Improving evaluation and optimization of MT systems against MEANT

Chi-kiu Lo*
NRC-CNRC
Multilingual Text Processing
National Research Council Canada
1200 Montreal Road, Ottawa,
Ontario K1A 0R6, Canada
ChiKiu.Lo@nrc-cnrc.gc.ca

Philipp C. Dowling*
TUM
Fakultät für Informatik
Technische Universität München
Boltzmannstraße 3,
85748 Garching bei München, Germany
dowling@cs.tum.edu

Dekai Wu
HKUST
Human Language Technology Center
HK University of Science and Technology
Clear Water Bay
Kowloon, Hong Kong
dekai@cs.ust.hk

Abstract

We show that, consistent with MEANT-tuned systems that translate into Chinese, MEANT-tuned MT systems that translate into English also outperforms BLEU-tuned systems across commonly used MT evaluation metrics, even in BLEU. The result is achieved by significantly improving MEANT's sentence-level ranking correlation with human preferences through incorporating a more accurate distributional semantic model for lexical similarity and a novel backoff algorithm for evaluating MT output which automatic semantic parser fails to parse. The surprising result of MEANT-tuned systems having a higher BLEU score than BLEU-tuned systems suggests that MEANT is a more accurate objective function guiding the development of MT systems towards producing more adequate translation.

1 Introduction

Lo and Wu (2013) showed that MEANT-tuned system for translating into Chinese outperforms BLEU-tuned system across commonly used MT evaluation metrics, even in BLEU. However, such phenomena are not observed in MEANT-tuned system for translating into English. In this paper, for the first time, we present MT systems for translating into English, which is tuned to a improved version of MEANT, also outperforms BLEU-tuned system across commonly used MT evaluation metrics, even in BLEU. The improvements in MEANT include incorporating more accurate distributional semantic model for lexical similarity and a novel backoff algorithm for evaluating MT output which the automatic semantic parser failed to parse. Empirical results show that the new version of MEANT is significantly improved in terms of sentence-level ranking correlation with human preferences.

The accuracy of MEANT relies heavily on the accuracy of the model that determines the lexical similarities of the semantic role fillers. However, the discrete context vector model based on the raw co-occurrence counts used in the original proposal of MEANT does not work well in predicting the similarity of the lexicons used in the reference and machine translations. Recent work by Baroni *et al.* (2014) shows that word embeddings trained by predict models outperforms the count based models in various lexical semantic tasks. Baroni *et al.* (2014) argues that *predict* models such as word2vec (Mikolov *et al.*, 2013) outperform count based models on a wide range of lexical semantic tasks. It is also common knowledge that raw co-occurrence counts do not work very well and performance can be improved when transformed by reweighing the counts for context informativeness and dimensionality reduction. In contrast to conventional word vector models, prediction based word vector models estimate the vectors directly as a supervised task, where the weights in a word vector are set to maximize the probability of the contexts in which the word is observed in the corpus (Bengio *et al.*, 2006; Collobert and Weston, 2008; Collobert *et al.*, 2011; Huang *et al.*, 2012; Mikolov *et al.*, 2013; Turian *et al.*, 2010).

In this paper, we show that MEANT's correlation with human adequacy judgments can be further improved by incorporating the word embeddings trained by the predict models. Subsequently, tuning MT system against the improved version of MEANT produce more adequate translations than tuning against BLEU.

2 The family of MEANT

MEANT and its variants (Lo *et al.*, 2012) measure weighted f-scores over corresponding seman-

*This work was completed at HKUST.

Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 434–441,
Lisboa, Portugal, 17-18 September 2015.

