flat, single level structure. Therefore, there is no frame precision/recall. The sentence precision/recall is the weighted sum of the number of correctly translated roles in all frames normalized by the weighted sum of the total number of roles in all frames in the MT/REF respectively. More precisely, again assuming the previous definitions of $C_{i,j}$, $P_{i,j}$, $M_{i,j}$, $R_{i,j}$ and w_{partial}, the sentence precision and recall are redefined as follows.

$$\begin{aligned} C_{\text{pred}} &\equiv \text{total \# correctly translated predicates} \\ M_{\text{pred}} &\equiv \text{total \# predicates in MT} \\ R_{\text{pred}} &\equiv \text{total \# predicates in REF} \end{aligned}$$

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

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

Note that there is a small modification of the definition of w_{pred} and w_j. Instead of the degree of contribution to the overall meaning of the semantic frame that the roles attached to, $w_{\text{pred}}$and w_j now represent the degree of contribution of the predicate and the arguments of type j to the overall meaning of the *entire* sentence.

It is worth noting that the semantic role features in the ULC metric proposed by Giménez and Màrquez (2008) also employ a flat feature-based representation of semantic roles. However, the definition of those semantic role features adopts a different methodology for determining the role fillers' translation accuracy, which prevents a controlled consistent environment for the comparative experiments that the present work focuses on.

3 Experimental setup

The evaluation data for our experiments consists of 40 sentences randomly drawn from the DARPA GALE program Phase 2.5 newswire evaluation corpus containing Chinese input sentence, English reference translations, and the machine translation from three different state-of-the-art GALE systems. The Chinese and the English reference translation have both been annotated with gold standard PropBank (Palmer *et al.*, 2005) semantic role labels. The weights w_{pred}, w_{core}, w_{adj}, w_j and w_{partial} can be estimated by optimizing correlation against human adequacy judgments, using any of the many standard optimization search techniques. In the work of Lo and

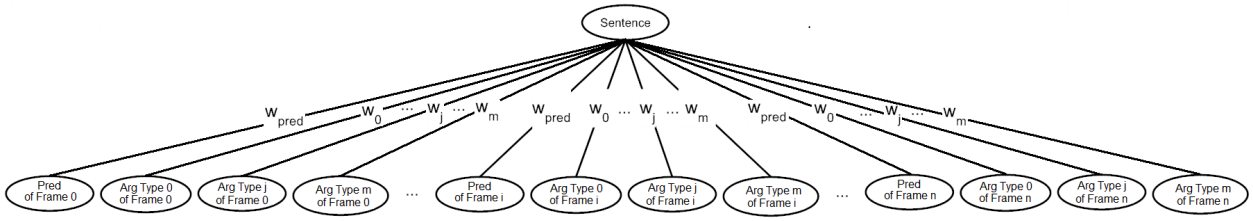

Figure 3: The flat role representation for the MEANT family of metrics as proposed in Lo and Wu (2011b) .

Wu (2011b), the correlations of all individual roles with the human adequacy judgments were found to be non-negative, therefore we found grid search to be quite adequate for estimating the weights. We use linear weighting because we would like to keep the metric's interpretation simple and intuitive.

Following the benchmark assessment in NIST MetricsMaTr 2010 (Callison-Burch *et al.*, 2010), we assess the performance of the semantic MT evaluation metric at the sentence level using the summed-diagonal-of-confusion-matrix score. The human adequacy judgments were obtained by showing all three MT outputs together with the Chinese source input to a human reader. The human reader was instructed to order the sentences from the three MT systems according to the accuracy of meaning in the translations. For the MT output, we ranked the sentences from the three MT systems according to their evaluation metric scores. By comparing the two sets of rankings, a confusion matrix is formed. The summed diagonal of confusion matrix is the percentage of the total count when a particular rank by the metric's score exactly matches the human judgments. The range of possible values of summed diagonal of confusion matrix is [0,1], where 1 means all the systems' ranks determined by the metric are identical with that of the human judgments and 0 means all the systems' ranks determined by the metric are different from that of the human judgment.

Since the summed diagonal of confusion matrix scores only assess the absolute ranking accuracy, we also report the Kendall's τ rank correlation coefficients, which measure the correlation of the proposed metric against human judgments with respect to their relative ranking of translation adequacy. A higher the value for τ indicates the more similar 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 in the same order as the human judgment and -1 means the systems are ranked in the reverse order as the human judgment.

