Proceedings Paper:

Fomicheva, M., Specia, L. orcid.org/0000-0002-5495-3128 and Guzmán, F. (2020) Multi-hypothesis machine translation evaluation. In: Jurafsky, D., Chai, J., Schluter, N. and Tetreault, J., (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 58th Annual Meeting of the Association for Computational Linguistics, 05-10 Jul 2020, Seattle, Washington. Association for Computational Linguistics , pp. 1218-1232. ISBN 9781952148255

10.18653/v1/2020.acl-main.113

Multi-Hypothesis Machine Translation Evaluation

Marina Fomincheva[1] **Lucia Specia**[1,2] **Francisco Guzmán**[3]
[1]Department of Computer Science, University of Sheffield, UK
[2]Department of Computing, Imperial College London, UK
[3]Facebook AI, Menlo Park, CA, USA
m.fomicheva@sheffield.ac.uk
l.specia@imperial.ac.uk
fguzman@fb.com

Abstract

Reliably evaluating Machine Translation (MT) through automated metrics is a long-standing problem. One of the main challenges is the fact that multiple outputs can be equally valid. Attempts to minimise this issue include metrics that relax the matching of MT output and reference strings, and the use of multiple references. The latter has been shown to significantly improve the performance of evaluation metrics. However, collecting multiple references is expensive and in practice a single reference is generally used. In this paper, we propose an alternative approach: instead of modelling linguistic variation in human reference we exploit the MT model uncertainty to generate multiple diverse translations and use these: (*i*) as surrogates to reference translations; (*ii*) to obtain a quantification of translation variability to either complement existing metric scores or (*iii*) replace references altogether. We show that for a number of popular evaluation metrics our variability estimates lead to substantial improvements in correlation with human judgements of quality by up 15%.

1 Introduction

Translation is an *open-ended* task with multiple valid solutions. There are often multiple equivalent translations for the same source sentence. This is due to inherent differences between languages and various sources of ambiguity, which is often impossible to solve without access to additional context. Furthermore, the source might suffer substantial changes in translation due to translator's need to adapt it to the target audience. With rare exceptions, translations are not literal, they can differ from the source text at any linguistic level – lexical, syntactic, semantic or even discourse – and still be considered correct. The ability to produce non-literal, more natural translations is one of the goals in the field of Machine Translation (MT). Neural MT (NMT) approaches have certainly made significant progress in this direction.

However, the diversity of possible outcomes makes it harder to evaluate MT models. Evaluation metrics (or humans in the case of monolingual manual evaluation) are given a single reference translation against which to compare the MT output. Fomicheva and Specia (2016) found differences of up to 1 point on a 1-5 point quality scale (i.e. 20%) between groups of annotators who use different references for manual evaluation. In automatic evaluation, which computes a similarity score between MT output and human reference, they found differences of up to 6 BLEU points depending on the reference used, showing that metrics strongly penalise perfectly correct translations that happen to be different from the reference provided.

Dreyer and Marcu (2012) showed that if multiple human translations are used, any automatic MT evaluation metric achieves a substantially higher correlation with human judgments. However, multiple translations are hardly ever available in practice due to the cost of collecting them.

Alternatives strategies for modelling linguistic variation in automatic MT evaluation include using paraphrasing, synonyms, or comparing linguistic structures of MT output and the reference translation (e.g. semantic role labels) instead of surface forms (§2). It is worth noticing that this line of work focuses on varying the reference translation. No existing work accounts for the diversity of possible MT outputs.

Instead of using multiple references or relaxing the string matching process, we use the MT system to generate multiple additional hypotheses representing potentially valid translation variations. We do so by exploring model uncertainty in output probability distributions.[1] To generate a diverse set

[1] We focus on sentence-level evaluation, as system-level

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1218–1232
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

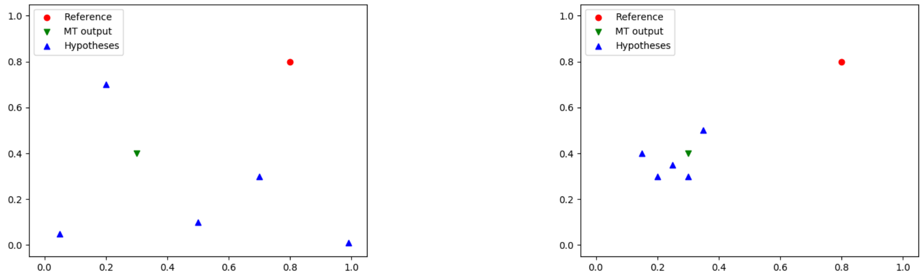

Figure 1: Hypothetical similarity space where a low-quality MT output (left) and a high-quality MT output (right) are equally distant from the reference but can be distinguished based on the similarity to additional MT hypotheses.

of hypotheses from neural MT (NMT) systems, we leverage recent work on uncertainty quantification for neural networks (§3).

The additional hypotheses produced for a given source sentence are then used for evaluation with or without human references. Intuitively, if some of the hypotheses match the reference, it is probable that the MT output under evaluation is also of high quality.[2] Furthermore, we posit that the differences between system hypotheses produced for the same source capture uncertainty. The more similar they are among themselves, the higher the confidence of the model. As illustrated in Figure 1, this could provide additional information for discriminating translation quality when measuring the distance to the reference translation does not suffice. We devise various new metrics based on this intuition and obtain large improvements in correlation with human judgments over traditional reference-based metrics.

Our main **contributions** are as follows: (1) We study different ways to generate additional MT hypotheses by exploring uncertainty in NMT models. We show that a light-weight Bayesian approximation method – *Monte Carlo Dropout*, which allows for uncertainty quantification by using dropout at inference time (Gal and Ghahramani, 2016) – works the best for the purpose of automatic MT evaluation; (2) We devise methods to effectively explore multiple MT hypotheses to better evaluate MT output quality with existing evaluation metrics. On two different datasets, we achieve a large improvement in correlation with human judgments over using both single reference and multiple references. To the best of our knowledge, this is the first work to leverage NMT model uncertainty for automatic MT evaluation.

