Just Ask!
Evaluating Machine Translation by Asking and Answering Questions

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 show that automatically-generated questions and answers can be used to evaluate the quality of Machine Translation systems. Building on recent work on the evaluation of abstractive text summarization, we propose a new metric for system-level Machine Translation evaluation, compare it with other state-of-the-art solutions, and show its robustness by conducting experiments for various translation directions.

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 latter is called Quality Estimation).

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 widely used BLEU (Papineni et al., 2002).

In this paper, we propose a new method for automatic evaluation of MT systems – MTEQA[1] (**M**achine **T**ranslation **E**valuation with **Q**uestion **A**nswering), building on previous works on evaluating abstractive summaries. We build 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. We leverage the recent progress in Question Generation (QG) and Question Answering (QA) to formulate and answer human readable questions about the MT system output. Our experiments show that the effectiveness of the proposed metric is comparable to performance of other automatic metrics, while considering only a certain amount of information from the whole translation. We also examine the robustness of the metric by considering several translation directions and target languages.

The remainder of this paper is structured as follows. In Section 2, we introduce relevant research on question-based evaluation. In Section 3, we describe our metric in detail. In Section 4, we present and discuss the results of our experiments including s the influence of different human scoring methods. Section 5 presents conclusions.

2 Related Work

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 dominant one, BLEU (Papineni et al., 2002), is a version of n-gram precision calculated by averaging over different values of n with penalization for overly short translations (brevity penalty). Another one, CHRF (Popović, 2015), considers the character-level n-grams, making it possible to reward partial token matches. The standardised implementation provided in the sacreBLEU[2] package takes care of pre-processing and enables direct comparison between MT outputs.

Recently, various works (e.g., Lo, 2019; Mathur et al., 2019; Bawden et al., 2020) explored the usage of contextualized word-level or sentence-level embeddings to compare the numerical representations of reference and hypothesis. Such metrics enable explicit regression towards the desired human-produced labels.

2.1 Evaluation of Summarization

The task of automatic text summarization is to produce a concise summary of a given document that

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

[2]https://github.com/mjpost/sacrebleu

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 495–506
November 10–11, 2021.

would preserve all the key information from the document. One of the most popular metrics used for evaluating summary quality is ROUGE (Lin, 2004), which compares overlapping n-grams between the model output and the reference summary.

To step beyond the n-grams comparison, Eyal et al. (2019) proposed the APES metric. 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 of questions that were answered correctly (using an Question Answering system), averaged over the whole test-set. The authors reported a higher correlation with the Pyramid method (Nenkova et al., 2007) for manual evaluation than the ROUGE metric. Scialom et al. (2019) extended their work into unsupervised settings by generating questions from the source document. Closest to our work are the metrics FEQA (Durmus et al., 2020) and QAGS (Wang et al., 2020), which automatically generate the natural language questions from the summary and/or document.

2.2 Question-based Evaluation of MT

Tomita et al. (1993) were the first to use the reading comprehension tests to measure the quality of MT systems. They translated several passages from TOEFL (Test of English as a Foreign Language) guide book into Japanese, using a selection of MT systems, while corresponding questions and answers were translated into Japanese by professional translators. The MT systems were evaluated by measuring the percentage of questions answered correctly by the Japanese speaking human annotators, using the MT output as a context.

Fuji et al. (2001) used the reading comprehension tests to examine the "usefulness" of machine-translated text. In their experiment, participants take the reading comprehension test in a foreign language (English), while also being presented with the text translated by the MT system into their mother language (Japanese). Authors claim that presenting the MT output yields a higher comprehension performance.

Castilho and Guerberof Arenas (2018) examine the user satisfaction when completing the comprehension type of test, using the context translated by the MT system. They collect the eye-tracking data to analyse the cognitive effort of the participants.

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. They use professional translators to translate the questions and answers to English. They examine the document-level translation quality by translating the documents by MT systems and asking the human annotators to complete the reading comprehension test using the MT output as a context. Forcada et al. (2018) used the same corpus to examine the usage of automatically generated gap-filling closure type of testing.

