Human Semantic MT Evaluation with HMEANT for IWSLT 2013

Chi-kiu Lo 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 present the results of large-scale human semantic MT evaluation with HMEANT on the IWSLT 2013 German-English MT and SLT tracks and show that HMEANT evaluates the performance of the MT systems differently compared to BLEU and TER. Together with the references, all the translations are annotated by annotators who are native English speakers in both semantic role labeling stage and role filler alignment stage of HMEANT. We obtain high inter-annotator agreement and low annotation time costs which indicate that it is feasible to run a large-scale human semantic MT evaluation campaign using HMEANT. Our results also show that HMEANT is a robust and reliable semantic MT evaluation metric for running large-scale evaluation campaigns as it is inexpensive and simple while maintaining the semantic representational transparency to provide a perspective which is different from BLEU and TER in order to understand the performance of the state-of-the-art MT systems.

1. Introduction

This paper presents the results from the human semantic MT evaluation with HMEANT on the IWSLT 2013 German-English MT and SLT tracks which show that HMEANT provides a perspective which is different from BLEU and TER in evaluating the performance of the MT systems. The IWSLT evaluation campaign has offered a variety of speech translation tasks over the past decade but none of them included evaluation of system performance using a semantic MT evaluation metric because of the inherent cost in evaluation in terms of both the (a) amount of time, and (b) the level of expertise needed by the human annotators. We choose HMEANT as a way around these challenges given substantial empirical evidence [1, 2] that HMEANT is a inexpensive, simple, and representationally transparent semantic MT evaluation metric that correlates with human translation adequacy judgements more highly than HTER [3] and other automatic MT evaluation metrics, such as BLEU [4], NIST [5], METEOR [6], PER [7], CDER [8], WER [9], and TER [3].

Although fast and inexpensive lexical n-gram based objective functions like BLEU have driven MT system development over the past decade, these metrics do not enforce translation utility adequately and often fail to preserve meaning [10, 11]. We believe that the system development should also be driven by semantic MT evaluation metrics which focus on getting the meaning right. Recent results [12, 13, 14] which indicate that more adequate translations are produced by tuning MT systems using the semantic evaluation metric MEANT, support us.

In this paper, we present the results of one of the largest semantic MT evaluations to date, in terms of both the number of systems and the number of translations evaluated, using HMEANT as the evaluation metric. The aims of this evaluation campaign are two-fold: (1) to demonstrate feasibility of running a large-scale semantic MT evaluation campaign using humans, and (2) to provide fine-grained statistics over a large number of systems that enable a fair comparison of semantic human MT evaluation metrics and other automatic metrics. While the former goal helps realize a practical semantically driven human MT evaluation metric in the place of expensive human MT evaluation metrics such as HTER or simple translation ranking which does not adequately reflect translation utility. The latter goal not only provides useful insights into the differences between metrics gauging semantic similarity and surface based metrics, but also quantifies the robustness

of HMEANT as an MT evaluation metric.

In the rest of the paper, we discuss the details of the evaluation campaign and provide results on the inter-annotator agreement on the tasks of semantic role annotation and alignment. We also provide an analysis of the time taken for annotation and the alignment of the semantic roles. We also report the results of different participating systems according to the criterion of our semantic evaluation metric HMEANT and its automatic variant, MEANT [15].

2. Participating tracks and systems

To perform a full-scale semantic MT evaluation, all the systems which participated in IWSLT 2013 German-English MT and SLT tracks were evaluated. There were 17 systems participating in the MT track and 3 systems participating in the SLT track.

The evaluation set consists of 136 sentences randomly drawn from the test set *(tst2013)*, which represents around 10% of the entire test set. The systems from the MT track are evaluated against the reference without disfluencies while the systems from the SLT track are evaluated against the reference with disfluencies. The details description of the tracks, the original test set and the participating systems can be found in the overview paper of IWSLT 2013 [16].

This is the largest scale semantic MT evaluation using HMEANT to date, in terms of both the number of systems and the number of translations evaluated.