2 Related Work

Meteor (Banerjee and Lavie, 2005) was the first MT evaluation metric to relax the exact match constraint between MT system output and reference translation by allowing matching of lemmas, synonyms or paraphrases. However, this requires linguistic resources which do not exist for most languages. Character-based metrics (Popović, 2015; Wang et al., 2016) also relax the exact word match constraint by allowing the matching of characters. However, ultimately they still assume a surface-level similarity between reference and MT.

A more recent direction compares MT and reference sentences in the embedding space. Chen and Guo (2015) extract word embedding representations for the two sentences and measure the (cosine) similarity between them. Similarly, in (Fomicheva et al., 2015; Servan et al., 2016; Tättar and Fishel, 2017) two words are considered to match if their cosine distance in the embedding space is above a certain threshold. The embeddings are thus used to provide a binary decision. MEANT 2.0 (Lo, 2017) and YISI (Lo, 2019) also relies on matching of words in the embedding space, but this is only used to score the similarity between pairs of words that have already been aligned based on their semantic roles, rather than to find the alignments between words. Finally, Chow et al. (2019) and Echizen'ya et al. (2019) perform the alignment in the embedding space using Earth Mover's Distance with some special treatment for word order. All of these metrics are however still limited to variance in the words used (even in the continuous space),

automatic evaluation can be by and large considered a solved problem (Ma et al., 2019a).

[2] The goal of this paper is not to evaluate the search space of the MT system, but to improve the evaluation of the given MT output by using additional hypotheses. Evaluating the NMT search space beyond the generated output could be an interesting direction to explore in future work.

rather than more general stylistic or structural variations which can only be captured with multiple references.

Another way of incorporating linguistic variation is pseudo-reference approach by Albrecht and Hwa (2007). They leverage various off-the-shelf MT systems to generate additional imperfect references and use them instead or alongside the original reference during evaluation. Evaluation scores obtained using each of the pseudo references and the available human references are combined as features by training a classifier to predict human judgments. Thus, this line of work implicitly learns the quality of the MT systems used to generate pseudo references. We revisit this idea in our paper by having pseudo-references as one type of diverse MT output.

3 Generating Diverse Hypotheses

We posit that using multiple MT hypotheses can help automatic MT evaluation in two ways. First, the difference between them may reflect model confidence and potential ambiguity or complexity of the source. Second, they provide an additional point of comparison with the reference, such that if the initial MT output is different from the provided reference due to acceptable linguistic variation, the risk of over-penalising this translation is lower.

3.1 Neural MT

Most recent work on NMT is based on the sequence-to-sequence approach with encoder and decoder networks (Bahdanau et al., 2014; Luong et al., 2015; Vaswani et al., 2017b). In these models probability of generating the output sequence $\vec{y}$ given the input sequence $\vec{x}$ is decomposed as follows:

$$p(\vec{y}|\vec{x},\theta) = \prod_{j=1}^{J} p(y_j|\vec{y}_{<j},\vec{x},\theta)$$

where θ represents model parameters. The decoder produces the probability distribution $p(y_j|\vec{y}_{<j},\vec{x},\theta)$ over system vocabulary at each time step using *softmax function*.

In this work we use state-of-art Transformer architecture proposed by Vaswani et al. (2017b), an encoder-decoder model that uses stacked self-attention and fully connected layers for both encoder and decoder.

3.2 Search Algorithm

One way to obtain multiple MT hypotheses is by taking top MT hypotheses resulting from the search algorithms used in NMT for decoding.

Beam Search. Hypotheses spaces in NMT are very large and it is not feasible to explore them exhaustively. Beam search is traditionally used for decoding in NMT by exploring the search space in a greedy left-to-right manner retaining the top-N candidates with the highest probability. While effective to select a likely translation, beam search tends to result in a list of N-best translations which lack linguistic diversity (Vijayakumar et al., 2016).

Diverse Beam Search. Vijayakumar et al. (2016) proposed the Diverse Beam Search algorithm to improve the diversity of top hypotheses. The algorithm promotes diversity by optimising a diversity-augmented objective.

3.3 Uncertainty

We propose that a better method for obtaining diverse MT hypotheses for automatic MT evaluation is by exploiting uncertainty in NMT. For the intuition, consider three different cases. First, if there is only one correct translation at each time step, the output probabilities will have "peakier" distributions with low entropy and a single word receiving a large portion of the probability mass. In this case, there is very little variation in the hypotheses space. Second, if there are various correct translation options at a given generation step, the output probability distribution will have higher entropy, with multiple target words receiving similar probabilities. In this case, generating hypotheses from the model will result in similar sentences containing synonyms or paraphrases. Finally, if the NMT model has not seen enough data during training for a given combination of words, we would expect output probabilities to exhibit high entropy, approximating a uniform probability distribution. In this case, generating MT hypotheses from the model should result in a highly diverse set with lower quality translations.

Below we explore various approaches to uncertainty quantification in neural networks in order to generate a set of additional hypotheses for MT evaluation.

