Improving Machine Translation Quality Estimation Using Named-Entity Masking and Assessment Scores

Anthony Reidy[a], Sean Cummins[a], Kian Sweeney[a], George Dockrell[a], Pintu Lohar[b] and Andy Way[b]

[a]School of Computing, Dublin City University, Dublin 9, Ireland
[b]ADAPT Centre, Dublin City University, Dublin 9, Ireland

Abstract

This paper reports our findings in quality estimation (QE) in machine translation (MT) using the data set of WMT-2020 shared task. We perform sentence-level direct assessment (DA) and focus on the English–Chinese, Romanian–English, and English–German language pairs. We build on the single XLM-R transformer model within the state-of-the-art *Transquest* system [1] through named entity (NE) masking and analysis of quality assessment scores. Our methodologies result in the improvement in *Transquest* system for all of our chosen language pairs by achieving a higher Pearson correlation. We also obtain a reduction in error for all of these language pairs.

Keywords

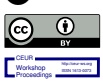
quality estimation, machine translation, TransQuest, named-entity masking

1. Introduction

The MT quality estimation frameworks attempt to estimate the quality of translation outputs at varying levels of granularity: word, phrase, sentence, and document, without access to 'gold-standard' human-generated reference translations [2]. It can greatly reduce the cost associated with this evaluation process and also has the added ability of determining whether a machine-generated translation can be published as is, or whether it requires human post-editing efforts [3]. In this work, we use the framework developed by the TransQuest team [4] that achieved the best results in the WMT-2020 shared task of quality estimation. However, they mentioned in their error analysis that the presence of NEs causes the largest number of errors between their predicted scores and expected scores within their system. Considering this problem, we propose an approach of NE masking (discussed in details later in Section 5.1) with an aim of improving the performance of the QE system.

In addition, we utilise the following assessment metrics in our experiments to further extend our contribution.

CERC 2021: Collaborative European Research Conference, September 09–10, 2021, Cork, Ireland
anthony.reidy3@mail.dcu.ie (A. Reidy); sean.cummins26@mail.dcu.ie (S. Cummins); kian.sweeney27@mail.dcu.ie (K. Sweeney); george.dockrell2@mail.dcu.ie (G. Dockrell); pintu.lohar@adaptcentre.ie (P. Lohar); andrew.way@dcu.ie (A. Way)
https://github.com/reidya3 (A. Reidy)



- Root mean square error (RMSE)
- Mean absolute error (MAE)
- Spearman correlation coefficient
- Pearson correlation coefficient

We place an emphasis on the Pearson correlation metric as this is the metric used to compare models in the WMT-2020 shared task.

For the purpose of this research, we focus on the first task which is sentence-level direct assessment. We examine the (i) English–German, (ii) Romanian–English, and (iii) English–Chinese language pairs because we are interested in investigating how our approach performs with language pairs of the same and different scripts.

The remainder of this paper is organised as follows. We discuss the related works in this field in Section 2. The description of the DA dataset is provided in Section 3. Section 4 outlines our system architecture. Section 5 explains our experiment where we discuss the methodologies we use in this work to improve the performance of *MonoTransQuest* (the name of the system developed by the *TransQuest* team). We provide our results in Section 6. Finally, we conclude our work and provide some possible directions for the future in Section 7.

2. Related Work

A considerable amount of work has been done in the area of MT quality estimation. Fomicheva et al. [5] apply both a glass box and black box approach that achieves impressive results across a number of language pairs. This glass box approach uses features extracted from the NMT model and is very cost effective. However, it is their black box model, which looks at pre-trained representations using source and target text, that tied for the winning submission in four of the seven language pairs. In a similar manner, Moura et al. [6] uses this newly available features of the NMT model to further extend the OpenKiwi system [7] by using a Kiwi glass box ensemble alongside an OpenKiwi-based submission. This glass box extracts these features and feeds them into the OpenKiwi system.

