Summary: Unsupervised Quality Estimation for Neural Machine Translation

Frederik Polachowski
TU Dortmund
Dortmund, Germany
frederik.polachowski@tu-dortmund.de

ACM Reference Format:
Frederik Polachowski. 2022. Summary: Unsupervised Quality Estimation for Neural Machine Translation. In *Proceedings of Seminar (Summary).* ACM, Dortmund, NRW, GER, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Machine translation services are widely used, and current improvements ensure a high perceived quality of translations for a wide variety of languages. To accurately estimate the quality of a translation system, though, remains challenging [7]. Considering the example translations in table 1. With the knowledge of the reference translation, the task of rating the quality of translation becomes trivial. Even without the reference, a human fluent in both languages is still capable of rating the quality. But in most real world scenarios, both information sources are rarely available. Each sentence can have multiple translation as well, further complicating this problem. Therefore, it is not straightforward to quantify the quality of translation by only looking at the source and target sentences [7]. Considering another example inspired by the work of Fomicheva et al. presented in table 2. Without being fluent in both languages and even without the reference translation, a human annotator could defer the quality of the individual translations. By comparing the different hypothesis and the dropout hypotheses, the similarity of these hypotheses could be a good indication of translation quality. Again, the problem to quantify this similarity and the express the quality based on these factors remains.

Fomicheva et al. present an unsupervised approach to quality estimation for neural machine translation. The authors explore different metrics which utilize the internal information of current SOTA machine translation models to estimate the quality of the produced translation. [7] propose multiple hypothesis, which are in turn evaluated in their paper. The main hypothesis is that the metrics presented in this paper are capable of reliably estimating the quality of neural machine translation hypothesis. The authors explore multiple other related hypothesis as well. Firstly, that by estimating the model uncertainty, the accuracy of quality estimation can be improved. Secondly, the internal information of the neural machine translation model can be exploited to defer the quality of the resulting translation. Lastly, by improving model output calibration, the performance of the presented metrics can be improved [7].

Original	Jackson pidas seal kõne, öeldes, et James Brown on tema suurim inspiratsioon.
Reference	Jackson gave a speech there saying that James Brown is his greatest inspiration.
Translation Hypothesis	Jackson gave a speech there, saying that his greatest inspiration is James Brown.

Table 1: Translation example of Estonian-English translation. It is noteworthy that the reference and translation hypothesis are almost paraphrases of one another. This indicates a high quality translation [7].

The reference methods used in this paper rely on current SOTA neural machine translation models, which incorporate the estimation of this quality. But these models require a large amount of annotated training data to produce reliable results [7]. For this reason and to evaluate the proposed methods, Fomicheva et al. created a translation quality dataset, containing translations with human annotated quality estimates.

2 DATASET

As in any machine learning setting, the output quality of the resulting model is dependent on the amount of training data [7]. To this end Fomicheva et al. introduced a dataset consisting of six different language pairs for the task of quality estimation for machine translation. The dataset is composed of English, German, Chinese, Romanian, Estonian, Sinhala, and Nepali source and target sentences. The language pairs can be further grouped in three different subcategories, namely high-, medium-, and low-resource language pairs. This grouping is based on the amount of available training data in existing datasets for the machine translation task. To further diversify the dataset, the authors changed the direction of translation to and from English [7].

The source sentences are extracted from Wikipedia. For this, a diverse set of Wikipedia articles in the intended source language were selected. For each language, all selected articles were further sorted based on multiple criteria. These criteria are composed of the ratio of original words in source language, the length of the sentence, and the exclusivity of the sentence relating to any other dataset [7]. This fact is most important, since inclusion in other datasets would be problematic when testing the performance of models on this test dataset, if it was trained with the same sentences. Finally, the top 100 articles were selected and from those articles

Summary, July 2022, Dortmund, NRM, GER

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

10,000 sentences were randomly sampled for each language to create the source sentences for this dataset [7].

After sampling the source sentences for each language, these sentences were translated using multiple implementations of SOTA Transformer models [7]. The models and their implementations are based on the work of Vaswani et al. [22] and Ott et al. [19]. For training, publicly available datasets such as Paracrawl [5] and Europarl [16] were used. The low-resource language models were trained using semi-supervised learning, as described in FLORES [12]. All models were optimized using cross-entropy loss [7].