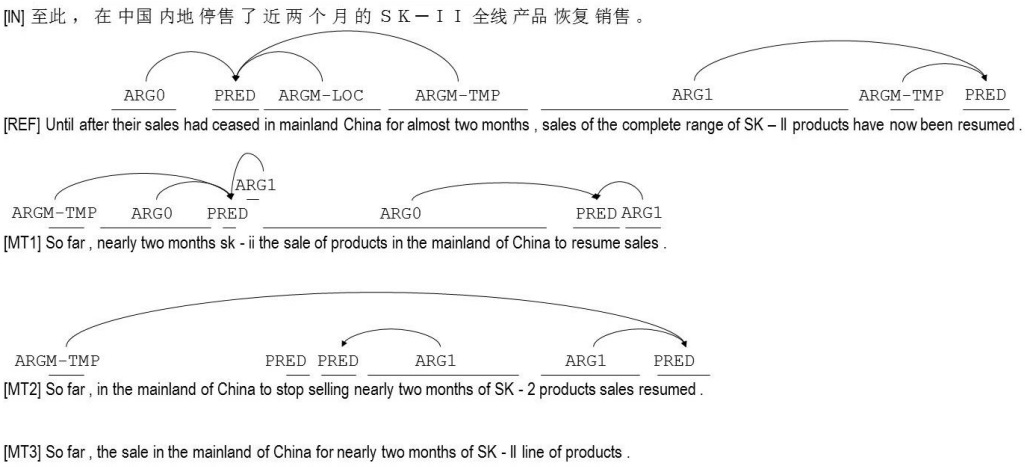

Figure 1: Examples of automatic shallow semantic parses. Both the reference and machine translations are parsed using automatic English SRL. There are no semantic frames for MT3 since there is no predicate in the MT output.

tic frames and role fillers in the reference and machine translations. MEANT typically outperforms BLEU, NIST, METEOR, WER, CDER and TER in correlation with human adequacy judgment, and is relatively easy to port to other languages, requiring only an automatic semantic parser and a monolingual corpus of the output language, which is used to train the discrete context vector model for computing the lexical similarity between the semantic role fillers of the reference and translation. Lo *et al.* (2014) describe a cross-lingual quality estimation variant, XMEANT, capable of evaluating translation quality without the need for expensive human reference translations, by utilizing semantic parses of the original foreign input sentence instead of a reference translation. MEANT is generally computed as follows:

1. Apply an automatic shallow semantic parser to both the reference and machine translations. (Figure 1 shows examples of automatic shallow semantic parses on both reference and MT.)

2. Apply the maximum weighted bipartite matching algorithm to align the semantic frames between the reference and machine translations according to the lexical similarities of the predicates.

3. For each pair of the aligned frames, apply the maximum weighted bipartite matching algorithm to align the arguments between the reference and MT output according to the lexical similarity of role fillers.

4. Compute the weighted f-score over the matching role labels of these aligned predicates and role fillers according to the following definitions:

$$
\begin{aligned}
q^0_{i,j} &\equiv \text{ARG } j \text{ of aligned frame } i \text{ in MT} \\
q^1_{i,j} &\equiv \text{ARG } j \text{ of aligned frame } i \text{ in REF} \\
w^0_i &\equiv \frac{\text{\#tokens filled in aligned frame } i \text{ of MT}}{\text{total \#tokens in MT}} \\
w^1_i &\equiv \frac{\text{\#tokens filled in aligned frame } i \text{ of REF}}{\text{total \#tokens in REF}} \\
w_{\text{pred}} &\equiv \text{weight of similarity of predicates} \\
w_j &\equiv \text{weight of similarity of ARG } j \\
\mathbf{e}_{i,\text{pred}} &\equiv \text{the pred string of the aligned frame } i \text{ of MT} \\
\mathbf{f}_{i,\text{pred}} &\equiv \text{the pred string of the aligned frame } i \text{ of REF} \\
\mathbf{e}_{i,j} &\equiv \text{role fillers of ARG } j \text{ of the aligned frame } i \text{ of MT} \\
\mathbf{f}_{i,j} &\equiv \text{role fillers of ARG } j \text{ of the aligned frame } i \text{ of REF} \\
s(e,f) &= \text{lexical similarity of token } e \text{ and } f \\
\text{prec}_{\mathbf{e},\mathbf{f}} &= \frac{\sum_{e \in \mathbf{e}} \max_{f \in \mathbf{f}} s(e,f)}{|\,\mathbf{e}\,|} \\
\text{rec}_{\mathbf{e},\mathbf{f}} &= \frac{\sum_{f \in \mathbf{f}} \max_{e \in \mathbf{e}} s(e,f)}{|\,\mathbf{f}\,|} \\
s_{i,\text{pred}} &= \frac{2 \cdot \text{prec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}} \cdot \text{rec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}}}{\text{prec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}} + \text{rec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}}} \\
s_{i,j} &= \frac{2 \cdot \text{prec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}} \cdot \text{rec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}}}{\text{prec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}} + \text{rec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}}}
\end{aligned}
$$