Nakamachi et al. [8] makes use of an ensemble model of four regression models based on XLM-R [9], adding a language token for each sentence while Hu et al. [10] also uses an ensemble model with transfer learning and multilingual pretrained models. Zhou et al. [11] exposes explicit cross lingual patterns to zero-shot models in order to augment BERT scores. Ranasinghe et al. [1], the winning submission for WMT-2020 uses crosslingual embeddings to remove the dependency on parallel data using a pre-trained XLM-R large transformer model. This simplifies the complex neural network architecture and hence reduces the computational cost. However, the authors of this winning team mentioned that one of the main problems in their system was caused by the presence of NEs. The proper handling of NEs in quality estimation task is still less explored. In this work, we address this problem by NE masking in combination with the analysis of quality assessment scores.

Table 1
English–German training set record

Column	Value
Original	The burning propellant generates inert gas which rapidly inflates the airbag in approximately 20 to 30 milliseconds.
Translation	Das brennende Treibmittel erzeugt inertes Gas, das den Airbag in etwa 20 bis 30 Millisekunden schnell aufblasen lässt.
Scores	[55, 62, 85]
Mean	67.3333333333333
Z_scores	[-1.5259978765984137, -0.3605398796887971, -0.4917681450113435]
Z_mean	-0.792768633766184
Model_scores	-0.374686628580093

3. Description of the DA Dataset

The organizers of the WMT-2020 shared task on quality estimation provided the participants with the data sets extracted from Wikipedia for six language pairs, including our language pairs of interest. The training sets consists of 7, 000 sentence pairs for each language pair whereas both the development and test sets contain 1, 000 sentence pairs each. Preliminary analysis of both the training and the test data using latent dirichlet allocation (LDA) and hierarchical latent dirichlet allocation (HLDA) produced inconclusive results. The data comes from a variety of Wikipedia articles and as such, there are no general themes. Table 1 shows the example of an English—German training set record with the following entities.

- **Original:** The source sentence from a Wikipedia article.
- **Translation:** The translation of the source sentence produced by the state-of-the-art transformer-based NMT models, built using the fairseq toolkit [12].
- **Scores:** A score denoting the perceived quality of a translation, ranging from 0-100. Professional translators followed the FLORES guidelines [13] when manually annotating the DA scores. The number of annotators range from three to six.
- **Mean:** The collective scores of the raters for the translation are averaged to obtain this value.
- **Z_scores**: A list representing the standardised scores.
- **Z_mean:** The mean of the z-scores. Our architecture seeks to predict the mean DA z-scores of the test sentence pairs.
- **Model_scores:** The baseline QE system for the shared task is an LSTM-based Predictor-Estimator approach [14], implemented in OpenKiwi [3, 15]. This feature is provided to foster improvements over the described baseline.