Monte Carlo Dropout. It has been shown that softmax function used in neural networks to generate output probability distribution does not properly capture uncertainty as it produces overconfident predictions (Gal and Ghahramani, 2016). Most of the work on uncertainty quantification in deep learning relies on Bayesian formalism (MacKay, 1992; Graves, 2011; Welling and Teh, 2011; Gal and Ghahramani, 2016; Tran et al., 2019). Representing uncertainty through Bayesian neural networks usually comes with prohibitive computational costs and various approximations have been developed to alleviate this issue. One such approximation by Gal and Ghahramani (2016) is called Monte Carlo (MC) dropout. Dropout is a method developed to reduce overfitting when training neural models Srivastava et al. (2014). It consists in randomly masking neurons to zero based on Bernoulli distribution. Gal and Ghahramani (2016) use dropout at test time before every weight layer. They perform N forward passes through the network and collect posterior probabilities generated by the model with parameters perturbed by dropout: $\{\hat{p}(\vec{y}|\vec{x},\hat{\theta})_{i=1}^{N}\}$ where $\hat{\theta}$ represents the perturbed parameters. They show that this is equivalent to an approximation to the probabilistic deep Gaussian process. Previous work has applied this method to quantify model uncertainty by taking the variance of the resulting probability distribution (Dong et al., 2018; Wang et al., 2019). We instead look at the *linguistic differences* between MT hypotheses generated as a result of N forward passes through the model with perturbed parameters. If the top MT output for a given source sentence is of high quality, it is probable that other hypotheses will be similar.

Ensembling. Ensemble model combination is another strategy commonly used for estimating predictive uncertainty (Lakshminarayanan et al., 2017; Pearce et al., 2018; Liu et al., 2019). We take an ensembling strategy typically applied in NMT to improve translation quality: we train four NMT models initialised with different random seeds. At decoding time, prediction distributions from the four models are combined by averaging. To generate additional hypotheses, the four models in the ensemble are used separately, each generating an independent set of translations.

Mixture of Experts. Shen et al. (2019) applied mixture of experts (MoE) framework to capture the inherent uncertainty of the MT task and generate diverse hypotheses. A mixture model introduces a multinomial latent variable $z \in 1, ..., K$. The marginal likelihood is then decomposed as:

$$\begin{aligned} p(\vec{y}|\vec{x};\theta) &= \sum_{z=1}^{K} p(\vec{y}, z|\vec{x};\theta) \\ &= \sum_{z=1}^{K} p(z|\vec{x};\theta)p(\vec{y}|z,\vec{x};\theta) \end{aligned}$$

The model is trained with the EM algorithm where the E-step estimates the responsibilities of each mixture component (“expert”) and M-step updates parameters θ with gradients weighted by their responsibilities. For our experiments, one of the mixture components was randomly selected to produce the MT output for human evaluation and the rest of them were used for the generation of additional hypotheses (§5).

3.4 Pseudo-Reference Translations

Here we revisit the approach previously used for statistical MT (Albrecht and Hwa, 2007) where outputs of other off-the-shelf MT systems are used as additional reference translations, with some differences. First, NMT outputs on average have substantially higher quality. Second, to avoid the need for labelled data, we do not rely on supervised training and treat the outputs of other MT systems in the same way we treat additional hypotheses that were produced using the methods described in the previous sections. We use publicly available online NMT systems (§5).

4 Scoring with Multiple Hypotheses

Using the methods described above we are able to produce a set of MT hypotheses for each given source segment. The final dataset which we use for evaluation contains a human reference translation (r), the top MT output (o) and this set of alternative N MT hypotheses ($H = \{h_1..h_N\}$). We devise the following ways of combining similarities between possible translations and between these and the reference to obtain more accurate evaluation. This accuracy will be measured by Pearson correlation with a direct assessment (DA) score collected for the o translation, as is common practice in the evaluation metrics field (Ma et al., 2019b).

4.1 Addressing Linguistic Variation

Here we compute the similarity against the reference translation for the set of all generated translation candidates, including the initial MT output and additional hypotheses, and take the average similarity score (micro-average). If the MT output is of high quality but does not match the provided human reference due to acceptable linguistic variation, other hypotheses may serve as paraphrases to match the reference.

However, it is important to assign a higher weight to the MT output that was *actually* evaluated (o), as compared to the alternative MT hypotheses. This is done using a simple variant of the above metric where we first take an average of the hypotheses-reference similarities, and then average this score with the MT output-reference similarity score (macro-average). This results in two metrics:

$$\text{hyp-ref}^{*}_{\text{micro}} = N^{-1} \sum_{i=1}^{N+1} sim(h'_i, r), h'_i \in H'$$

$$\text{hyp-ref}^{*}_{\text{macro}} = \frac{N^{-1} \sum_{i=1}^{N} sim(h_i, r) + sim(o, r)}{2}$$

where $H' = \{h'_1..h'_N, o\}$ is a set including additional hypotheses and the MT output, and sim corresponds to a similarity function of choice (§4.3). The $*$ represents different ways of combining hypotheses-reference similarities: average (as shown in the equations above), minimum (i.e. choosing the score for the most distant hypothesis) and maximum (i.e. choosing the score of the closest hypotheses).

4.2 Incorporating Model Uncertainty

As discussed in Section §3.3, similarity between translation hypotheses capture model confidence and could thus be indicative of translation quality. We propose two metric variants to capture this idea. First, we compute the similarity between all translations candidates including the additional hypotheses and the MT output:

$$\text{hyp-self}^{*} = \frac{1}{C} \sum_{i=1}^{|H'|} \sum_{j=1}^{|H'|} sim(h'_i, h'_j)$$

where $h'_i \in H', i \neq$ j and $C = 2^{-1}|H'|(|H'| - 1)$ is the number of pairwise comparisons for H' hypotheses. As before, $*$ corresponds to different ways of combining similarity scores: average, minimum and maximum.

Second, as before, we give a higher weight to the MT output whose quality we wish to evaluate (o). To that end we compare the MT output against additional generated hypotheses. This comparison indirectly captures the similarity between MT hypotheses themselves:

$$\text{hyp-mt}^{*} = N^{-1} \sum_{i=1}^{N} sim(h_i, o)$$

Both of these variants can be used with and without reference translation. Interestingly, as will be shown in §6.2, they perform comparably to other methods even without the reference, putting into question the need for human reference in MT evaluation. As in the previous section, to add human reference translations into the mix, we average the results as follows:

$$\text{hyp-mt}^{*}\text{-ref} = \frac{N^{-1} \sum_{i=1}^{N} sim(h_i, o) + sim(o, r)}{2}$$

Figure 2 summarises the methods discussed above.