$$
\text{precision} = \frac{\sum_i w^0_i \frac{w_{\text{pred}} s_{i,\text{pred}} + \sum_j w_j s_{i,j}}{w_{\text{pred}} + \sum_j w_j |q^0_{i,j}|}}{\sum_i w^0_i} \quad (1)
$$

$$
\text{recall} = \frac{\sum_i w^1_i \frac{w_{\text{pred}} s_{i,\text{pred}} + \sum_j w_j s_{i,j}}{w_{\text{pred}} + \sum_j w_j |q^1_{i,j}|}}{\sum_i w^1_i} \quad (2)
$$

$$
\text{MEANT} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (3)
$$

where w_{pred} and w_j are the weights of the lexical similarities of the predicates and role fillers of the arguments of type j of all frame between the reference translations and the MT output. There is a total of 12 weights for the set of semantic role labels in MEANT as defined in Lo and Wu (2011b). The value of these weights are determined in supervised manner using a simple grid search to optimize the correlation with human adequacy judgments (Lo and Wu, 2011a) for MEANT and in unsupervised manner using relative frequency of each semantic role label in the references for UMEANT (Lo and Wu, 2012). Thus UMEANT is useful when human judgments on adequacy of the development set are unavailable. $s_{i,\text{pred}}$ and $s_{i,j}$ are the phrasal similarities of the predicates and role fillers of the arguments of type j between the reference translations and the MT output. Lo *et al.* (2012) and Tumuluru *et al.* (2012) described how the lexical similarities, $s(e, f)$, are computed using a discrete context vector model and how the phrasal similarities are computed by aggregating the lexical similarities via various heuristics. In the latest version of MEANT (Lo *et al.*, 2014), as shown in above, it uses f-score to aggregate individual token similarities into the phrasal similarities of semantic role fillers. Another MEANT's variant, IMEANT (Wu *et al.*, 2014), which uses ITG to constrain the token alignments between the semantic role fillers of the reference and the machine translations and is shown outperforming MEANT (Lo *et al.*, 2014).

3 Improvements to MEANT

We improve the performance of MEANT by incorporating a word embedding model for more accurate evaluation of the semantic role filler similarity and a novel backoff algorithm for evaluating translations when the automatic semantic parser fails to reconstruct the semantic structure of the translations. Our evaluation results show that the new version of MEANT is significantly improved in correlating with human ranking preferences at both the sentence-level and the document-level.

3.1 Discrete context vectors vs. word embeddings

MEANT's discrete context vector model is very sparse because of the extremely high dimension of the discrete context vector model. The number of dimensions of a vector in the discrete context vector model is the total number of token types in the training corpus. The vector sparsity issue makes the lexical similarity highly sensitive of exact token matching and thus hurts the accuracy of MEANT. We aim at tackling the sparse vector issue by replacing the discrete context vector model with the continuous word embeddings in order to further improve the accuracy of MEANT.

We first train the word embeddings on the same monolingual corpus as the discrete context vector model, i.e. Gigaword, for a fair comparison. However, since the memory consumption of the word embeddings is significantly reduced when comparing with the discrete context vector model due to the reduced dimension in the vectors, it is now possible to increase the size of the training corpus of the word embeddings so as to improve the token coverage of the lexical similarity model. We compare the in-house Gigaword word embeddings which covers 1.2 million words and phrases with the Google pretrained word embeddings (Mikolov *et al.*, 2013) that is trained on a 100 billion tokens news dataset and covers 3 million words and phrases. We show that the high portability of MEANT is preserved when replacing the discrete context vector model with word embeddings as the size of the monolingual training data for the word embeddings does not significantly affect the correlation of MEANT with human adequacy judgments.

Another interesting property of the word embeddings is the compositionality of words vectors into phrases. As described in Mikolov *et al.* (2013), for example, the result of linear vector calculation vec("Madrid")-vec("Spain")+vec("France") is closer to vec("Paris") than to any other vectors. It seems to be natural that phrasal similarity of the semantic role fillers could be more accurately computed using the composite phrase vector than using the align-and-aggregate approach because the vector composition approach is not affected by the errors of token misalignment. However, we show that surprisingly, the align-and-aggregate approach outperforms the naive linear word vector composition in computing the phrasal similarities of the semantic role fillers.

