Semantic vs. Syntactic vs. N-gram Structure for Machine Translation Evaluation

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

Abstract

We present results of an empirical study on evaluating the utility of the machine translation output, by assessing the accuracy with which human readers are able to complete the semantic role annotation templates. Unlike the widely-used lexical and n-gram based or syntactic based MT evaluation metrics which are fluency-oriented, our results show that using semantic role labels to evaluate the utility of MT output achieve higher correlation with human judgments on adequacy. In this study, human readers were employed to identify the semantic role labels in the translation. For each role, the filler is considered an accurate translation if it expresses the same meaning as that annotated in the gold standard reference translation. Our SRL based f-score evaluation metric has a 0.41 correlation coefficient with the human judgement on adequacy, while in contrast BLEU has only a 0.25 correlation coefficient and the syntactic based MT evaluation metric STM has only 0.32 correlation coefficient with the human judgement on adequacy. Our results strongly indicate that using semantic role labels for MT evaluation can be significantly more effective and better correlated with human judgement on adequacy than BLEU and STM.

1 Introduction

In this paper, we show that evaluating machine translation quality by assessing the accuracy of human performance in reconstructing the semantic frames from the MT output has a higher correlation with human judgment on translation adequacy than (1) the widely-used lexical n-gram precision based MT evaluation metric, BLEU (Papineni *et al.*, 2002), as well as (2) the best-known syntactic tree precision based MT evaluation metric, STM (Liu and Gildea, 2005). At the same time, unlike some highly labor intensive evaluation metrics such as HTER (Snover *et al.*, 2006), our proposed semantic metric only requires simple and minimal instructions to the human judges involved in the evaluation cycle.

We argue that neither n-gram based metrics, like BLEU, nor syntax-based metrics, like STM, adequately capture the similarity in meaning between the machine translation and the reference translation—which, ultimately, is essential for translations to be *useful*.

First, n-gram based metrics assume that "good" translations share the same lexical choices with the reference translation. While BLEU score performs well in capturing the translation *fluency*, Callison-Burch *et al.* (2006) and Koehn and Monz (2006) report cases where BLEU strongly disagrees with human judgment on translation quality. The underlying reason is that lexical similarity does not adequately reflect the similarity in meaning.

Second, just like n-gram based metrics such as BLEU, syntax-based metrics are still more fluency-oriented than adequacy/accuracy-oriented. While STM addresses the failure of BLEU in evaluating the translation *grammaticality*, a grammatical translation can nonetheless achieve a high STM score even if contains errors arising from confusion of semantic roles. Syntactic structure similarity still inadequately reflects similarity of meaning.

As MT systems improve, the shortcomings of

Proceedings of SSST-4, Fourth Workshop on Syntax and Structure in Statistical Translation, pages 52–60,
COLING 2010, Beijing, August 2010.