The resulting translations were evaluated by six different professional translators from two different service providers [7]. The evaluation was done based on a rating system called Direct Assessment (DA) [11]. In this rating system, each annotator is given the source sentence and is asked to rate the translation hypothesis on a scale from 0-100. In this case, 0 corresponds to a wrong translation and 100 to a perfect translation of the source sentence. The final rating of the translation was done by normalizing the individual annotator ratings based on the mean and standard deviation of each annotator. Additionally, mean scores were provided for both the standard rating and the normalized ratings [7].

3 METHODOLOGY

The following methods proposed by Fomicheva et al. assume that a sequence-to-sequence NMT architecture consisting of encoder-decoder models using attention mechanism is used to create a translation hypothesis [7]. This sequence-to-sequence model maps an input $x = x_1, ..., x_I$ to an output $y = y_1, ..., y_T$. Assuming further that a softmax function has been used to create the output, the result resembles a probability distribution [7]. Therefore, the estimated output probability of y can be described as:

$$p(y|x, \theta) = \prod_{t=1}^{T} p(y_t|y_{<t}, x, \theta) \tag{1}$$

where θ represents the parameters of the model and $p(y_t|y_{<t}, x, \theta)$ the calculated probability of token y_t being correct [7]. This distribution reflects the confidence of the model in the correctness of the output sentence.

3.1 Exploiting the Softmax Distribution

The idea behind the first proposed method for QE is to estimate the quality of the translation based on the confidence of the model. This is done by calculating the average confidence of the model over each individual token of the translated sentence. The assumtion is the more confident the model is in the correctness of each token, the better the overall quality [7]. For numerical stability, the sum of log-probabilities is used instead of the product of probabilities:

$$\text{TP} = \frac{1}{T} \sum_{t=1}^{T} \log p(y_t|y_{<t}, x, \theta) \tag{2}$$

This method is however limited by only consider the 1-best translation hypothesis for each token. Furthermore, this method relies on the calibration of the output distribution to reflect the true confidence of the network [7]. Therefore, the model has to be well calibrated to produce reliable information for quality estimation [7]. If for example the model is overconfident, the estimation would indicate a perfect quality translation even though receiving an unreliable MT output.

To expand on the 1-best prediction approach, Fomicheva et al. created additional metrics considering the entire vocabulary at each translation step. The entropy over the entire vocabulary V is calculated for each token as follows:

$$\text{Softmax-Ent} = -\frac{1}{T} \sum_{t=1}^{T} \sum_{v=1}^{V} p(y_t^v) \log p(y_t^v) \tag{3}$$

where $p(y_t^v)$ represents the previous examined probability $p(y_t|y_{<t}, x, \theta)$. Assuming the probability mass is concentrated on a few tokens, the resulting entropy will be low, and the translation can be assumed to be correct [7]. As with equation 2, this expresses the confidence of the network in the translation. Additionally, this considers the overall confidence in each individual vocabulary token. By contrast, if the probability mass is dispersed evenly, the model would consider each token to be equally likely. The resulting entropy would be high and would express a low quality translation.

Considering the example of the entropy for two tokens of $[0.9, 0.1]$ and $[0.5, 0.5]$. Both entropy sets would create the same mean, but the information expressed in these sets of values expresses a deeper meaning [7]. For this reason Fomicheva et al. introduced a third metric which examines the standard deviation of the token probability:

$$\text{Sent-Std} = \sqrt{\text{E}[P^2] - \text{E}[P]^2} \tag{4}$$

where $P = p(y_1, ..., y_T)$. is the standard deviation of the word-level probabilities.

3.2 Quantifying Uncertainty

The previous metrics focused on the output probability calculated by the underlying model. For the next metrics Fomicheva et al. proposed similar approaches which utilized uncertainty quantification to improve on the previous metrics. Using Monte Carlo Dropout [8], the uncertainty of a model can be estimated by producing multiple translation hypothesis with perturbed model parameters [7]. In this case, the parameters of the model are perturbed during testing and the resulting hypothesis and their probability distributions are compared and evaluated. The first three methods focus on the probability distributions of the hypothesis, whereas the final metric examines the different translation hypothesis themselves.