3.2 Backoff algorithm for evaluating translations without semantic parse

MEANT fails to evaluate the quality of the translations if the automatic semantic parser fails to reconstruct the semantic structure of the transla-

tions. According to the error analysis in Lo and Wu (2013) , the two main reasons for the automatic shallow semantic parser failing to identify the semantic frames are the failure to identify the semantic frames for copula or existential senses of "be" in a perfectly grammatical sentence and the absence of any predicate verb at all in the sentence. They showed that manually reconstructing the "be" semantic frames for MEANT yields significantly higher correlation with human adequacy judgment. Thus, we present a novel backoff algorithm for MEANT to reconstruct the "be" semantic frame and evaluate the whole sentence using the lexical similarity function and weigh it according to the ratio of unlabeled tokens in the MT/REF.

The reconstruction of the "be" semantic frame is triggered when the automatic shallow semantic parser fails to find a semantic frame in the sentence. It utilizes the syntactic parse of the sentence and labels the verb-to-be as the predicate. Then, it labels the constituent of the NP subtree sibling immediate left to the predicate as the "who" role, the constituent of the NP subtree sibling immediate right to the predicate as the "what" role and any constituent of other subtree siblings of the predicate as "other" role. The reconstructed "be" frame is then evaluated the same way as other semantic frames using MEANT.

When there is no predicate verb in the whole sentence, we evaluate the whole sentence using the lexical similarity function and weighted according to the amount of unlabeled tokens in the MT/REF. Thus, equation (1), (2) and (3) are replaced by equation (4), (5) and (6).

$$
\begin{aligned}
w^0_{\text{nf}} &\equiv \frac{\text{\#tokens that are not fillers of any role in MT}}{\text{total \#tokens in MT}} \\
w^1_{\text{nf}} &\equiv \frac{\text{\#tokens that are not fillers of any role in REF}}{\text{total \#tokens in REF}} \\
\mathbf{e}_{\text{sent}} &\equiv \text{the whole sentence string of MT} \\
\mathbf{f}_{\text{sent}} &\equiv \text{the whole sentence string of REF} \\
s_{\text{sent}} &= \frac{2 \cdot \text{prec}_{\mathbf{e}_{\text{sent}},\mathbf{f}_{\text{sent}}} \cdot \text{rec}_{\mathbf{e}_{\text{sent}},\mathbf{f}_{\text{sent}}}}{\text{prec}_{\mathbf{e}_{\text{sent}},\mathbf{f}_{\text{sent}}} + \text{rec}_{\mathbf{e}_{\text{sent}},\mathbf{f}_{\text{sent}}}}
\end{aligned}
$$

$$
\text{precision} = \frac{\sum_i w^0_i \frac{w_{\text{pred}} s_{i,\text{pred}} + \sum_j w_j s_{i,j}}{w_{\text{pred}} + \sum_j w_j |q^0_{i,j}|} + w^0_{\text{nf}} s_{\text{sent}}}{\sum_i w^0_i + w^0_{\text{nf}}} \quad (4)
$$

$$
\text{recall} = \frac{\sum_i w^1_i \frac{w_{\text{pred}} s_{i,\text{pred}} + \sum_j w_j s_{i,j}}{w_{\text{pred}} + \sum_j w_j |q^1_{i,j}|} + w^1_{\text{nf}} s_{\text{sent}}}{\sum_i w^1_i + w^1_{\text{nf}}} \quad (5)
$$

$$
\text{MEANT} = \frac{\text{precision} \cdot \text{recall}}{\alpha \cdot \text{precision} + (1-\alpha) \cdot \text{recall}} \quad (6)
$$

Note that we have also introduced the weight α for the precision and recall. Later, we show that optimal value of α for MT evaluation is different from that for MT optimization.

3.3 Results

Table 1 shows the document-level Pearson's score correlation and table 2 shows the sentence-level Kendall's rank correlation with human preferences of the improved version of MEANT with the previous version of MEANT (Lo *et al.*, 2014) on WMT2014 metrics task test set (Macháček and Bojar, 2014). For the sake of stable performance across all the tested language pairs, the weights of the semantic role labels are estimated in unsupervised manner.