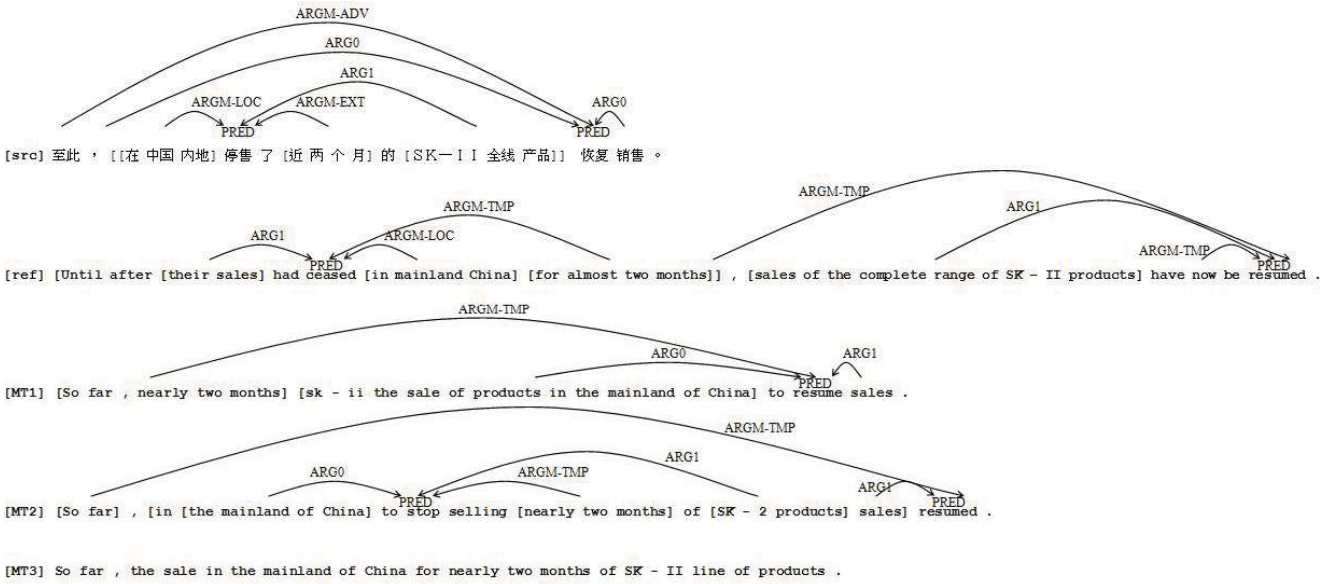

Figure 1: Example of semantic frames in Chinese input, English reference translation and MT output.

lexical n-gram based and syntax-based evaluation metrics are becoming more apparent. State-of-the-art MT systems are often able to output translations containing roughly the correct words and being almost grammatical, but not expressing meaning that is close to the source input. We adopt the outset of the principle that *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 important meaning of the source utterances. Our objective is to evaluate how well the most essential semantic information is being captured by the machine translation systems from the user's point of view.

In this paper, we describe in detail the methodology that underlies the new semantic machine translation evaluation metrics we are developing. We present the results of the study on evaluating machine translation utility by measuring the accuracy with which human readers are able to complete the semantic role annotation templates. Last but not the least, we show that our proposed evaluation metric has a higher correlation with human judgments on adequacy than BLEU and STM.

2 Related Work

2.1 Semantic models in SMT

Numerous recent works has been done on applying different semantic models to statistical machine translation. Word sense disambiguation (WSD) models combine a wide range of context features into 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), a generalization of the WSD approach, 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).

Another recent research direction on semantic SMT is applying semantic role labeling models. Semantic role labeling (SRL) is the task of identifying the semantic predicate-argument structures within a sentence. Semantic role labels represent an abstract level of understanding in meaning. There is an increasing availability of large parallel corpora annotated with semantic role information, in particular, in the work of Palmer *et al.* (2005) and Xue and Palmer (2005). As a result, the accuracy of automatic SRL task is also rising.

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 cross-lingual semantic verb frame argument mappings with accuracy of 89.3% as reported in the same work.

The example in Figure 1 is labeled with semantic roles in the Propbank convention. **src** shows a fragment of a typical Chinese source sentence that is drawn from newswire genre of the evaluation corpus. **ref** shows the corresponding fragment of the English reference translation. **MT1**, **MT2** and **MT3** show the three corresponding fragments of the machine translation output from three different MT systems.

A relevant subset of the semantic roles and predicates has been annotated in these fragments, using the PropBank convention of OntoNotes. In the Chinese source sentence, there are two main verbs marked PRED. The first verb "停售" (cease of sales) has three arguments: one in ARG1 experiencer role, "ＳＫ－ＩＩ 全线 产品" (the complete range of SK-II products); one in ARGM-LOC location role, "在 中国 内地" (in mainland China), and one in ARGM-EXT extent role, "近 两 个 月" (for almost two months). The second verb "恢复" (resumed) also has three arguments: two in ARG0 agent roles, "在 中国 内地 停售 了 近 两 个 月 的 ＳＫ－ＩＩ 全线 产品 " (the complete range of SK-II products which sales had ceased in mainland China for almost two months) and "销售" (sales), and one in ARGM-ADV role, "至此" (until then).

