YiSi - A unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources

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

Abstract

We present YiSi, a unified automatic semantic machine translation quality evaluation and estimation metric for languages with different levels of available resources. Underneath the interface with different language resources settings, YiSi uses the same representation for the two sentences in assessment. Besides, we show significant improvement in the correlation of YiSi-1's scores with human judgment is made by using contextual embeddings in multilingual BERT–Bidirectional Encoder Representations from Transformers to evaluate lexical semantic similarity. YiSi is open source and publicly available.

1 Introduction

A good automatic MT quality metric is one that closely reflect the usefulness of the translation, in terms of assisting human readers to understand the meaning of the input sentence. BLEU (Papineni et al., 2002) has long been shown not to correlate well with human judgment on translation quality (Machacek and Bojar, 2014; Stanojević et al., 2015; Bojar et al., 2016, 2017; Ma et al., 2018). However, it is still the most commonly used metric for reporting quality of machine translation systems. One of the major reasons is that BLEU is ready-to-deploy to all languages due to its simplicity. Semantic MT evaluation metrics, such as METEOR (Denkowski and Lavie, 2014) and MEANT (Lo, 2017), require additional linguistic resources to more accurately evaluate the meaning similarity between the MT output and the reference translation. The lower portability hinders the wide adoption of these metrics.

We, therefore, propose a unified framework, YiSi, for MT quality evaluation and estimation that take advantage of both metric paradigms by providing options to fallback to surface-level lexical similarity when semantic models are not available for the languages in assessment.

YiSi were first used in WMT 2018 metrics shared task (Ma et al., 2018) and performed well and consistently at segment-level across the tested language pairs in correlating with human judgment. An YiSi based system successfully served in WMT2018 parallel corpus filtering task (Lo et al., 2018).

This year, instead of using `word2vec` (Mikolov et al., 2013) to evaluate lexical semantic similarity in YiSi, we use BERT –Bidirectional Encoder Representation from Transformers (Devlin et al., 2018). YiSi is open source and publicly available.[1]

2 YiSi

YiSi[2] is a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. Inspired by MEANT (Lo, 2017), YiSi-1 is a MT quality evaluation metric that measures the similarity between a machine translation and human references by aggregating the weighted distributional lexical semantic similarities and optionally incorporating shallow semantic structures. YiSi-0 is the degenerate version of YiSi-1 that is ready-to-deploy to any languages. It uses longest common character substring to measure the lexical similarity. YiSi-2 is the bilingual, reference-less version, which uses bilingual embeddings to evaluate crosslingual lexical semantic similarity between the input and MT output. Like YiSi-1, YiSi-2 can exploit shallow semantic structures as well.

YiSi-0 and YiSi-1 were first used in WMT 2018 metrics shared task (Ma et al., 2018) and performed well and consistently at segment-level

[1] http://chikiu-jackie-lo.org/home/index.php/yisi

[2] YiSi is the romanization of the Cantonese word 意思 ('meaning').

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers (Day 1) pages 507–513
Florence, Italy, August 1-2, 2019.