First and the most importantly, the document-level score correlation with human preferences of all versions of MEANT consistently outperforms all the submitted metrics in Macháček and Bojar (2014). While the variations on document-level correlation with human preferences of different versions of MEANT are not significant, we focus on discussing about the sentence-level results.

On sentence-level ranking, MEANT with Gigaword word embeddings correlates significantly better with human preference than MEANT with Gigaword discrete context vectors. Although the Google pretrained word embeddings covers more than twice as many token types as the Gigaword word embeddings, our results show that MEANT incorporated with the Google pretrained word embeddings only marginally better that incorporated with the Gigaword word embeddings. Our results show that MEANT's portability to languages with lower resources is preserved as MEANT with Gigaword word embeddings achieves comparable accuracy without using huge amount of resources.

While the linear vector composition property of word embeddings receive a lot of attention recently, our results show that, surprisingly, MEANT with word embeddings using the align-and-aggregate approach in computing the phrasal similarities significantly outperforms that using the simple linear vector composition across all language pairs in the test set. Our results suggest that more investigation on using word embeddings is necessary for it to be useful for efficient evaluation of phrasal similarities.

Our results also show that MEANT with an α value of 1, i.e. recall only, significantly outperforms that with balanced precision and recall weighting, in correlation with human preferences. This could be due to the fact that MT sys-

Table 1: System-level Pearson's score correlation with human preferences of MEANT on WMT2014 metrics track test set

metric	cs-en	de-en	fr-en	hi-en	ru-en	ave.
MEANT (Lo *et al.*, 2014) (i.e. α=0.5)						
+ Gigaword discrete context vectors & fillers alignment	0.975	0.973	0.972	0.957	0.877	0.951
+ Gigaword word embeddings & fillers alignment	0.939	0.967	0.979	0.948	0.912	0.949
+ Google pretrained word embeddings & vector composition	0.919	0.955	0.981	0.941	0.940	0.947
+ Google pretrained word embeddings & fillers alignment	0.948	0.970	0.979	0.950	0.922	0.954
MEANT (α=1)						
+ Google pretrained word embeddings & fillers alignment	0.990	0.965	0.977	0.921	0.909	0.952
+backoff	0.986	0.970	0.981	0.947	0.915	0.960

Table 2: Sentence-level Kendall's rank correlation with human preferences of MEANT on WMT2014 metrics track test set

metric	cs-en	de-en	fr-en	hi-en	ru-en	ave.
MEANT (Lo *et al.*, 2014) (i.e. α=0.5)						
+ Gigaword discrete context vectors & fillers alignment	0.188	0.209	0.235	0.229	0.193	0.211
+ Gigaword word embeddings & fillers alignment	0.192	0.235	0.252	0.230	0.206	0.223
+ Google pretrained word embeddings & vector composition	0.195	0.222	0.242	0.231	0.201	0.218
+ Google pretrained word embeddings & fillers alignment	0.206	0.229	0.253	0.236	0.214	0.228
MEANT (α=1)						
+ Google pretrained word embeddings & fillers alignment	0.229	0.257	0.285	0.243	0.239	0.251
+ backoff	0.267	0.301	0.336	0.324	0.266	0.299

tems tend to under-generate (i.e. missing meaning in the translation output) rather than over-generate. This also explains why the precision-oriented metrics, such as BLEU, usually correlate poorly with human adequacy judgments.

Lastly, our results show that the novel backoff algorithm significantly improves MEANT's correlation with human preferences.

4 Tuning against the new MEANT

Lo *et al.* (2013b) show that for MT system translating into Chinese, tuning against MEANT outperforms the common practice of tuning against BLEU or TER across commonly used MT evaluation metrics, i.e. beating BLEU-tuned systems in BLEU and TER-tuned systems in TER. However, for MT system translating into English, previous work (Lo *et al.*, 2013a; Lo and Wu, 2013) show that tuning against MEANT only achieves balanced performance in both n-gram based metrics and edit distance based metrics, without overfitting to either type of metrics. We argue with the significant improvement in sentence-level correlation with human preferences in evaluating translations in English, the performance of MT system tuned against the newly improved MEANT would also improved.