3. HMEANT

HMEANT is the weighted f-score over matching semantic roles between the reference and the MT output, where the labeling and alignment of frames and role fillers is performed manually by minimally trained annotators. HMEANT, which can be driven by low-cost monolinguals of the output language, not only outperforms the commonly used automatic MT evaluation metrics, such as, BLEU, NIST, METEOR, WER, CDER and TER, but also outperforms HTER in correlating with human adequacy judgment at much lower labor cost.

HMEANT is computed as follows:

1. Human annotators annotate the shallow semantic structures of both the reference and the MT output (Figure 1 shows examples of human shallow semantic parses on both reference and MT output.)
2. Human judges align the semantic frames between the references and the MT output by judging the correctness of the predicates.
3. For each pair of aligned semantic frames,
   (a) Human judges determine the translation correctness of the semantic role fillers.
   (b) Human judges align the semantic role fillers between the reference and the MT output according to the correctness of the semantic role fillers.
4. Compute the weighted f-score over the matching role labels of these aligned predicates and role fillers according to the mathematical definitions in the following.

$$
\begin{aligned}
M_{i,j} &\equiv \text{total \# ARG j of aligned frame i in MT} \\
R_{i,j} &\equiv \text{total \# ARG j of aligned frame i in REF} \\
C_{i,j} &\equiv \text{\# correct ARG j of aligned frame i} \\
P_{i,j} &\equiv \text{\# partially correct ARG j of aligned frame i} \\
w_{\text{pred}} &\equiv \text{weight of similarity of predicates} \\
w_j &\equiv \text{weight of similarity of ARG j} \\
w_{\text{partial}} &\equiv \text{weight of the partially correct translated ARG}
\end{aligned}
$$

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

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

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

where m_i and r_i are the weights for frame, i, in the MT/REF respectively. These weights estimate the degree of contribution of each frame to the overall meaning of the sentence. $M_{\text{i,j}}$ and $R_{\text{i,j}}$ are the total counts of argument of type j in frame i in the MT and REF respectively. $C_{i,j}$ and $P_{i,j}$ are the count of the correctly and partially correct translated argument j in frame i in the MT output. The weights w_{pred} and w_j are the weights of the predicates and role fillers of the arguments of type j between the reference translations and the MT output.

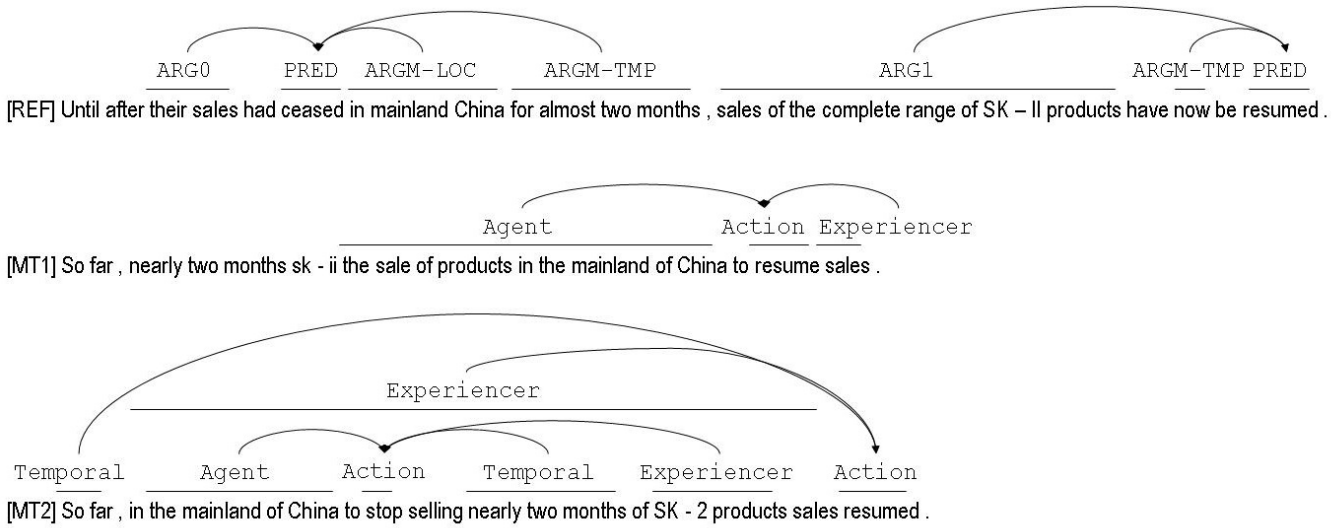

Figure 1: Examples of human semantic role labeling. There are no semantic frames for MT3 since there is no predicate.