The following metrics are similar to the metric in equation 2. For the first metric, the TP score for the output distributions are averaged over the different parses utilizing dropout [7]. This creates a similar output as the aforementioned metric, but is more concise [7]:

$$\text{D-TP} = \frac{1}{N} \sum_{n=1}^{N} \text{TP}_{\theta' n} \tag{5}$$

The second metric reflects the variance of the output probabilities over the multiple dropout passes:

$$\text{D-Var} = \text{E}[TP_{\theta'}^2] - \text{E}[TP_{\theta'}]^2 \tag{6}$$

Lastly, Fomicheva et al. created a combination of the dropout based metrics to combine the average and variance of probability distributions:

$$\text{D-Combo} = (1 - \frac{\text{D-TP}}{\text{D-Var}}) \tag{7}$$

Different from the previous metrics, the next proposed metric uses the translation hypothesis itself. Similar to the initial example in table 2, the quality of a translation can be estimated examining multiple translation hypothesis. The more similar the hypothesis are, the more likely it is, that the hypothesis are correct [7]. The previous methods do not consider synonyms or different ending of words with the same stem [7]. The following metric examines the similarity of all translation hypothesis by using the Meteor similarity method [4]. This method calculates a similarity score for two sentences by counting the number of matching criteria. These criteria examine individual token and include exact matches, stemming matches, synonym, and paraphrases [4].

Considering H to be the set of translation hypothesis, the metric can be expressed as follows:

$$\text{D-Lex-Sim} = \frac{2}{|H| \times (|H| - 1)} \sum_{i=1}^{|H|} \sum_{j=1}^{|H|} sim(h_i, h_j) \tag{8}$$

where $h_i, h_j \in H, i \neq j$. The final score is the average over all hypothesis pairs.

3.3 Attention weights

Attention weights represent the strength of connection between source and target tokens [7]. Similar to the output distribution of the model, the internally used attention weights can be used to estimate translation quality [7]. In equation 2 and 3 the estimated quality was expressed by the confidence of the model in each token. By using the attention weights, the quality can be estimated, examining the correlation between the source and target tokens [7]. To this end, Fomicheva et al. propose a method calculating the entropy of the attention weights similar to equation 3:

$$\text{Att-Ent} = -\frac{1}{I} \sum_{i=1}^{I} \sum_{j=1}^{J} \alpha_{ji} \log \alpha_{ji} \tag{9}$$

where α_{ji} represent the attention weight of the source token i and the target token j.

Current SOTA transformer models use multiple attention heads and therefore there exist multiple head/layer combinations [7]. The previous metric cannot be used when dealing with multiple attention heads. To this end, Fomicheva et al. propose two metrics to deal with this problem. Firstly, they calculate the minimal entropy over all head/layer combinations:

$$\text{AW:Ent-Min} = min_{\{hl\}}(\text{Att-Ent}_{hl}) \tag{10}$$

and secondly, the average entropy over all head/layer combinations:

$$\text{AW:Ent-Avg} = \frac{1}{H \times L} \sum_{h=1}^{H} \sum_{l=1}^{L} \text{Att-Ent}_{hl} \tag{11}$$

(1)	Original	Tanganjikast püütakse niiluse ahvenat ja kapentat.
	Reference	Nile perch and kapenta are fished from Lake Tanganyika.
	Hypothesis	There is a silver thread and candle from Tanzeri.
		There will be a silver thread and a penny from Tanzer.
	Dropout	There is an attempt at a silver greed and a carpenter from Tanzeri.
		There will be a silver bullet and a candle from Tanzer.
		The puzzle is being caught in the chicken's gavel and the coffin.
(2)	Original	Siis aga võib tekkida seesmise ja välise vaate vahele lõhe.
	Reference	This could however lead to a split between the inner and outer view.
	Hypothesis	Then there may be a split between internal and external viewpoints.
		Then, however, there may be a split between internal and external viewpoints.
	Dropout	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.

Table 2: Trivial examples of a dropout based hypothesis creation. These examples include a low quality example (1) and a high quality example (2). From the accumulated translation hypotheses, it is obvious that in the first example, all hypotheses have little similarity between them. This together with the information available through the reference demonstrates a high uncertainty in MT output. The opposite is true for the second example. [7]

After the initial findings of this paper, a third method relating to attention weights was proposed. The *AW:best head/layer* metric examines the best DA score correlating head/layer combination for estimating translation quality. But this metric requires a test dataset for initial evaluation and is therefore not unsupervised.