In the corresponding English target, there are also two main verbs marked PRED. The first verb (ceased) has three arguments: one in an ARG1 experiencer role, "their sales"; one in an ARGM-LOC role, "in mainland China", and one in ARGM-TMP temporal role, "for almost two months". The second verb (resumed) also has three arguments: two in ARGM-TMP temporal roles, "until after their sales ceased in mainland China for almost two months" and "now", and one in ARG1 experiencer role, "sales of the complete range of SK-II products".

Similarly, the first two MT outputs are also annotated with semantic roles in the PropBank convention. Since there is no verb appeared in the third MT output, no predicate-argument structure is annotated.

Recent work by Wu and Fung (2009a) and Wu and Fung (2009b) has begun to apply SRL to statistical machine translation using a semantic reordering model based on SRL that successfully returns a better translation with fewer semantic role confusion errors.

With recent rise of work applying semantic model to statistical machine translation, there is a high demand for MT evaluation metrics that are directly sensitive to the semantic improvement made. We believe evaluating machine translation utility based on semantic roles should reflect semantic improvement better than current widely-used automated n-gram precision based MT evaluation metrics, like BLEU or fluency-oriented syntactic MT evaluation metrics, like STM.

2.2 STM: syntax-based MT evaluation

Liu and Gildea (2005) proposed to use syntactic features in MT evaluation and developed subtree metric (STM) which based on the similarity of syntax tree of the MT output and that of the reference. It is the first proposed metric that incorporates syntactic features in MT evaluation and underlies all the other recently proposed syntactic MT evaluation metrics.

STM is a precision based metric that captures the fractions of the subtree in a specific depth of the MT output syntax tree which also appear in the reference syntax tree. The fractions of different depths are then average in arithmetic mean.

$$\mathrm{STM} = \frac{1}{D}\sum_{n=1}^{D}\frac{\sum_{t\in \mathrm{subtree}_n(\mathrm{hyp})}\mathrm{count}_{\mathrm{found}}(\mathrm{t})}{\sum_{t\in \mathrm{subtree}_n(\mathrm{hyp})}\mathrm{count}\,(t)}$$

where D is the maximum depth of subtress considered, $\mathrm{count}\,(t)$ denotes the number of times subtree t appears in the MT output's syntax tree, and $\mathrm{count}_{\mathrm{found}}\,(t)$ denotes the found number of times t appears in the references' syntax tree, each subtree in reference will only be found once.

Figure 2 shows the syntax tree of a reference translation and that of the corresponding MT output. For example, we set the maximum depth of subtree considered to 4. There are seven 1-depth subtrees in the MT output (S, NP, VP, PRP, V,

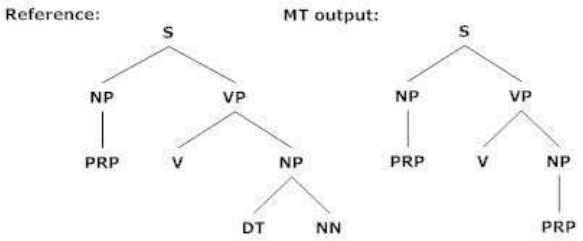

Figure 2: Example for the computation of STM

NP and PRP) in which only six of them appear in the references (S, NP, VP, PRP, V and NP). Note that the found count of PRP should be 1 rather than 2 because there is only one PRP in the reference translation syntax tree. For 2-depth, there are four subtrees in the MT output (S→NP VP, NP→PRP, VP→V NP and NP→PRP) in which three of them appear in the reference (S→NP VP, NP→PRP and VP→V NP). Similarly, there are one out of two 3-depth subtrees and zero out of one 4-depth subtrees in the MT output found in the reference. Therefore, the final STM score for this example is (6/7+3/4+1/2+0/1)/4=0.527.

2.3 MT evaluation metric based on semantic role overlap

Giménez and Màrquez (2008) introduced ULC, a new automatic MT evaluation metric in which a series of linguistic features are combined together. One of those linguistic features is shallow semantic similarity on semantic role overlap. The semantic role overlap metric calculates the lexical overlapping between semantic roles of the same type in the machine translation output and the corresponding reference translations and then considers the average lexical overlapping over all semantic role types.

