MTEQA at WMT21 Metrics Shared Task

Mateusz Krubiński[1], **Erfan Ghadery**[2], **Marie-Francine Moens**[2], **and Pavel Pecina**[1]

[1]Charles University, Faculty of Mathematics and Physics
{krubinski,pecina}@ufal.mff.cuni.cz
[2]KU Leuven, Department of Computer Science
{erfan.ghadery,sien.moens}@kuleuven.be

Abstract

In this paper, we describe our submission to the WMT 2021 Metrics Shared Task. We use the automatically-generated questions and answers to evaluate the quality of Machine Translation (MT) systems. Our submission builds upon the recently proposed MTEQA framework. Experiments on WMT20 evaluation datasets show that at the system-level the MTEQA metric achieves performance comparable with other state-of-the-art solutions, while considering only a certain amount of information from the whole translation.

1 Introduction

The goal of automatic Machine Translation (MT) evaluation is to automatically evaluate the output quality produced by MT systems. Metrics used for this task assign a score by comparing the MT output to either a reference translation or to the source sentence. The main indicator that is used to assess the performance of a specific metric is the correlation with human judgement computed for outputs from several systems. It was recently shown that metrics based on contextualized embeddings, such as YiSi (Lo, 2019) or ESIM (Mathur et al., 2019), are able to achieve better performance than the most common BLEU (Papineni et al., 2002).

In this paper, we describe application of the recently proposed metric – MTEQA (Krubiński et al., 2021) for the task of evaluating the quality of MT outputs in the context of the WMT21 Metric task.

The MTEQA[1] framework is inspired by previous works on evaluating abstractive summaries. It builds upon the fact that state-of-the-art (neural) MT systems tend to produce a fluent output but sometimes fail in adequacy of the translation. It leverages the recent progress in Question Generation (QG) and Question Answering (QA) to formulate and answer questions based on the MT output.

2 Related Work

2.1 MT Evaluation

Metrics that are most widely used for automatic evaluation of MT outputs produce a score by comparing surface-level forms of hypothesis and reference translation. The most common one, BLEU, is a modified version of n-gram precision calculated by averaging over different values of n with penalization for too short translations (brevity penalty). The recently proposed CHRF (Popović, 2015) considers the character-level n-grams, making it possible to reward partially matched tokens. Recently, various works (e.g., Lo, 2019; Mathur et al., 2019; Bawden et al., 2020) explored the usage of contextualized word- or sentence-level embeddings to compare the numerical representations of reference and hypothesis. Such metrics enable explicit regression towards the desired human-produced labels.

2.2 Question-based Evaluation

Previous works examined the usage of reading comprehension tests to measure the quality and "usefulness" of MT systems (Tomita et al., 1993; Fuji et al., 2001; Castilho and Guerberof Arenas, 2018). Berka et al. (2011) were the first to use the *yes/no* type of questions for manual evaluation of MT systems, examining the English-to-Czech direction. Scarton and Specia (2016) approached the problem of document-level Quality Estimation (QE) by extending the CREG corpus (Ott et al., 2012) of German documents designed for reading comprehension exercises.

More work on the questions-based evaluation was done in the context of text summarization. Eyal et al. (2019) proposed the APES metric for the task of evaluating abstractive text summarization. They used the reference summary to produce fill-in-the-blank type of questions, by finding all possible entities using a NER system. The APES score for a given summarization model is the percentage

[1] https://github.com/ufal/MTEQA

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 1024–1029
November 10–11, 2021. ©2021 Association for Computational Linguistics

Reference	Extracted Answers	Generated Questions	MT output	Test Answers
The 56-year-old Macura studied at Prague University of Economics.	56	How old is Macura?	Fifty-six-year-old Macura graduated from the University of Economics in Prague.	Fifty-six
	Prague University of Economics	Where did Macura study?		University of Economics
Um 19 Uhr haben wir das Auto gepackt und sind an Bord der Fähre nach Portsmouth gegangen.	Portsmouth	Wo sind wir hin, nachdem wir das Auto gepackt haben?	Gegen 19 Uhr haben wir das Auto gepackt und die Fähre nach Portsmouth bestiegen.	Portsmouth
	19 Uhr	An welchem Datum haben wir das Auto gepackt?		19 Uhr

Figure 1: Example of the Extracted Answers, Generated Questions and corresponding Test Answers from a *newstest2021* reference file.

of questions that were answered correctly (using a Question Answering system), averaged over the whole test-set. Scialom et al. (2019) extended their work into unsupervised settings by generating questions from the source document. The FEQA (Durmus et al., 2020) and QAGS (Wang et al., 2020) metrics further extend the idea by automatically generating human-readable questions.