4 RESULTS

To evaluate the performance of the proposed metrics, Fomicheva et al. calculated the Pearson correlation between estimated quality and human assigned DA scoring of the individual translations [7]. The evaluation can be further broken down into three sections based on the general approach used in the corresponding set of metrics. To compare the performance of the all metrics, two reference methods are used. These reference methods use the currently best performing supervised approaches for translation quality estimation available with open source implementations [7]. Namely, the PredEst [15] and the BERT-BiRNN [1] have been selected as comparison methods. These methods are supervised Transformer based models, which create translation hypothesis along with quality

estimations. To compute the significance of the findings, Hotelling-Williams tests with a p-value ≤ 0.05 have been used [7]. The results are visualized in table 3.

Overall, the sequence-level probability TP already performs competitively in medium-resource language pairs compared to supervised approaches, but is outperformed by D-TP [7]. Fomicheva et al. contribute this behavior to the bad calibration of output probabilities. The strong performance of dropout based metrics supports the hypothesis that estimating model uncertainty improves estimated translation quality.

All methods achieve overall lowest correlation for the En-DE language pair. Fomicheva et al. argue this is due to the overall high quality of translations for this language pair. Because of this, the quality distribution is centered around high quality example, with few low quality translations. The low correlation could indicate that capturing quality differences is more subtle than DA scores can express [7].

All explored approaches drop significantly in performance for low-resource language pairs. For the supervised methods, this drop is even more significant than for the unsupervised metrics [7]. Fomicheva et al. argue this is due to the models overfitting on the small training corpus. This leads the authors to state, that the unsupervised metrics are better suited compared to supervised approaches for low-resource scenarios [7].

	Low-resource		**Mid-resource**		**High-resource**	
Method	Si-En	Ne-En	Et-En	Ro-En	En-De	En-Zh
TP	0.399	0.482	0.486	0.647	0.208	0.257
Softmax-Ent	0.457	0.528	0.421	0.613	0.147	0.251
Sent-Std	0.418	0.472	0.471	0.595	**0.264**	0.301
D-TP	0.460	0.558	**0.642**	0.693	**0.259**	**0.321**
D-Var	0.307	0.299	0.356	0.332	0.164	0.232
D-Combo	0.286	0.418	0.475	0.383	0.189	0.225
D-Lex-Sim	**0.513**	**0.600**	0.612	0.669	0.172	**0.313**
AW:Ent-Min	0.097	0.265	0.329	0.524	0.000	0.067
AW:Ent-Avg	0.10	0.205	0.377	0.382	0.090	0.112
AW:best hl	0.255	0.381	0.416	0.636	0.241	0.168
PredEst	0.374	0.386	0.477	0.685	0.145	0.190
BERT-BiRNN	**0.473**	**0.546**	**0.635**	**0.763**	**0.273**	**0.371**

Table 3: Pearson (r) correlation between all evaluated approaches and the human assigned DA scores. Bold values are methods which are not significantly outperformed by any other method. [7]

4.1 Correlation with human judgement

4.1.1 Group I. The TP metric acts as a baseline for all other evaluated approaches. Its performance is already competitively compared to the supervised approaches, but is outperformed by the other two metrics of this group in four language pairs [7]. Fomicheva et al. contribute this performance to two features of these approaches. Firstly, the Softmax-Ent considers a more holistic view of uncertainty in translation output probability due to the examination of the entire vocabulary. Secondly, Sent-Std is capable to distinguish patterns in the variation of output probabilities [7].

4.1.2 Group II. The dropout based quality metrics perform overall best [7]. The D-TP and D-Lex-Sim achieve performances rivaling the current SOTA supervised approaches. Fomicheva et al. argue this performance is due to the accurate estimation of model uncertainty, improving the quality estimation. Especially, D-Lex-Sim reflects the inherent ambiguity of translation hypothesis by directly exploiting the similarity of translation hypothesis [7]. This again directly supports the stated hypothesis of the authors, in which estimating model uncertainty should improve the accuracy of the metrics.