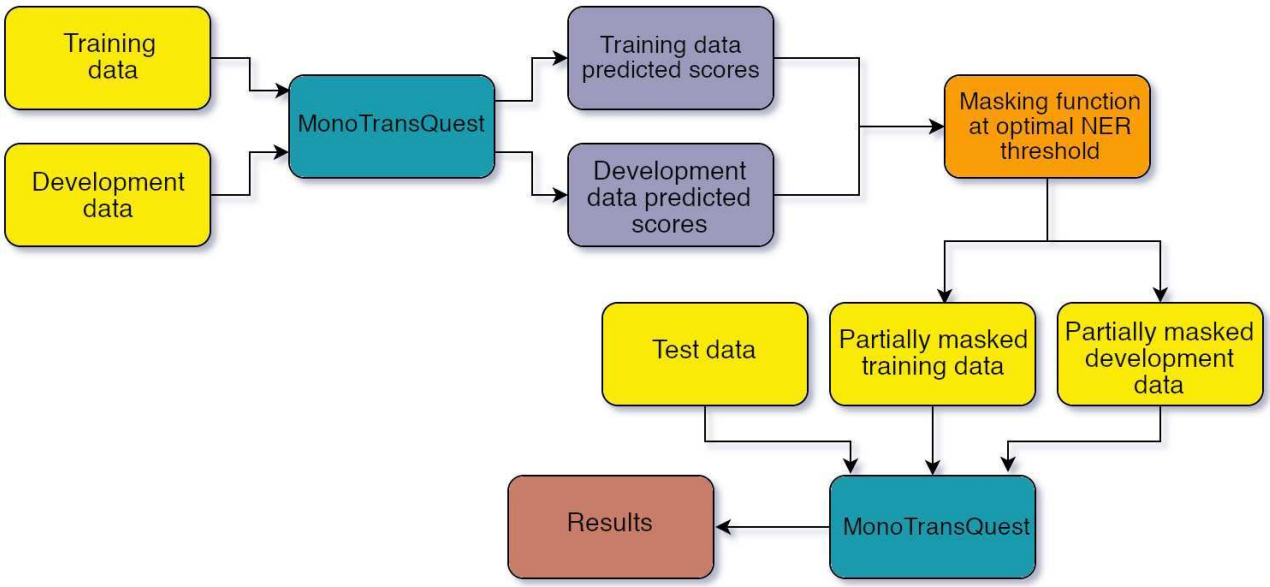


Figure 1: Architecture of our MTQE system

4. System Architecture

Our system architecture, shown in Figure 1, focuses on improving the results from the baseline *MonoTransQuest* system by incorporating our proposed approach.

Firstly, both the training and development data for a chosen language pair are passed through a standard run of *MonoTransQuest* where each tuple is assigned a predicted z-mean score. We calculate the absolute error (AE) for each tuple as the absolute difference between the predicted and actual z-mean scores.

We experiment with various AE thresholds (discussed later in Section 5). Once the masking process is complete for both the training and development data sets, the data is passed through *MonoTransQuest* again to obtain the final results. The improvements resulting from this architecture are highlighted and discussed in Section 6.

5. Experiments

In this section we present the methods we used in our experiments in chronological order.

5.1. NE masking

The TransQuest team documented their system's difficulties with NEs in their paper. The occurrence of NEs in the source and target language sentences causes a large proportion of errors. The source-language sentences containing NEs regularly have translations that contain slight misspellings. *TransQuest* seems to penalise such occurrences greatly. In order to address this problem, we use *spaCy* [16], an open-source software for advanced natural language processing (NLP). In addition to *spaCy*, we use Stanford NLP's *Stanza* [17], the Python equivalent of the originally Java-based Stanford NLP. *spaCy* and *Stanza* offer different language models

Table 2
An example of a Romanian–English instance after NE masking

Language	Input	Output
Romanian	În urmă explorărilor Căpitanului James Cook, Australia și Noua Zeelandă au devenit ținte ale colonialismului britanic.	În urmă explorărilor Căpitanului NE8734, NE27 și NE4612 au devenit ținte ale colonialismului britanic
English	Following the explorations of Captain James Hook, Australia and New Zealand became targets of British colonialism.	Following the explorations Captain NE5123, NE78113 and NE892 became targets of British colonialism.

meaning our choice of toolkit is dependant on the language pairs of our interest. We apply the following approaches for NE masking.

1. *The Regex*[1] *method* allows us to specifically target equivalent entities between the source- and target-language texts. This method is only practical for the English–German language pair for the reasons outlined in the '*The Above MAE method*'.

   The Regex package provides a fuzzy-match functionality that allows users to find similarities between texts that may be less than 100% perfect. For example, in the German sentence *'Er regierte unterdrückerisch und fast bankrott Mali'*, a search for *'bankrupt'* will return *'bankrott'*. Regex does this by allowing a certain amount of letter substitutions, deletions, and insertions. These operations contribute to an overall edit-distance.

   For every NE found in an English sentence, the corresponding German translation is searched for sub-strings that are similar to the English NE within a specified amount of text alterations. We then mask the NE with the same '**NEXXX**' value that is used for that NE in the source sentence. After experimentation, we deduce that allowing our matching function a total of 5 letter deletions and 5 letter replacements produces the best quality matches for this language pair. We also experiment with masking corresponding NEs with different values as opposed to the same value method mentioned previously. Masking with different values achieves better results.

