Evaluating Machine Translation Utility via Semantic Role Labels

Chi-kiu LO Dekai WU

Human Language Technology Center
Department of Computer Science and Engineering,
Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
{jackielo,dekai}@cs.ust.hk

Abstract

We present the methodology that underlies new metrics for semantic machine translation evaluation that we are developing. Unlike widely-used lexical and n-gram based MT evaluation metrics, the aim of semantic MT evaluation is to measure the utility of translations. We discuss the design of empirical studies to evaluate the utility of machine translation output by assessing the accuracy for key semantic roles. Such roles can be annotated using Propbank-style PRED and ARG labels. Recent work by Wu and Fung (2009) introduced methods based on automatic semantic role labeling into statistical machine translation, to enhance the quality of MT output. However, semantic SMT approaches have so far still only been evaluated using lexical and n-gram based SMT evaluation metrics such as BLEU, which are not aimed at evaluating the utility of MT output. Direct data analysis is still needed to understand how semantic models can be leveraged to evaluate the utility of MT output. In this paper, we discuss a new methodology for evaluating the utility of the machine translation output, by assessing the accuracy with which human readers are able to match the Propbank annotation frames.

1. Introduction

A good or useful translation is one from which human readers can successfully and accurately understand the essential meanings of the original input language sentences. After years of research, today's machine translation systems still often make many glaring errors of meaning. Recent work has started introducing semantic models into SMT so as to improve the semantic adequacy. Lexical semantics has been successfully applied to improve SMT by adapting word sense disambiguation, for example in work by Carpuat and Wu (2007), Chan *et al.* (2007), and Giménez and Màrquez (2007a). More recent work of integrating semantic role labeling (SRL) into SMT model begun to adapt even more complex types of lexical semantics from semantic parsing to the translation task, as in Wu and Fung (2009a) and Wu and Fung (2009b).

However, all the above mentioned semantic approaches to SMT are still being evaluated by lexically-oriented, n-gram based automatic evaluation metrics such as BLEU (Papineni *et al.*, 2002) , which tend to reward fluency than adequacy. The problem is that lexical similarity of two sentences does not necessarily indicate that the two sentences have similar meaning. Semantic similarity measures that capture the similarities over named entities, semantic roles and semantic tree were aggregated within rich sets of linguistic features being employed in a recent automatic evaluation metric design by Giménez and Màrquez (2007b) and Giménez and Màrquez (2008). This approach was reported to produced more reliable MT system evaluation scores, although the high complexity and the cost of running the evaluation are factors that have hindered its popularity as an automatic evaluation metric.

We take a different approach here toward assessing the utility of machine translation output, adopting from the outset the principle that *a useful 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) which represent the most important meaning of the source utterances.

We leverage work in progress taking place within the DARPA GALE program subsequent to Phase 2.5, in which both a subset of the Chinese source sentences from the evaluation data set, as well as their English reference translations, are being annotated with Propbank semantic role labels. Our objective is to assess how well the most essential semantic information, from a utility standpoint, is being captured by translation systems.

The semantic role labeling (SRL) task identifies the predicate verb and their corresponding arguments within a sentence. With the increasing availability if large parallel corpora annotated with semantic role label information. in particular Palmer *et al.* (2005) and Xue and Palmer (2005), the best monolingual shallow semantic parser by Fung *et al.* (2006) achieved an F-score of 82.01 in Chinese semantic role labeling, while the best crosslingual semantic verb frame argument mappings with accuracy of 89.3% (Parton *et al.*)as reported in the same work. We believe it is fea-