Table 1: Sentence-level correlations against human adequacy judgments as measured by Kendall's τ and summed diagonal of confusion matrix as used in MetricsMaTr 2010. "SRL - blueprint" is the blueprint model described in section 2.1. "HMEANT (structured)" is HMEANT using the structured role representation described in section 2.2. "HMEANT (flat)" is HMEANT using the flat role representation described in section 2.3.

Metric	**Kendall**	**MetricsMaTr**
HMEANT (flat)	**0.4685**	**0.5583**
HMEANT (structured)	0.4324	0.5083
SRL - blueprint	0.3784	0.4667

4 Round 1: Flat beats structured

Our first round of comparative results quantitatively assess whether a structured role representation (that properly preserves the semantic frame structure, which is typically hierarchically nested in compositional fashion) outperforms the simpler (but less intuitive, and certainly less linguistically satisfying) flat role representation.

As shown in table 1, disturbingly, HMEANT using flat role representations yields higher correlations against human adequacy judgments than using structured role representations, regardless of whether role types are aggregated into core and adjunct classes. The results are consistent for both Kendall's tau correlation coefficient and MetricsMaTr's summed diagonal of confusion matrix. HMEANT using a flat role representation achieved a Kendall's tau correlation coefficient and summed diagonal of confusion matrix score of 0.4685 and 0.5583 respectively, which is superior to both

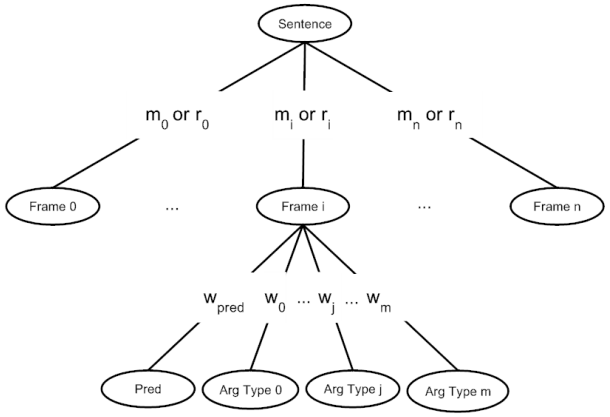

Figure 4: The new proposed structured role representation, incorporating a weighting scheme reflecting the degree of contribution of each semantic frame to the overall sentence.

HMEANT using a structured role representation (0.4324 and 0.5083 respectively) and the blueprint model (0.3784 and 0.4667 respectively).

Error analysis, in light of these surprising results, strongly suggests that the problem lies in the design which uniformly averages the frame precision/recall over all frames in a sentence when computing the sentence precision/recall. This essentially assumes that each frame in a sentence contributes equally to the overall meaning in the sentence translation. Such an assumption is trivially wrong and could well hugely degrade the advantages of using a structured role representation for semantic MT evaluation. This suggests that the structured role representation could be improved by also capturing the degree of contribution of each frame to the overall sentence translation.

5 Capturing the importance of each frame

To address the problem in the previous models, we introduce a weighting scheme to reflect the degree of contribution of each semantic frame to the overall sentence. However, unlike the contribution of each role to a frame, the contribution of each frame to the overall sentence cannot be estimated across sentences. This is because unlike semantic roles, which can be identified by their types, frames do not necessarily have easily defined types, and their construction is also different from sentence to sentence so that the positions of their predicates in the sentence are the only way to identify the frames. However, the degree of contribution of each frame does not depend on the position of the predicate in the sentence. For example, the two sentences I met Tom when I was going home and When I was walking home, I saw Tom have similar meanings. The verbs met and saw are the predicates of the key event frames which contribute more to the overall sentences, whereas going and walking are the predicates of the minor nested event frames (in locative manner roles of the key event frames) and contribute less to the overall sentences. However, the two sentences are realized with different surface constructions, and the two key frames are in different positions. Therefore, the weights learned from one sentence cannot directly be applied to the other sentence.