2. *The double model method* involves using separate language models for the source- and the target-language texts. We use the source-language NE model to detect NEs within the source-language text, which once found, are replaced by an '**NEXXX**' string, where '**XXX**' is a uniformly random distributed number between 0 and 8,000 as rarely did we identify more than 8,000 NEs. An example of NE masking can be found in Table 2.

   The target-language NE model is used to mask entities in the target-language text in an identical fashion. *spaCy* is used for the Romanian–English language pair whilst *Stanza* is used for the English-Chinese and English–German language pairs.

[1]Python Regex - https://pypi.org/project/regex/

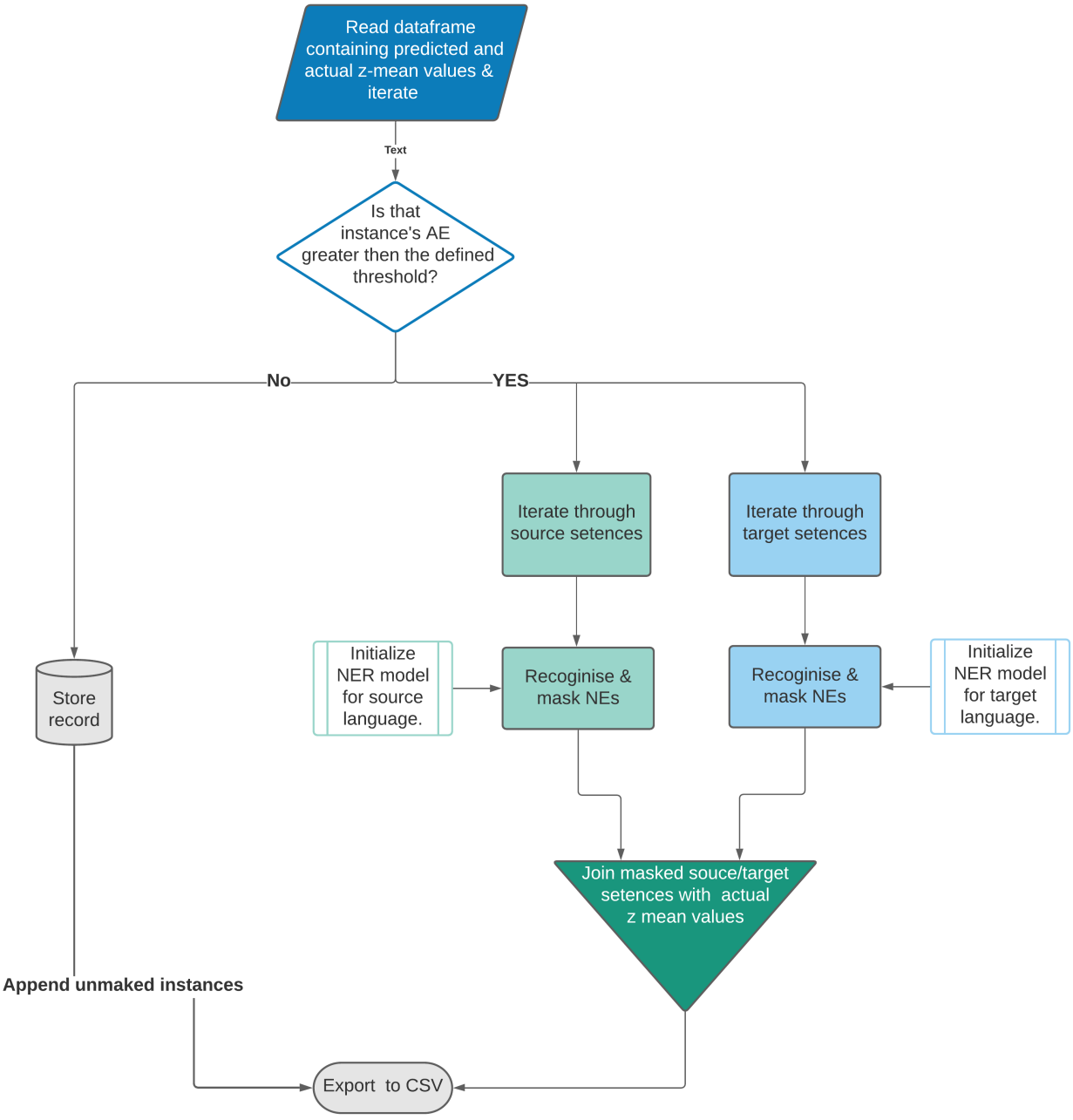


Figure 2: Workflow of the NE masking system

3. *The Above MAE method* is a less naive adaptation of the previous Regex method for the English–German language pair and the double model method for the Romanian–English and English-Chinese language pairs. It is impractical to use the Regex method for the two aforementioned language pairs due to large fundamental differences between the languages. English and Romanian are not part of the same language family whilst English and Chinese have different alphabets.

5.2. Analysis of the assessment scores

The TransQuest team indicates that they performed error analysis where the difference between the predicted score and actual score is the largest. We propose a feedback mechanism in order to retrieve the predicted z-mean scores of both the training and development data. We then calculate the value of AE between the expected and predicted z-mean value for each record in the data set. This is shown in Equation (1)

$$AE_i = |y_i - x_i| \tag{1}$$

where AE_i = the absolute error of an instance, y_i = the predicted z-mean value by *MonoTransQuest* of that instance , and x_i = the actual z-mean value of that instance calculated from the professional translators' scores. All the found NEs, in sentence pairs that have an AE greater than the MAE of standard *MonoTransquest* for that language pair are masked. Instances less than the MAE are left unchanged.