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

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

The weight w_{partial} is the weight of the partially correct translated arguments. There is a total of 12 weights for the set of semantic role labels in MEANT as defined in [17] and a weight for the partially correct translated arguments. These weights can be determined using supervised estimation via a simple grid search to optimize the correlation with human adequacy judgments [1] or like UMEANT, estimated in an unsupervised manner using relative frequency of each semantic role label in the reference translations. U(H)MEANT can thus be used when when human judgments on adequacy of the development set are unavailable [18].

Figure 1 shows examples of human judges' decisions for semantic frame annotation on the reference and the MT output. Table 1 shows examples of the human judges' decisions for semantic frame alignment

Table 2: Inter-annotator agreement for the human semantic role labeling task.

	reference	MT output
IAA	80.86%	72.69%

and translation correctness for each semantic role for the "MT2" output in Figure 1.

4. Human annotation

HMEANT consists of two human annotation steps: (1) human semantic role labeling, which labels semantic frames within the translations and (2) human role filler alignment that determines the correctness of the translation according to the captured meaning structures. We run the human annotation using HKUST's efficient and user friendly HMEANT web-based user interface workflow [19].

4.1. Semantic role labeling

Human semantic role labeling was carried out on the references and all the submitted German-English systems in the MT track and the SLT track to capture the meaning of the translation into the "**who** *did* **what** to **whom**, **when**, **where**, **why** and **how**" structure.

4.1.1. Task description and setup

As opposed to HTER which is driven by professional bilingual translators, the semantic role labeling task in HMEANT is driven by monolinguals with minimal training of 15 minutes. To increase the robustness of the human semantic role labeling, we increased the training time for the annotators from 15 minutes to 20 minutes. The additional 5 minutes contribute to showing more annotated examples that demonstrate how to annotate the ungrammatical MT output.

Each system was annotated by two annotators who are native English speakers to support estimation of the annotation reliability. In addition, each annotator labeled the sentences from the evaluation set only once to prevent them from getting extra out-of-context information in understanding the meaning of the translation.

4.1.2. Inter-annotator agreement and time efficiency

Table 2 shows that the IAA is over 80% for labeling the semantic roles manually in the reference translation and over 72% in the MT output. The high IAA shows that the human semantic role labeling is robust and reliable.

Previous work shows that it takes the minimally trained annotators approximately 1.5 minutes to finish labeling the semantic roles of one translation output. In this evaluation, the average time needed to label the semantic roles for one translation output is significantly decreased to 50 seconds due to the fact that the semantic structures of the TED talk sentences are simpler than formal newswire text.

Table 3: Inter-annotator agreement for the human role filler alignment task.

	alignment
IAA	63.23%

4.2. Semantic role filler alignment

Human semantic role filler alignment was carried out between the references and all the submitted German-English systems in the MT track and the SLT track to determine the translation correctness according to the captured semantic structures in the previous human semantic role labeling step.

4.2.1. Task description and setup

Similar to human semantic role labeling task, we increased the training time for the native English speaking annotators by 5 minutes for showing more examples that demonstrate how to align the ungrammatical MT output to the reference.

To support the reliability analysis of the evaluation, each system was annotated by two annotators. Since the annotators are constrained to determine the phrasal translation correctness of the labeled role fillers only, it is less likely that they could be contaminated by the out-of-context information acquired due to seeing translations of the same sentence more than once. Therefore, a single annotator allowed to align translations of the same sentence from different systems.

4.2.2. Inter-annotator agreement and time efficiency

Table 3 shows that the IAA is over 63% for aligning the semantic role fillers between the reference and the MT output. The high IAA shows that the human semantic role filler alignment task is robust and reliable.