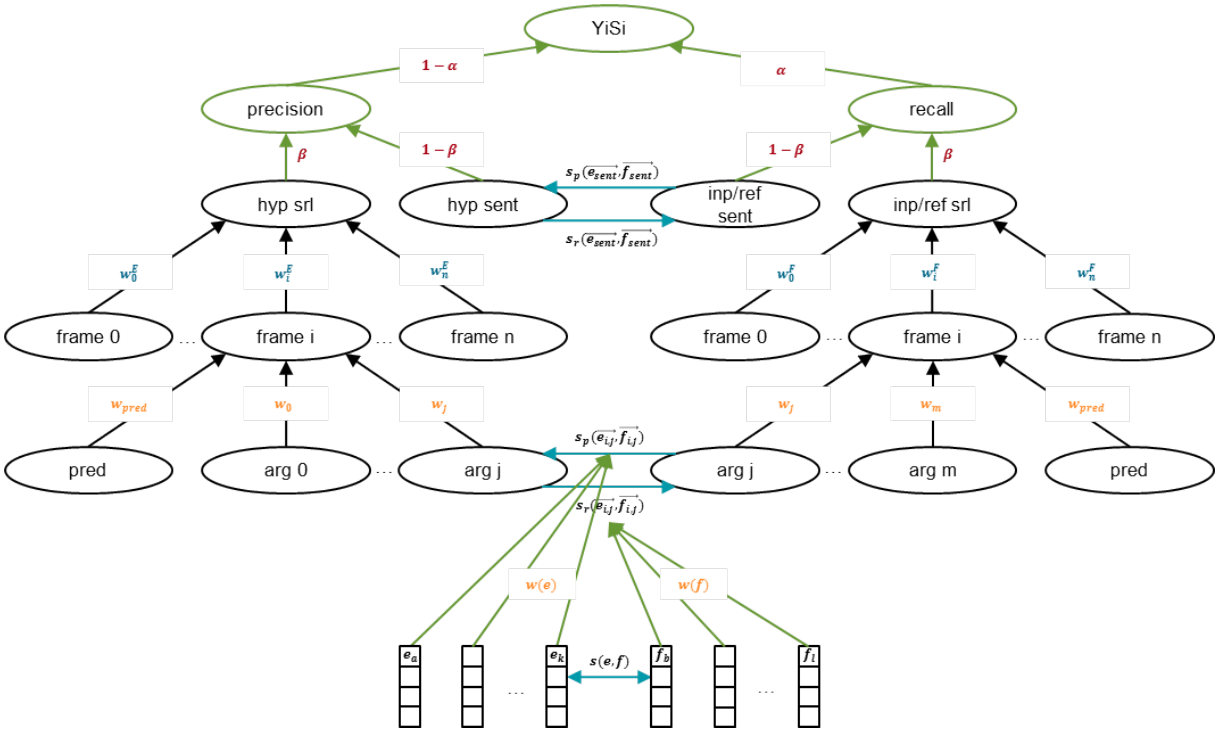

Figure 1: Graphical representation of the computation of YiSi.

across the tested language pairs in correlating with human judgment. While YiSi-1 also successfully served in WMT2018 parallel corpus filtering task, YiSi-2 showed comparable accuracy in our internal experiments (Lo et al., 2018).

2.1 Overview

Following the guiding principle that a good MT quality metric reflects how well human readers understand the meaning of the input sentence, YiSi is the weighted f-scores over corresponding semantic frames and role fillers in the two sentences E and F in assessment. The procedure of computing YiSi is described as follow:

1. Apply a shallow semantic parser to both E and F.

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

3. For each pair of aligned frames, apply the maximum weighted bipartite matching algorithm to align the arguments between E and F 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: (Figure 1 is the graphical representation of the following computation.)

$$\begin{aligned} w\left(e\right) &= \text{lexical weight of } e \\ s(e,f) &= \text{lexical similarity of } e \text{ and } f \end{aligned}$$

where $s(e, f)$ is the lexical similarity and it is weighted by $w(e)$ and $w(f)$ for computing phrasal precision and recall respectively. Different variants of YiSi have different definition of lexical similarities and weights depend on the resources available for the assessment settings. By aggregating the weighted lexical similarities into n-gram similarities, we then align the bag of n-grams in the two sentences using maximum alignment on the n-gram similarities. The phrasal similarity precision, s_p, and recall, s_r, (as defined below) are the weighted average of the similarities of the aligned n-gram.

$$s_p(\vec{e}, \vec{f}) = \frac{\sum_a \max_b \sum_{k=0}^{n-1} w\left(e_{a+k}\right) \cdot s\left(e_{a+k}, f_{b+k}\right)}{\sum_a \sum_{k=0}^{n-1} w\left(e_{a+k}\right)}$$

$$s_r(\vec{e}, \vec{f}) = \frac{\sum_b \max_a \sum_{k=0}^{n-1} w\left(f_{b+k}\right) \cdot s\left(e_{a+k}, f_{b+k}\right)}{\sum_b \sum_{k=0}^{n-1} w\left(f_{b+k}\right)}$$