Berka et al. (2011) used the *yes/no* type of questions for manual evaluation of MT systems, examining the English-to-Czech direction. The authors prepared a set of English texts from various domains and used human annotators to come up with three content-based question-answer pairs in Czech for each of the texts. In the next step, the annotators were given the outputs from MT systems (in Czech) and were tasked to answer the questions using the corresponding translation as the context. For each system, the percentage of properly answered questions was measured.

We believe no prior work examines the usage of automatically generated questions and answers to assess the quality of MT systems.

2.3 Keyphrase Extraction

Keyphrases are representative and characteristic phrases from a text that express the key aspects of its content (Papagiannopoulou and Tsoumakas, 2020). In our work, keyphrases play the role of answers, i.e., the pieces of information which we test to be preserved in translation.

In recent years, a wide range of supervised and unsupervised keyphrase extraction methods have been proposed. Unsupervised methods normally perform two main steps to extract keyphrases: 1) select candidate phrases based on some heuristics such as matching with a specific part-of-speech pattern; 2) rank the candidates and select the top ones. Various approaches have been proposed to address this problem such as statistics-based (Won et al., 2019), graph-based (Mihalcea and Tarau, 2004), topic models-based (Liu et al., 2010), and language model-based (Tomokiyo and Hurst, 2003) methods.

On the other hand, supervised methods are relying on labeled data in which keyphrases are annotated in the documents. Supervised methods

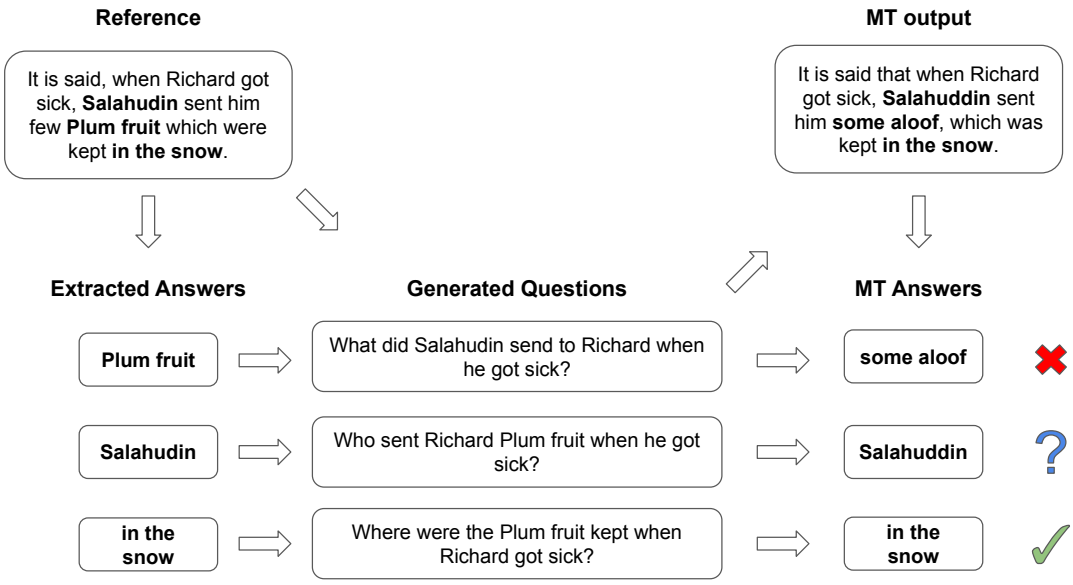

Figure 1: An illustration of the MTEQA pipeline. One of the MT answers is clearly wrong, one is correct but the other differs with just a single character, raising a question about the choice of the answer-comparison metric.

generally model the keyphrase extraction problem as binary classification to predict whether a candidate phrase is a keyphrase or not (Wang and Li, 2017), learning to rank to learn a ranking function that sorts the candidate phrases based on their score (Zhang et al., 2017), and sequence labeling problem (Zhang et al., 2016).

3 MTEQA

Our idea of evaluating MT quality by asking and answering questions is based on the assumption that a good translation should preserve all of the key information that one can extract from the reference. We propose to use a question answering framework as the proxy to measure this.

