

Regressive Ensemble for Machine Translation Quality Evaluation

Michal Štefánik and Vít Novotný and Petr Sojka

Masaryk University, Faculty of Informatics, MIR group <https://mir.fi.muni.cz>
{stefanik.m,witiko}@mail.muni.cz, sojka@fi.muni.cz

Abstract

This work introduces a simple regressive ensemble for evaluating machine translation quality based on a set of novel and established metrics. We evaluate the ensemble using a correlation to expert-based MQM scores of the WMT 2021 Metrics workshop. In both monolingual and zero-shot cross-lingual settings, we show a significant performance improvement over single metrics. In the cross-lingual settings, we also demonstrate that an ensemble approach is well-applicable to unseen languages. Furthermore, we identify a strong reference-free baseline that consistently outperforms the commonly-used *BLEU* and *METEOR* measures and significantly improves our ensemble’s performance.

1 Introduction

Automated evaluation of text generation is challenging due to many orthogonal qualitative aspects that the user expects from a text generation system. In machine translation, we can observe errors of the so-called critical category (Wulczyn et al., 2017) such as hallucinating (Lee et al., 2019), omitting parts of the input from translation, or the negation of meaning (Matusov, 2019).

Consider an example of the last category:

- *Reference*: “I never wrote this article, I just edited it.”
- *Hypothesis 1*: “It is not my article, I just edited it.”
- *Hypothesis 2*: “I never wrote this article, I never edited it.”

In this example, all *BERTScore* (Zhang et al., 2019), *BLEURT* (Sellam et al., 2020), and *Prism* (Thompson and Post, 2020b) metrics rank *Hypothesis 2* higher than *Hypothesis 1*. In *BLEURT* and *Prism*, this can be due to a known vulnerability of Transformers, which rely on a lexical intersection (McCoy et al., 2019) if such a heuristic fits the prob-

lem sufficiently well. Trivially, just counting the negations can easily remedy this specific problem. However, such a heuristic would fail in many other cases, such as when we adjust *Hypothesis 1* to “It is an article of somebody else, I just edited it.”

Such cases motivate our ensemble approach that aims to expose both surface and deeper semantic properties of texts and subsequently learn to utilize these for the specific task of translation evaluation. Even though the objective might constrain the particular metrics, data set, or systematically fail in some cases, another metric or their combination in the ensemble allows the flaw to be corrected.

2 Metrics for machine translation evaluation

This section reviews the related work, focusing on the metrics we used in our ensemble.

The standard and still widely-used surface-level metrics for the evaluation of machine translation quality are *BLEU* (Papineni et al., 2002), *ROUGE* (Lin, 2004), and *TER* (Snover et al., 2006). Surface-level metrics are not able to capture the proximity of meaning in cases where one text paraphrases the other, which is an ability commonly observed in deep neural language models (Lewis et al., 2020). One metric that addresses this flaw is *METEOR* (Banerjee and Lavie, 2005), which utilizes WordNet to account for synonymy, word inflection, or token-level paraphrasing.

Evaluation of semantic text equivalence is closely related to a problem of accurate textual representations (embeddings). The traditional method that we identify as relevant for the evaluation of segment-sized texts is FastText (Bojanowski et al., 2017). FastText learns representations of character *n*-grams from which it creates a unified representation of tokens by averaging. Additionally, a distance of a pair of texts can be computed directly from the token-level embeddings using methods such as the soft vector space model (the soft co-

sine measure, *SCM*) (Novotný, 2018), or by solving a minimum-cost flow problem (the word mover’s distance, *WMD*) (Kusner et al., 2015).

Similar matching is performed by *BERTScore* (Zhang et al., 2019), which uses internal token embeddings of a selected BERT layer optimal for the task. Although the token representations are multilingual in some models (Devlin et al., 2018), which makes *BERTScore* usable without references, we are not aware of prior work evaluating it as such. A possible drawback of cross-lingual alignment using a max-reference matching scheme of *BERTScore* lies in a possibility of a significant mismatch of sub-word tokens in source and target text. In contrast, the metric that we refer to as *WMD-contextual* uses the same embeddings as *BERTScore* but uses the network-flow optimization matching scheme of *WMD*.