With the phrasal semantic precision and recall, we compute the structural semantic precision and recall as follow:

$$
\begin{aligned}
q^E_{i,j} &= \text{argument } j \text{ of aligned frame } i \text{ in } E \\
q^F_{i,j} &= \text{argument } j \text{ of aligned frame } i \text{ in } F \\
w^E_i &= \frac{\text{\#units filled in aligned frame } i \text{ of } E}{\text{total \#units in } E} \\
w^F_i &= \frac{\text{\#units filled in aligned frame } i \text{ of } F}{\text{total \#units in } F}
\end{aligned}
$$

$$
\begin{aligned}
w_j &= \text{count}\,(\text{argument } j \text{ in } \mathbb{F}) \\
w_t &= 0.25 * \text{count}\,(\text{predicate in } \mathbb{F}) \\
\text{srl}_p &= \frac{\sum_i w^e_i \frac{w_t s_p(\overrightarrow{e_{i,t}}, \overrightarrow{f_{i,t}}) + \sum_j w_j s_p(\overrightarrow{e_{i,j}}, \overrightarrow{f_{i,j}})}{w_t + \sum_j w_j |q^e_{i,j}|}}{\sum_i w^e_i} \\
\text{srl}_r &= \frac{\sum_i w^f_i \frac{w_t s_r(\overrightarrow{e_{i,t}}, \overrightarrow{f_{i,t}}) + \sum_j w_j s_r(\overrightarrow{e_{i,j}}, \overrightarrow{f_{i,j}})}{w_t + \sum_j w_j |q^f_{i,j}|}}{\sum_i w^f_i}
\end{aligned}
$$

where w_t is the weight of the lexical similarities of the aligned predicates in step 2. w_j is the weight of the phrasal similarities of the role fillers of the arguments of role type j of the aligned frames between the reference translations and the MT output in step 3 if their role types are matching. As in (Lo, 2017), we merge the semantic role labels into 8 role types (who, did, what, whom, when, where, why, how) for more robust performance. Thus, there is a total of 8 weights for the set of semantic role types in YiSi estimated by type counts in the document $\mathbb{F}$. The frame precision/recall is the weighted sum of the phrasal precision/recall of the aligned role fillers. The token coverage w^e_i and w^f_i estimate the importance of frame i in the sentence E and F. The structural semantic precision and recall is the weighted average of all the aligned frames in sentence E and F respectively.

Now, the overall precision and recall is the weighted sum of the phrasal precision and recall of the whole sentence of $\overrightarrow{e_{\text{sent}}}$ and $\overrightarrow{f_{\text{sent}}}$, like in the following:

$$
\begin{aligned}
\text{precision} &= \beta \cdot \text{srl}_p + (1-\beta) \cdot s_p(\overrightarrow{e_{\text{sent}}}, \overrightarrow{f_{\text{sent}}}) \\
\text{recall} &= \beta \cdot \text{srl}_r + (1-\beta) \cdot s_r(\overrightarrow{e_{\text{sent}}}, \overrightarrow{f_{\text{sent}}})
\end{aligned}
$$

It is important to note that the weight β should *NOT* be interpreted as the importance of the structural semantic similarity in YiSi because there is a huge overlap in the structural semantic similarity and the phrasal semantic similarity. Instead, we should pay attention to the significant difference in the performance of YiSi with and without structural semantic similarity, especially in YiSi-2, the crosslingual variant. In this experiment, β is set to 0.1.

