KoBE: Knowledge-Based Machine Translation Evaluation

Zorik Gekhman   **Roee Aharoni**   **Genady Beryozkin**
Markus Freitag   **Wolfgang Macherey**
Google Research
{zorik,roeeaharoni}@google.com

Abstract

We propose a simple and effective method for machine translation evaluation which does not require reference translations. Our approach is based on (1) grounding the entity mentions found in each source sentence and candidate translation against a large-scale multilingual knowledge base, and (2) measuring the recall of the grounded entities found in the candidate vs. those found in the source. Our approach achieves the highest correlation with human judgements on 9 out of the 18 language pairs from the WMT19 benchmark for evaluation without references, which is the largest number of wins for a single evaluation method on this task. On 4 language pairs, we also achieve higher correlation with human judgements than BLEU. To foster further research, we release a dataset containing 1.8 million grounded entity mentions across 18 language pairs from the WMT19 metrics track data.

1 Introduction

Reliable and accessible evaluation is an important catalyst for progress in machine translation (MT) and other natural language processing tasks. While human evaluation is still considered the gold-standard when done properly (Läubli et al., 2020), automatic evaluation is a cheaper alternative that allows for rapid development cycles. Today's prominent automatic evaluation methods like BLEU (Papineni et al., 2002) or METEOR (Banerjee and Lavie, 2005) rely on n-gram matching with reference translations. While these methods are widely adopted, they have notable deficiencies:

- Reference translations cover a tiny fraction of all relevant input sentences or domains, and non-professional translators yield low-quality results (Zaidan and Callison-Burch, 2011).

- Different words in the candidate and reference translations that share an identical meaning
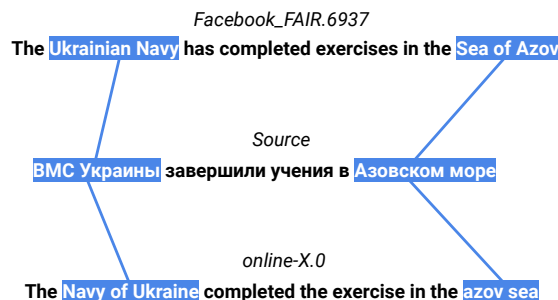

Figure 1: Example annotations and entity matches using our method.

will be penalized by simple n-gram matching, and multiple references are rarely used to alleviate this (Qin and Specia, 2015).

- Human translations have special traits ("Translationese", Koppel and Ordan, 2011) and reference-based metrics were shown to be biased to produce higher scores for translationese MT outputs than for valid, alternative MT outputs (Freitag et al., 2020).

- N-gram matching enables measurement of relative improvements, but does not provide an interpretable quality signal (Lavie, 2010).

To alleviate these issues, we propose Knowledge-Based Evaluation (KoBE), an evaluation method based on a large-scale multilingual knowledge base (KB). In our approach, we first ground each source sentence and candidate translation against the KB using entity linking (McNamee et al., 2011; Rao et al., 2013; Pappu et al., 2017; Gillick et al., 2019; Wu et al., 2019). We then measure the recall for entities found in the candidate vs. entities found in the source for all sentence pairs in the test set. Matching entities are ones linked to the same KB entry in both the source and the candidate. Figure 1 shows our entity matches for two candidate translations vs. the source, where different surface

forms that convey the same meaning are properly matched.

Our approach does not require reference translations, as it is based on linking entity mentions to the KB. This also makes it language-pair agnostic as long as the KB and entity linking systems cover the languages of interest. Since different words that share the same meaning should be resolved to the same entry in the KB, our method will not penalize different valid translations of the same entity. As our method measures the recall of the entities found in the source sentence, it is useful as an absolute quality signal and not just as a relative one. Finally, we can perform fine-grained error analysis using entity metadata to better understand where a system fails or succeeds in terms of entity types and domains.

To test our approach, we experiment with the "Quality Estimation as a Metric" benchmark (also named "Metrics Without References") from the WMT19 shared task on quality estimation (Fonseca et al., 2019). KoBE performs better than the other participating metrics on 9 language pairs, and obtains better correlation with human raters than BLEU on 4 language pairs, even though BLEU uses reference translations and KoBE is reference-agnostic. This demonstrates that KoBE is a promising step towards MT evaluation without reference translations.