D-Var and therefore the combination D-Combo have a much lower correlation to human annotated quality estimations [7]. Fomicheva et al. contribute this to the inherent problem of only relying on variance of probabilities. The actual output probability of the translation hypothesis is ignored [7]. Considering an evenly distributed output probability over the entire vocabulary for each output token. If the model is producing this even distribution reliably, D-Var and D-Combo would estimate the translation to be of high quality. Though in actual fact, the model is uncertain about each token of the translation hypothesis.

4.1.3 Group III. Lastly, the attention based metrics perform overall worse [7]. Fomicheva et al. contribute this behavior to the lack of a direct mapping between attention weight entropy and translation quality. But in further analysis of this, the additional metric *AW:best head/layer* supports the hypothesis that there is information to be gained from exploring attention weights for the purpose of quality estimation. This leads the authors to state that this should be used over simplified combination of different heads and layers [7]. Since this is no longer unsupervised and requires further analysis, the authors leave this for future work.

4.1.4 Comparison. As stated before, the dropout based quality metrics perform overall best compared to the supervised references. D-Lex-Sim rivals the best performing supervised approach, Bert-BiRNN in four of the six language pairs [7]. This supports the authors' hypothesis, that the unsupervised approaches presented in this paper can be effectively be used to estimate the quality of neural machine translation. This also supports the hypothesis that useful information can be extracted, through the consideration of MT systems as a glass-box, for the task of neural machine translation.

4.2 Further evaluation

All approaches relying on the output probability distribution face the problem of bad calibrated models [7]. To further evaluate this behavior and explore solutions to this problem, Fomicheva et al. explored different factors influencing this calibration. Namely, domain shift, different underlying NMT systems, and the duration of training.

4.2.1 Domain Shift. Providing a trained model with a sentence which topic domain differs significantly from all sentences used during training, the model would most likely produce a low-quality translation. A good calibrated model provided with such an input sentence would produce a low confidence in the translation correctness, whereas a badly calibrated model would be overconfident in the correctness of the translation. In domain shift, such scenarios

are deliberately explored to test the influence of model calibration on the performance of the metric [7].

The test section of the provided dataset was used as test data for the in domain section. The out of domain examples were extracted from the Wikipedia documents not considered during dataset selection. To retrieve the most-out-of domain examples, the distance metric based on Niehues et al. [18] was used. The distance scores are calculated using the encoder hidden states of the input sentences. The furthest examples were selected for the purpose of this experiment [7].

The standard token-level probability TP and dropout based token-level probability D-TP metric were evaluated on both outputs for in- and out of domain examples [7].

The results show, there is no significant difference for the standard TP metric. This indicates a less robust quality measure when dealing with unreliable probability output calibration [7]. The D-TP metric on the other hand produces significantly different results. This indicates that better estimation model uncertainty improves the accuracy of quality estimation [7]. This again support the hypothesis of the authors that estimating model uncertainty improves quality estimation performance.

4.2.2 NMT Systems. As a second set of experiments, the underlying neural translation system have been evaluated. This experiment again evaluates the influence of model output calibration on the quality estimation performance of TP. The experiment can be broken down in three sections. Firstly, the underlying NMT architecture was changed. These architectures include RNN-based NMT [3], Mixture of Experts [13], and model ensemble [9]. The model ensemble consist of four Transformer models initialized by different random seeds. For the mixture of experts, a hard mixture model with five components was used, where the translation hypothesis were generated by randomly chosen components with standard beam search [7]. Secondly, two different approaches than standard beam search were evaluated, namely diverse beam search [23] and sampling [7]. The D-TP metric was used as a reference method for evaluating overall improvements.

According to [7] the performance of the metric can be improved using different systems by improving output calibration. Model ensemble provides the best correlation results for standard TP metric. Fomicheva et al. relate this achievement to the better uncertainty quantification of the NMT model. The D-TP metric still performs comparatively to the best performing TP variations [7]. Fomicheva et al. finally argue, supported by the experiment data, that the correlation between output probability and DA is not necessarily related to the quality of MT outputs [7].