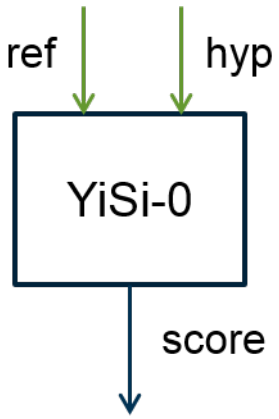

Figure 2: Resources used in YiSi-0.

Finally, the weight α for the precision and recall is introduced for different usages of YiSi. α should be set to 0.7 to make YiSi more recall-oriented when it is used for MT evaluation. When used for MT system optimization, α should be set to 0.5 to balance precision and recall.

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

In the following, we describe how we estimate the lexical similarity $s(e, f)$ and lexical weight $w(e)$ under different resource conditions.

2.1.1 YiSi-0: quality evaluation metric for extremely low resource languages

YiSi-0 is the degenerate resource-free variant of YiSi for MT quality evaluation, where sentence E is the MT output and sentence F is the reference. Figure 2 shows the resources used in YiSi-0.

YiSi-0 uses the longest common character substring accuracy to evaluate lexical similarity between the MT output and human reference. Since the MT output and the human reference are both in the same language, the lexical weight $w(e)$ of word e in the translation and the lexical weight $w(f)$ of word f in the reference are both estimated by the inverse-document-frequency of those words in the reference document $\mathbb{F}$. Thus, formally YiSi-

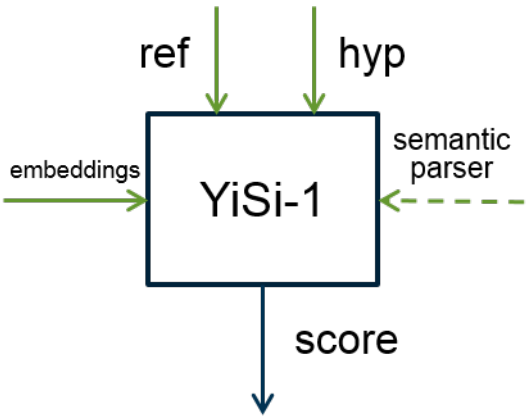

Figure 3: Resources used in YiSi-1. The dash arrow means that the semantic parser is optional.

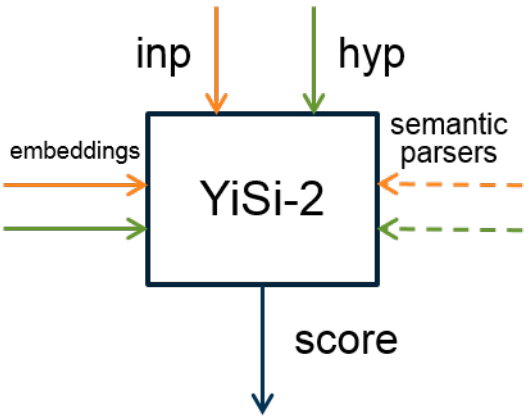

Figure 4: Resources used in YiSi-2. Arrows in green depict resources in target language and arrows in orange depict resources in source language. The dash arrows mean that the semantic parsers are optional.

0 is defined as follow:

$$
\begin{aligned}
l(e,f) &= \text{longest common substring of } e \text{ and } f \\
s_0(e,f) &= \frac{2 * l(e,f)}{|e| + |f|} \\
w(e) &= idf(e) = log(1 + \frac{|\mathbb{F}| + 1}{|\mathbb{F}_{\exists e}| + 1}) \\
\text{YiSi-0} &= \text{YiSi}(s{=}s_0, \beta{=}0.0, E{=}\text{MT}, F{=}\text{REF})
\end{aligned}
$$

2.1.2 YiSi-1: quality evaluation metric with access to an embedding model

YiSi-1 is the monolingual variant of YiSi for MT quality evaluation, where sentence E is the MT output and sentence F is the reference. Figure 3 shows the resources used in YiSi-1.