Instead of estimating the weight of each frame using optimization techniques, we make an assumption that *a semantic frame filled with more word tokens expresses more concepts and thus contributes more to the overall sentence*. Following this assumption, we determine the weights of each semantic frame by its span coverage in the sentence. In other words, the weight of each frame is the percentage of word tokens it covers in the sentence.

Figure 4 depicts the structured role representation with the proposed new frame weighting scheme. The significant difference between this representation and the structured role representation in the MEANT variants proposed in Lo and Wu (2011a) is that each frame is now assigned an independent weight, which is its span coverage in the MT/REF when obtaining the frame precision/recall respectively.

As in Lo and Wu (2011a), each sentence consists of a number of frames, and each frame consists of a predicate and a number of arguments of type j. Each type of argument is assigned an independent weight to represent its degree of contribution to the overall meaning of the semantic frame they attached to. The frame precision/recall is the weighted sum of the number of correctly translated roles in a frame normalized by the weighted sum of the number of all roles in that frame in the MT/REF. The sentence precision/recall is the weighted sum of the frame precision/recall for all frames normalized by the weighted sum of the total number of frames in MT/REF respectively. More precisely, again assuming the ear-

lier definitions of $C_{i,j}$, $P_{i,j}$, $M_{i,j}$, $R_{i,j}$, w_{pred} and w_{partial} in section 2, the sentence precision and recall are redefined as follows.

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

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

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

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.

6 Round 2: Structured beats flat

We now assess the performance of the new proposed structured role representation, by comparing it with the previous models under the same experimental setup as in section 4. We have also run contrastive experiments against BLEU and HTER under the same experimental conditions. In addition, to investigate the consistency of results for the automated variants of MEANT, we also include comparative experiments where shallow semantic parsing (ASSERT) replaces human semantic role labelers for each model of role representation.

Figure 5 shows an example where HMEANT with the frame weighting scheme outperforms HMEANT using other role representations in correlation against human adequacy judgments. **IN** is the Chinese source input. **REF** is the corresponding reference translation. **MT1**, **MT2** and **MT3** are the three corresponding MT output. The human adequacy judgments for this set of translation are that MT1>MT3>MT2. HMEANT with the proposed frame weighting predicts the same ranking order as the human adequacy judgment, while HMEANT with the flat role representation and HMEANT with the structured role representation without frame weighting both predict MT3>MT1>MT2. There are four semantic frames in IN while there are only three semantic frames in the REF. This is because the predicate 造成 in IN is translated in REF as had which is not a predicate. However, for the same frame, both MT1 and MT2 translated ARG1 不利影响 into the predicate affect, while MT3 did not translate the predicate 造成 and translated the ARG1 不利影响 into the noun phrase adverse impact. Therefore, using the flat role representation or the previous structured role representation which assume all frames have an identical degree of contribution to the overall sentence translation, MT1's and MT2's sentence precision is greatly penalized for having one more extra frame than the reference. In contrast, applying the frame weighting scheme, the degree of contribution of each frame is adjusted by its token coverage. Therefore, the negative effect of the less important extra frames is minimized, allowing the positive effect of correctly translating more roles in more important frames to be more appropriately reflected.

Table 2 shows that HMEANT with the proposed new frame weighting scheme correlates more closely with human adequacy judgments than HMEANT using the previous alternative role representations. The results from Kendall's tau correlation coefficient and MetricsMaTr's summed diagonal of confusion matrix analysis are consistent. HMEANT using the frame-weighted structured role representation achieved a Kendall's tau correlation coefficient and summed diagonal of confusion matrix score of 0.2865 and 0.575 respectively, bettering both HMEANT using the flat role representation (0.4685 and 0.5583) and HMEANT using the previous un-frame-weighted structured role representation (0.4324 and 0.5083).

HMEANT using the improved structured role representation also outperforms other commonly used MT evaluation metrics. It correlates with human adequacy judgments more closely than HTER (0.4324 and 0.425 in Kendall's tau correlation coefficient and summed diagonal of confusion matrix, respectively). It also correlates with human adequacy judgments significantly more closely than BLEU (0.1982 and 0.425).