Task-agnostic methods have recently been outperformed by methods that fine-tune a pre-trained model for a related objective: *BLEURT* (Sellam et al., 2020) fine-tunes a *BERT* (Devlin et al., 2018) model directly on Direct Assessments of submissions to WMT to predict the judgements using a linear head over contextual embeddings of a classification [*CLS*] token. *Comet* (Rei et al., 2020) learns to predict Direct Assessments from tuples of source, reference, and translation texts with the triplet objective or the standard MSE objective.

Some of the most recent work incorporates latent objectives and/or data sets. For instance, *Prism* (Thompson and Post, 2020b,a) learns a language-agnostic representation from multilingual paraphrasing in 39 languages, thus being one of the few well-performing reference-free metrics. The orthogonality of its training objective might lower its correlation to other methods that use contextual embeddings.

3 Methodology

Our methodology aims to answer the following major question with additional supporting questions:

1. **Can an ensemble of surface, syntactic, and semantic-level metrics significantly improve the performance of single metrics?**
2. Can such an approach be applied cross-lingually, i.e., on languages that it has not been trained on?
3. Can surface-level metrics in reference-free

configuration achieve results comparable to the reference-based ones?

4. Are contextual token representations important for evaluating semantic equivalence, or can these be replaced with pre-inferred token representations?

3.1 Experimental setup

We perform our primary evaluation on Multidimensional Quality Metrics (MQM) data set (Freitag et al., 2021), where we use averaged judgement scores as our gold standard. Where multiple judgements are available for the given pair of a source and a hypothesis, we average the scores over the judgements and consider this average as our gold standard. We split the samples into train (80%) and test (20%) subsets based on unique source texts.

In our experimental framework, which we release as an open-source Python library and Docker image for ease of reproduction¹², we implement a selected set of the metrics based on their guidelines, together with a bunch of novel metrics, introduced in Section 3.2 aiming to provide additional, orthogonal insight of textual equivalence.

Subsequently, we train a regressive ensemble on the *standardized* metric features of the whole train set, intending to predict the averaged MQM expert judgements. We evaluate the ensemble, together with all other selected metrics using pairwise Spearman’s rank correlation (Spearman’s ρ) with the MQM judgements on the held-out 20% test split.

In addition to our primary evaluation on the MQM data set, we perform our experiments on the Direct Assessments (DA) from WMT 2015 and 2016, and a dev set of Catastrophic errors from the Post-editing Dataset (Fomicheva et al., 2020a) of the Multilingual Quality Estimation Dataset (MLQE-PE) (Fomicheva et al., 2020b) used for evaluation at the Quality Estimation workshop of WMT 2021. Refer to Section 3.6 for a detailed description of our experiments on DA and MLQE-PE.

We performed all our evaluations on segment-level judgements. To minimize the impact of calibration for each of the specific metrics to evaluation, we report Spearman’s rank correlation coefficient, reflecting the mutual qualitative ordering rather than particular values of the judgements.

¹<https://github.com/MIR-MU/regemt>

²<https://hub.docker.com/r/miratmu/regemt>

3.2 Novel metrics

In addition to a selected set of metrics based on a literature review, we implement a set of novel metrics that allows our ensemble to reflect a wider variance of properties of the evaluated texts.

3.2.1 Soft Cosine Measure

The soft cosine measure (*SCM*) (Novotný, 2018) is the cosine similarity of texts in the soft vector space model, where the axes of terms are at an angle corresponding to their cosine similarity S in a token embedding space:

$$SCM(\vec{x}, \vec{y}) = \frac{\vec{x}^T \cdot S \cdot \vec{y}}{\sqrt{\vec{x}^T \cdot S \cdot \vec{x}} \cdot \sqrt{\vec{y}^T \cdot S \cdot \vec{y}}} \quad (1)$$

where \vec{x} is a weighted bag-of-words (BoW) vector of a reference (or a source in a reference-free setting), and \vec{y} is a weighted BoW vector of a hypothesis.