To check whether a piece of information is preserved, we automatically generate pairs of a question and its (gold-standard) answer from the reference translation and employ 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.

We assume that if it was possible to answer a question looking only at the reference, it should also be possible to answer this question looking only at the MT output and that the two answers should be identical or very similar.

In principle, the proposed MTEQA metric requires solving the following tasks:

1) Answer extraction identifies the key information in a sentence (keyphrases) which should be also present in the MT output. This extraction can be treated in a hierarchical/nested manner. For instance, given the sentence *"Today for dinner I had an organic pasta with garlic."*, the question *"What did you have for dinner today?"* can be correctly answered by all the following phrases *pasta*, *organic pasta* and *organic pasta with garlic*. Thus, answer extraction is performed first and the questions are generated afterwards for each of the answers independently. The same question can be paired with multiple (nested) answers which allows capturing a partial correspondence.

2) Question generation, given a reference translation, produces a human readable question, for which a given keyphrase is the correct answer. For each of the extracted answers, each question is generated independently from the other answers.

3) Question answering generates an answer, given a natural language question and a sentence used as a context. Since we assume that the MT output should carry enough information to answer any question asked based on the reference, we do not consider the non-answerable questions.

4) Answer comparison assesses to what extent the generated answer is correct, given the gold-

Pattern	Extracted Answer	Sentence
NOUN	Coldplay	... the British rock group Coldplay with special guest performers ...
ADJ NOUN	natural grass	As is customary for Super Bowl games played at natural grass stadiums ...
DET NOUN	a fumble	... including a fumble which they recovered for a touchdown ...
NUM NOUN	10 times	The South Florida/Miami area has previously hosted the event 10 times ...
PROPN PROPN	Carolina Panthers	... the National Football Conference (NFC) champion Carolina Panthers ...
DET ADJ NOUN	A professional fundraiser	A professional fundraiser will aid in finding business sponsors ...
DET VERB NOUN	a broken arm	... went down with a broken arm in the NFC Championship Game ...
NUM PUNCT NUM	15–1	The Panthers finished the regular season with a 15–1 record ...
DET NOUN ADP NOUN	the application of electricity	Tesla theorized that the application of electricity to the brain ...

Table 1: Examples of the most frequent POS patterns of gold-standard answers in the XQuAD dataset.

	BLEU	ROUGE-L	F1
Question Answering	-	-	90.27
Question Generation	21.01	43.25	-

Table 2: Performance of the baseline model used in our experiments on the development set of SQuADv1.

standard answer extracted from the reference. Metrics based on exact match should be avoided because they are too strict. For example, given the gold-standard answer *"Tchaikovsky"*, both the *"Tchaikovski"* and *"Beethoven"* would get the same score.

3.1 Scoring Procedure

The entire procedure of MTEQA is illustrated in Figure 1. Formally, for a given segment s_i, reference translation r_i and MT system output t_i, it proceeds as follows:

1. Generate the gold-standard answers $a_{i1}, a_{i2}, \ldots, a_{ik}$ from the reference r_i
2. For each answer a_{ij} and reference r_i, generate a natural language question q_{ij}
3. Answer each question q_{ij} using the MT output t_i as a context, obtaining answer $\tilde{a}_{ij}$
4. The final score for a given translation of a segment s_i, is the average over all generated questions:

$$MTEQA(t_i) = \frac{\sum_1^k D(a_{ij}, \tilde{a}_{ij})}{k},$$

where $D(\cdot, \cdot)$ is a string-comparison metric used to compare the two answers and k is the number of gold-standard answers extracted from the reference.

For the task of comparing MT systems on the entire test-set (i.e. system-level comparison) or at the document-level, we simply report the average of the segment-level scores. When more than one reference $\hat{r}_i$ is available for a given segment, we can use it to generate additional questions and answers.

3.2 Baseline Implementation

Our implementation of the proposed 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[3]. Performance on the development set of SQuADv1 in Table 2. We report word-level F1 for question answering and BLEU and ROUGE-L for question generation.