To make our findings reproducible and useful for future work, we release the annotations we used together with scripts to reproduce our results. These entity linking annotations span over 425k sentences in 18 language pairs from 262 different MT systems, and contain 1.8 million entity mentions of 28k distinct entities.[1]

To summarize, this work includes the following contributions:

- We introduce KoBE, a novel knowledge-based, reference-less metric for machine translation quality estimation.

- We show this approach outperforms previously published results on 9 out of 18 language pairs from the WMT19 benchmark for evaluation without references.

- We release a data set with 1.8 million grounded entity mentions for the WMT19 benchmark to foster further research on knowledge-based evaluation.

2 Method

To obtain a system-level score, we first annotate all source sentences $s_i \in S$ and candidate translations $t_i \in T$ from a test set of n sentence pairs using entity linking pipelines.[2] As a knowledge base, we used the publicly available Google Knowledge Graph Search API[3] which offers entities from various domains. Unfortunately, we are not aware of any open-source multilingual KB and entity linking systems that we could rely on for the same purpose. We then count the matches for each sentence pair; matches are all candidate entities that are linked to the same record in the KB as source entities. Entities mentioned several times are counted as individual matches, and matches are clipped by the number of appearances of each entity in the source. As a pre-processing step, we ignore entity mentions in the candidate that are not in the target language using an in-house language identification tool, which we found to improve results in early experiments. We then compute recall by summing the number of matching entities across all sentence pairs and dividing by the the number of entities mentioned in all source sentences:

$$\text{recall} = \frac{\sum_{i=0}^{n} |\text{matches}(\text{entities}(s_i), \text{entities}(t_i))|}{\sum_{i=0}^{n} |\text{entities}(s_i)|}$$

Our decision to ignore candidate entities that are not in the correct language came from an observation that for some low-resource language pairs, MT systems fail to translate the input and instead copy most of its content to the output – see Ott et al. (2018) for a similar observation. As our entity linking system is language agnostic, it was detecting the copied entities, which resulted in false matches.

We found precision to have weaker correlation on most language pairs, as it rewards systems producing a lower number of entities – systems that usually produced lower quality translations. Recall is more stable as the number of entities in the source is constant for all evaluated systems, and only the match count is changing. Since recall may give inflated scores when over-producing entities, we introduce an entity count penalty (ECP), inspired by BLEU's brevity penalty. ECP penalizes

	de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en
BLEU	0.849	0.982	0.834	0.946	0.961	0.879	0.899
LASIM	0.247	–	–	–	–	-0.31	–
LP	-0.474	–	–	–	–	-0.488	–
UNI	0.846	**0.93**	–	–	–	0.805	–
UNI+	0.85	0.924	–	–	–	0.808	–
YiSi-2	0.796	0.642	-0.566	-0.324	0.442	-0.339	0.94
YiSi-2_srl	0.804	–	–	–	–	–	**0.947**
KoBE	**0.863**	0.538	**0.828**	**0.899**	**0.704**	**0.928**	0.907

Table 1: System-level Pearson correlation with human judgements for language pairs into English from the WMT19 metrics-without-references shared task. Best QE results are marked in bold.

	en-cs	en-de	en-fi	en-gu	en-kk	en-lt	en-ru	en-zh
BLEU	0.897	0.921	0.969	0.737	0.852	0.989	0.986	0.901
LASIM	–	0.871	–	–	–	–	-0.823	–
LP	–	-0.569	–	–	–	–	-0.661	–
UNI	0.028	0.841	**0.907**	–	–	–	**0.919**	–
UNI+	–	–	–	–	–	–	0.918	–
USFD	–	-0.224	–	–	–	–	0.857	–
USFD-TL	–	-0.091	–	–	–	–	0.771	–
YiSi-2	0.324	0.924	0.696	**0.314**	0.339	**0.055**	-0.766	-0.097
YiSi-2_srl	–	**0.936**	–	–	–	–	–	-0.118
KoBE	**0.597**	0.888	0.521	-0.34	**0.827**	-0.049	0.895	**0.216**