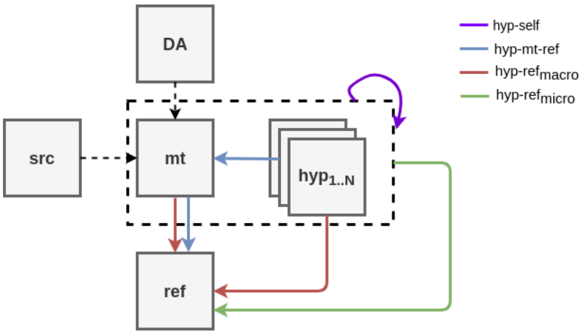

Figure 2: Methods to explore similarities between MT output, system hypotheses and references.

4.3 Similarity Functions

To measure similarity amongst hypotheses and against the reference(s), we experiment with the following standard MT evaluation metrics:[3]

sentBLEU (Papineni et al., 2002). BLEU measures the similarity between MT and the reference translation based on the number of matching n-grams. We use a smoothed version of BLEU as described by Lin and Och (2004) with N = 4.

[3] We use these metrics out of the box. Better results could possibly be achieved by adapting them to our settings, e.g. by changing the weight of precision and recall depending on the direction of the comparison between MT output, hypotheses and the reference. For instance, when using BLEU as similarity function for computing hyp-mt*, we are evaluating recall on the MT output, whereas BLEU is designed as a precision-oriented metric. But the choice of similarity function is orthogonal to the goal of this paper, and we leave further refinements in this direction to future work.

TER (Translation Edit Rate) (Snover et al., 2006). TER computes the edit distance defined as the minimum number of word substitutions, deletions, insertions and shifts that are needed to convert MT into the reference.

ChrF (Popović, 2015). ChrF calculates the F-score of character n-grams of maximum length 6.

Meteor (Denkowski and Lavie, 2014). Meteor aligns MT output to the reference translation using synonyms and paraphrases besides exact word matching. The similarity is based on the proportion of aligned words in the candidate and in the reference and a fragmentation penalty.

BERTScore. (Zhang et al., 2019). We also looked at this very recent metric (published after the submission of this paper), which uses powerful pre-trained embeddings. BERTScore computes a cosine similarity score for each token in the MT output with each token in the reference sentence using contextual embeddings from BERT (Devlin et al., 2019), which can generate different vector representations for the same word depending on the context, thus better capturing meaning. Maximum similarity values for MT and reference words are then used to compute a soft F1-score. We use the implementation available at https://github.com/Tiiiger/bert_score.

5 Experimental Settings

To test whether our methods improve correlation with human judgments, we need to have access to the NMT model and human judgments for the translations generated by this model. This data is not generally readily available in evaluation campaigns such as Metrics Task at WMT conferences. Below we describe two datasets that satisfy these conditions. They cover two different language pairs and two different domains.

News English-Czech dataset. We use available data from the WMT19 News Translation Task. We focus on the University of Edinburgh's submission (Bawden et al., 2019) to the English-Czech translation task, since its NMT model is available. The system was trained using the MarianMT toolkit with a standard Transformer architecture (Vaswani et al., 2017a). Details on model training and architecture are described in (Bawden et al., 2019). For producing pseudo-references, we use all five "online" systems whose submissions were provided as part of the WMT19 Translation Task.

Human judgments were collected in the form of Direct Assessments (DA) following the methodology proposed by Graham et al. (2015), which suggests that 15 segment-level DA judgements are required for trustworthy correlation analysis. However the number of DA judgements in the WMT19 Metrics Task was much smaller. We select segments with at least two DA annotations (795 segments with an average DA score of 80.22) to minimise this issue, but the results reported here for English-Czech should be interpreted with caution.

Wikipedia Estonian-English dataset. This is a new dataset we collected which contains 1K sentences randomly selected from Wikipedia articles in Estonian and translated into English. Two human reference translations were generated independently by two professional translators.

All the NMT models were trained using the Fairseq toolkit based on the standard Transformer architecture (Vaswani et al., 2017a) and the training settings described in Ott et al. (2018). We used publicly available parallel datasets for training the models: the Rapid corpus of EU press releases (Rozis and Skadiņš, 2017) and Europarl (Koehn, 2005), which amount to around 4M parallel sentences in total.

A set of 400 segments were translated by the model variants described in §3 to assess the impact of uncertainty types. The following settings were used for model variants. For MC dropout we use dropout rate of 0.3, same as for training the basic Transformer model. Additional hypotheses were produced by performing N stochastic forward passes through the network with dropout, as described in §3. For this analysis we use $N = 30$, which was shown to perform well for uncertainty quantification (Dong et al., 2018). We also test how the number of hypotheses affects the results (see Appendix B). For MoE we use hard mixture model with uniform prior and $K = 5$ mixture components. To produce the translations we generate from a randomly chosen component with greedy search following the settings in Shen et al. (2019). For generating additional hypotheses with beam search the top-K sentences K $\in$ [2..5] from the beam were used ($K = 1$ corresponds to the initial MT output). For pseudo-reference approach we use three online systems: Systran, Google and Bing.

Human judgements were given by professional translators following the FLORES setup (Guzmán et al., 2019) which presents a form of DA judgements (Graham et al., 2013). The annotators were asked to rate each sentence from 0–100 according to the perceived translation quality. Specifically, the 0–10 range represents an incorrect translation; 11–29, a translation with few correct keywords, but the overall meaning is different from the source; 30–50, a translation with major mistakes; 51–69, a translation which is understandable and conveys the overall meaning of the source but contains typos or grammatical errors; 70–90, a translation that closely preserves the semantics of the source sentence; and 90–100, a perfect translation. Each segment was annotated by up to 6 translators. Raw scores were converted into *z-scores*, i.e. standardised according to each individual annotator's overall mean and standard deviation. The scores collected for each segment were averaged to obtain the final score.