The SQuAD dataset was created manually by tasking the crowd-workers to create up to five questions-answer pairs from a single paragraph from Wikipedia. While the crowd-workers were encouraged to formulate the questions in their own words, the answers were restricted to be continuous sub-sequences of words from the given paragraph. In MTEQA, the answers generated by this model are also continuous sub-sequences of words from the reference and test translations.

The same system is also used for question answering and question generation by prompting the model with a different initial token in the input – for Question Answering:

```
"question: {question_text}
 context: {context_text}"
```

for Question Generation:

```
"answer: {answer_text}
 context: {context_text}".
```

3.3 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 formulate a question about.

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

	cs-en 12	de-en 12	zh-en 16	avg	en-de 14	en-cs 12
MTEQA F1	0.782*	0.997*	0.952*	0.893*	0.946*	0.845*
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 3: 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).

Considering the whole predictive power of our metric is based on questions, we propose two methods of generating additional questions.

1) We exploit the MT output as an additional source of question/answer pairs. After following the standard procedure, we swap the roles of MT output and reference – we generate gold-standard answers and questions from the MT output, and use reference as a context to answer it. As a final score we take the sum of the two scores.

2) We add keyphrases extracted by linguistic processing of the sentences based on Part-of-Speech (POS) pattern matching and Named Entity Recognition (NER). 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. Table 1 shows some examples of the extracted POS patterns. 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.

3.4 Choice of the $D(\cdot,\cdot)$ Metric

As already pointed, selection of the $D(\cdot,\cdot)$ might be crucial for optimal performance of the proposed metric and thus we consider several options. Motivated by QA evaluation, we employ the word-level F1 (Rajpurkar et al., 2016; Trischler et al., 2017; Chen et al., 2019; Durmus et al., 2020). Motivated by MT evaluation we also consider the BLEU (Papineni et al., 2002) metric and the CHRF (Popović, 2015) metric. Finally we also employ "exact match" (Rajpurkar et al., 2016) score, mainly for comparison. All of the metrics we use operate on a surface level and assign a similarity score for a pair of strings. In the future, it may be worth to explore e.g. cosine similarity between word embeddings.

4 Experiments

We evaluate the proposed MTEQA metric using the submissions to the WMT20 News translation task (Barrault et al., 2020) and their (direct) human assessments (DA). For each of the MT systems participating in the task, we compute a single score as the average of segment-level scores and report the system-level Pearson correlation with the human assessment. We report individual results for selected translation directions into English plus aggregated results (averages) for all to-English directions which were part of the WTM20 Metric Task (Mathur et al., 2020b) evaluation campaign[4].

4.1 Baseline

The baseline implementation is described in Section 3. It 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 (Section 3.4).

The results of this system are shown in Table 3 labeled as MTEQA F1 together with other metrics for comparison. We experiment with the to-English direction, since the SQuADv1 dataset used for fine-tuning is in English. On average, the baseline

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

	cs-en	de-en	zh-en	ja-en	ru-en	ps-en	avg	en-de 14	en-cs 12
MTEQA F1	0.782	0.997	0.952	0.982	0.908	0.982	0.893	0.946	0.845
MTEQA CHRF	0.796	0.996	**0.959**	0.982	0.901	0.980	0.887	0.950	0.815
MTEQA BLEU	0.762	0.998	0.954	**0.983**	0.925	0.985	0.894	**0.957**	0.840
MTEQA EXACT	0.762	0.998	0.954	0.966	0.910	0.986	0.883	0.950	0.874
MTEQA F1 OUT	0.808	0.998	0.949	0.980	0.917	0.984	0.891	-	-
MTEQA CHRF OUT	0.835	0.997	0.957	0.979	0.910	0.986	0.891	-	-
MTEQA BLEU OUT	0.809	0.998	0.950	0.981	0.929	0.984	0.896	-	-
MTEQA EXACT OUT	0.827	**0.999**	0.948	0.969	0.902	0.983	0.884	-	-
MTEQA F1 KEYPHRASE	0.851	0.998	0.944	0.978	0.930	0.986	0.896	0.941	0.877
MTEQA CHRF KEYPHRASE	**0.890**	0.998	0.951	0.978	0.927	0.981	**0.905**	0.952	0.859
MTEQA BLEU KEYPHRASE	0.844	0.998	0.939	0.973	**0.945**	0.991	0.900	0.943	0.873
MTEQA EXACT KEYPHRASE	0.858	0.997	0.938	0.959	0.936	0.990	0.893	0.948	**0.915**
MTEQA F1 OUT KEYPHRASE	0.831	0.998	0.942	0.978	0.914	**0.992**	0.893	-	-
MTEQA CHRF OUT KEYPHRASE	0.851	0.998	0.947	0.977	0.917	0.990	0.902	-	-
MTEQA BLEU OUT KEYPHRASE	0.842	0.998	0.938	0.971	0.913	0.990	0.895	-	-
MTEQA EXACT OUT KEYPHRASE	0.838	0.998	0.936	0.960	0.918	**0.992**	0.887	-	-