Table 2: System-level Pearson correlation with human judgements for from-English language pairs from the WMT19 metrics-without-references shared task. Best QE results are marked in bold.

systems producing c entities if c is more than twice the number of entities in the source, s:

$$\text{ECP} = \begin{cases} 1 & \text{if} \quad c < 2s \\ e^{(1-c/2s)} & \text{if} \quad c \geq 2s \end{cases}$$

Finally:

$$\text{KoBE} = \text{ECP} \cdot \text{recall}$$

3 Experimental Setup

The WMT conference holds a Quality Estimation track (QE) that aims to predict the quality of MT systems given the source sentences and candidate translations (without reference translations). While this was usually done at the word or sentence level, one of the novelties in WMT19 was introducing a new task for using QE as a metric at the corpus level, testing the generalization ability of QE approaches in a massive multi-system scenario (Fonseca et al., 2019). To test our approach, we used the same setting as in this shared task. For every language pair of the 18 evaluated pairs, we use KoBE to score the MT systems participating in same years news translation task. We then measure the Pearson correlation of our scores for each system with its human direct-assessment (DA) scores. To ensure a fair comparison, we recompute the correlations for the other participating metrics and confirm that we reproduce the reported scores.[4]

---

[4]More implementation details for reproducing our results are available in the supplemental material.

	de-cs	de-fr	fr-de
BLEU	0.941	0.891	0.864
ibm1-morpheme	0.355	-0.509	-0.625
ibm1-pos4gram	–	0.085	**-0.478**
YiSi-2	0.606	**0.721**	-0.53
KoBE	**0.958**	0.485	-0.785

Table 3: Pearson correlation results on language pairs excluding English from the WMT19 metrics-without-references task. Best QE results are marked in bold.

4 Results

We compare KoBE with all participating metrics in the shared task. We refer the reader to Fonseca et al. (2019) for more details about the different metrics. We also compare our results with BLEU to have a benchmark for a reference-based metric.

The results for into-English language pairs are available in Table 1. KoBE outperforms all other submissions for German-to-English, Gujarati-to-English, Kazakh-to-English, Lithuanian-to-English and Russian-to-English, making it the best system in this section in terms of the number of wins. Results for from-English language pairs are available in Table 2. In this case KoBE outperforms the submitted systems for English-to-Czech and English-to-Kazakh with Pearson correlations of 0.597 and 0.827, and also obtains high correlations for English-to-German and English-to-Russian with 0.888 and 0.895, respectively. For

English-to-Chinese we also obtain the highest correlation, but it is very low overall. Table 3 describes the results on language pairs not involving English (German-to-Czech, German-to-French and French-to-German). In this case KoBE obtains the best result for German-to-Czech with Pearson correlation of 0.958. For 4 language pairs (German-to-English, Russian-to-English, Chinese-to-English and German-to-Czech), KoBE outperforms BLEU in terms of the correlation with human judgements. This is encouraging given that KoBE does not use reference translations while BLEU does.

In Table 4 we perform additional experiments to test whether our method can also be used as a reference-based metric, by measuring the recall of entities mentioned in the candidate translations vs. entities mentioned in the references. KoBE indeed correlates well with human judgements and outperforms BLEU on 5 out of 7 language pairs, which we find impressive given that it only considers unordered entity mentions and not on all n-grams as in BLEU. Figure 2 shows a comparison of our scores vs. BLEU and human direct-assessment on Russian-to-English. In addition to the higher correlation with human judgements (0.928 vs. 0.879), our metric produces scores which are closer to the human scores on an absolute scale.

5 Discussion and Analysis

Summarizing the above findings, our method obtains the best results on 9 out of 18 language pairs, which makes it the method with the largest number of wins on the WMT19 metrics-without-references benchmark. This shows that knowledge-based evaluation is a promising path towards MT evaluation without references. In comparison, the next best method is YiSi-2 (Lo, 2019) which is based on token-level cosine-similarity using context-aware token representations from multilingual BERT (Devlin et al., 2019). We believe that combining our knowledge-based approach with such methods may result in even better correlation with human judgements, but leave this for future work.