YiSi-1 requires an embedding model to evaluate lexical semantic similarity and optionally requires a semantic role labeler in the output language for evaluating structural semantic similarity. The lexical semantic similarity is the cosine similarity of the embeddings from the lexical representation model. Similar to YiSi-0, the lexical weight $w(u)$ of word unit u in the MT and the reference are estimated by the inverse-document-frequency of that word in the reference document $\mathbb{F}$. Thus, formally YiSi-1 is defined as follow:

$$
\begin{aligned}
v(u) &= \text{embedding of unit } u \\
s_1(e,f) &= cos(v(e), v(f)) \\
w(u) &= idf(u) = log(1 + \frac{|\mathbb{F}| + 1}{|\mathbb{F}_{\exists u}| + 1}) \\
\text{YiSi-1} &= \text{YiSi}(s{=}s_1, \beta{=}0.0, E{=}\text{MT}, F{=}\text{REF}) \\
\text{YiSi-1\_srl} &= \text{YiSi}(s{=}s_1, \beta{=}0.1, E{=}\text{MT}, F{=}\text{REF})
\end{aligned}
$$

2.1.3 YiSi-2: quality estimation metric for languages with access to a bilingual embedding model

YiSi-2 is the cross-lingual variant of YiSi for MT quality estimation, where sentence E is the MT output and sentence F is the input. Figure 4 shows the resources used in YiSi-2.

YiSi-2 requires a cross-lingual embedding model for evaluating cross-lingual lexical semantic similarity and optionally requires a semantic role labeler in both the input and the output languages for evaluating structural semantic similarity. The lexical semantic similarity is the cosine similarity of the embeddings from the cross-lingual lexical representation model. The lexical weight $w(e)$ of word unit e in the MT is estimated by the inversion-document-frequency of the word in the MT document $\mathbb{E}$ while the lexical weight $w(f)$ of word unit f in the MT is estimated by the inversion-document-frequency of the word in the MT document $\mathbb{F}$. Thus, formally YiSi-2 is defined as follow:

$$
\begin{aligned}
v(u) &= \text{embedding of unit } u \\
s_2(e,f) &= cos(v(e), v(f)) \\
w(e) &= idf(e) = log(1 + \frac{|\mathbb{E}| + 1}{|\mathbb{E}_{\exists e}| + 1}) \\
w(f) &= idf(f) = log(1 + \frac{|\mathbb{F}| + 1}{|\mathbb{F}_{\exists f}| + 1}) \\
\text{YiSi-2} &= \text{YiSi}(s{=}s_2, \beta{=}0.0, E{=}\text{MT}, F{=}\text{IN}) \\
\text{YiSi-2\_srl} &= \text{YiSi}(s{=}s_2, \beta{=}0.1, E{=}\text{MT}, F{=}\text{IN})
\end{aligned}
$$

2.2 Using BERT for lexical unit semantic similarity

In WMT 2018 metrics shared task, YiSi-1 uses `word2vec` (Mikolov et al., 2013) to evaluate lexical semantic similarity between the MT output and the human reference at word level. The shortcomings of this kind of static embedding models (also including but not limited to `GloVe` (Pennington et al., 2014)) is that they provide the same embedding representation for the same word without reflecting context of different sentences. In contrast, BERT (Devlin et al., 2018) uses a bidirectional transformer encoder (Vaswani et al., 2017) to capture the sentence context in the output embeddings (at subword unit level), such that the embedding for the same word/subword unit in different sentences would be different and better represented in the embedding space. Zhang et al. (2019) provided an extensive study on the performance of the output embeddings of difference layers of BERT model in correlation with human adequacy. Following the recommendation from their studies, we use embeddings extracted from BERT models with the following settings:

- the 18th layer of the pretrained English cased BERT-Large model to represent the subword units in the reference and MT output in English for computing YiSi-1;
- the 9th layer of the pretrained Chinese BERT-Base model to represent the characters in the reference and MT output in Chinese for computing YiSi-1; and
- the 9th layer of the pretrained multilingual cased BERT-Base model to represent the subword units in the reference and MT output in languages other than Chinese and English for computing YiSi-1 and to represent the subword units in the original input and MT output in all language pairs for computing YiSi-2.