Similar to the human semantic role labeling task,

Table 4: HMEANT, MEANT, BLEU and TER scores of all the systems participating in the IWSLT 2013 German-English MT track on the evaluation set randomly drawn from tst2013 where the BLEU and TER scores are the results of the official case insensitive, without disfluencies evaluation[16]. Italicized scores indicate systems that are ranked differently from HMEANT by the corresponding metrics.

system	HMEANT	MEANT	BLEU	TER
KIT.primary	**56.55**	48.90	**27.16**	57.41
KIT.contrastive1	55.99	48.36		
EU-BRIDGE.primary	55.89	***48.97***	27.14	***56.38***
EU-BRIDGE.contrastive1	55.62	47.28		
KIT.contrastive2	55.11	*46.87*		
UEDIN.primary	54.84	47.13	25.87	*60.08*
RWTH.primary	54.63	46.51	25.86	59.51
RWTH.contrastive	54.46	46.44		
NTT-NAIST.primary	54.01	46.02	*26.45*	59.82
HDU.primary	53.99	45.99	24.07	*59.11*
HDU.contrastive2	52.47	45.37		
HDU.contrastive1	51.54	44.96		
NTT-NAIST.contrastive1	51.35	44.09		
NTT-NAIST.contrastive2	50.29	42.78		
NTT-NAIST.contrastive3	49.74	42.04		
Baseline	49.12	41.91	19.55	65.11
KLE.primary	44.53	*43.91*	*21.65*	68.04

Table 5: HMEANT and MEANT scores of all the systems participating in the IWSLT 2013 German-English SLT track on the evaluation set randomly draw from tst2013 where the BLEU and TER scores are the results of the official case insensitive, with disfluencies evaluation[16]. Italicized scores indicate systems that are ranked differently from HMEANT by the corresponding metrics.

system	HMEANT	MEANT	BLEU	TER
KIT.primary	**45.96**	**37.54**	**19.80**	**61.34**
UEDIN.primary	40.05	35.39	15.39	67.28
UEDIN.contrastive1	37.18	33.55		

in this evaluation the average time taken by minimally trained annotators to align the semantic roles between the reference and the MT output significantly decreases from 1.5 minutes to 42 seconds because the semantic structures of the TED talk sentences are simpler compared to formal newswire text. HMEANT scores are calculated by averaging the scores obtained from the two different annotations in each of the annotation task.

5. Results

From the results one can observe how HMEANT and MEANT provide different rankings compared to BLEU and TER. All four metrics HMEANT, MEANT, BLEU and TER rate KIT primary and EU-BRIDGE primary systems as closely tied in the first place according to the numbers in Table 4. On the other hand, while BLEU claims that NTT-NAIST primary system significantly outperforms UEDIN, RWTH, and HDU, both HMEANT and MEANT indicate that all four teams in the middle actually achieved comparable results. Surprisingly, HDU which is ranked the best system according to TER is ranked worst according to BLEU. These differences in the ranking of different systems between HMEANT, BLEU and TER indicates that HMEANT does offer a different perspective compared to BLEU and TER. Further, the evidence for the high correlation of HMEANT with human adequacy judgement makes HMEANT an ideal candidate for human semantic MT evaluation. Table 5 reports the scores of all four met-

rics for the three systems in the SLT track. Between KIT.primary and UEDIN.primary, all the metrics agree that KIT is better by a wide margin.

From Table 4, we can also notice that the HMEANT score of KLE.primary system is significantly smaller than the other systems. This is because the annotators failed to understand the translation output of the KLE.primary system due to excessive amounts of punctuations and symbols in the translations. Unlike BLEU score which is better than the baseline, this dearth of adequacy is appropriately represented by a sharp decrease in the HMEANT score (compared to the baseline) indicating that HMEANT reflects the translation adequacy when traditional evaluation metrics like BLEU fail to do so.

6. Conclusion

We presented the results of human semantic MT evaluation with HMEANT on the IWSLT 2013 German-English MT and SLT tracks. We also showed that rankings provided by HMEANT are different compared to BLEU and TER thereby offering a different perspective on evaluating MT system performance. The empirical evidence for HMEANT's high correlation with human judgement on translation adequacy, its semantic motivation and representational transparency makes HMEANT a viable human semantic MT evaluation metric. Further, the high inter-annotator agreement and low annotation time cost as demonstrated in this evaluation indicate that HMEANT is robust, reliable and efficient to run a large scale human semantic MT evaluation. Given our results, we believe that it would be essential to include HMEANT in evaluation campaigns so as to provide a different semantically motivated view of the state-of-the-art MT system performance to the research community.