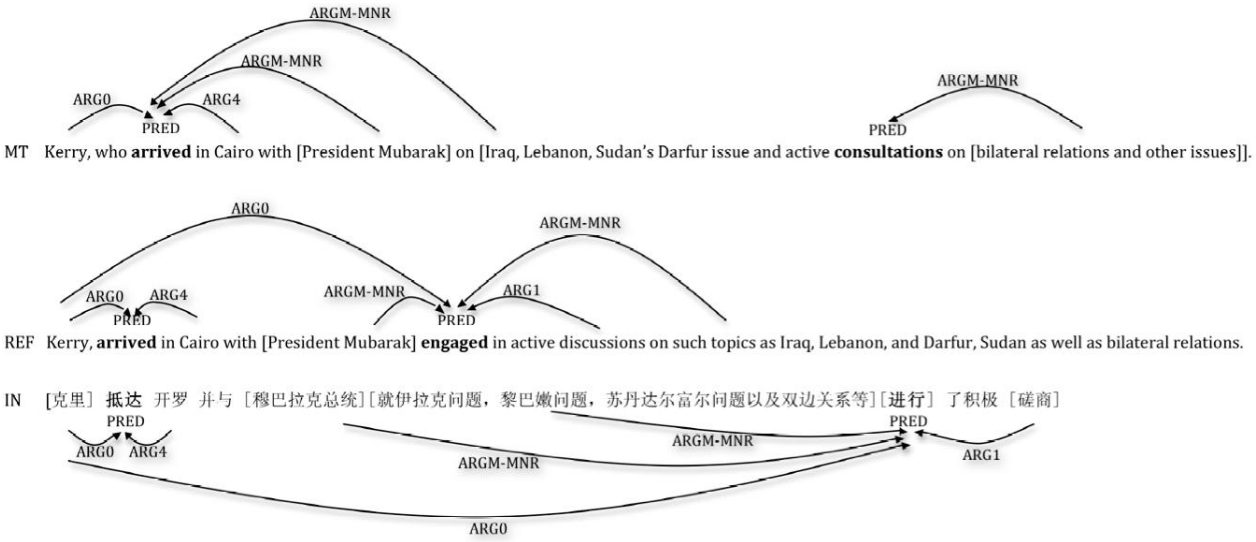

Figure 1: Example of semantic frames in Chinese input and English MT output.

sible to further develop translation evaluation methods based on roles.

In this paper, we describe methodology that evaluate the utility of the machine translation output by measuring the accuracy with which human judges are able to determine semantic roles in the machine translation output, by completing Propbank-style frames.

2. Related Work

The idea of evaluating the utility of machine translation output via human semantic role filling ability is a new research direction. The background relevant to the experiments proposed here thus consists of indirectly related work, and includes (1) works on adding semantic information into SMT, (2) machine translation evaluation via templates filling, and (3) new evaluation metrics for semantic similarity.

2.1. Semantic information in SMT

When applied to statistical machine translation, word sense disambiguation (WSD) models combine a wide range of context features in to a single lexical choice prediction, as in the work of Carpuat and Wu (2007), Chan *et al.* (2007), and Giménez and Màrquez (2007a). In particular, Phrase Sense Disambiguation (PSD) is a generalized WSD approach that automatically acquires fully phrasal translation lexicons and provides a context-dependent probability distribution over the possible translation candidates for any given phrasal lexicon (Carpuat and Wu, 2007).

At the next level of lexical semantics, semantic role labeling (SRL) is a task of identifying the semantic predicate-argument structures within a sentence – "who did what to whom, when, where, why, and how" (Pradhan *et al.*, 2004). Semantic role labels represent a more abstract level of understanding in meaning than syntactic parsing, which only performs analysis of the surface structure of a sentence. The annotated example in Figure 1 shows, from bottom to top, (IN) a fragment of a typical Chinese input source sentence that is drawn from newswire text, (REF) the corresponding fragment from its English reference sentence, and (MT) the corresponding fragment of the output sentence from a typical state-of-the-art statistical machine translation system, that achieves high BLEU and METERO scores.

A relevant subset of the semantic roles and predicates has been annotated in these fragments, using the PropBank convention of OntoNotes. In the Chinese input and its corresponding English reference, there are two main verbs marked PRED. The first (arrived has two arguments: one in an ARG0 agent role (Kerry); and another in an ARG4 destination role (Cairo). The second verb (engaged) has four arguments: one in an ARG0 agent role, again Kerry; one in an ARG1 role (discussions); and two others in ARGM-MNR manner roles (with Mubarak and on topics).

In contrast, in the machine translation output, a very different set of predicates and arguments is see. While the PRED arrived still has the same correct ARG0 Kerry and ARG4 Cairo, now the ARGM-MNR manner role with President Mubarak is incorrectly modifying the arrived, instead of an engaged predicate. In fact, the engaged predicate has erroneously been completely dropped by the machine translation system, so there is no verb to which the arguments of engaged can be attached.