MAE is chosen as RMSE gives a higher weight to large errors. In addition, RMSE does not just simply describe the average errors. It also describes other characteristics which are often difficult to understand and comprehend. Willmott and Matsuura [18] suggest that the RMSE is not a good indicator of average model performance, and might be a misleading indicator of average error. Thus, the MAE is a better metric for this purpose. This method negates the possibility of our masking function reducing the score of sentence pairs that *MonoTransQuest* can predict relatively well.

4. *The Specific Threshold method* is an improvement of the above MAE method. Instead of masking all instances that have an AE value above their respective MAE value, we experiment with different thresholds. In this method, the training and development data are sorted by decreasing AE values.

   In order to find the optimal setting for this method, we experiment with different threshold values. Through rigorous testing in the range of 5% to 1% (i.e 5% of training and development sentence pairs with the highest AE values are masked), we find the value that results in the highest increase in Pearson correlation.

6. Results

We first evaluate our framework with *the Above MAE method* followed by the *the Specific Threshold method*. Although Pearson correlation is the most commonly used evaluation metric in WMT quality estimation shared tasks, we decide to incorporate other metrics in our evaluation in order to gain a complete understanding of our architecture's performance and limitations. We choose *MonoTransQuest* as our baseline as this was the strongest system in WMT-2020 and what our architecture is built on. The results of our methods are shown in Figure 3 and also in Tables 3, 4, and 5.

Table 3 shows the results for the English–Chinese language pair. We achieve our highest Pearson correlation boost of 0.2204 (this is a 40% improvement on the *MonoTransquest* system) using *the Above MAE method*.