We use *SCM* with two token representations:

1. We use the static token representations of FastText (Grave et al., 2018). We refer to the resulting metric as *SCM*.
2. We use the contextual token representations of BERT (Devlin et al., 2018) using the methodology of BERTScore (Zhang et al., 2019). We collect representations of all tokens segmented by the WordPiece (Wu et al., 2016) tokenizer, and we treat each unique (token, context) pair as a single term in our vocabulary.

Subsequently, we *decontextualize* these representations as follows: For each WordPiece token, we average the representations of all (token, context) pairs in the training corpus. We refer to the resulting metric as *SCM-decontextualized*. Due to the multilingual character of the learned BERT token representations, this metric is applicable both in reference-based and source-based approaches.

In addition to two token representations, we also use two different SMART weighting schemes of Salton and Buckley (1988) for the BoW vectors \vec{x}, \vec{y} and the construction of the term similarity matrix S :

1. We use raw term frequencies as *weights* in the BoW vectors, the $n \times n$ SMART weighting scheme, and we construct the term similarity matrix in the vocabulary order. We refer to the resulting metrics as *SCM* and *SCM-decontextualized*.

2. We use term frequencies discounted by inverse document frequencies as weights in the BoW vectors, the nfx SMART weighting scheme, and we construct the term similarity matrix in the decreasing order of inverse document frequencies (Novotný, 2018, Section 3). We refer to the resulting metrics as *SCM-tfidf* and *SCM-decontextualized-tfidf*.

3.2.2 Word Mover’s Distance

The Word Mover’s Distance (*WMD*) (Kusner et al., 2015) finds the minimum-cost flow F between vector space representations of two texts:

$$WMD(\vec{x}, \vec{y}) = \text{minimum cumulative cost } F^T \cdot S$$

$$\text{subject to } \sum_j F_{ij} = x_i, \sum_i F_{ij} = y_j, \quad (2)$$

where \vec{x} is an ℓ_1 -normalized weighted BoW vector of a reference (or a source in reference-free setting), \vec{y} is an ℓ_1 -normalized weighted BoW vector of a hypothesis, and S is a term similarity matrix.

Similar to *SCM* described in the previous section, we experiment with two token representations: FastText embeddings of whole tokens (*WMD*) and decontextualized embeddings of WordPiece tokens (*WMD-decontextualized*). Additionally, we also use the contextual embeddings of WordPiece tokens (*WMD-contextual*) to show the impact of decontextualization on the metric performance: If the impact is negligible, future work could avoid the costly on-the-fly inference of BERT representation and significantly reduce the vocabulary size.

Similarly to *SCM*, we also use two different weighting schemes: raw term frequencies (*WMD-**) and term frequencies discounted by inverse document frequencies (*WMD-*-tfidf*).

3.2.3 Compositionality

Our custom metric that we refer to as *Compositionality* constructs a transition graph of an arbitrary text G_t based on directed, pairwise transitions of the tokens’ part-of-speech (PoS) categories. As the models for PoS tagging are language-dependent, we use the compliant—though not always systematically aligned—schemes of tagging used for training the taggers in English (Weischedel et al., 2013), German (Brants et al., 2002), Chinese (Weischedel et al., 2013), and Norwegian (Unhammer and Trosterud, 2009).

Subsequently, we row-normalise the values of matrix G_t and we define a distance metric of *Compositionality* C for $x = G_{t_1}$ and $y = G_{t_2}$:

$$C(x, y) = \ell_1\text{-norm}(x_{ii} - y_{ii}), \quad (3)$$

where the PoS tags $i \in x$ and $i \in y$.

In our submission, we apply this metric only if the language belongs to a set of the languages for which we have a tagger: English, German, Chinese, or Norwegian. In reference-based evaluations, this constraint applies to the target language; in the case of reference-free evaluations, it applies to both source and target languages.

3.3 Ensemble

We ensemble the aforementioned metrics as predictors in a regression model, minimizing the residual between the average segment-level MQM scores and predicted targets.