Recent work by Wu and Fung (2009a) and Wu and Fung (2009b) has begun to apply SRL to statistical

machine translation for the first, using a semantic reordering model based on SRL that successfully returns a better translation with fewer semantic role confusion errors.

2.2. Task-based MT evaluation via templates filling

Task-based evaluation of machine translation developed by Voss and Tate (2006) aimed at grading how well MT engines can assist human to extract who, when, where elements of information from the MT outputs in document level. Human subjects read the MT outputs with all who-, when- and where-item high-lighted in the whole document and fill in the corresponding templates that summarize the document. The filled templates were then classified into correct, incorrect or non-response.

This evaluation approach is not only costly, time-consuming and labor-intensive. With such heavy duty on human decision, it limits the possibilities of extending to wards a fully automatic evaluation metric. Thus, we believe that the human decision in the evaluation cycle have to be as simple as possible.

2.3. MT evaluation metrics based on semantic role overlap

New MT evaluation metrics that consider the semantic similarity are introduced and refined by Giménez and Màrquez (2007b) and Giménez and Màrquez (2008). These similarity measures capture the similarities over named entities, semantic roles, and sematic trees. These features are included as part of a rich compilation of linguistic features employed in automated evaluation metrics.

Despite the fact that these metrics exhibit an improved correlation with human judgement of translation quality (Giménez and Màrquez, 2007b, 2008), they are not commonly used in large-scale MT evaluation campaign. The reasons may lie in their high complexity and cost in semantic parsing. At the same time, semantic role labels are difficult to annotate automatically and deterministically. Thus, we believe it is important to first focus on developing simple measures to evaluate MT translation utility, that make use of *human* extraction of role information. It is necessary to first understand the upper bounds of human performance on this task, as a foundation for better design of automated metrics.

3. Semantic role translation accuracy

To evaluate the semantic utility of machine translation output, we propose to conduct a comparative analysis on Propbank frames as labeled by the human readers in the machine translation output versus the reference translation or the original source input.

3.1. Evaluation corpus

The evaluation corpus is distributed across four genres: newswire, broadcast news, broadcast conversation and web text. For each Chinese sentence, there is one corresponding English reference translation and three machine translation outputs. The machine translation outputs are semantic role labeled manually in the Propbank annotation. For this evaluation, we are focusing on the newswire genre. The evaluation set shall be formed by randomly selecting 40 sentences of newswire sentences from the corpus.

3.2. Methodology

Seven human readers are employed to conduct the analysis. Four of them are bilingual Chinese/English readers and the other three of them are monolingual English readers.

One of the bilingual Chinese English readers is given the reference translation and the original source input. This sanity check serves as the control condition of the analysis.

For the other three bilingual Chinese/English readers, each of them is given one set of machine translation system output, consisting of 40 sentences together with the corresponding original source input. Similarly, each monolingual English readers is given one set of machine translation system output of 40 sentences as well as the corresponding reference translation.

For each predicate in the source or reference sentence, the human readers are required to judge whether there is a match predicate annotated in the machine translation output. If there is a match, the human readers are required to judge whether each of the argument associated with that matched predicate in the machine translation output is translated: Correct, Partial or Incorrect. Translations of the arguments are judged Correct if they express the same meaning as that of the reference translations or the original source input. Translations of the arguments may also be judged Partial if only part of the meaning is correctly matched with the reference translations or the original source input. In both categories, extra meaning in the argument translation will not be penalize unless the extra meaning includes meaning from another argument. In that case, the translations of the arguments are judged Incorrect.