The judgments were collected for the 1K segments translated by the standard Transformer model and for the 400 segments produced by four MT model variants in §3, resulting in a total of $1000 + 4 * 400 = 2600$ source-MT pairs annotated with DA judgments. The distribution of DA scores for English-Czech and 1K Estonian-English datasets is shown in the Appendix A.[4]

6 Results

In this section, we present the results of our experiments for generating additional MT hypotheses (§6.1) and the methods for exploiting similarities between them (§6.2).

6.1 Diverse MT Generation Approaches

We start by comparing the different strategies for generating multiple MT hypotheses described in §3 for the Estonian-English dataset. Note that some variants also produce different top MT outputs (o), as they were trained using different architectures or decoding algorithms. As a result we have four sets of DA annotations collected for 400 segments for system variants with different MT outputs: standard Transformer, Transformer with diverse beam search, MoE and ensembling. MT outputs for beam search and MC dropout variants correspond to the same underlying NMT model.

Table 1 presents the results. First, beam search performs the poorest. This is in line with the well known fact that beam suffers from low diversity of produced hypotheses (Vijayakumar et al., 2016). As expected, diverse beam search results in a higher difference in correlation compared to *mt-ref*. However, it is still outperformed by all other methods that capture model uncertainty, with MC dropout achieving the highest difference in correlation against mt-ref. We note that this is not related to the number of generated hypotheses (see Appendix B for details). We suggest that this is due to the fact that linguistic differences between additional hypotheses for high vs. low-quality MT outputs is more discriminating when the hypotheses are generated using MC dropout for representing model uncertainty (see example in Table 3). The difference in correlation observed between different system variants is not related with the quality of MT outputs, as demonstrated by the average DA scores in Table 1. Pseudo-references also perform very well, potentially due to the high quality of the MT systems used to generate them. We select MC dropout and pseudo-references as the two best performing options to conduct a more detailed analysis below.

6.2 Scoring Approaches

Table 2 shows the results for the 1K Estonian-English dataset and for English-Czech dataset.[5] *mt-ref* stands for the standard reference-based evaluation. The remaining methods correspond to those described in §4. The methods *pseudo-mt-max* and *pseudo-mt-max-ref* are equivalent to the *hyp-mt-** and *hyp-mt-*-ref* but instead of dropout-based hypotheses, the outputs of other MT systems are used.

For Estonian-English, since we have two human references we compute the correlation for each of them separately (*mt-ref-1* and *mt-ref-2*), as well as in a multi-reference scenario (*mt-ref-multi*).[6] We use *mt-ref-1* to calculate all the remaining methods that involve a reference translation.

Significance of the differences in correlation for the proposed methods with respect to *mt-ref-1* and *mt-ref-multi* is assessed using Hotelling-Williams

[4] The dataset and the NMT models required to reproduce our results are available at https://github.com/facebookresearch/mlqe/tree/master/data-multi-hyp.

[5] For the full set of results see the Appendix C.

[6] In the multi-reference scenario, BLEU score is computed by counting the n-gram matches between the MT output and all references as in (Papineni et al., 2002). For the rest of the metrics, the closest reference is used for each segment to compute the score, as in (Denkowski and Lavie, 2014).

	Top-5 Beam	Top-5 Diverse Beam	MC-dropout	MoE	Ensemble	Pseudo
mt-ref	0.316	0.340	0.316	0.286	0.312	0.316
hyp-ref-avg$_{\text{macro}}$	0.325	0.345	0.323	0.310	0.354	0.167
hyp-ref-avg$_{\text{micro}}$	0.327	0.345	0.319	0.316	0.372	0.223
hyp-mt-avg-ref	0.275	0.380	**0.438**	0.371	0.413	**0.408**
hyp-mt-avg	0.022	0.340	**0.433**	0.352	0.388	**0.424**
Average DA	58.88	55.12	58.88	51.20	61.19	58.88

Table 1: Pearson correlation with human judgments for single reference (mt-ref) and the metrics based on MT hypotheses in §4 for the different MT systems generating diverse MT in §3 for the Estonian-English dataset, using BLEU as similarity function. The last row shows the average absolute DA scores for each model variant.

	Estonian-English					English-Czech				
	BLEU	TER	ChrF	Meteor	BERT	BLEU	TER	ChrF	Meteor	BERT
mt-ref-1	0.417	-0.413	0.508	0.550	0.653	0.312	-0.335	0.346	0.380	0.422
mt-ref-2	0.432	-0.436	0.521	0.521	0.662	-	-	-	-	-
mt-ref-multi	0.494	-0.497	0.554	0.547	0.688	-	-	-	-	-
pseudo-mt-max	0.526	-0.350	**0.589**	0.550	0.672	0.271	-0.219	0.285	0.279	0.350
pseudo-mt-max-ref	**0.539**	-0.453	**0.600**	**0.607**	0.699	0.352	-0.346	0.398	0.388	0.473
hyp-ref-avg$_{macro}$	0.448	-0.445	0.532	**0.580**	0.671	0.319	-0.334	0.348	0.382	0.423
hyp-ref-avg$_{micro}$	0.455	-0.454	0.532	**0.583**	0.658	0.320	-0.328	0.346	0.380	0.419
hyp-mt-avg-ref	**0.553**	**-0.543**	**0.601**	**0.638**	0.700	0.378	-0.382	0.389	0.431	0.464
hyp-mt-avg	**0.562**	**-0.548**	**0.597**	**0.610**	0.665	0.296	-0.281	0.288	0.334	0.326
hyp-self-avg	**0.557**	**-0.567**	**0.589**	**0.614**	0.660	0.313	-0.306	0.309	0.351	0.355