We experiment with a wide range of simple regressors and observe a superior performance of simple approaches of fully connected, two-layer perceptron with 100-dimensional hidden layer and ReLu activation and linear regression with squared residuals. We report the results for *RegEMT* as the best-performing one of these classifiers picked on a 20% held-out validation subset of the train data set.

In addition to the ensemble of all available metrics, we evaluate a baseline regressive ensemble *Reg-base* using solemnly two surface-level features: character-level source and target length according to WordPiece (Wu et al., 2016) tokens.

3.4 Cross-lingual experiments

As expert judgments are incredibly costly to obtain, it is unrealistic to expect that the training data for the trained systems will be available in the future for a vast majority of language pairs containing under-resourced languages. To estimate the performance of all metrics on uncovered language pairs, we perform a cross-lingual evaluation on average MQM judgements of two available language pairs: *zh-en* and *en-de*.

Where applicable, we fit the metric parameters on the train split of the non-reported language pair. Subsequently, we evaluate and report the results on a test split of the reported pair to MQM judgements.

3.5 Ablation study

To understand the impact of individual metrics in their roles as predictors for our ensemble, we use

their pairwise correlations for systematic feature elimination.

In our ablation study, we iteratively select the metric with the highest Spearman’s ρ to any other metric. We eliminate the selected metric from our ensemble by fitting a new regression model on the remaining features. We continue until all metrics are eliminated and evaluate the ensemble at each step of the process.

3.6 Additional evaluations

To allow for additional insight into the consistency of the results to other relevant evaluation sources, together with an evaluation of the metrics in the novel application of critical error recognition, we perform the experiments analogically on a DA data set of the WMT submissions from years 2015 and 2016, as well as to the Critical Errors dev set of MLQE-PE data set for reference-free metrics.

In the case of DA judgements, we use the assessments from the year 2015 as a training split and assessments from the year 2016 as a test split.

In the case of MLQE-PE, we split the data analogically to MQM by splitting the unique source texts in an 80:20 ratio. In this case, we consider as gold judgements the mean severity of error assigned by three annotators to each of the translations.

4 Results

Correlations to MQM judgements. Table 1 lists correlations to MQM for source-based, i.e., reference-free metrics (upper) and reference-based metrics (middle). Results reported for *RegEMT* fit a selected regression model on the estimates of all the other metrics available for a given evaluation scheme. As described in Section 3.3, we pick the evaluated regression model based on its performance on a held-out portion of the train set: a two-layer perceptron for the source-based *zh-en* pair and a simple linear regression in all other cases, with negligible mutual differences between regression models in performance on the validation set (below 2%). In the reports suffixed with X , we fit the regression model on the other language pair than the one used for the evaluation.

The results suggest that a simple regressive ensemble can benefit from the variance of the predictors in a majority of the evaluated configurations, including the cross-lingual settings and other evaluated datasets. We observe the highest margins in correlations in the case of MQM judgements.

	RegEMT	Prism	BERTScr	WMD-cont	WMD-dec	WMD-dec-tf	SCM-dec	SCM-dec-tf	Compos	Reg-base	Comet	SCM	SCM-tf	WMD	WMD-tf	BLEUrt	BLEU	METEOR
MQM-src <i>zh-en</i>	.59	.36	.44	.44	.29	.17	.19	.13	.13	.34								
MQM-src <i>zh-en-X</i>	.49	.36	.44	.44	.29	.17	.19	.13	.13	.34								
MQM-src <i>en-de</i>	.36	.09	.14	.06	.04	.07	.03	.02	.23	.28								
MQM-src <i>en-de-X</i>	.31	.09	.14	.06	.04	.07	.04	.02	.23	.28								
MQM-ref <i>zh-en</i>	.62	.45	.45	.43	.27	.21	.10	.26	.01	.35	.51	.19	.35	.29	.27	.48	.25	.28
MQM-ref <i>zh-en-X</i>	.62	.45	.45	.43	.27	.21	.09	.25	.01	.31	.51	.19	.35	.29	.27	.48	.25	.28
MQM-ref <i>en-de</i>	.60	.32	.22	.25	.32	.28	.33	.18	.12	.27	.48	.06	.14	.13	.07	.10	.13	.20
MQM-ref <i>en-de-X</i>	.38	.32	.22	.25	.32	.28	.34	.17	.12	.29	.48	.06	.14	.13	.07	.10	.13	.20
DA 2016-src	.84	.72	.74	.73	.57	.51	.37	.51	.29	.18	.82	.39	.45	.44	.42	.81	.42	.50
DA 2016-tgt	.68	.70	.34	.25	.09	.10	.10	.24	.13	.04								
catastrophic-src	.29	.26	.13	.11	.12	.15	.13	.09	.10	.09								