2.3 MTEQA

MTEQA is the first MT metric based on the principles of question answering.

The automatically generated pairs of a question and its (gold-standard) answer from the reference translation are used by a question answering system to provide a new (test) answer given the question and the MT output (translation) used as the context.

The generated (test) answer is then compared to the gold-standard answer, using the string-comparison metric. The final score for a given MT output is the average taken over all of the question/answer pairs generated for a corresponding reference.

3 Experiments

Our implementation of the MTEQA metric is based on the state-of-the-art system capable of solving the initial three tasks of the procedure: answer extraction, question generation, question answering. It is the T5 model (Raffel et al., 2020) fine-tuned on the SQuADv1 dataset (Rajpurkar et al., 2016) by Patil (2020) and available from GitHub[2]. The limitation of the T5 model is that it was trained on English data and most importantly tuned on the SQuADv1 dataset which is in English. Thus, this model only allows evaluation of MT systems translating from any language to English.

To overcome that, we used the multilingual mT5 model (Xue et al., 2021) and fine-tuned it on machine translation of SQuADv1 dataset. We exploited the existing translations into German (Lewis et al., 2020) and into Czech (Macková and Straka, 2020) which allows score translations into German (xx-de) and into Czech (xx-cs) directions. Due to time constraints we were not able to train QA and QG systems in other languages.

Figure 1 presents examples of extracted answers and generated questions.

3.1 Baseline

The baseline implementation is based on the T5 model tuned on the SQuADv1 dataset and used to generate: 1) the gold-standard answers from the reference translations, 2) a question for each gold-standard answer, 3) a test answer for each question and MT output (context) pair. The test answers are compared by the word-level F1 score commonly used for QA evaluation (Rajpurkar et al., 2016; Trischler et al., 2017; Chen et al., 2019; Durmus et al., 2020).

For each of the MT systems participating in WMT20 News translation task (Barrault et al., 2020), we compute both segment-level scores and a single system-level score, an average of segment-level scores. We report the system-level Pearson correlation with a DA human assessment using the *newstest2020* references We report correlation for a English → German, English → Czech and a few to-English directions (see Table 1, row MTEQA F1). We also include an average over all of the

[2] https://github.com/patil-suraj/question_generation

	cs-en	de-en	zh-en		en-de	en-cs
	12	12	16	avg	14	12
MTEQA F1	0.782*	0.997*	0.952*	0.893*	0.946*	0.845*
MTEQA F1 Keyphrase	0.851*	**0.998***	0.944*	0.896*	0.941*	0.877*
MTEQA chrF Keyphrase	**0.890***	**0.998***	0.951*	**0.905***	0.952*	0.859*
sentBLEU	0.844	0.978	0.948	0.859	0.934	0.840
BLEU	0.851	0.985	0.956	0.854	0.928	0.825
PRISM	0.818	**0.998**	0.957	0.880	**0.958**	**0.949**
YiSi-2	0.764	0.988	**0.964**	0.821	0.899	0.714

Table 1: System-level Pearson correlation for selected metrics used for measuring MT quality with DA human assessment over MT systems using the *newstest2020* references. Average (avg) is computed over all to-English directions available. Number below the language pair indicates the number of systems considered. Figures without * are taken from Mathur et al. (2020a).

to-English directions, which were part of WMT20 Metric Task (Mathur et al., 2020b) evaluation campaign[3]. Other metrics are included for a comparison. At the segment-level, we report the Kendall's Tau correlation of segment-level metric scores with daRR human assessment scores, see Table 2. We use the same Kendall's Tau-like formulation which was used by Mathur et al. (2020b) in WMT20 evaluation campaign.