7. Acknowledgments

This material is based upon work supported in part by the European Union under the FP7 grant agreement no. 287658; by the Defense Advanced Research Projects Agency (DARPA) under BOLT contract no. HR0011-12-C-0016, and GALE contract nos. HR0011-06-C-0022 and HR0011-06-C-0023; and by the Hong Kong Research Grants Council (RGC) research grants GRF620811, GRF621008, and GRF612806. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA, the EU, or RGC. We are grateful to all the annotators for running the evaluation in the short period of time. Thanks to Karteek Addanki for assistance with editing the paper.

8. References

[1] Chi-kiu Lo and Dekai Wu, "MEANT: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles," in *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*, 2011.

[2] ——, "Structured vs. flat semantic role representations for machine translation evaluation," in *Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-5)*, 2011.

[3] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul, "A study of translation edit rate with targeted human annotation," in *7th Biennial Conference Association for Machine Translation in the Americas (AMTA 2006)*, Cambridge, Massachusetts, August 2006, pp. 223–231.

[4] 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)*, Philadelphia, Pennsylvania, July 2002, pp. 311–318.

[5] George Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics," in *The second international conference on Human Language Technology Research (HLT '02)*, San Diego, California, 2002, pp. 138–145.

[6] Satanjeev Banerjee and Alon Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, Michigan, June 2005, pp. 65–72. [Online]. Available: http://www.aclweb.org/anthology/W/W05/W05-0909

[7] Christoph Tillmann, Stephan Vogel, Hermann Ney, Arkaitz Zubiaga, and Hassan Sawaf, "Accelerated DP based search for statistical translation," in *Fifth European Conference on Speech Communication and Technology (EUROSPEECH 1997)*, 1997.

[8] Gregor Leusch, Nicola Ueffing, and Hermann Ney, "CDer: Efficient MT evaluation using block movements," in *11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, 2006.

[9] Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney, "A evaluation tool for machine translation: Fast evaluation for MT research," in *The Second International Conference on Language Resources and Evaluation (LREC 2000)*, 2000.

[10] Chris Callison-Burch, Miles Osborne, and Philipp Koehn, "Re-evaluating the role of BLEU in machine translation research," in *11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, 2006, pp. 249–256.

[11] Philipp Koehn and Christof Monz, "Manual and automatic evaluation of machine translation between european languages," in *Workshop on Statistical Machine Translation (WMT-06)*, 2006, pp. 102–121.

[12] Chi-kiu Lo, Karteek Addanki, Markus Saers, and Dekai Wu, "Improving machine translation by training against an automatic semantic frame based evaluation metric," in *51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, 2013.

[13] Chi-kiu Lo and Dekai Wu, "Can informal genres be better translated by tuning on automatic semantic metrics?" in *14th Machine Translation Summit (MT Summit XIV)*, 2013.

[14] Chi-kiu Lo, Meriem Beloucif, and Dekai Wu, "Improving machine translation into Chinese by tuning against Chinese MEANT," in *International Workshop on Spoken Language Translation (IWSLT 2013)*, 2013.

[15] Chi-kiu Lo, Anand Karthik Tumuluru, and Dekai Wu, "Fully automatic semantic MT evaluation," in *7th Workshop on Statistical Machine Translation (WMT 2012)*, 2012.

[16] Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico, "Report on the 10th IWSLT evaluation campaign," in *International Workshop on Spoken Language Translation (IWSLT 2013)*, 2013.

[17] Chi-kiu Lo and Dekai Wu, "SMT vs. AI redux: How semantic frames evaluate MT more accurately," in *Twenty-second International Joint Conference on Artificial Intelligence (IJCAI-11)*, 2011.

[18] ——, "Unsupervised vs. supervised weight estimation for semantic MT evaluation metrics," in *Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-6)*, 2012.

[19] ——, "A radically simple, effective annotation and alignment methodology for semantic frame based SMT and MT evaluation," in *International Workshop on Using Linguistic Information for Hybrid Machine Translation (LIHMT-2011)*, 2011.