Based on the comparative matrices collected from the human judges, a precision-recall analysis of semantic role correctness, reflecting the utility of each machine

translation system could be done.

$$
\begin{aligned}
N_{\text{ci}} &= \text{no. of Correct ARG of PRED i in MT} \\
N_{\text{pi}} &= \text{no. of Partial ARG of PRED i in MT} \\
N_i &= \text{total no. of ARG of PRED i in MT} \\
N_c &= \sum_{\text{all matched predicates}} \frac{N_{\text{ci}}}{N_i} \\
N_p &= \sum_{\text{all matched predicates}} \frac{N_{\text{pi}}}{N_i} \\
P &= \frac{N_c + (0.5 * N_p)}{\text{total no. of predicates in reference}} \\
R &= \frac{N_c + (0.5 * N_p)}{\text{total no. of predicates in MT output}} \\
F - \text{measure} &= \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}
\end{aligned}
$$

4. Conclusions and Future Work

To our best knowledge, this is a first methodology in MT evaluation aimed at assessing the *utility* of machine translation output via how well it facilitates semantic role template filling. Our methodology can be seen as indirectly assessing the accuracy of semantic role template information in machine translation output. This first ground breaking data analysis could provide us the true picture of the upper bounds of human performance on this task.

With the improving performance of shallow semantic parsers, we believe our methodology could be further developed in to automatic MT evaluation metric that assessing the accuracy of the translation in the semantic role templates; and such semantic MT metric would outrank those lexically-oriented n-gram based metric in terms of correlation with human judgements. The semantic role manually annotated evaluation corpus is in Chinese and English. Chinese and English often realize semantic frames using very different surface forms. The effect is seen in the fact that state-of-the-art machine accuracy remains low for Chinese-English MT. We conjecture that similar experiments proposed in this paper can be conducted in most other language pairs, especially for those having closer semantic frame correspondences that would be easier for the human judges to determine the correctness of the semantic role template translations.

5. Acknowledgments

This material is based upon work supported in part by the Defense Advanced Research Projects Agency (DARPA) under GALE Contract No. 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

Marine Carpuat and Dekai Wu. Improving Statistical Machine Translation using Word Sense Disambiguation. In *Conference on Empirical Methods in Natural Language Processing and Conference on Computational Natural Language Learning (EMNLP-CoNLL 2007)*, Prague, Jun 2007.

Y.S. Chan, H.T. Ng, and D. Chiang. Word sense disambiguation improves statistical machine translation. In *45th Annual Meeting of the Association for Computational Linguistics*, Prague, Czech, 2007.

Pascale Fung, Zhaojun Wu, Yongsheng Yang, and Dekai Wu. Automatic Learning of Chinese English Semantic Structure Mapping. In *IEEE Spoken Language Technology Workshop, 2006*, pages 230–233, 2006.

Jesús Giménez and Lluís Màrquez. Discriminative Phrase Selection for Statistical Machine Translation. *Learning Machine Translation*, 2007.

Jesús Giménez and Lluís Màrquez. Linguistic features for automatic evaluation of heterogenous MT systems. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 256–264, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

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

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*, pages 311–318, 2002.

K. Parton, K.R. McKeown, R. Coyne, M. Diab, R. Grishman, D. Hakkani-Tür, M. Harper, H. Ji, W.Y. Ma, A. Meyers, et al. Who, What, When, Where,

Why? Comparing Multiple Approaches to the Cross-Lingual 5W Task. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP*, Suntec, Singapore, August.

Sameer Pradhan, Wayne Ward, Kadri Hacioglu, James H. Martin, and Dan Jurafsky. Shallow Semantic Parsing Using Support Vector Machines. In *Proceedings of NAACL-HLT 2004*, 2004.

Clare R. Voss and Calandra R. Tate. Task-based Evaluation of Machine Translation (MT) Engines: Measuring How Well People Extract Who, When, Where-Type Elements in MT Output. In *Proceedings of 11th Annual Conference of the European Association for Machine Translation (EAMT-2006)*, pages 203–212, Oslo, Norway, June 2006.

Dekai Wu and Pascale Fung. Can Semantic Role Labeling Improve SMT? In *Proceedings of the 13th Annual Conference of the EAMT*, pages 218–225, Barcelona, Spain, May 2009.

Dekai Wu and Pascale Fung. Semantic Roles for SMT: A Hybrid Two-Pass Model. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 13–16. Association for Computational Linguistics, 2009.

Nianwen Xue and Martha Palmer. Automatic Semantic Role Labeling for Chinese Verbs. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence, Edinburgh, Scotland*, 2005.