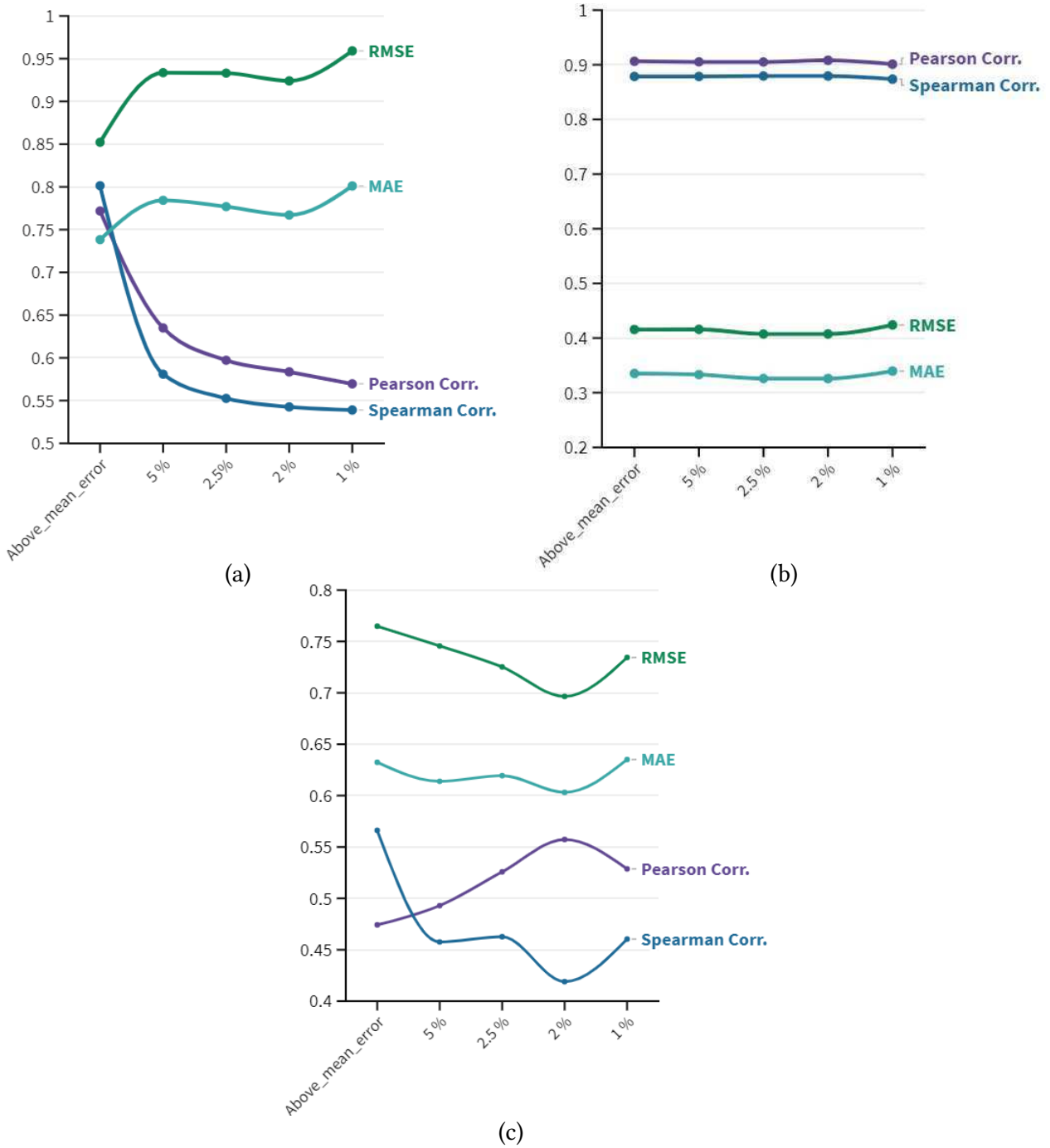


Figure 3: Figures displaying results for the (a) English–Chinese, (b) Romanian–English and (c) English–German language pairs.

This is also our highest Pearson correlation boost of any language pair. In addition, we achieve a 0.2525 Spearman correlation boost (an improvement of 46% on *MonoTransquest*) as well as a reduction in both RMSE and MAE. This suggests that the presence of NEs in the English–Chinese language pair hinders *MonoTransQuest*'s performance significantly.

The results for Romanian–English are highlighted in Table 4. As can be seen in this table, we achieved our highest Pearson correlation boost (0.0061) over *MonoTransQuest* when using *the Specific Threshold method* at the 2% level.