Turning to the variants that replace human SRL with automated SRL, table 2 shows that MEANT

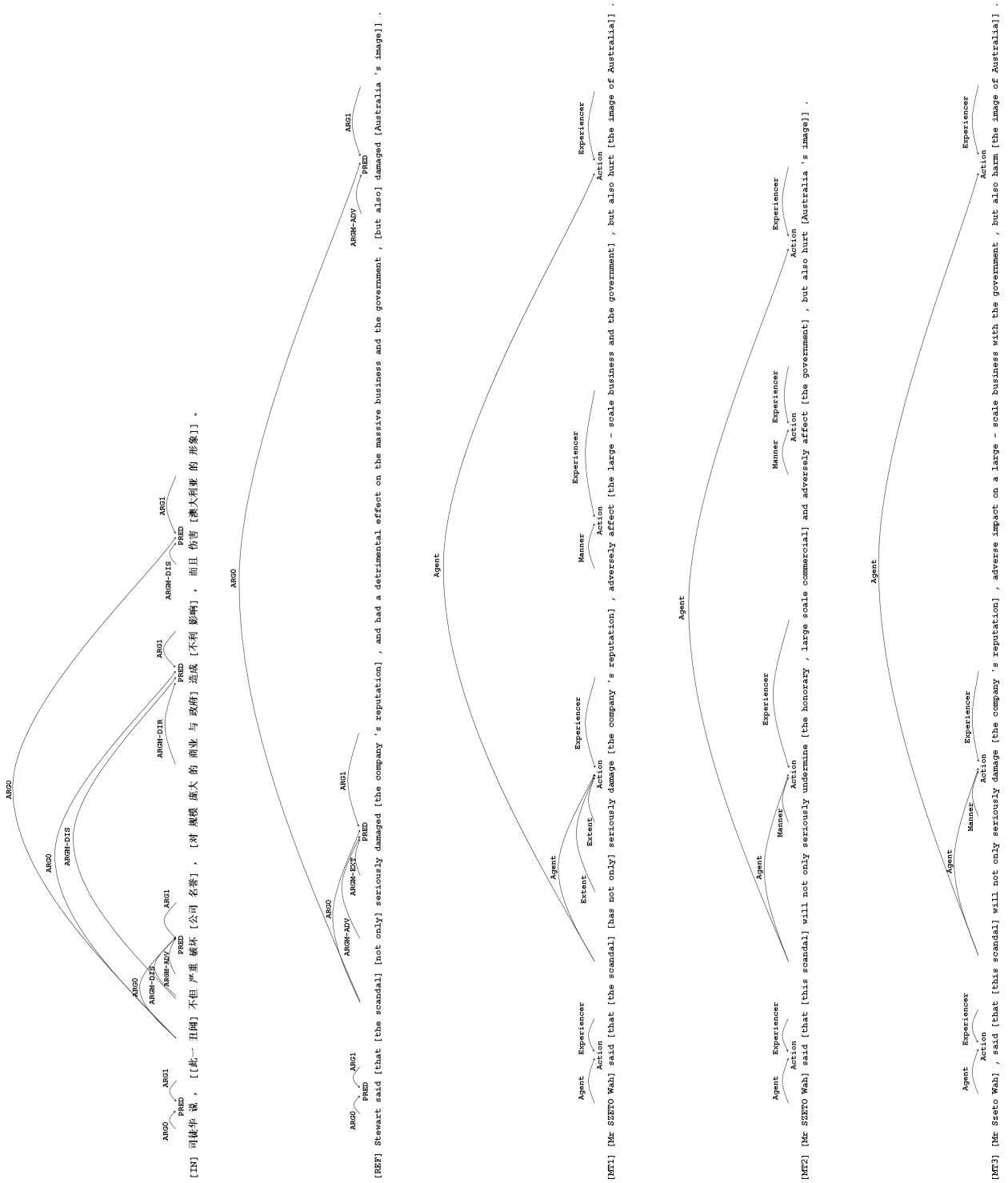

Figure 5: Example input sentence along with reference and machine translations, annotated with semantic frames in Propbank format. The MT output is annotated with semantic frames by minimally trained humans. HMEANT with the new frame-weighted structured role representation successfully ranks the MT output in an order that matches with human adequacy judgments (MT1>MT3>MT2), whereas HMEANT with a flat role representation or the previous un-frame-weighted structured role representation fails to rank MT1 and MT3 in an order that matches with human adequacy judgments. See section 6 for details.