Table 2: Pearson correlation with human judgments for single reference (mt-ref-1/2), multiple references (mt-ref-multi) and the metrics based on MT hypotheses in §4 for the Estonian-English 1K dataset and English-Czech dataset. Results significantly improving on single-reference evaluation are underlined, and those significantly improving on multi-reference evaluation are marked in bold. We use Hotelling-Williams test (Williams, 1959) for significant differences in correlation.

	type	**text**	**info**
high-quality	**Source**	Siis aga võib tekkida seesmise ja välise vaate vahele lõhe.	DA = 75
	Reference	This could however lead to a split between the inner and outer view.	ME = 0.262
	MT Output	Then there may be a split between internal and external viewpoints.	
	Dropout	Then, however, there may be a split between internal and external viewpoints. Then, however, there may be a gap between internal and external viewpoints. Then there may be a split between internal and external viewpoints. Then there may be a split between internal and external viewpoints.	HY = 0.532
	Beam	Then there may be a split between internal and external viewpoints. Then there may be a gap between internal and external viewpoints. Then there may be a split between internal and external viewpoints. Then there may be a gap between internal and external viewpoints.	
low-quality	**Source**	Kant on see, kellele kuulub see teene, et ta täiustas mateeria käsitust seeläbi, et ta vaatles seda tõukumise ja tõmbumise ühtsusena.	DA = 3
	Reference	It is Kant who has the merit of refining the concept of matter by seeing it as a unity of pushing and pulling.	ME = 0.304
	MT Output	It is the person who owes it to the merit of pardoning it by looking at it as a unity of push and withdrawal.	
	Dropout	It is the person who owes it to the merit of pardoning this approach by looking at it as a unity of push and withdrawal. It is Mrs Kant who owes to the fact that he has perfected his approach by looking at it as a unity of impetus and resignation. It is the one who owes the service that he has perfected his approach by seeing it as a catalyst and a stand-alone. It is the person who owes it to the merit of perfecting this approach by looking at it as a means of push and withdrawal.	HY = 0.182
	Beam	It is the person who owes it to the merit of pardoning it by seeing it as a unity of push and withdrawal. It is the person who owes it to the merit of pardonating the concept by looking at it as a unity of push and withdrawal. It is the person who owes it to the merit of pardoning it by looking at it as a unity of impetus and withdrawal. It is the person who owes it to the merit of pardoning it by seeing it as a unity of impetus and withdrawal.	

Table 3: Example of MC dropout and beam search hypotheses for a high-quality and a low-quality MT output. The last column shows the DA score for these two translations, as well as the Meteor score (ME) and our hyp-mt-avg-ref (HY) score obtained for them.

test (Williams, 1959), as described in Graham et al. (2015).

First, we observe that the methods based on the similarity against the reference (*hyp-ref-**) do not perform as well as those relying more on the relation between MT hypotheses (*hyp-mt-**). As discussed in §4, the latter capture the uncertainty of NMT models when generating the output for a given source sentence. Overall, *hyp-mt-avg-ref* consistently outperforms all the other variants by a large margin, for all automatic evaluation metrics considered. Logically, the improvement is larger for exact-matching metrics, but also significant for Meteor, ChrF and BERTScore, which attempt to capture linguistic variation.

Surprisingly, *hyp-mt-avg-ref* performs better than the *mt-ref-multi*. Reasons may be that it can potentially cover a larger number of paraphrases than one additional reference translation, and that besides computing similarity to a reference translation, it incorporates information on model uncertainty.

Interestingly, our reference-free metric *hyp-mt-avg*, which only compares the MT output against additional generated hypotheses and does not rely on human references, also performs competitively. This result confirms the important role played by the model confidence component in measuring MT quality. Note that for Estonian-English dataset it performs better than the evaluation with single reference, indicating that model confidence alone can be more reliable for assessing MT quality than using a single reference translation.

Finally, we observe that using translations from online MT systems also outperforms reference-based evaluation. The differences are larger for Estonian-English. This could be because for into-English translation the quality of pseudo-references is higher, making them as good as actual reference translations, while yet closer to the MT output under evaluation. For English-Czech, pseudo-references are closer to *mt-ref* and generally worse than *hyp-mt-avg-ref*.

Table 3 illustrates the advantage of our uncertainty-aware evaluation over standard reference-based scoring. We show MC dropout and top beam hypotheses for a high quality and for a low quality MT output. First, note that MC dropout hypotheses are very different for a low-quality MT output and fairly similar for good-quality translation. By contrast, beam hypotheses are similar or the same in both cases. Second, the evaluation scores obtained using MC dropout hypotheses result in a large difference between low-quality and high-quality MT outputs, whereas Meteor assigns a higher score to the low-quality example due to surface word and synonym matches that are in this case not indicative of MT quality.

The proposed approach has some limitations. First, it requires access to the NMT system that was used to generate the translations. Second, we note that this idea works better if the NMT model is reasonably well trained, as additional hypotheses could be less informative otherwise. Finally, it is not clear how the methods presented here would work for comparing the output quality of different MT systems, but this is a different application of our proposed approach and we leave this question to future work.

7 Conclusions

We proposed to explore NMT model uncertainty to generate additional hypotheses for MT evaluation. We showed that by exploiting similarities in the space of translation hypotheses generated by the model, along with methods to effectively combine information from these multiple hypotheses, we can achieve more accurate estimation on the quality of MT output than standard reference-based comparison, including cases with multiple references. This suggests that model uncertainty alone can be more reliable for assessing MT quality than standard reference-based evaluation.

This work can be extended in numerous ways. First, we plan to test whether similar observations will hold for more language pairs and text domains. Second, the score combination strategies could be improved by learning weights for each component. Finally, we would like to test this approach for comparing different MT systems.

Acknowledgments

Marina Fomicheva and Lucia Specia were supported by funding from the Bergamot project (EU H2020 Grant No. 825303). We thank Mark Fishel (University of Tartu) for collecting one set of Estonian-English references.