Despite the fact that the metric shows an improved correlation with human judgment of translation quality (Giménez and Màrquez, 2007b, 2008; Callison-Burch *et al.*, 2007, 2008), it is not commonly used in large-scale MT evaluation campaign. The reason may lie in the high time cost.

We believe it is important to first focus on developing simple measures to evaluate machine 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 efficient automated metrics.

2.4 HTER: non-automated MT evaluation metric

Human-targeted Translation Edit Rate (HTER) in the work of Snover *et al.* (2006) is a non-automatic machine translation evaluation metric based on the number of edits required to correct the translation hypotheses. A human annotator edits each MT hypothesis so that it is meaning-equivalent with the reference translation. It emphasizes on making the minimum possible number of edits. The Translation Edit Rate (TER) is then calculated using the human-edited translation as a targeted reference for the MT hypothesis.

The HTER is highly labor intensive in the evaluation process. The human annotators are not only required to understand the meaning expressed in the reference translation and the machine translation, but are also required to propose minimum possible number of edits to the translation hypotheses. With such heavy-duty human decision requirements, the cost in evaluation is enormously increased, bottlenecking the evaluation cycle. Instead, we believe that any human decisions in the evaluation cycle should be reduced to be as simple as possible.

3 Semantic role translation accuracy

To evaluate the semantic utility of machine translation output, we conduct a comparative analysis on the Propbank annotation templates completed by the human readers in the machine translation output versus the reference translation.

3.1 Evaluation corpus

The sentences of the evaluation corpus are randomly drawn from the newswire genre of the DARPA GALE program Phase 2.5 evaluation. For each Chinese input sentence, there are one corresponding English reference translation and three state-of-the-art machine translation systems' outputs. The Chinese source and the English reference are annotated with gold standard semantic role labels in Propbank style.

South Korea 's Ministry of Agriculture and Forestry said this evening that an Asian City duck farm reported to the relevant department on the 11th that since the 5th of this month , the number of egg production of over 9,000 ducks in the duck farm had fallen sharply .

Agent 1:	South Korean 's Ministry of Agriculture and Forestry
Action 1:	said
Experiencer1:	an Asian City duck farm reported to the relevant department on the 11th that since the 5th of this month , the number of egg production of over 9,000 ducks in the duck farm had fallen sharply
Temporal 1:	this evening
Agent 2:	an Asan City duck farm
Action 2:	reported
Experiencer 2:	since the 5th of this month , the number of egg production of over 9,000 ducks in the duck farm had fallen sharply
Patient 2:	the relevant department
Temporal 2:	on the 11th
Agent 3:	the number of egg production of over 9,000 ducks in the duck farm
Action 3:	fallen
Temporal 3:	since the 5th of this month
Manner 3:	sharply

Figure 3: Example given to human annotators demonstrating how to label the semantic frames.

Table 1: List of semantic roles that human judges are requested to label.

Label	Event	Label	Event
Actor	who	Temporal	when
Action	did	Location	where
Experiencer	what	Other adverbial arg.	why / how
Patient	whom		

3.2 Reconstruction of semantic frames in MT output

Four groups of bilingual Chinese English human annotators are employed to conduct the analysis. One group of them is given the reference translation. This sanity check serves as the control condition of the analysis. The other three groups of them is given one set of the machine translation system output. The four groups are all disjoint such that no annotators annotate more than one sentence from a MT-reference set to avoid contamination in annotators' judgments. To reduce the effect of personal bias on annotations, each sentence is annotated by at least two human annotators. The results are reported as the average among the annotators.

With the aim of evaluating machine translation utility from a user standpoint, we have simplified the Propbank annotation into a more intuitive event structure, i.e. "who did what to whom, when, where, why and how". Since the layman annotators find that it is difficult to distinguish between the "why" and "how" events type, we have combined the "why" and "how" events in to one "Other adverbial argument" label. Human annotators are given simple and minimal instructions on what to label and two examples demonstrating how to label. Table 1 shows the list of labels annotators are requested to annotate. Figure 3 shows the example shown to the human annotators on how to label semantic frames.