2.3 Using MATE/MATEPLUS for structural semantic similarity

There are a handful of shallow semantic parsers available publicly. `mate-tools` (Björkelund et al., 2009) is an SVM classifier based on features extracted from a dependency parse. Its successor `mateplus` (Roth and Woodsend, 2014) also uses features extracted from distributional word embeddings. `mate-tools` and `mateplus` are integrated into YiSi because of their support for languages other than English. We use `mateplus` for German's and English's semantic role labeling and `mate-tools` for Chinese's semantic role labeling.

3 Experiments and results

We use WMT 2018 metrics task evaluation set (Ma et al., 2018) for our development experiments.

The official human judgments of translation quality in WMT 2018 were collected using direct assessment. The direct assessment evaluation protocol in WMT2018 gave the annotators the reference and a MT output and asked them to evaluate the translation adequacy of the MT output on an absolute scale.

Due to space limitations, we only report the results of YiSi, chrF (Popović, 2015), BLEU and the best correlation in each of the individual language pairs. Since we use exactly the same correlation analysis as the official task for each of the test sets, our reported results are directly comparable with those reported in the task's overview paper. We summarize our observations in the following sections.

3.1 Correlation with human judgment at system-level

Table 1 shows the Pearson's correlation with WMT 2018 official aggregated human direct assessment of translation adequacy at system-level.

YiSi-0 performs more stably than chrF and BLEU in correlating with human on translation quality across all translation directions. YiSi-0 achieves comparable results with chrF and BLEU in most of the translation directions while significantly outperforms chrF and BLEU in correlating with human in evaluating Turkish-English and English-Turkish translations.

YiSi-1 beats all the WMT2018 participants in correlation with human at system level for evaluating Czech-English, German-English, Chines-English, English-German, English-Estonian, English-Finnish and English-Russian translations. In addition, YiSi-1_srl further improves YiSi-1's correlation with human at system level for evaluating German-English, Chinese-English translations.

For the quality estimation variants, YiSi-2 achieves reasonably good results (with less than

input lang.	cs	de	et	fi	ru	tr	zh	en	en	en	en	en	en	en
output lang.	en	en	en	en	en	en	en	cs	de	et	fi	ru	tr	zh
individual best	.981	.997	**.991**	**.996**	**.995**	**.958**	.982	**.999**	.991	.984	.974	.992	**.990**	**.983**
chrF	.966	.994	.981	.987	,990	.452	.960	.990	.990	.981	.969	.989	.948	.944
BLEU	.970	.971	.986	.973	.979	.657	.978	.995	.981	.975	.962	.983	.826	.947
YiSi-0	.962	.995	.982	.986	.985	.857	.972	.984	.989	.984	.954	.989	.980	.956
YiSi-1	**.990**	.998	.986	.994	.993	.830	.988	.993	**.995**	**.988**	**.979**	**.993**	.929	.977
YiSi-1_srl	.989	**.999**	.987	.993	.993	.793	**.989**	–	**.995**	–	–	–	–	.976
Quality estimation as a metric														
YiSi-2	.919	.946	.865	.927	.566	.061	.797	.710	.862	.156	.475	.204	.389	.417
YiSi-2_srl	–	.948	–	–	–	–	.781	–	.902	–	–	–	–	.472

Table 1: Pearson's correlation of the metric scores with WMT 2018 aggregated human direct assessment scores at system-level.