4.2.3 Calibration across training Epochs. Finally, the training of neural machine translation models is evaluated. This is done to observe the effect of the amount of training on the correlation between translation probability or the calibration of the output and translation quality [7]. In this experiment, the model has been trained for 60 epochs, where the translations were evaluated after each epoch. Fomicheva et al. find that while the test quality stabilizes with continuing training, the correlation between the probabilities and quality decreases. This leads the authors to argue that the continued training does not affect output quality, but damages the calibration of the output probability [7]. This supports the hypothesis of the authors that well calibrated model outputs can be reliably used for quality estimation.

5 DISCUSSION

In this section, the previous explored work is discussed. An outlook of the presented methods and suggested future work are presented, as well as some criticism of the paper.

5.1 Future work

As by the authors suggested, the strong correlation between human judgement and attention weights for selected head/layer combination indicates some information contained in the attention weights. This is an interesting concept which could not only help quality measure for translation task. Therefore, this approach should be investigated further.

The work by Fomicheva et al. focuses solely on epistemic uncertainty quantification for improving quality estimation [7]. Approaches related to modelling aleatoric uncertainty can be explored as well. For this, the noise inherent to the observation could be used to further improve on the existing metrics presented in the work.

On a more general note, the presented metrics focus mainly on sentence-level quality estimation. This can be extended to different levels of translation quality estimation, like discribed in [14]. The scope of quality estimation can either be reduced to word-level quality estimation, which should be rather straightforward [7] or the scope can be extended to sections or entire documents.

The presented metrics can also be used to extend the existing supervised quality estimation approaches. Since both a good overall performance and a benefit compared to supervised methods in special cases has been shown, the information could improve existing approaches. This poses an interesting combination of both general concepts for future work.

Currently, only machine translation scenarios have been considered, but the proposed metrics can be easily used in different problem domains. These could be more closely related to the original domain, like machine transcription. Other than that, different tasks like classification, where there are no sequences involved, but the output resembles that of the translation task can also be considered.

Fomicheva et al. already explore semi-supervised training of models in their experiments. The presented metrics could be used as a quality estimator for the teacher-network in such models. This could improve the performance of supervised models in low-resource scenarios, since the authors already provided evidence of the benefit of their unsupervised approaches in these scenarios.

The metrics could also be used to directly support human annotators during the annotation process. Either by providing an initial indication of quality or by annotating some data based on the previous human annotations. This together with the semi-supervised learning can also be done in different problem domains, other than machine translation. Especially, the D-Lex-Sim approach could be used for regression problems, where the similarity between hypothesis can be easily measured by the euclidean distance between regression hypothesis. Therefore, these unsupervised quality estimations can not only be used to estimate the quality of machine translation but a much bigger set of different problem domains.

This would be an interesting use case for future work in different problem domains.

When dealing with heterogeneity systems, the metrics can be used to decide which model should be used to create the translations. Consider two models, one model which creates translation with few resource costs but overall low average quality and the second model, which creates high quality translations but with a higher resource cost. As long as the measured quality from the low-quality model is sufficient, the model can be used. When the measured quality drops below a certain threshold, the high-quality model is used to create a better, although more costly translation.

On a more simplistic note, the metrics can directly be incorporated in current machine translation services to inform the user about the quality of the currently provided translation. This would help to inform the user whether to trust the provided translation.

5.2 Conclusion

Although Fomicheva et al. state, that the presented set of reference methods is by no means extensive, the chosen approaches only account for the supervised methods. Furthermore, the set of reference methods is rather small. The following section describes related work, which could have been used as further reference for the presented metrics.

[17] worked on sentence quality measurement by comparing a sentence to a well-formed sentence database, where the similarity measure is based on TF-IDF. The provided dataset could have been adapted to provide a well-formed sentence database, which would have allowed the authors to use this method for comparison.

[2] worked on a supervised approach of estimating the confidence of a NLP system by evaluating the input and output of the system. For evaluating their results, they used negative log-likelihood between their probabilistic model and the test corpus of translation quality. The test corpus used in [2] is similar to the dataset that was created for this paper.

[6] developed multiple feature extractors for quality estimation. These feature extractors focused on different aspects of the translation task, such as the word alignment scores, the lexical translation probabilities, and measuring the fluency of word sequences in a given language. This method is in principle similar to the D-Lex-Sim metric proposed in [7] and could be a good reference for evaluating the use of Meteor similarity compared to other methods such as in [6].