After reconstruction of the semantic frames, the annotated machine translation outputs are distributed to another disjoint group of three monolingual human judges. The human judges are required to match each predicate in the reference translation with those annotated in the MT output. Then, for each matched predicate, they are required to judge whether each of the associated argument in the reference translation is translated and annotated in the MT output: Correct, Incorrect or Partial. Translations of the semantic frames are judged Correct if they express the same meaning as that of the reference translations or the original source input. Translations of the semantic frames are judged Incorrect if they express meaning(s) that belongs in other arguments. Translation of the semantic frames may also be judged Partial if only part of the meaning is correctly expressed. Extra meaning in the semantic frames will not be penalized unless it belongs in another argument. The partially correct category is designed to facilitate a finer-grained measurement of the translation utility.

3.3 SRL based evaluation metric

Based on the comparative matrices collected from the human judges, a precision-recall analysis of accuracy with the reconstructed semantic frames, reflecting the utility of each machine translation system could be done.

$$
\begin{aligned}
C_{\text{core i}} &= \text{no. of Correct core ARG of PRED i in MT} \\
C_{\text{argm i}} &= \text{no. of Correct ARGM of PRED i in MT} \\
P_{\text{core i}} &= \text{no. of Partial core ARG of PRED i in MT} \\
P_{\text{argm i}} &= \text{no. of Partial ARGM of PRED i in MT} \\
\text{MT}_{\text{core i}} &= \text{total no. of core ARG of PRED i in MT} \\
\text{MT}_{\text{argm i}} &= \text{total no. of ARGM of PRED i in MT} \\
\text{Ref}_{\text{core i}} &= \text{total no. of core ARG of PRED i in ref.} \\
\text{Ref}_{\text{argm i}} &= \text{total no. of ARGM of PRED i in ref.}
\end{aligned}
$$

$$
\begin{aligned}
C_{\text{precision}} &= \sum_{\text{all matched}} \frac{w_0 + w_1 C_{\text{core i}} + w_2 C_{\text{argm i}}}{w_0 + w_1 \text{MT}_{\text{core i}} + w_2 \text{MT}_{\text{argm i}}} \\
C_{\text{recall}} &= \sum_{\text{all matched}} \frac{w_0 + w_1 C_{\text{argn i}} + w_2 C_{\text{argmi}}}{w_0 + w_1 \text{Ref}_{\text{core i}} + w_2 \text{Ref}_{\text{argm i}}} \\
P_{\text{precision}} &= \sum_{\text{all matched}} \frac{w_1 P_{\text{core i}} + w_2 P_{\text{argm i}}}{w_0 + w_1 \text{MT}_{\text{core i}} + w_2 \text{MT}_{\text{argm i}}} \\
P_{\text{recall}} &= \sum_{\text{all matched}} \frac{w_1 P_{\text{core i}} + w_2 P_{\text{argm i}}}{w_0 + w_1 \text{Ref}_{\text{core i}} + w_2 \text{Ref}_{\text{argm i}}}
\end{aligned}
$$

$$
\begin{aligned}
\text{Precision} &= \frac{C_{\text{precision}} + (w_{\text{partial}} \times P_{\text{precision}})}{\text{total no. of predicates in MT}} \\
\text{Recall} &= \frac{C_{\text{recall}} + (w_{\text{partial}} \times P_{\text{recall}})}{\text{total no. of predicatesin ref.}} \\
\text{F} - \text{score} &= \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}
\end{aligned}
$$

$C_{\text{core i}}$ and $C_{\text{argm i}}$ represent the number of correctly translated core arguments and adjunct arguments of a matched predicate i respectively while $P_{\text{core i}}$ and $P_{\text{argm i}}$ represent the number of partially translated core arguments and adjunct arguments of a matched predicate i. $\text{MT}_{\text{core i}}$ and $\text{MT}_{\text{argm i}}$ represent the total number of core arguments and adjunct arguments of the matched predicate i in the MT output and $\text{Ref}_{\text{core i}}$ and $\text{Ref}_{\text{argmi}}$ represent the total number of core arguments and adjunct arguments of the matched predicate i in the reference.