Table 1: Results for Spearman’s correlations with selected gold standards. From top: correlation of source-based metrics (top) and reference-based metrics (middle) to averaged scores of MQM expert judgements for specified languages. Results suffixed with *X* are evaluated cross-lingually: both ensemble metrics are trained on other language pairs than evaluated. (Bottom): Results for other data sets: Direct assessments of WMT 2016 submissions and dev set of Catastrophic translations from MLQE-PE data set; reported values are an average of correlations over all available language pairs.

Baseline ensemble. Table 1 shows that *Reg-base*, using only the counts of reference and hypothesis Word-pieces, demonstrates its consistent superiority over the standard surface-level metrics of *BLEU* and *METEOR*, even in the cross-lingual vs. monolingual comparison. With respect to the MQM judgements, the correlations of *Reg-base* are reasonably consistent; hence, in the reference-free cases of *en-de* language pair, the correlation of *Reg-base* is very close to the correlations of *RegEMT*.

Importance of contextualization. The results in Table 1 are inconsistent concerning the importance of contextualization in token-level metrics (*WMD-cont** vs. *WMD-dec** and *SCM-cont** vs. *SCM-dec**). We observe a significant (15–16%) decrease of correlation between a contextualized and decontextualized versions of WMD in *all* cases of *zh-en* language pair. The situation differs in *en-de* pair, where for the reference-based case, the correlation of decontextualized version of WMD is superior by 7%.

Metrics correlations. Table 2 demonstrates mutual correlations of the evaluated metrics. We see the strong pairwise correlations among the metrics based on contextualized representations, such as between *Comet*, *Prism*, *BERTScore* and *BLEUrt*; all of these are higher than 0.79. The situation is similar among the metrics based on static token

representations of *SCM* and *WMD*, both with and without TF-IDF.

In contrast, we observe a low correlation of *BLEU* and *METEOR* to *Reg-base* forming a cluster of surface-level metrics.

Ablation study. Figure 1 displays performance development in Spearman’s ρ of the regressive ensemble when we incrementally eliminate the metrics from the set of ensembled predictors. Following the methodology described in Section 3.5, the exact ordering of the metrics in ablation for *zh-en* pair is shown in Table 2, and we observe it to be similar also for the other language pair of MQM.

In ensembles of reference-based metrics (left), we observe a high consistency throughout the removal of most of the metrics. A longer consistency in *zh-en* case is attributed to a consistent performance in ensembling *BLEUrt* (removed in step 14) and *METEOR* (removed in step 15). These metrics only reach the correlation of 0.48 and 0.28, respectively, when evaluated independently. In *en-de* case, the most significant drops can be attributed to a removal of the best-performing *Comet* (step 9) and *Prism* (step 11).

Ensemble of source-based metrics (right) shows significant drops in *zh-en* pair after removing *Prism* (step 3) and *WMD-contextual* (step 4). In *en-de* language pair, the correlation is relatively low throughout the whole ablation process. The least corre-