For WMT2015 tuning task, we tuned the basic Czech-English baseline system against the newly improved MEANT using the official development set and k-best MERT (with 100-best hypothesis list). Unfortunately, there is a bug in the integration of MEANT and Moses k-best MERT in the submitted system. Table 3 and 4 shows the results of both the submitted buggy system and the debugged version of the experiments on the official dev and test test.

In the previous section, MEANT with an α value of 1, i.e. 100% recall, has the highest correlation with human preferences on the test set. However, surprisingly, our tuning experiment results show that tuning against a balanced precision-recall version of MEANT yields better scores across the commonly used MT evaluation metrics. This is because the optimization algorithm needs the guidance from precision to avoid blindly generating too many words which would achieve high recall.

More importantly, our results show that MT system tuning against the improved MEANT beats the BLEU-tuned system across the commonly used MT evaluation metrics, even in BLEU.

5 Related Work

Most of the common used MT evaluation metrics like BLEU (Papineni *et al.*, 2002), NIST (Doddington, 2002), CDER (Leusch *et al.*, 2006), WER (Nießen *et al.*, 2000), and TER (Snover *et al.*, 2006) rely heavily on the exact match of the surface form of the tokens in the reference and the MT output. Thus, they do not only fail to capture the

Table 3: Translation quality of MT system tuned against MEANT and BLEU on WMT15 tuning task dev set. MEANT reported here is the version using Google pretrained word embeddings with α=1 and backoff algorithm.

system	BLEU	NIST	WER	PER	CDER	TER	MEANT
BLEU-tuned	19.38	6.48	67.63	50.48	58.17	63.57	42.77
MEANT-tuned (official submitted buggy system)	18.20	6.27	70.09	51.84	59.93	65.53	42.23
MEANT-tuned (α=1)	18.96	6.44	68.41	50.77	58.74	64.30	43.43
MEANT-tuned (α=0.5)	**19.74**	**6.62**	**66.31**	**49.22**	**57.20**	**62.28**	**43.62**

Table 4: Translation quality of MT system tuned against MEANT and BLEU on WMT15 tuning task test set. MEANT reported here is the version using Google pretrained word embeddings with α=1 and backoff algorithm.

system	BLEU	NIST	WER	PER	CDER	TER	MEANT
BLEU-tuned	17.06	5.99	69.67	52.86	59.85	65.71	40.10
MEANT-tuned (official submitted buggy system)	15.89	5.80	71.82	53.93	61.43	67.59	39.34
MEANT-tuned (α=1)	16.75	5.95	70.19	53.05	60.29	66.25	40.12
MEANT-tuned (α=0.5)	**17.15**	**6.08**	**68.53**	**52.03**	**59.07**	**64.65**	**40.23**

meaning similarities of lexicons that do not share the same surface form, but also ignore the meaning structures of the translations.

METEOR (Banerjee and Lavie, 2005; Denkowski and Lavie, 2014) evaluates lexical similarities beyond surface-form by incorporating a large collection of linguistic resources, like synonym table from hand-crafted WordNet and paraphrase table learned from large parallel corpus. Another trend of improving MT evaluation metrics is incorporating the evaluation of meaning structure of the translations. Owczarzak *et al.* (2007a,b) improved the correlation with human *fluency* judgments by using LFG to extend the approach of evaluating syntactic dependency structure similarity in Liu and Gildea (2005), but did not improve the correlation with human *adequacy* judgments when comparing to METEOR. Similarly, TINE, an automatic recall-oriented basic meaning event structured based evaluation metric (Rios *et al.*, 2011) correlated with human adequacy judgment comparable to that of BLEU but not as high as that of METEOR. ULC (Giménez and Màrquez, 2007, 2008) incorporates several semantic similarity features and shows improved correlation with human judgement of translation quality (Callison-Burch *et al.*, 2007; Giménez and Màrquez, 2007; Callison-Burch *et al.*, 2008; Giménez and Màrquez, 2008) but no work has been done towards tuning an MT system using a pure form of ULC perhaps due to its expensive run time.

By incorporating word embeddings into MEANT, translations are evaluated via both the structural and lexical semantics accurately and thus, MT system tuned against the improved MEANT beats BLEU-tuned system across commonly used metrics, even in BLEU.