$C_{\text{precision}}$ and $P_{\text{precision}}$ are the sum of the portions of correctly or partial correctly translated predicate-argument structures in the MT output. They can be viewed as the true positive for precision. C_{recall} and P_{recall} are the sum of the portion of correctly or partial correctly translated predicate-argument structures in the reference. They can be viewed as the true positive for recall. Note that w_0, w_1 and w_2 are the weights for the matched predicate, core arguments and adjunct arguments. These weights can be viewed as the importance of meanings in the different categories of semantic roles. In this very first preliminary study, we have set them all to 1 and we expect tuning these weights can further increase the correlation of the evaluation metric with human judgment of translation utility.

The precision, recall and f-score of the SRL based MT evaluation metric are defined in terms of the translation accuracy of predicate-argument structures. Note that w_{partial} is the weights for the partially correct translated arguments. In this experiment, we have arbitrarily set it to 0.5.

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 express the same meaning as the corresponding arguments in the reference translations, the f-score of the SRL based MT evaluation metric will be equal to 1.

Table 2: SRL annotation of example 1 MT1 output in figure 1 and the human judgement on translation correctness of each argument.

SRL	reference	MT1	decision
PRED	ceased	–	not match
PRED	resumed	resume	match
ARG0	–	sk - ii the sale of products in the mainland of China	incorrect
ARG1	sales of complete range of SK - II products	sales	partial
TMP	Until after , their sales had ceased in mainland China for almost two months	So far , nearly two months	partial
TMP	now	–	incorrect

3.4 Experiment and Results

Table 2 shows the SRL annotation of MT1 by one of the annotators of example 1 in figure 1

Table 3: SRL based MT evaluation average on all annotators and all sentences.

System	Precision	Recall	F-score
Reference	0.75	0.73	0.73
MT1	0.39	0.35	0.36
MT2	0.37	0.31	0.33
MT3	0.34	0.30	0.30

and the human judgement on translation correctness of each argument. The predicate "ceased" in the reference translation did not match with any predicate annotated in MT1 while the predicate "resumed" matched with the predicate "resume" annotated in MT1. The ARGM-TMP argument, "Until after their sales had ceased in mainland China for almost two months", in the reference translation is partially translated to ARGM-TMP argument, "So far , nearly two months", in MT1; the ARG1 argument, "sales of the complete range of SK - II products", in the reference translation is partially translated to ARG1 argument, "sales", in MT1 and the ARGM-TMP argument, "now" in the reference translation is missing in MT1. The SRL based f-score of this example is 0.33. The final sentence-level SRL based MT evaluation metric of MT1 is the f-score averaged on all annotators. Table 3 shows the results of the SRL based MT evaluation metric averaged on all annotators and all sentences. Our results show that the evaluation metric can successfully distinguish the translation utility of the human translation and the three MT systems; and on system level, MT1 provides the most accurate translation.

4 Inter-annotator Agreement

We measured the inter-annotator agreement in two tasks: role identification and role classification. The standard f-score is used to measure the agreement on SRL annotation as in Brants (2000).

For role identification, the agreement is counted on the matching of word span in the annotated arguments with a tolerance of ± 1 word in mismatch. The tolerance is designed for the fact that annotators are not consistent in handling the articles or punctuations at the beginning or the end of the annotated arguments. The agreement rate on SRL annotations in role identification of reference translation is 76%, and that on MT output is 72%.

For role classification, in addition to the requirement of matching of word span in role identification task, the agreement is counted on the matching of the semantic role labels within two aligned word spans. The agreement rate on SRL annotations of reference translation and that on MT output are 69% and 65% respectively.

The results show that with such minimal training, the layman annotators perform consistently in identifying the semantic structure in both the reference translation and the MT output. The results suggest that the layman annotators also having problem in role confusion and we believe that a slightly more detailed explanation on the role labels may help to clear the confusion.

5 Correlation with human judgments on translation adequacy

We used the Spearman's rank correlation coefficient ρ to measure the correlation of the evaluation metrics with the human judgment on adequacy at sentence-level and took average on the whole data set. The human judgment on adequacy was 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 the raw scores of the evaluation metrics. The STM scores are calculated based on the syntax tree of the reference and MT output parsed by the Charniak parser (Charniak, 2001). Table 4 shows the raw scores of example 1 under the our proposed SRL based evaluation metric, sentence-level BLEU, sentence-level STM and the corresponding ranks assigned to each of the systems, together with the human ranks on adequacy.