[20] and [25] focus on the extraction of information from attention weights. While [20] focuses on different measures to calculate the strength of attention connection between source and target sentences. The absentmindedness penalty method described in their work closely resembles the equation 9. [25] focuses on the similarity between input and output segments. This was done by using a method called *BLEU2VEC* to calculate an embedding score cross-lingual [25]. Both methods are interesting reference approaches for the attention based methods presented in [7].

Although, as [21] states, most of these methods do not rival the performance of current neural based approaches. The previously presented methods are mostly outperformed by [24] which in turn is outperformed by [1]. The choice of [1] as the main reference approach is therefore reasonable. It is still possible to infer the performance of these unsupervised approaches to the previous works. Still, the number of reference methods could be increased to provide a more reliable and meaningful analysis of the presented metrics.

Although the dataset is designed with sentence-level translation quality in mind, the translated sentences originate from the same Wikipedia articles. Sentences extracted from the same article will inevitably contain information from other sentences. The neural translation model and the human annotator lack the knowledge of this information. This in turn could both impact the quality of the translation and the ability to accurately rate the quality of the translation.

The selected service providers pose a source of uncertainty in the quality of the quality ratings. The evaluation of the metrics directly rely on the quality of the human annotation on each translation hypothesis. Other works like [10] use a far lager amount of annotators for their dataset. But both works point out the inherent problems that come with an unsupervised human annotation task, which cannot be evaluated. But Fomicheva et al. state that the variance of ratings is lower than stated in [10], which could indicate a higher quality of annotation. Even tough, this is not definite and also not further discussed in this paper.

When manually evaluating the annotations, all annotations which have a difference in rating by more than 30 points are redone with an additional annotator [7]. This in of itself is problematic since this does not reflect the true annotated quality and influences the annotation directly. The influence of this process can not be evaluated since all differing annotations are discarded and can no longer be tracked. Furthermore, there exists no information on the translation for which ratings had to be redone.

The additional evaluation section provides important information about the hypothesis of the authors regarding output calibration. All system, which were evaluated, were trained using different data. The attributes of the used sets of training data can not be believed to be equivalent, and similar problems as stated by [7] can occur, like for the En-De language pair. Therefore, the meaning of the results drawn from the comparison between the experiments attained from each method is limited.

The provided dataset consists of a highly complex ensemble of different language pairs. The authors not only provide the translation with the estimated translation quality, but additionally provide different information for recreating these findings. Fomicheva et al. provide sentence-level probabilities for all translations for evaluating TP scores and multiple translation hypothesis for each translation for evaluating D-Lex-Sim. These additional information help to validate the findings of the authors for the two aforementioned metrics.

The data from the additional evaluation section provide a deeper insight into different aspects which influence the performance of the metrics. Especially, the reduction in output calibration over longer training duration is interesting and could be useful in considering the number of training epochs for various problem domains and their influence on the achieved results.

The analysis of the findings in the main experiments are extensive and important findings and their related concepts are clearly explained, explored, and visualized to provide an easier and better understanding of the underlying causes.

The hypothesis of the authors are well discussed, supported and analyzed, especially with the additional experiments conducted.

REFERENCES

[1] Frederic Blain, Nikolaos Aletras, and Lucia Specia. 2020. Quality In, Quality Out: Learning from Actual Mistakes. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation, Lisboa, Portugal, 145–153. https://aclanthology.org/2020.eamt-1.16

[2] John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence Estimation for Machine Translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*. COLING, Geneva, Switzerland, 315–321. https://aclanthology.org/C04-1046

[3] Jaemin Cho, Min Joon Seo, and Hannaneh Hajishirzi. 2019. Mixture Content Selection for Diverse Sequence Generation. *CoRR* abs/1909.01953 (2019). arXiv:1909.01953 http://arxiv.org/abs/1909.01953

[4] Michael Denkowski and Alon Lavie. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Baltimore, Maryland, USA, 376–380. https://doi.org/10.3115/v1/W14-3348

[5] Miquel Esplà, Mikel Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. 2019. ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*. European Association for Machine Translation, Dublin, Ireland, 118–119. https://aclanthology.org/W19-6721

[6] Thierry Etchegoyhen, Eva Martínez Garcia, and Andoni Azpeitia. 2018. Supervised and Unsupervised Minimalist Quality Estimators: Vicomtech's Participation in the WMT 2018 Quality Estimation Task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Association for Computational Linguistics, Belgium, Brussels, 782–787. https://doi.org/10.18653/v1/W18-6461

[7] Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised Quality Estimation for Neural Machine Translation. *CoRR* abs/2005.10608 (2020). arXiv:2005.10608 https://arxiv.org/abs/2005.10608

[8] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48* (New York, NY, USA) *(ICML'16)*. JMLR.org, 1050–1059.