Table 4: System-level Pearson correlation for various variants of the proposed metric with DA human assessment over MT systems using the *newstest2020* references. Average is computed over all to-English directions available.

outperforms the traditional MT evaluation metrics (SENTBLEU, BLEU) as well as the recently proposed ones that performed very well in the WTM20 Metric Task (PRISM (Thompson and Post, 2020), YISI-2), though for some of the translation directions (e.g. Czech-English) MTEQA F1 is much worse (but for Czech-English, YISI-2 also does not beat BLEU).

4.2 Variants of the $D(\cdot, \cdot)$ metric

To assess the effect of choice of the $D(\cdot, \cdot)$ metric, we modified the baseline to exploit other options (see Section 3.4). The results are shown in the first section of Table 4. Unsurprisingly, the worst results are achieved by MTEQA EXACT which requires exact match of the test answer and the gold-standard one. But overall, the differences here are not large.

4.3 Generating Additional Answers

In general, the T5 model fine-tuned on the SQuADv1 dataset does not generate plentiful question/answer pairs. In fact, the average number of such pairs that are generated for an English sentence is only around two. Table 5 (row *baseline*) presents exact figures from our experiments, i.e., the average numbers of questions generated from a single segment of the *newstest2020* reference files for selected translation directions and the average computed for all directions into English.

To increase the number of question/answer pairs, we implemented the two methods described in Section 3.3 and present the results in Table 4. The systems denoted as OUT exploit question/answer pairs extracted from the references and MT outputs and the systems denoted as KEYPHRASE extract the pairs by POS pattern matching and NER.

The average correlation obtained using the MT output to generate questions (denoted as OUT) was very similar, but slightly worse than the one using just the questions from the reference. However, the method based on POS pattern matching and NER (denoted as KEYPHRASE) yielded improvements over various translation directions and answer comparison methods. The average numbers of question/answer pairs obtained by this method is shown in Table 5. It increased by the factor of 4 (approximately). Together with the CHRF metric used for answer comparison, it forms the best-performing configuration of the proposed metric. We also include its results in Table 3 . From now on, we will report our results using this variant. See Appendix A for examples of usage of different evaluation methods.

4.4 Non-English Reference

So far, all the experiments were conducted for the translations directions into English. This is given by the limitation of the T5 model which was trained on English data and most importantly by the SQuADv1 dataset which was used for fine-tuning and which is in English.

To overcome that, we used the multilingual mT5 model (Xue et al., 2021) and fine-tuned it on machine translation of SQuADv1 dataset into German by Lewis et al. (2020) and into Czech by (Macková and Straka, 2020). The results for English-Czech and English-German are included in both

	cs-en	de-en	ja-en	pl-en	zh-en	avg	en-cs	en-de
BASELINE	2.87	2.75	1.74	1.36	1.65	1.76	1.66	1.41
KEYPHRASE	13.36	12.01	6.66	5.10	8.79	6.98	9.45	8.71

Table 5: Average number of questions generated from a single segment in the *newstest2020* reference file by the baseline system (fine-tuned T5) and the keyphrase extraction method (POS pattern matching and NER). The average is computed over all to-English directions.