Table 3
Results for English–Chinese

Experiments	RMSE	MAE	Spearman	Pearson
MonoTransQuest	0.9511	0.7804	0.5489	0.5514
Above MAE	**0.8522**	**0.7383**	**0.8014**	**0.7718**
Worst 1% Masked	0.9591	0.8011	0.539	0.5696
Worst 2% Masked	0.924	0.7671	0.5426	0.5836
Worst 2.5% Masked	0.9331	0.7769	0.5526	0.5972
Worst 5% Masked	0.9336	0.7843	0.5809	0.635

Table 4
Results for Romanian–English

Experiments	RMSE	MAE	Spearman	Pearson
MonoTransQuest	0.4209	0.3375	0.872	0.9021
Above MAE	0.4156	0.3352	0.8786	0.9064
Worst 1% Masked	0.424	0.3396	0.8737	0.9011
Worst 2% Masked	**0.4074**	**0.3259**	**0.8794**	**0.9082**
Worst 2.5% Masked	0.4131	0.333	0.8758	0.9051
Worst 5% Masked	0.4159	0.3331	0.8777	0.9053

Table 5
Results for English–German

Experiments	RMSE	MAE	Spearman	Pearson
MonoTransQuest	0.7757	0.6444	**0.4807**	0.461
Above MAE	0.7495	0.6259	0.3695	0.4309
Worst 1% Masked	0.7683	0.6588	0.3603	0.5085
Worst 2% Masked	**0.6965**	**0.6031**	0.4419	**0.5573**
Worst 2.5% Masked	0.7252	0.6194	0.4627	0.5258
Worst 5% Masked	0.7563	0.6365	0.3856	0.4271

This setting also achieves an increased Spearman value and a reduction in RMSE and MAE. Table 5 highlights the results for the English–German language pair. The *Specific Threshold method* produces the largest improvement in Pearson correlation (0.0963) at the 2% level. It also results in improvements in the RMSE and MAE over the baseline *MonoTransQuest* system.

7. Conclusions and Future Work

In this paper, we discussed the background of the WMT-2020 shared task, detailing the scientific topics that this benchmark intends to progress. We focused on the task of sentence-level direct assessment from WMT-2020. We explained our methodologies behind planned improvements on the *TransQuest* baseline system (*MonoTransQuest*), showing both our successful and less successful approaches.

The experimental results revealed that our proposed system outperformed the state-of-the-art *TransQuest* team. However, there are several possibilities to extend this work in future. One of them is to apply the NE masking system to other language pairs. Another possibility is to explore other medium and high resource language pairs in order to further test the robustness of our system. In addition, it is also possible to conduct experiments with data augmentation by extending the training data by using additional data that is similar to the test data set. It can be done by using a text similarity method based on word embeddings and other state-of-the-art sentence similarity methods in order to extract data that are semantically equivalent to the test data set.

Acknowledgements

This research has been supported by the ADAPT Centre for Digital Content Technology which is funded under the SFI Research Centres Programme (Grant 13/RC/2106).

References

[1] T. Ranasinghe, C. Orasan, R. Mitkov, TransQuest at WMT2020: Sentence-level direct assessment, in: Proceedings of the Fifth Conference on Machine Translation, Online, 2020, pp. 1049–1055. URL: https://www.aclweb.org/anthology/2020.wmt-1.122.

[2] L. Specia, F. Blain, V. Logacheva, R. F. Astudillo, A. F. T. Martins, Findings of the WMT 2018 shared task on quality estimation, in: Proceedings of the Third Conference on Machine Translation: Shared Task Papers, Belgium, Brussels, 2018, pp. 689–709. URL: https://www.aclweb.org/anthology/W18-6451. doi:10.18653/v1/W18-6451.

[3] F. Kepler, J. Trénous, M. Treviso, M. Vera, A. F. T. Martins, OpenKiwi: An Open Source Framework for Quality Estimation, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Florence, Italy, 2019, pp. 117–122. doi:10.18653/v1/P19-3020.