References

Joshua Albrecht and Rebecca Hwa. 2007. Regression for sentence-level mt evaluation with pseudo references. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 296–303.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

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.

Rachel Bawden, Nikolay Bogoychev, Ulrich Germann, Roman Grundkiewicz, Faheem Kirefu, Antonio Valerio Miceli Barone, and Alexandra Birch. 2019. The university of edinburgh's submissions to the wmt19 news translation task. *arXiv preprint arXiv:1907.05854*.

Boxing Chen and Hongyu Guo. 2015. Representation based translation evaluation metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 150–155, Beijing, China. Association for Computational Linguistics.

Julian Chow, Lucia Specia, and Pranava Madhyastha. 2019. Wmdo: Fluency-based word moverâ$^{\text{TM}}$s distance for machine translation evaluation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 494–500, Florence, Italy.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

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.

Li Dong, Chris Quirk, and Mirella Lapata. 2018. Confidence Modeling for Neural Semantic Parsing. *arXiv preprint arXiv:1805.04604*.

Markus Dreyer and Daniel Marcu. 2012. Hyter: Meaning-equivalent semantics for translation evaluation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 162–171. Association for Computational Linguistics.

Hiroshi Echizen'ya, Kenji Araki, and Eduard Hovy. 2019. Word embedding-based automatic MT evaluation metric using word position information. 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 1874–1883, Minneapolis, Minnesota. Association for Computational Linguistics.

Marina Fomicheva, Núria Bel Rafecas, Iria Da Cunha Fanego, and Anton Malinovskiy. 2015. Upf-cobalt submission to wmt15 metrics task. In *Bojar O, Chatterjee R, Federmann C, Haddow B, Hokamp C, Huck M, Logacheva V, Pecina P, editors. Proceedings of the Tenth Workshop on Statistical Machine Translation; 2015 Sep 17-18; Lisboa, Portugal: Association for Computational Linguistics; 2015. p. 373-9.* ACL (Association for Computational Linguistics).

Marina Fomicheva and Lucia Specia. 2016. Reference bias in monolingual machine translation evaluation. In *54th Annual Meeting of the Association for Computational Linguistics*, pages 77–82. Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *international conference on machine learning*, pages 1050–1059.

Yvette Graham, Timothy Baldwin, and Nitika Mathur. 2015. Accurate evaluation of segment-level machine translation metrics. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1183–1191.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41.

Alex Graves. 2011. Practical variational inference for neural networks. In *Advances in neural information processing systems*, pages 2348–2356.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. 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 6097–6110, Hong Kong, China. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. Citeseer.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6413.

Chin-Yew Lin and Franz Josef Och. 2004. Orange: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation. In *Proceedings of the 20th international conference on Computational Linguistics*, page 501. Association for Computational Linguistics.

Jeremiah Liu, John Paisley, Marianthi-Anna Kioumourtzoglou, and Brent Coull. 2019. Accurate uncertainty estimation and decomposition in ensemble learning. In *Advances in Neural Information Processing Systems*, pages 8950–8961.

Chi-kiu Lo. 2017. Meant 2.0: Accurate semantic mt evaluation for any output language. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, 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.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019a. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019b. Results of the wmt19 metrics shared task: Segment-level and strong mt systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90.

David JC MacKay. 1992. *Bayesian methods for adaptive models*. Ph.D. thesis, California Institute of Technology.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 1–9.

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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Tim Pearce, Mohamed Zaki, Alexandra Brintrup, and Andy Neel. 2018. Uncertainty in neural networks: Bayesian ensembling. *arXiv preprint arXiv:1810.05546*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 372–375, Lisboa, Portugal.

Roberts Rozis and Raivis Skadiņš. 2017. Tilde model-multilingual open data for eu languages. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 263–265. Tilde Model RAPID Corpus and EU data therein licensed under CC-BY 4.0.

Christophe Servan, Alexandre Berard, Zied Elloumi, Hervé Blanchon, and Laurent Besacier. 2016. Word2Vec vs DBnary: Augmenting METEOR using Vector Representations or Lexical Resources? *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1159–1168.

Tianxiao Shen, Myle Ott, Michael Auli, and Marc'Aurelio Ranzato. 2019. Mixture models for diverse machine translation: Tricks of the trade. *arXiv preprint arXiv:1902.07816*.

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*, volume 200.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Dustin Tran, Mike Dusenberry, Mark van der Wilk, and Danijar Hafner. 2019. Bayesian layers: A module for neural network uncertainty. In *Advances in Neural Information Processing Systems*, pages 14633–14645.

Andre Tättar and Mark Fishel. 2017. bleu2vec: the painfully familiar metric on continuous vector space steroids. *Proceedings of the Second Conference on Machine Translation*, page 619–622.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017a. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017b. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Shuo Wang, Yang Liu, Chao Wang, Huanbo Luan, and Maosong Sun. 2019. Improving Back-Translation with Uncertainty-based Confidence Estimation. *arXiv preprint arXiv:1909.00157*.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. Character: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 505–510.

Max Welling and Yee W Teh. 2011. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688.

Evan James Williams. 1959. *Regression Analysis*, volume 14. Wiley, New York, USA.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Distribution of DA Scores

Figure 3 shows the distribution of DA scores for Estonian-English 1K and English-Czech datasets.

B Number of MC Dropout Hypotheses

Figure 4 Pearson correlation with human judgments for hyp-mt-avg-ref with sentBLEU metric as a function of the number of stochastic forward passes with MC dropout. The improvements in correlation become small after 10 hypotheses for both Estonian-English and English-Czech.

C Combination Methods

Tables 4 and 5 show a full set of Pearson correlation results for the scoring approaches described in Section 3.2.