6 Conclusion

In this paper we presented the first results of using word embeddings to improve the correlation with human adequacy judgments of MEANT, the state-of-the-art semantic MT evaluation metric. We also showed that using a smaller and easy-to-obtain monolingual corpus (e.g., Gigaword, Wikipedia) for training the word embeddings does not significantly affect the accuracy of MEANT. We showed that the align-and-aggregate approach outperforms the naive linear word vector composition, although the compositional property is highly advertised as the advantage of using word embeddings. We also described a novel backoff algorithm in MEANT for evaluating the meaning accuracy of the MT output when automatic shallow semantic parser fails to parse the sentence. In this tuning shared task, we successfully integrate MEANT with the Moses framework. This enable further investigation into tuning MT system against MEANT using newer tuning techniques and features. Most importantly, we show that tuning MT system against the improved version of MEANT outperforms BLEU-tuned system across all commonly used MT evaluation metrics, even in BLEU.

Acknowledgements

This material is based upon work supported in part by the Defense Advanced Research Projects Agency (DARPA) under BOLT contract nos. HR0011-12-C-0014 and HR0011-12-C-0016, and GALE contract nos. HR0011-06-C-0022 and HR0011-06-C-0023; by the European Union under the FP7 grant agreement no. 287658; 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.

References

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.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, 2014.

Yoshua Bengio, Holger Schwenk, Jean-Sébastien Senécal, Fréderic Morin, and Jean-Luc Gauvain. Neural probabilistic language models. In *Innovations in Machine Learning*, pages 137–186. Springer, 2006.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. (meta-) evaluation of machine translation. In *Second Workshop on Statistical Machine Translation (WMT-07)*, 2007.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. Further meta-evaluation of machine translation. In *Third Workshop on Statistical Machine Translation (WMT-08)*, 2008.

Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537, 2011.

Michael Denkowski and Alon Lavie. METEOR universal: Language specific translation evaluation for any target language. In *9th Workshop on Statistical Machine Translation (WMT 2014)*, 2014.

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.

Jesús Giménez and Lluís Màrquez. Linguistic features for automatic evaluation of heterogenous MT systems. In *Second Workshop on Statistical Machine Translation (WMT-07)*, pages 256–264, Prague, Czech Republic, June 2007.

Jesús Giménez and Lluís Màrquez. A smorgasbord of features for automatic MT evaluation. In *Third Workshop on Statistical Machine Translation (WMT-08)*, Columbus, Ohio, June 2008.

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882. Association for Computational Linguistics, 2012.

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.

Ding Liu and Daniel Gildea. Syntactic features for evaluation of machine translation. In *Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, Michigan, June 2005.

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.

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.

Chi-kiu Lo and Dekai Wu. Unsupervised vs. supervised weight estimation for semantic MT evaluation metrics. In *Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-6)*, 2012.

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.

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

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.

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.

Chi-kiu Lo, Meriem Beloucif, Markus Saers, and Dekai Wu. XMEANT: Better semantic MT evaluation without reference translations. In *52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, 2014.

Matouš Macháček and Ondřej Bojar. Results of the WMT14 metrics shared task. In *Ninth Workshop on Statistical Machine Translation (WMT 2014)*, Baltimore, Maryland USA, June 2014.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

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.

Karolina Owczarzak, Josef van Genabith, and Andy Way. Dependency-based automatic evaluation for machine translation. In *Syntax and Structure in Statistical Translation (SSST)*, 2007.

Karolina Owczarzak, Josef van Genabith, and Andy Way. Evaluating machine translation with LFG dependencies. *Machine Translation*, 21:95–119, 2007.

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, Philadelphia, Pennsylvania, July 2002.

Miguel Rios, Wilker Aziz, and Lucia Specia. TINE: A metric to assess MT adequacy. In *Sixth Workshop on Statistical Machine Translation (WMT 2011)*, 2011.

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)*, pages 223–231, Cambridge, Massachusetts, August 2006.

Anand Karthik Tumuluru, Chi-kiu Lo, and Dekai Wu. Accuracy and robustness in measuring the lexical similarity of semantic role fillers for automatic semantic MT evaluation. In *26th Pacific Asia Conference on Language, Information, and Computation (PACLIC 26)*, 2012.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394. Association for Computational Linguistics, 2010.

Dekai Wu, Chi-kiu Lo, Meriem Beloucif, and Markus Saers. Better semantic frame based mt evaluation via inversion transduction grammars. 2014. SSST.