input lang.	cs	de	et	fi	ru	tr	zh	en	en	en	en	en	en	en
output lang.	en	en	en	en	en	en	en	cs	de	et	fi	ru	tr	zh
individual best	.347	.498	.368	.273	.311	.259	.218	.518	.696	.573	.525	.407	**.418**	.323
chrF	.288	.479	.328	.229	.269	.210	.208	.516	.677	.572	.520	.383	.409	.328
sentBLEU	.233	.415	.285	.154	.228	.145	.178	.389	.320	.414	.355	.330	.261	.311
YiSi-0	.308	.480	.330	.210	.284	.213	.216	.454	.670	.530	.468	.396	.362	.316
YiSi-1	.391	**.544**	**.397**	.299	**.352**	**.301**	**.254**	**.548**	**.734**	**.599**	**.549**	**.427**	.402	**.371**
YiSi-1_srl	**.396**	.543	.390	**.303**	.351	.297	.253	–	.719	–	–	–	–	.368
Quality estimation as a metric														
YiSi-2	.014	.279	.186	.151	.088	.066	.091	-.043	.359	.106	.172	.061	.103	.101
YiSi-2_srl	–	.281	–	–	–	–	.085	–	.380	–	–	–	–	.103

Table 2: Kendall's correlation of metric scores with the rankings at segment-level human direct assessment in WMT 2018.

0.1 degradation in correlation with human) in evaluating Czech-English , German-English, Finnish-English translation without using the human translation as reference. At the same time, YiSi-2_srl improves YiSi-2's correlation with human at system level for evaluating English-German, English-Chinese translations.

3.2 Correlation with human judgment at segment-level

Table 2 shows the Kendall's correlation with the rankings at segment-level human direct assessment obtained in the WMT 2018.

YiSi-0 achieves comparable results with chrF and BLEU for evaluating all translation directions at segment level. YiSi-1 beats all the WMT2018 participants in correlation with human at segment level for evaluating almost all translation directions, except English-Turkish. In addition, YiSi-1_srl further improves YiSi-1's correlation with human at segment level for evaluating Czech-English and Finnish-English translations.

For the quality estimation variants, YiSi-2 performs significantly worse than YiSi-1 due to the lacking of a reference translation in the same language for evaluating fluency. Therefore, We can see that as shown by the significant improvement in YiSi-2_srl for evaluating English-German translation without reference translation, using semantic parsers to extract the semantic frames of the input sentence and machine translation become very helping in evaluating translation fluency.

4 Conclusion

We have presented the on-going work in developing a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. Initial experiment results show that the improved variants of YiSi that use BERT contextual embeddings correlate with human judgment significantly better than other trained metrics.

References

Anders Björkelund, Love Hafdell, and Pierre Nugues. 2009. Multilingual semantic role labeling. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 43–48, Boulder, Colorado. Association for Computational Linguistics.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 489–513, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016. Results of the WMT16 Metrics Shared Task. In *Proceedings of the First Conference on Machine Translation*, pages 199–231, Berlin, Germany. Association for Computational Linguistics.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Chi-kiu Lo. 2017. MEANT 2.0: Accurate semantic MT evaluation for any output language. In *Proceedings of the Second Conference on Machine Translation*, pages 589–597, Copenhagen, Denmark. Association for Computational Linguistics.

Chi-kiu Lo, Michel Simard, Darlene Stewart, Samuel Larkin, Cyril Goutte, and Patrick Littell. 2018. Accurate semantic textual similarity for cleaning noisy parallel corpora using semantic machine translation evaluation metric: The NRC supervised submissions to the parallel corpus filtering task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 908–916, Belgium, Brussels. Association for Computational Linguistics.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 671–688, Belgium, Brussels. Association for Computational Linguistics.

Matous Machacek and Ondrej Bojar. 2014. Results of the WMT14 Metrics Shared Task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 293–301, Baltimore, Maryland, USA. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Michael Roth and Kristian Woodsend. 2014. Composition of word representations improves semantic role labelling. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 407–413, Doha, Qatar. Association for Computational Linguistics.

Miloš Stanojević, Amir Kamran, Philipp Koehn, and Ondřej Bojar. 2015. Results of the WMT15 Metrics Shared Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 256–273, Lisbon, Portugal. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.