As our metric is based on the recall of entities in the target with respect to the source, it is important that the entities will be properly detected in the target. A failure to detect an entity in the source will just lead KoBE to use less entities, while a failure to detect an entity in the target will lead KoBE to penalize an entity that is actually present. Our entity linking pipelines work best in English, which

results in much higher correlations with human judgements when English is the target language (Table 1) vs. the correlations when English is the source (Table 2). We believe that as entity linking systems will improve for languages other than English, our metric will improve accordingly. Another possible concern may be regarding the evaluation of sentences which do not contain any detected entities – our analysis shows that was the case for less than 8% of the sentences, so it did not have a large effect on the corpus-level metric.[5]

Figure 3 shows matching statistics for different MT systems across several entity categories from the KB. We can see that our scores vary across different categories between and within different systems, which can give an interpretable signal for system developers regarding where improvement efforts should be invested.

Our reproduction of the correlation results raises an issue with the current evaluation methodology in the shared task. In the published results (Fonseca et al., 2019), in order to support both lower-is-better metrics (e.g. TER Snover et al., 2006) and higher-is-better metrics (e.g. BLEU), the absolute values of the Pearson correlations are reported. However, when looking in Table 2 and Table 3 we see that the same metric may be correlated with different signs in different language pairs. This may result in wrong ranking of evaluation metrics, as the absolute value may "cover up" such cases. We hope future evaluations will take this detail into account.

A possible drawback of our approach is that it only relies on entities, which do not fully cover the sentence semantics. However, in the quality estimation setting, we only have access to the source and candidate translation, which are in different languages. As different languages use different syntactic structures and vocabulary, it is hard to employ other structural cues - for example, the order of the entities may be different due to the grammatical differences between the languages. The strong correlation between our metric and human judgements shows that knowledge-based comparison is a strong indicator of translation quality in this challenging setting. This is in line with the results of Freitag et al. (2020) who showed that BLEU with extensively paraphrased references correlates better with human judgements than BLEU with vanilla references – our method is "paraphrasing" or "stripping

---

[5]See Figure 4 in the supplemental material for a histogram of entity counts per sentence.

	de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en	mean
BLEU	0.849	**0.982**	0.834	0.946	**0.961**	0.879	0.899	0.907
KoBE	**0.906**	0.961	**0.85**	**0.961**	0.901	**0.954**	**0.947**	**0.926**

Table 4: Comparison of the Pearson correlation with human judgements for BLEU and KoBE, on the into-English language pairs from the WMT19 metrics shared task. Best results are marked in bold.
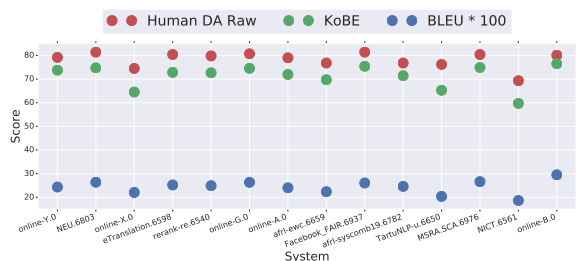

Figure 2: Comparison of BLEU, KoBE and human judgements for Russian-to-English newstest2019.

Figure 3: Statistics for entity categories for different Chinese-to-English MT systems and the source side.

down" the candidate and reference to only contain the mentioned entities during evaluation.

6 Related Work

Quality estimation for MT has been studied extensively in recent years – see Specia et al. (2018) for a thorough overview. Most work has been on the sentence or word level, using supervised approaches e.g. Open-Kiwi (Kepler et al., 2019). Using semantic knowledge for MT evaluation was proposed in different approaches: METEOR (Denkowski and Lavie, 2014) used paraphrase tables for reference-based evaluation; YiSi (Lo, 2019) and MEANT (Lo, 2017) used semantic role labeling (SRL) annotations; Birch et al. (2016) used the UCCA semantic annotations (Abend and Rappoport, 2013) for human evaluation of MT; Li et al. (2013) proposed a name-aware BLEU score giving more weight to named entities. Babych and Hartley (2004) conducted a comparative evaluation of named entity recognition (NER) from MT outputs, concluding

that the success rate of NER does not strongly correlate with human or automatic evaluation scores. We show contradicting results, which may stem from the better NER and MT systems available today, and from the entity linking step we add. To the best of our knowledge, our work is the first to introduce a reference-less MT evaluation method based purely on entity linking against a multilingual knowledge-base.