	0)	1)	2)	3)	4)	5)	6)	7)	8)	9)	10)	11)	12)	13)	14)	15)	16)	17)	18)
0) <i>MQM avg scr</i>	1.	.62	.51	.45	.45	.43	.27	.21	.1	.26	.19	.35	.29	.27	.48	.25	.28	.01	.35
1) <i>RegEMT</i>	.59	1.	.56	.5	.44	.42	.25	.23	.4	.28	.15	.36	.3	.3	.55	.26	.32	.11	.51
2) <i>Comet</i>			1.	.79	.82	.82	.71	.61	.54	.62	.57	.64	.64	.6	.83	.64	.63	.44	.4
3) <i>Prism</i>	.36	.5		1.	.9	.9	.85	.77	.63	.71	.71	.77	.79	.77	.8	.81	.77	.54	.19
4) <i>BERTScr</i>	.44	.48		.13	1.	.13	.06	.01	.11	.1	.13	.04	.76	.73	.82	.77	.74	.56	.18
5) <i>WMD-cont</i>	.44	.5		.53	.88	1.	.92	.8	.68	.69	.71	.74	.77	.73	.82	.81	.75	.58	.18
6) <i>WMD-dec</i>	.29	.42		.22	.4	.6	1.	.89	.81	.7	.77	.72	.77	.77	.75	.81	.71	.66	.37
7) <i>WMD-dec-tf</i>	.17	.3		.08	.12	.32	.69	1.	.71	.66	.73	.71	.74	.8	.63	.74	.64	.6	.32
8) <i>SCM-dec</i>	.19	.26		.03	.21	.35	.71	.49	1.	.51	.67	.49	.55	.58	.57	.61	.53	.65	.58
9) <i>SCM-dec-tf</i>	.13	.26		.2	.32	.43	.47	.26	.29	1.	.58	.72	.64	.61	.62	.62	.58	.48	.19
10) <i>SCM</i>											1.	.78	.9	.86	.54	.75	.74	.67	.37
11) <i>SCM-tf</i>												1.	.85	.84	.6	.71	.72	.52	.13
12) <i>WMD</i>													1.	.94	.6	.81	.83	.6	.17
13) <i>WMD-tf</i>														1.	.58	.78	.78	.58	.22
14) <i>BLEUrt</i>															1.	.6	.58	.43	.11
15) <i>BLEU</i>																1.	.85	.66	.27
16) <i>METEOR</i>																	1.	.56	.15
17) <i>Compos</i>	.13	.11		.13	.06	.01	.11	.1	.13	.04								1.	.52
18) <i>Reg-base</i>	.34	.44		.37	.07	.07	.24	.19	.28	.02									1.

Table 2: Pairwise Spearman’s correlations of the evaluated metrics and their correlations to averaged MQM judgements for *zh-en* language pair. Top-right triangle: mutual correlations of reference-based metrics, bottom-left triangle: correlations of metrics supporting multilingual source-based evaluation.

lated and hence the last ones eliminated are *WMD-decontextualized-tfidf* (step 6) and *Prism* (step 7).

5 Discussion

Regressive ensemble. Following the objectives that we set in Section 3, we empirically confirm that an ensemble can push the quality of modeling the expert judgements in most of the configurations while performing close-to-the-best metrics on the others. Additionally, we demonstrate that such ensemble is transferable to new language pairs and that its use is motivated by qualitative gains even in cross-lingual settings.

At the same time, one must acknowledge the limitations that an ensemble system exposes compared to single and unsupervised metrics. An ensemble might inherit the systematic biases of each of its metrics. This problem is observable in the results of the source-based *en-de* pair of MQM in Table 1, where the ensemble follows the low correlations of its ensembled metrics. Further, relying entirely on the metrics’ consistency, the ensemble will inevitably expose errors in domains where some metrics behave markedly out of their usual range.

On the other hand, we argue that this might rarely be the case with the surface-level metrics that are mainly unsupervised. We suspect it to be unlikely with learnable metrics, too, having their output

space constrained by the range of their imitated metrics.

Values of correlations in Table 2 and partially also the threshold metrics in Figure 1 suggest that our ensemble relies primarily on trained contextualized metrics with regards to their correlation with the target as summarized in Table 1. We suspect that oversampling of under-represented categories of errors would increase the significance of other types of metrics, as the under-represented error categories would be the ones where the fine-tuned metrics perform worse.

Baseline ensemble. Surprisingly, our baseline ensemble *Reg-base* consistently outperforms other standard surface-based metrics such as *BLEU* (Papineni et al., 2002). This suggests possible applicability of surface-level metrics also in reference-free evaluation.