Tables 3 and 4. Overall, MTEQA still performs very well. It is better than the traditional metrics (SENTBLEU, BLEU) and also YISI-2 and comparable with PRISM for English-German. However, it is substantially worse than PRISM for English-Czech. Given the fact, that the system is multilingual and fine-tuned on machine-translated data, the results are encouraging and open doors for a cross-lingual setting which would not require reference translations.

4.5 Comparison with 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.

To provide a more complete picture of the performance of the proposed MTEQA metric, we also report correlation with the MQM assessments. Table 6 presents the system-level Pearson correlation of the proposed 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[5].

The results are surprising and to a large extent unintuitive. Metrics performing well in comparison with MQM are bad in comparison with DA. This issue was already discussed by Freitag et al. (2021) and we leave deeper analysis of the difference for the future when MQM labels will be available for more data and for more translation directions.

5 Conclusions

In this paper we introduced a new metric for automatic evaluation of Machine Translation systems. 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 proved that our metric is robust by conducting experiments over multiple translation directions.

We examined a linguistically motivated way of extracting key phrases from the sentence and showed that it boosts the final performance. We checked the influence of various word-level comparison metrics used to compare the test and gold-standard answers, and reported how it affects the correlation with human scores. In our work, we focused on translation directions into English. The only limiting factor in applying our metric to other translation directions is the availability of Question Generation and Question Answering systems in a given language. However, automatic translation of SQuAD can be an effective way to obtain data for training such systems.

Finally, we examined the performance against the MQM labels and compared the performance against the DA labels. While for the DA labels our metric performs close to state-of-the-art solutions, for the MQM labels there is a noticeable drop in performance.

In the future, we plan to examine the cross-lingual approach – instead of generating questions and answers from the reference, one may instead use the source directly.

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

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.

	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 6: 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).

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.

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.

Mikel L. Forcada, Carolina Scarton, Lucia Specia, Barry Haddow, and Alexandra Birch. 2018. Exploring gap filling as a cheaper alternative to reading comprehension questionnaires when evaluating machine translation for gisting. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 192–203, Brussels, Belgium. 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.

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.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Zhiyuan Liu, Wenyi Huang, Yabin Zheng, and Maosong Sun. 2010. Automatic keyphrase extraction via topic decomposition. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 366–376.

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.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. 2007. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Trans. Speech Lang. Process.*, 4(2):4–es.

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.

Eirini Papagiannopoulou and Grigorios Tsoumakas. 2020. A review of keyphrase extraction. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(2):e1339.

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.

Brian Thompson and Matt Post. 2020. Paraphrase generation as zero-shot multilingual translation: Disentangling semantic similarity from lexical and syntactic diversity. In *Proceedings of the Fifth Conference on Machine Translation*, pages 561–570, Online. Association for Computational Linguistics.

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.

Takashi Tomokiyo and Matthew Hurst. 2003. A language model approach to keyphrase extraction. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment*, pages 33–40.

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*.

Liang Wang and Sujian Li. 2017. Pku_icl at semeval-2017 task 10: Keyphrase extraction with model ensemble and external knowledge. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 934–937.

Miguel Won, Bruno Martins, and Filipa Raimundo. 2019. Automatic extraction of relevant keyphrases for the study of issue competition. In *Proceedings of the 20th international conference on computational linguistics and intelligent text processing, Berkeley, La Rochelle, France*.

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.

Qi Zhang, Yang Wang, Yeyun Gong, and Xuan-Jing Huang. 2016. Keyphrase extraction using deep recurrent neural networks on twitter. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 836–845.

Yuxiang Zhang, Yaocheng Chang, Xiaoqing Liu, Sujatha Das Gollapalli, Xiaoli Li, and Chunjing Xiao. 2017. Mike: keyphrase extraction by integrating multidimensional information. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1349–1358.

A Appendix

A.1 Answer extraction

Below we show the difference in the answer extraction process using the baseline approach as opposed to the proposed method based on POS patterns and NER tags. In both cases the same system is used for question generation.