7 Conclusions and Future Work

We proposed KoBE, a method for reference-less machine translation evaluation using entity linking to a multilingual knowledge base. We demonstrated the applicability of our method by achieving strong results on the WMT19 benchmark for reference-less evaluation across 9 language pairs, where in 4 cases it also outperforms the reference-based BLEU. Our method is simple, interpretable and produces scores closer to human judgements on an absolute scale, while enabling more fine-grained analysis which can be useful to find weak spots in the evaluated model.

In future work, we would like to combine knowledge-based signals with unsupervised approaches like YiSi (Lo, 2019) and XMoverScore (Zhao et al., 2020) that use contextualized representations from cross-lingual LMs like multilingual BERT (Devlin et al., 2019). As our method does not require reference translations, we would like to explore scaling it to use much larger or domain specific monolingual datasets. Our knowledge-based approach can also be applied to other text generation tasks like summarization or text simplification where BLEU was shown to be problematic (Sulem et al., 2018). Finally, performing outlier-aware meta-evaluation which was recently shown to be important in such settings (Mathur et al., 2020) could be beneficial.

Acknowledgements

We thank the reviewers for their valuable comments and suggestions for improving this work.

References

Omri Abend and Ari Rappoport. 2013. Universal conceptual cognitive annotation (UCCA). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 228–238, Sofia, Bulgaria. Association for Computational Linguistics.

Bogdan Babych and Anthony Hartley. 2004. Comparative evaluation of automatic named entity recognition machine translation output. In *IJCNLP*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Alexandra Birch, Omri Abend, Ondřej Bojar, and Barry Haddow. 2016. HUME: Human UCCA-based evaluation of machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1264–1274, Austin, Texas. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the WMT 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. Bleu might be guilty but references are not innocent. *arXiv preprint arXiv:2004.06063*.

Daniel Gillick, Sayali Kulkarni, Larry Lansing, Alessandro Presta, Jason Baldridge, Eugene Ie, and Diego Garcia-Olano. 2019. Learning dense representations for entity retrieval. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, Hong Kong, China. Association for Computational Linguistics.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. 2019. OpenKiwi: An open source framework for quality estimation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 117–122, Florence, Italy. Association for Computational Linguistics.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, Portland, Oregon, USA. Association for Computational Linguistics.

Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. A set of recommendations for assessing human–machine parity in language translation. *Journal of Artificial Intelligence Research*, 67:653–672.

Alon Lavie. 2010. Evaluating the output of machine translation systems. *Conference of the Association for Machine Translation in the Americas*.

Haibo Li, Jing Zheng, Heng Ji, Qi Li, and Wen Wang. 2013. Name-aware machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 604–614, Sofia, Bulgaria. Association for Computational Linguistics.

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. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Paul McNamee, James Mayfield, Dawn Lawrie, Douglas Oard, and David Doermann. 2011. Cross-language entity linking. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 255–263, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Myle Ott, Michael Auli, David Grangier, and MarcAurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *International Conference on Machine Learning*, pages 3956–3965.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Aasish Pappu, Roi Blanco, Yashar Mehdad, Amanda Stent, and Kapil Thadani. 2017. Lightweight multilingual entity extraction and linking. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 365–374.

Ying Qin and Lucia Specia. 2015. Truly exploring multiple references for machine translation evaluation. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, pages 113–120, Antalya, Turkey.

Delip Rao, Paul McNamee, and Mark Dredze. 2013. Entity linking: Finding extracted entities in a knowledge base. In *Multi-source, multilingual information extraction and summarization*, pages 93–115. Springer.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*.

Lucia Specia, Carolina Scarton, and Gustavo Henrique Paetzold. 2018. *Quality estimation for machine translation*. Morgan & Claypool Publishers.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. BLEU is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium. Association for Computational Linguistics.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2019. Zero-shot entity linking with dense entity retrieval. In *arXiv:1911.03814*.