On average, the baseline outperforms the traditional MT evaluation metrics (sentBLEU, BLEU) as well as the recently proposed ones that performed very well in the WMT20 Metric Task (PRISM, YiSi-2), though for some of the translation directions (e.g. cs-en) MTEQA F1 is much worse (but for cs-en YiSi-2 also does not beat BLEU). The segment-level correlation is much lower, even negative for some directions (e.g. zh-en) .

3.2 Generating Additional Answers

Since the QG system generates a single question for each sub-sequence of words marked as an extracted answer, the limit factor is the number of gold-standard answers we extract. To generate more questions we need more keyphrases to be asked about.

Considering the whole predictive power of the MTEQA metric is based on questions, we used linguistic processing of the sentence based on Part-of-Speech (POS) pattern matching and Named Entity Recognition (NER) to extract more keyphrases.

Given a sentence as the input, first, we parse the sentence using UDPipe (Straka et al., 2016) to extract part of speech (POS) tags. Then, we extract phrases that are matched with one of the patterns in our POS pattern bank. The POS pattern bank is created by parsing the sentences from XQuAD (Artetxe et al., 2020) dataset, extracting the POS patterns corresponding to the gold-standard answers, and taking the most frequent patterns. This dataset contains professional translations of the development set of SQuADv1, translated into various languages from different language families and using different scripts. Second, we extract named entities mentioned in the input sentence using a combination of two multilingual NER models, POLYGLOT-NER (Al-Rfou et al., 2015), and Stanza (Qi et al., 2020). Finally, we output the union of the extracted phrases and named entities as the potential answers. At both system- and segment-level using the MTEQA F1 Keyphrase variant yields improvements for most of the translation directions.

3.2.1 Tuning the Answer Comparison Metric

The choice of the Answer Comparison Metric can have a considerable impact on the final performance. Using the word-level F1 metric, given the gold-standard answer *"Tchaikovsky"*, both the *"Tchaikovski"* and *"Beethoven"* would get the same score. In the context of MT, it may be worth to consider a more fine-grained comparison.

We decided to use the chrF (Popović, 2015) metric, since it operates on the level of characters, and enables scoring even partial matches. Using the MTEQA chrF Keyphrase variant yields further improvements at both system- and segment-level.

For the WMT21 Metrics Shared Task we submit this variant of the metric – the gold-standard answers are extracted by POS pattern matching and NER, and the chrF metric is used for answer comparison (MTEQA chrF Keyphrase).

[3] cs, de, ja, pl, ru, ta, zh, iu, km, ps → en

	cs-en	de-en	zh-en	en-de	en-cs
MTEQA F1	-0.422^*	0.041^*	-0.430^*	-0.581^*	-0.480^*
MTEQA F1 Keyphrase	-0.108^*	0.273^*	-0.058^*	-0.016^*	0.100^*
MTEQA chrF Keyphrase	0.017^*	0.327^*	0.030^*	0.159^*	0.227^*
sentBLEU	0.068	0.413	0.093	0.303	0.432
PRISM	**0.143**	**0.475**	**0.167**	**0.447**	**0.619**
YiSi-2	0.068	0.413	0.116	0.296	0.187

Table 2: Segment-level Kendall's Tau correlation for a few metrics used for measuring MT quality with DARR human assessment scores, over MT systems using the *newstest2020* references. Numbers without * are taken from (Mathur et al., 2020a).

4 MQM scores

Recently, Freitag et al. (2021) demonstrated that the WMT DA method traditionally used for human evaluations has actually lower correlation with expert-based labels than the Multidimensional Quality Metrics (MQM) scoring method developed in the EU QTLaunchPad and QT21 projects. Following their findings, the WMT21 Metric Task will report the correlation with MQM labels in the official results.

To provide a more complete picture of the performance of the MTEQA metric, we also report correlation with the MQM assessments. Table 3 presents the system-level Pearson correlation of the metric with both the MQM and DA labels for 8 systems that were re-annotated by Freitag et al. (2021) and are available from GitHub[4].

The results are surprising and to a large extent unintuitive. Metrics performing well in comparison with MQM are often bad in comparison with DA.

5 Conclusions

In this paper we described our submission to the WMT21 Metrics Shared Task. We showed that the degree to which the MT output can be used to answer questions about the reference can be used as a proxy to evaluate the translation quality.