Table 2: Sentence-level correlations against human adequacy judgments as measured by Kendall's τ and summed diagonal of confusion matrix as used in MetricsMaTr 2010. "SRL - blueprint", "HMEANT (structured)" and "HMEANT (flat)" are the same as in table 1. "MEANT (structured)" and "MEANT (flat)" use automatic rather than human SRL. "MEANT (frame)" and "HMEANT (frame)" are MEANT/HMEANT using the structured role representation with the frame weighting scheme described in section 5.

Metric	**Kendall**	**MetricsMaTr**
HMEANT (frame)	**0.4865**	**0.575**
HMEANT (flat)	0.4685	0.5583
HMEANT (structured)	0.4324	0.5083
HTER	0.4324	0.425
SRL - blueprint	0.3784	0.4667
MEANT (frame)	0.3514	0.4333
MEANT (structured)	0.3423	0.425
MEANT (flat)	0.3333	0.425
BLEU	0.1982	0.425

using the new frame-weighted structured role representation yields an approximation that is about 81% as closely correlated with human adequacy judgment as HTER, and is better than all previous MEANT variants using alternative role representations. All results consistently confirm that using a structured role representation with the new frame weighting scheme, which captures the event structure and an approximate degree of contribution of each frame to the overall sentence, outperforms using a flat role representation for SRL based MT evaluation metrics.

7 Conclusion

We have shown how the MEANT family of SRL based MT evaluation metrics is significantly improved beyond the state-of-the-art for both HTER and previous variants of MEANT, through the introduction of a simple but well-motivated weighting scheme to reflect the degree of contribution of each semantic frame to the overall sentence translation. Following the assumption that a semantic frame filled with more word tokens tends to express more concepts, the new model weight each frame by its span coverage. Consistent experimental results have been demonstrated under conditions utilizing both human and automatic SRL. Under the new frame weighted representation, properly nested structured semantic frame representations regain an empirically preferred position over the less intuitive and linguistically unsatisfying flat role representations.

One future direction of this work will be to compare MEANT against the feature based and string based representations of semantic relations in ULC. Such a comparison could yield a more complete credit/blame perspective on the representation model when operating under the condition of using automatic SRL.

Another interesting extension of this work would be to investigate the discriminative power of the MEANT family of metrics to distinguish distances in translation adequacy. In this paper we confirmed that the MEANT family of metrics are stable in correlation with human ranking judgments of translation adequacy. Further studies could focus on the correlation of the MEANT family of metrics against human scoring. We also plan to experiment on meta-evaluating MEANT on a larger scale in other genres and for other language pairs.

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 *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 *13th Confer-*

ence of the European Chapter of the Association for Computational Linguistics (EACL-06), pages 249–256, 2006.

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, 15-16 July 2010.

G. Doddington. Automatic Evaluation of Machine Translation Quality using N-gram Co-occurrence Statistics. In *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 *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 *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 *Workshop on Statistical Machine Translation*, pages 102–121, 2006.

Gregor Leusch, Nicola Ueffing, and Hermann Ney. CDer: Efficient MT Evaluation Using Block Movements. In *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 *ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, page 25, 2005.

Chi-Kiu Lo and Dekai Wu. Evaluating Machine Translation Utility via Semantic Role Labels. In *7th International Conference on Language Resources and Evaluation (LREC-2010)*, 2010.

Chi-Kiu Lo and Dekai Wu. Semantic vs. Syntactic vs. N-gram Structure for Machine Translation Evaluation. In *Proceedings of the 4th Workshop on Syntax and Structure in Statistical Translation (SSST-4)*, 2010.

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

Chi-Kiu Lo and Dekai Wu. SMT vs. AI redux: How semantic frames evaluate MT more accurately. In *To appear in 22nd International Joint Conference on Artificial Intelligence*, 2011.

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

Sebastian Pado, Michel Galley, Dan Jurafsky, and Chris Manning. Robust Machine Translation Evaluation with Entailment Features. In *Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP-09)*, 2009.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. The Proposition Bank: an Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106, 2005.

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

Sameer Pradhan, Wayne Ward, Kadri Hacioglu, James H. Martin, and Dan Jurafsky. Shallow Semantic Parsing Using Support Vector Machines.

In *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 *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 *5th European Conference on Speech Communication and Technology (EUROSPEECH-97)*, 1997.