The Spearman's rank correlation coefficient ρ can be calculated using the following simplified equation:

$$\rho = 1 - \frac{6 \sum {d_i}^2}{n(n^2 - 1)}$$

where d_i is the difference between the ranks of the

Table 4: Sentence-level SRL based f-score evaluation metrics average on annotators, sentence-level BLEU, sentence-level STM, their corresponding rank assigned and the human rank on adequacy for example 1.

System	MT output	SRL		BLEU		STM		Human
		score	rank	score	rank	score	rank	rank
Src	至此， 在 中国 内地 停售 了 近 两 个 月 的 ＳＫ－ＩＩ 全线 产品 恢复 销售 。	-	-	-	-	-	-	-
Ref	Until after their sales had ceased in mainland China for almost two months , sales of the complete range of SK － II products have now be resumed .	-	-	-	-	-	-	-
MT1	So far , nearly two months sk - ii the sale of products in the mainland of China to resume sales .	0.167	2	0.012	3	0.364	1	2
MT2	So far , in the mainland of China to stop selling nearly two months of SK - 2 products sales resumed .	0.317	1	0.013	2	0.303	3	1
MT3	So far , the sale in the mainland of China for nearly two months of SK - II line of products .	0.000	3	0.124	1	0.344	2	3

Table 5: Average sentence-level correlation for the evaluation metrics.

Metric	Correlation with human
SRL based evaluation	0.41
BLEU	0.25
STM	0.32

evaluation metrics and the human judgment over of system i and n is the number of systems. 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. The higher the value for ρ indicates the more similar the ranking by the evaluation metric to the human judgment.

Our results show that the proposed SRL based evaluation metric has a higher correlation with the human judgment on adequacy than either the BLEU or STM metrics. Table 5 compares the average sentence-level ρ for our proposed SRL based evaluation metric, BLEU, and STM. The correlation coefficient for the proposed SRL based evaluation metric is 0.41, while that for BLEU is 0.25. The correlation coefficient for STM is 0.32, significantly better than BLEU, but still far short of our SRL based metric.

6 Conclusions and Future Work

We presented results of an empirical study on evaluation the utility of MT output, by assessing the accuracy with which human reader are able to complete the SRL templates. The SRL based f-score evaluation metric we proposed provided an intuitive picture on how much information of the original source input the machine translation users can extract by reading the MT output. Comparing to HTER where the human decision is heavy and requires advance knowledge in how to fix the translation with minimum change, only minimal instructions is necessary to be given to the human readers in our proposed metric. The human readers may not necessarily be translation experts.

Our results show that using SRL in semantic MT evaluation is a highly promising direction for further research. We evaluated the proposed SRL based metric with human judgment on adequacy using Spearman's correlation coefficient. The proposed SRL based evaluation metric was found to be significantly better correlated to human judgment on adequacy than either BLEU or the syntax-based evaluation metric STM.

Our current direction is to discriminatively tune the weights within the SRL based evaluation metric, so as to further increase the correlation of the metric with human judgment.

Our other main avenue of current work is to construct automated metrics approximating the

evaluation method described here (which provides an upper bound for automated SRL-based metrics). With the improving performance of shallow semantic parsers, we believe that the proposed evaluation metric could be further developed into inexpensive automatic MT evaluation metrics.

7 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

T. Brants. Inter-annotator agreement for a German newspaper corpus. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC-2000)*. Citeseer, 2000.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluating the role of BLEU in machine translation research. In *Proceedings of EACL 2006*, pages 249–256, 2006.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. (Meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158. Association for Computational Linguistics, 2007.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106. Association for Computational Linguistics, 2008.

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.

E. Charniak. Immediate-head parsing for language models. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, page 131. Association for Computational Linguistics, 2001.

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.

Philipp Koehn and Christof Monz. Manual and automatic evaluation of machine translation between european languages. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 102–121. Association for Computational Linguistics, 2006.

D. Liu and D. Gildea. Syntactic features for evaluation of machine translation. *ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, page 25, 2005.

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.

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.

M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231, 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.