Omar F. Zaidan and Chris Callison-Burch. 2011. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1220–1229, Portland, Oregon, USA. Association for Computational Linguistics.

Wei Zhao, Goran Glavaš, Maxime Peyrard, Yang Gao, Robert West, and Steffen Eger. 2020. On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

A Supplemental Material

The data used in this paper is taken from the WMT19 results.[6] We downloaded the news translation task submissions[7] and annotated them using entity linking pipelines. We make our annotations publicly available to reproduce our results. We downloaded the Metrics task data[8] and obtained the submitted metrics scores, together with the standardized human direct assessment (DA) scores, from the `results/sys-level_scores_metrics.csv` file. We recalculated the Pearson correlations for all metrics and made sure we got the same results as reported in the WMT19 official results (Fonseca et al., 2019).

Our submission contains a copy of the `sys-level_scores_metrics.csv` file, containing the submitted metrics scores, together with the human direct assessment (DA) scores. In addition, we publish the annotations for all WMT19 news translation task submissions. The published data contains a file for each system in each language pair, as well as the annotations for the source text and reference translations. Our annotations are in json format and contain all the entities that were detected in each sentence. Each entity has an id and a start and end positions in the sentence. In addition, we publish a python script that, given the `sys-level_scores_metrics.csv` file and the annotations, first calculates our score

for all language pairs and all systems and then calculates the Pearson correlations with human DA scores. This script and data can be used to exactly reproduce the results reported in the paper. We also hope that the large annotated data set will help researchers who wish to further explore multilingual knowledge-based evaluation methods.

We also calculate the entity statistics for each language pair using the source and the reference sentences. Table 5 shows statistics for into-English language pairs, Table 6 shows statistics for from-English language pairs and Table 7 shows statistics for language pairs excluding English. Those tables can be also obtained by running the provided script. Note that the numbers here denote the entities that were detected by the entity linking system. Figure 4 shows a histogram of entities count per sentence on the Russian-to-English source corpus.
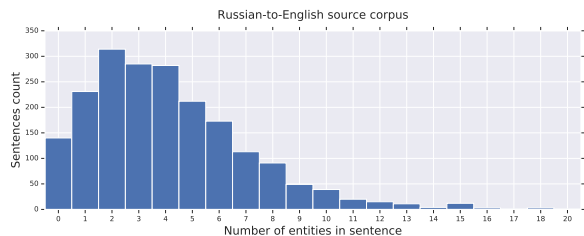

Figure 4: Histogram of number of entities in each sentence for the Russian-to-English source corpus.

	de-cs	de-fr	fr-de
sentence count	1997	1701	1701
source sentences with entities	1878	1586	1634
source entities count	8649	6794	9102
reference entities count	5820	6437	4810
source distinct entities count	2643	1571	1917
reference distinct entities count	1445	1450	1152
common distinct entities count	910	739	737

Table 7: Statistics for language pairs excluding English from the WMT19 metrics-without-references task.

	de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en
sentence count	2000	1996	1016	1000	1000	2000	2000
source sentences with entities	1795	1672	796	751	934	1860	1958
source entities count	5831	4645	1911	1932	4320	8230	15339
reference entities count	6582	7070	3650	4103	5140	8413	18088
source distinct entities count	2244	1525	523	661	1241	2404	3312
reference distinct entities count	2270	2141	1276	1329	1616	2506	3474
common distinct entities count	1184	920	320	371	740	1446	1969

Table 5: Statistics for into English language pairs from the WMT19 metrics-without-references shared task.

	en-cs	en-de	en-fi	en-gu	en-kk	en-lt	en-ru	en-zh
sentence count	1997	1997	1997	998	998	998	1997	1997
source sentences with entities	1870	1870	1870	934	934	934	1870	1870
source entities count	9845	9846	9845	4711	4710	4710	9846	9845
reference entities count	5824	5345	5113	2163	1219	2807	7563	10646
source distinct entities count	3150	3149	3149	1941	1941	1941	3149	3149
reference distinct entities count	1446	1528	1238	572	330	847	1899	2739
common distinct entities count	971	1006	899	364	202	555	1238	1679

Table 6: Statistics for from-English language pairs from the WMT19 metrics-without-references shared task.