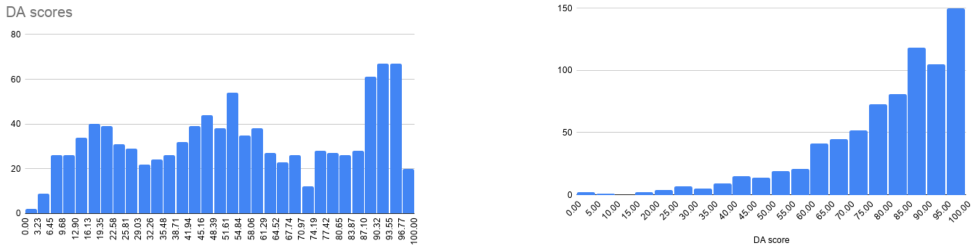

Figure 3: Distribution of DA scores for Estonian-English 1K and English-Czech datasets

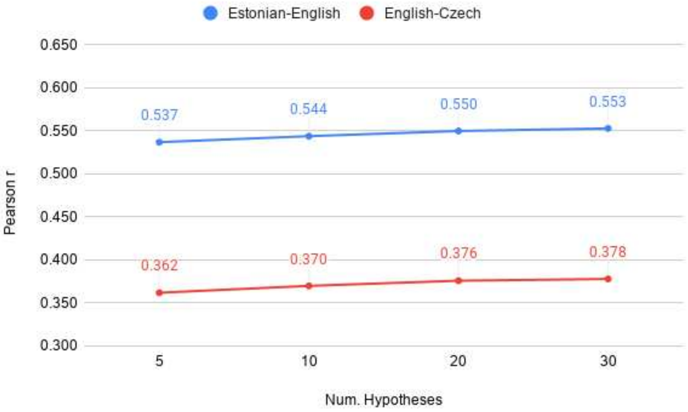

Figure 4: Pearson correlation with human judgments for hyp-mt-avg-ref with sent-BLEU metric as a function of the number of stochastic forward passes with MC dropout

	BLEU	TER	ChrF	Meteor	BERT
mt-ref$_1$	0.417	-0.413	0.508	0.550	0.653
mt-ref$_2$	0.432	-0.436	0.521	0.521	0.662
mt-ref-multi	0.494	-0.497	0.554	0.547	0.688
pseudo-mt-avg	0.494	-0.485	0.541	0.545	0.547
pseudo-mt-min	0.363	-0.531	0.396	0.400	0.289
pseudo-mt-max	0.526	-0.350	0.589	0.550	0.672
pseudo-mt-avg-ref	0.508	-0.506	0.570	0.604	0.657
pseudo-mt-min-ref	0.445	-0.531	0.505	0.547	0.493
pseudo-mt-max-ref	0.539	-0.453	0.600	0.607	0.699
hyp-ref-avg$_{\text{macro}}$	0.448	-0.445	0.532	0.580	0.671
hyp-ref-min$_{\text{macro}}$	0.438	-0.443	0.514	0.586	0.667
hyp-ref-max$_{\text{macro}}$	0.446	-0.440	0.536	0.548	0.648
hyp-ref-avg$_{\text{micro}}$	0.455	-0.454	0.532	0.583	0.658
hyp-ref-min$_{\text{micro}}$	0.388	-0.429	0.460	0.555	0.622
hyp-ref-max$_{\text{micro}}$	0.427	-0.415	0.525	0.497	0.591
hyp-mt-avg-ref	0.553	-0.543	0.601	0.638	0.700
hyp-mt-min-ref	0.479	-0.536	0.550	0.615	0.676
hyp-mt-max-ref	0.577	-0.505	0.617	0.614	0.686
hyp-mt-avg	0.562	-0.548	0.597	0.610	0.665
hyp-mt-min	0.421	-0.546	0.479	0.550	0.593
hyp-mt-max	0.549	-0.423	0.571	0.528	0.595
hyp-self-avg	0.557	-0.567	0.589	0.614	0.660

Table 4: Pearson correlation with human judgments for single reference (mt-ref), multiple references (mt-ref-multi) and the metrics based on MT hypotheses described in Section 3.2 for Estonian-English dataset

	BLEU	TER	ChrF	Meteor	BERT
mt-ref	0.312	-0.335	0.346	0.380	0.422
pseudo-mt-avg	0.283	-0.293	0.304	0.304	0.375
pseudo-mt-min	0.214	-0.272	0.221	0.258	0.280
pseudo-mt-max	0.271	-0.219	0.285	0.279	0.350
pseudo-mt-avg-ref	0.356	-0.385	0.394	0.404	0.476
pseudo-mt-min-ref	0.325	-0.382	0.347	0.381	0.425
pseudo-mt-max-ref	0.352	-0.346	0.398	0.388	0.473
hyp-ref-avg$_{\mathrm{macro}}$	0.319	-0.334	0.348	0.382	0.423
hyp-ref-min$_{\mathrm{macro}}$	0.322	-0.334	0.344	0.391	0.429
hyp-ref-max$_{\mathrm{macro}}$	0.326	-0.339	0.346	0.379	0.412
hyp-ref-avg$_{\mathrm{micro}}$	0.320	-0.328	0.346	0.380	0.419
hyp-ref-min$_{\mathrm{micro}}$	0.310	-0.320	0.324	0.382	0.421
hyp-ref-max$_{\mathrm{micro}}$	0.321	-0.328	0.333	0.362	0.391
hyp-mt-avg-ref	0.378	-0.382	0.389	0.431	0.464
hyp-mt-min-ref	0.329	-0.357	0.339	0.371	0.448
hyp-mt-max-ref	0.359	-0.373	0.375	0.422	0.438
hyp-mt-avg	0.296	-0.281	0.288	0.334	0.326
hyp-mt-min	0.225	-0.218	0.199	0.246	0.295
hyp-mt-max	0.240	-0.248	0.228	0.260	0.204
hyp-self-avg	0.313	-0.306	0.309	0.351	0.355

Table 5: Pearson correlation with human judgments for single reference (mt-ref) and the metrics based on MT hypotheses described in Section 3.2 for English-Czech dataset