Answer	Question
Answers extracted using the method based on POS sequences and NER tags	
the stadium	Where did the cat fall from?
an American football match	At what event did spectators catch a cat?
upper deck	What part of the stadium did the cat fall from?
A cat	What animal was caught by spectators at an American football match in Miami Gardens?
Florida	Where is Miami Gardens located?
spectators	Who caught a cat at an American football match in Miami Gardens?
Miami Gardens	Where was a cat caught by spectators at an American football match?
Answers extracted using the baseline model	
cat	What animal was caught by spectators at a football match in Miami Gardens?
Miami Gardens	Where was a cat caught by spectators at an American football match?

Table 7: Extracted keyphrases and generated corresponding questions for the sentence: `A cat was caught by spectators at an American football match in Miami Gardens, Florida, after it fell from the stadium's upper deck.`

Answer	Question
Answers extracted using the method based on POS sequences and NER tags	
Liberal	What party did Ed Davey belong to?'
vaccine passports	What did Ed Davey call 'divisive, unworkable and expensive'?
opposition	What type of opposition was there on the Covid Recovery Group?
the Covid Recovery Group	What group did Tory MPs oppose?
Ed Davey	Which Liberal Democrat leader called vaccine passports 'divisive, unworkable and expensive'?
Tory	What political party opposed vaccine passports?
leader	Who is Ed Davey?
Democrats	Along with Tory MPs, what party opposed vaccine passports?
Answers extracted using the baseline model	
Ed Davey	Which Liberal Democrat leader called vaccine passports 'divisive, unworkable and expensive'?
vaccine passports	What did Ed Davey call 'divisive, unworkable and expensive'

Table 8: Extracted keyphrases and generated corresponding questions for the sentence: `There had been opposition from Tory MPs on the Covid Recovery Group as well as the Liberal Democrats, whose leader Ed Davey called vaccine passports 'divisive, unworkable and expensive'.`

Answer	Question
Answers extracted using the method based on POS sequences and NER tags	
russischen	Welche Nationalität sind die Pelmeni?
Pelmeni	Wie ist der russische Name für Pirggen?
Piroggen	Was wird manchmal mit gebrannten Zwiebeln angerichtet?
gebratenen Zwiebeln	Mit welchen Arten von Zwiebeln werden die russischen Pelmeni angerichtet?
Answers extracted using the baseline model	
-	-

Table 9: Extracted keyphrases and generated corresponding questions for the sentence: `Ähnlich wie die russischen Pelmeni werden Piroggen manchmal mit gebratenen Zwiebeln angerichtet.`

A.2 Answer comparison

Below we show the difference between gold-standard answers extracted from the reference and test answers obtained with the Question Answering system, using the MT output as context.

Question	Gold-standard Answer	Test Answer
MT Output: The men‘s 100 metres semi-final begins at Sunnybrown Haquim (left).		
In what distance is Sani Brown Hakim in the men‘s semifinals?	100m	100 metres
Who is Sani Brown Hakim in the 100m semifinals?	the men	Sunnybrown Haquim
Who started in the men‘s 100m semifinals?	Sani Brown Hakim	Sunnybrown Haquim
MT Output: Sani Brown Hakeem (left) will start the men‘s 100 metres semi-final.		
In what distance is Sani Brown Hakim in the men‘s semifinals?	100m	100 metres
Who is Sani Brown Hakim in the 100m semifinals?	the men	Sani Brown Hakeem
Who started in the men‘s 100m semifinals?	Sani Brown Hakim	Sani Brown Hakeem

Table 10: Extracted keyphrases, generated corresponding questions and answers extracted from MT output for the reference: `Sani Brown Hakim (left) starting in the men‘s 100m semifinal.`

Question	Gold-standard Answer	Test Answer
MT Output: Recently I flew from Moscow, where I was trained," Andrei Borovikoff said.		
Who said that he flew from Moscow to study?	Andrei Borovikov	Andrei Borovikoff
Where was I studying?	Moscow	Moscow
MT Output: Recently, I flew from Moscow, where he was trained ", Andrey Borovikov told.		
Who said that he flew from Moscow to study?	Andrei Borovikov	Andrey Borovikov
Where was I studying?	Moscow	Moscow

Table 11: Extracted keyphrases, generated corresponding questions and answers extracted from MT output for the reference: `Recently I flew from Moscow where I was studying," said Andrei Borovikov.`