We showed a gradual improvement of our submission. We examined a linguistically motivated way of extracting keyphrases from the sentence, and showed that it boosts both the segment- and system-level correlation with DA human judgments. We were able to further boost the final performance by using the chrF metric to compare the reference and test answers.

Finally, we examined the performance against the MQM labels and compared the performance against the DA labels.

Acknowledgements

This work was supported by the European Commission via its H2020 Program (contract no. 870930) and CELSA (project no. 19/018), and has been using data and tools provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (project no. LM2018101).

References

Rami Al-Rfou, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2015. Polyglot-ner: Massive multilingual named entity recognition. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 586–594. SIAM.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (wmt20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Rachel Bawden, Biao Zhang, Andre Tättar, and Matt Post. 2020. ParBLEU: Augmenting metrics with automatic paraphrases for the WMT'20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 887–894, Online. Association for Computational Linguistics.

Jan Berka, Martin Černý, and Ondřej Bojar. 2011. Quiz-based evaluation of machine translation. *The Prague Bulletin of Mathematical Linguistics*, 95.

[4] `https://github.com/google/wmt-mqm-human-evaluation`

	zh-en		en-de	
	MQM	DA	MQM	DA
MTEQA CHRF KEYPHRASE	0.630	**0.818**	0.761	0.394
PRISM	0.778	0.351	**0.989**	0.607
COMET	**0.889**	0.188	0.965	**0.628**
PARBLEU	0.380	0.565	0.722	0.218
CHRF	0.523	0.579	0.853	0.576
TER	0.352	0.511	0.810	0.477

Table 3: System-level Pearson correlation for selected metrics used for measuring MT quality with the DA and MQM labels, computed for the *newstest2020* references and the 8 MT systems re-annotated by Freitag et al. (2021).

Sheila Castilho and Ana Guerberof Arenas. 2018. Reading comprehension of machine translation output: What makes for a better read? In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 79–88, Alacant/Alicante, Spain. European Association for Machine Translation.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Evaluating question answering evaluation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 119–124, Hong Kong, China. Association for Computational Linguistics.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Matan Eyal, Tal Baumel, and Michael Elhadad. 2019. Question answering as an automatic evaluation metric for news article summarization. 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 3938–3948, Minneapolis, Minnesota. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *arXiv preprint arXiv:2104.14478*.

Masaru Fuji, Hatanaka N, Ito E, Kamei S, Kumai H, Sukehiro T, Yoshimi T, and Isahara Hitoshi. 2001. Evaluation method for determining groups of users who find mt useful. In *MT Summit VIII: Machine Translation in the Information Age*, pages 103–108.

Mateusz Krubiński, Erfan Ghadery, Marie-Francine Moens, and Pavel Pecina. 2021. Just ask! evaluating machine translation by asking and answering questions. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. 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.

Kateřina Macková and Milan Straka. 2020. Reading comprehension in czech via machine translation and cross-lingual transfer. In *23rd International Conference on Text, Speech and Dialogue*, pages 171–179, Cham, Switzerland. Springer.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2019. Putting evaluation in context: Contextual embeddings improve machine translation evaluation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2799–2808, Florence, Italy. Association for Computational Linguistics.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020a. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020b. Results of the wmt20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

Niels Ott, Ramon Ziai, and Detmar Meurers. 2012. Creation and analysis of a reading comprehension exercise corpus. *Multilingual corpora and multilingual corpus analysis*, 14:47.

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.

Suraj Patil. 2020. Question generation. https://github.com/patil-suraj/question_generation.

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.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Carolina Scarton and Lucia Specia. 2016. A reading comprehension corpus for machine translation evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3652–3658, Portorož, Slovenia. European Language Resources Association (ELRA).

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3246–3256, Hong Kong, China. Association for Computational Linguistics.

Milan Straka, Jan Hajic, and Jana Straková. 2016. Udpipe: trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297.

Masaru Tomita, Shirai Masako, Tsutsumi Junya, Matsumura Miki, and Yoshikawa Yuki. 1993. Evaluation of mt systems by toefl. In *Proceedings of the Theoretical and Methodological Implications of Machine Translation (TMI-93)*, pages 252–265.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.