We suspect that the baseline features of length based on a multilingual WordPiece tokenizer (Wu et al., 2016) might reflect on the missing or inappropriately added segments more *strictly* than other surface-level metrics. At the same time, these errors are usually highly weighted in the overall score.

Table 2 shows considerable orthogonality of *Reg-base* to other metrics. This motivates the inclusion

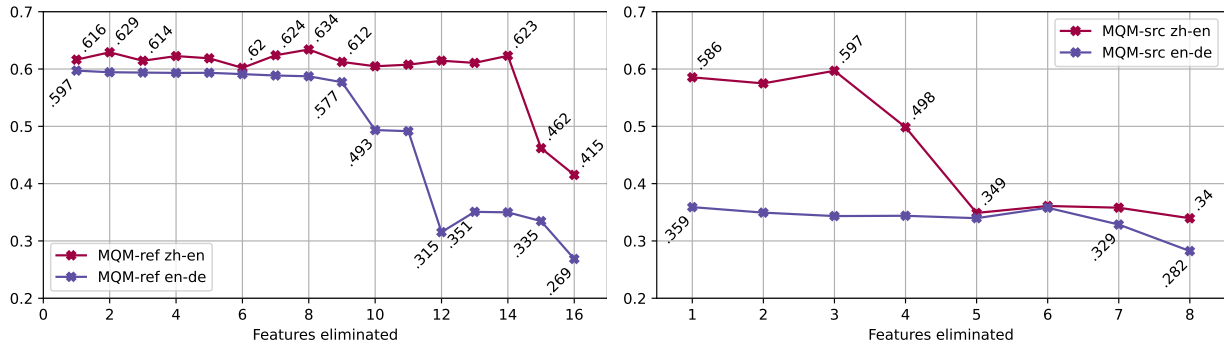


Figure 1: Ablation study: correlation of ensemble estimator to averaged MQM judgements during incremental elimination of the most-correlated metric from the ensembled predictors. Left: ablation of reference-based metrics, right: ablation of source-based metrics.

of the other weak surface metrics into the baseline ensemble to alleviate some of its apparent flaws.

Impact of contextualization. Based on the results of *WMD*-* described in Section 3.2.2, one can not draw a consistent conclusion regarding the impact of contextualization. On average, decontextualization has decreased the performance of *WMD* by 8%, but the original motivation of a significant improvement in the usability of estimators might compensate. On the other hand, *WMD-decontextualized* and *WMD-decontextualized-tfidf* reached a considerable improvement of 16–18% as compared to *WMD* and *WMD-tfidf* using FastText embeddings, while losing none of their flexibility.

“It is the harmony of the diverse parts, their symmetry, their happy balance; in a word it is all that introduces order, all that gives unity, that permits us to see clearly and to comprehend at once both the ensemble and the details.”

Henri Poincaré

6 Conclusion

This work evaluates the potential of ensembling multiple diverse metrics (*RegEMT*) for an evaluation of machine translation quality and offers a new simple baseline metric *Reg-base* that achieves better results than *BLEU* and *METEOR* by using just the source and reference lengths. We measure significant gains in Spearman’s correlation to MQM with *RegEMT* compared to standalone metrics and we demonstrate that even simple linear estimators can benefit from the expressivity that the methods of all levels of representation provide. Additionally, as we demonstrate, the ensemble based on metrics supporting multilingualism can push the quality further even on unseen language pairs.

We recognize the inherent limitations of the regressive ensemble, which is inevitably slower, resource-heavier, and prone to inherit latent inductive biases of underlying metrics or their combinations. However, *RegEMT* shows the agility of the simple ensemble approach, which is in contrast to attempts to learn the full complexity of quality estimation through a single objective and allows the quality estimator to avoid the blind spots of particular metrics. We hope that our results will motivate future work in the ensemble evaluation.