[4] T. Ranasinghe, C. Orasan, R. Mitkov, TransQuest: Translation quality estimation with cross-lingual transformers, in: Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 5070–5081. URL: https://www.aclweb.org/anthology/2020.coling-main.445. doi:10.18653/v1/2020.coling-main.445.

[5] M. Fomicheva, S. Sun, L. Yankovskaya, F. Blain, V. Chaudhary, M. Fishel, F. Guzmán, L. Specia, Bergamot-latte submissions for the wmt 20 quality estimation shared task, in: Proceedings of the Fifth Conference on Machine Translation, Online, 2020, pp. 1008–1015.

[6] J. Moura, M. Vera, D. van Stigt, F. Kepler, A. F. T. Martins, IST-unbabel participation in the WMT20 quality estimation shared task, in: Proceedings of the Fifth Conference on Machine Translation, Online, 2020, pp. 1029–1036. URL: https://www.aclweb.org/anthology/2020.wmt-1.119.

[7] F. Kepler, J. Trénous, M. Treviso, M. Vera, A. F. T. Martins, OpenKiwi: An open source framework for quality estimation, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics–System Demonstrations, Association for Com-

putational Linguistics, Florence, Italy, 2019, pp. 117–122. URL: https://www.aclweb.org/anthology/P19-3020.

[8] A. Nakamachi, H. Shimanaka, T. Kajiwara, M. Komachi, Tmuou submission for wmt20 quality estimation shared task, in: Proceedings of the Fifth Conference on Machine Translation, Online, 2020, pp. 1035–1039.

[9] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 2020, pp. 8440–8451. doi:10.18653/v1/2020.acl-main.747.

[10] C. Hu, H. Liu, K. Feng, C. Xu, N. Xu, Z. Zhou, S. Yan, Y. Luo, C. Wang, X. Meng, T. Xiao, J. Zhu, The niutrans system for the wmt20 quality estimation shared task, in: Proceedings of the Fifth Conference on Machine Translation, Online, 2020, pp. 1016–1021.

[11] L. Zhou, L. Ding, K. Takeda, Zero-shot translation quality estimation with explicit cross-lingual patterns, in: Proceedings of the Fifth Conference on Machine Translation, Online, 2020, pp. 1066–1072.

[12] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, M. Auli, fairseq: A fast, extensible toolkit for sequence modeling, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 48–53. URL: https://www.aclweb.org/anthology/N19-4009. doi:10.18653/v1/N19-4009.

[13] F. Guzmán, P.-J. Chen, M. Ott, J. Pino, G. Lample, P. Koehn, V. Chaudhary, M. Ranzato, 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, 2019, pp. 6098–6111. URL: https://www.aclweb.org/anthology/D19-1632. doi:10.18653/v1/D19-1632.

[14] H. Kim, J.-H. Lee, S.-H. Na, 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, 2017, pp. 562–568. URL: https://www.aclweb.org/anthology/W17-4763. doi:10.18653/v1/W17-4763.

[15] L. Specia, F. Blain, M. Fomicheva, E. Fonseca, V. Chaudhary, F. Guzmán, A. F. T. Martins, Findings of the WMT 2020 shared task on quality estimation, in: Proceedings of the Fifth Conference on Machine Translation, Association for Computational Linguistics, Online, 2020, pp. 743–764. URL: https://www.aclweb.org/anthology/2020.wmt-1.79.

[16] M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, spacy: Industrial-strength natural language processing in python, 2020. URL: https://doi.org/10.5281/zenodo.1212303. doi:10.5281/zenodo.1212303.

[17] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, C. D. Manning, Stanza: A python natural language processing toolkit for many human languages, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Online, 2020, pp. 101–108. URL: https://www.aclweb.org/anthology/2020.acl-demos.14. doi:10.18653/v1/2020.acl-demos.14.

[18] C. J. Willmott, K. Matsuura, Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance, Climate research 30 (2005) 79–82.