[9] Ekaterina Garmash and Christof Monz. 2016. Ensemble Learning for Multi-Source Neural Machine Translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, 1409–1418. https://aclanthology.org/C16-1133

[10] 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*. Association for Computational Linguistics, Sofia, Bulgaria, 33–41. https://aclanthology.org/W13-2305

[11] YVETTE GRAHAM, TIMOTHY BALDWIN, ALISTAIR MOFFAT, and JUSTIN ZOBEL. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering* 23, 1 (2017), 3–30. https://doi.org/10.1017/S1351324915000339

[12] 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)*. Association for Computational Linguistics, Hong Kong, China, 6098–6111. https://doi.org/10.18653/v1/D19-1632

[13] Xuanli He, Gholamreza Haffari, and Mohammad Norouzi. 2018. Sequence to Sequence Mixture Model for Diverse Machine Translation. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Brussels, Belgium, 583–592. https://doi.org/10.18653/v1/K18-1056

[14] Julia Ive, Frédéric Blain, and Lucia Specia. 2018. deepQuest: A Framework for Neural-based Quality Estimation. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 3146–3157. https://aclanthology.org/C18-1266

[15] Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. Predictor-Estimator using Multilevel Task Learning with Stack Propagation for Neural Quality Estimation. In *Proceedings of the Second Conference on Machine Translation*. Association for Computational Linguistics, Copenhagen, Denmark, 562–568. https://doi.org/10.18653/v1/W17-4763

[16] Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of Machine Translation Summit X: Papers*. Phuket, Thailand, 79–86. https://aclanthology.org/2005.mtsummit-papers.11

[17] Erwan Moreau and Carl Vogel. 2012. Quality Estimation: an experimental study using unsupervised similarity measures. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Montréal, Canada, 120–126. https://aclanthology.org/W12-3114

[18] Jan Niehues and Ngoc-Quan Pham. 2019. Modeling Confidence in Sequence-to-Sequence Models. *CoRR* abs/1910.01859 (2019). arXiv:1910.01859 http://arxiv.org/abs/1910.01859

[19] 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*. Association for Computational Linguistics, Brussels, Belgium, 1–9. https://doi.org/10.18653/v1/W18-6301

[20] Matiss Rikters and Mark Fishel. 2017. Confidence through Attention. *CoRR* abs/1710.03743 (2017). arXiv:1710.03743 http://arxiv.org/abs/1710.03743

[21] Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón F. Astudillo, and André F. T. Martins. 2018. Findings of the WMT 2018 Shared Task on Quality Estimation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Association for Computational Linguistics, Belgium, Brussels, 689–709. https://doi.org/10.18653/v1/W18-6451

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *CoRR* abs/1706.03762 (2017). arXiv:1706.03762 http://arxiv.org/abs/1706.03762

[23] Ashwin K. Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David J. Crandall, and Dhruv Batra. 2016. Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models. *CoRR* abs/1610.02424 (2016). arXiv:1610.02424 http://arxiv.org/abs/1610.02424

[24] Jiayi Wang, Kai Fan, Bo Li, Fengming Zhou, Boxing Chen, Yangbin Shi, and Luo Si. 2018. Alibaba Submission for WMT18 Quality Estimation Task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Association for Computational Linguistics, Belgium, Brussels, 809–815. https://doi.org/10.18653/v1/W18-6465

[25] Elizaveta Yankovskaya, Andre Tättar, and Mark Fishel. 2018. Quality Estimation with Force-Decoded Attention and Cross-lingual Embeddings. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Association for Computational Linguistics, Belgium, Brussels, 816–821. https://doi.org/10.18653/v1/W18-6466