References

- Satanjeev Banerjee and Alon Lavie. 2005. *METEOR: An automatic metric for MT evaluation with improved correlation with human judgments*. In *Proc. of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, USA. ACL.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. *Enriching Word Vectors with Subword Information*. *Transactions of the ACL*, 5:135–146.
- Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. *The TIGER Treebank*. In *Proc. of the workshop on Treebanks and Linguistic Theories*, pages 24–41.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. *BERT: Pre-training of deep bidirectional transformers for language understanding*. ArXiv:1810.04805v2.
- Marina Fomicheva, Shuo Sun, Erick Fonseca, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. 2020a. *MLQE-PE: A Multilingual Quality Estimation and Post-Editing Dataset*. ArXiv:2010.04480.

- Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020b. [Unsupervised Quality Estimation for Neural Machine Translation](#). *Transactions of the ACL*, 8:539–555.
- 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:2104.14478.
- Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomáš Mikolov. 2018. [Learning word vectors for 157 languages](#). ArXiv:1802.06893v2.
- Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. [From Word Embeddings To Document Distances](#). In *Proc. of International Conference on Machine Learning*, volume 37, pages 957–966, Lille, France. PMLR.
- Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fanjiang, and David Sussillo. 2019. [Hallucinations in Neural Machine Translation](#). Accepted ICLR 2019 Conference Blind Submission.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension](#). In *Proc. of the 58th Annual Meeting of the ACL*, pages 7871–7880.
- Chin-Yew Lin. 2004. [ROUGE: A Package for Automatic Evaluation of Summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. ACL.
- Evgeny Matusov. 2019. [The Challenges of Using Neural Machine Translation for Literature](#). In *Proc. of the Qualities of Literary Machine Translation*, pages 10–19, Dublin, Ireland. European Association for Machine Translation.
- Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. [Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference](#). In *Proc. of the 57th Annual Meeting of the ACL*, pages 3428–3448, Florence, Italy. ACL.
- Vít Novotný. 2018. [Implementation Notes for the Soft Cosine Measure](#). In *Proc. of the 27th ACM International Conference on Information and Knowledge Management, CIKM '18*, pages 1639–1642, New York, NY, USA. ACM.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [BLEU: a method for automatic evaluation of machine translation](#). In *Proc. of the 40th Annual Meeting on ACL*, pages 311–318, USA. ACL.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. [COMET: A Neural Framework for MT Evaluation](#). In *Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702. ACL.
- Gerard Salton and Christopher Buckley. 1988. [Term-weighting approaches in automatic text retrieval](#). *Information Processing & Management*, 24(5):513–523.
- Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. [BLEURT: Learning Robust Metrics for Text Generation](#). In *Proc. of the 58th Annual Meeting of the ACL*, pages 7881–7892.
- Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. [A Study of Translation Edit Rate with Targeted Human Annotation](#). In *Proc. of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.
- Brian Thompson and Matt Post. 2020a. [Automatic Machine Translation Evaluation in Many Languages via Zero-Shot Paraphrasing](#). In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121. ACL.
- Brian Thompson and Matt Post. 2020b. [Paraphrase Generation as Zero-Shot Multilingual Translation: Disentangling Semantic Similarity from Lexical and Syntactic Diversity](#). In *Proc. of the 5th Conference on Machine Translation (Vol. 1)*, pages 561–570. ACL.
- Kevin Unhammer and Trond Trosterud. 2009. [Reuse of Free Resources in Machine Translation between Nynorsk and Bokmål](#). In *Proc. of the First International Workshop on Free/Open-Source Rule-Based Machine Translation*, pages 35–42, Alicante. Universidad de Alicante.
- Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. [OntoNotes Release 5.0](#).
- Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Norouzi, Macherey, Krikun, Cao, Gao, Macherey, Klingner, Shah, Johnson, Liu, Kaiser, Gouws, Kato, Kudo, Kazawa, Stevens, Kurian, Patil, Wang, Young, Smith, Riesa, Rudnick, Vinyals, Corrado, Hughes, and Dean. 2016. [Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation](#).
- Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. [Wikipedia Talk Labels: Personal Attacks](#).
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. [BERTScore: Evaluating text generation with BERT](#). ArXiv:1904.09675v3.