EPOQUE: An English-Persian Quality Estimation Dataset

Mohammad Hossein Jafari Harandi, Fatemeh Azadi, Mohammad Javad Dousti, and Heshaam Faili

School of Electrical and Computer Engineering, College of Engineering, University of Tehran

Tehran, Iran

{hosein.jafari.h, ft.azadi, mjdousti, hfaili}@ut.ac.ir

Abstract

Translation quality estimation (QE) is an important component in real-world machine translation applications. Unfortunately, human labeled QE datasets, which play an important role in developing and assessing QE models, are only available for limited language pairs. In this paper, we present the first English–Persian QE dataset, called EPOQUE, which has manually annotated direct assessment labels. EPOQUE contains 1000 sentences translated from English to Persian and annotated by three human annotators. It is publicly available, and thus can be used as a zero-shot test set, or for other scenarios in future work. We also evaluate and report the performance of two state-of-the-art QE models, i.e., Transquest and CometKiwi, as baselines on our dataset. Furthermore, our experiments show that using a small subset of the proposed dataset containing 300 sentences to fine-tune Transquest, can improve its performance by more that 8% in terms of the Pearson correlation with a held-out test set.

Keywords: Quality Estimation, Machine Translation, Direct Assessment, English–Persian dataset

1. Introduction

Translation quality estimation (QE) is the task of evaluating a translation system's quality without the need of gold reference translations (Blatz et al., 2004; Specia et al., 2018b). It has many applications in computer-aided translation systems and other real-time scenarios, where we do not have reference translations and we need to decide if the MT output is reliable enough or to flag sentences with unreliable translations or potential critical mistakes. It can also be used to indicate the sentences that need human revision and estimate the effort needed for it.

The WMT shared task on QE has been conducted each year since 2012, to facilitate and encourage the research on QE tasks (Zerva et al., 2022). One of the key components which facilitates the research in this area is the available QE datasets. In sentence level QE, these datasets consist of source sentences, their automatic translations, and the gold QE labels for those translations. As a result of the WMT QE shared tasks, several QE datasets in various language pairs have been created during these years (Bojar et al., 2014; Specia et al., 2020, 2021; Zerva et al., 2022). Fomicheva et al. (2022) also created and published a multilingual QE dataset for 11 language pairs. However, these datasets are still only available for a limited number of language pairs, and many language pairs do not even have a small test set with QE labels yet, making it impossible to study QE for them. English-Persian is one of those low-resource language pairs, which we target in this paper.

published a QE test dataset for English to Persian translation with 1000 sentences, their MT outputs, and human post-edited translations, while computing the *Human-mediated Translation Edit Rate* (HTER) (Snover et al., 2006) metric as the QE label. This dataset was also used in the word level subtask of WMT23 QE shared task(Blain et al., 2023)¹. Although using HTER is a common way to assess the performance of QE systems, it has been shown to be unreliable for this purpose in Graham et al. (2016). They proposed using *Direct Assessment* (DA), where the translation is scored on a 0 to 100 scale by a human, as a more reliable alternative to HTER.

In this paper, we prepare DA labels for the aforementioned English-Persian test dataset, to present a more reliable QE test set for English-Persian language pair. Our dataset, named EPOQUE, contains the 1000 sentence dataset from Azadi et al. (2023), while providing three DA annotations for each sentence and its translation. We also assess the performance of two well-known QE models, Transquest and CometKiwi, as baselines on our dataset. Furthermore, by considering a small portion of this dataset to fine-tune the Transquest model, we show that although this baseline is multilingual, it is not sufficiently trained for English-Persian and adding a very small-scale training dataset can significantly improve its performance. EPOQUE is publicly available². In the rest of this paper, we first briefly discuss the related work in Section 2. Next in Section 3, we intro-

²https://huggingface.co/datasets/ universitytehran/EPOQUE

¹https://wmt-qe-task.github.io/

duce our dataset, while describing its preparation process and statistics. Finally, we present our experiments and discussing the results in Section 4.

2. Related Work

In this section, we briefly discuss some of the previous work done on creating datasets for estimating the quality of machine translation in different language pairs. One of the very first QE datasets was published by Specia et al. (2010) for English-Spanish language pair. It consists of 4,000 English sentences, their corresponding reference translations, and four alternative translations for each of them into Spanish, produced by four statistical MT systems, amounting to 16,000 source-targetreference triples. To get QE labels, these translations were given to professional translators to assign a score from 1 to 4 based on the amount of post-editing needed to make them ready for publishing. In another work, Fujita and Sumita (2017) prepared a dataset for Japanese to English, Chinese and Korean translation QE and automatic post-editing. It contains more than 10,000 segments for each language pair with their post-edited translations. For QE labels, they used both HTER labels, and manual grades from human evaluators based to their 5-level grading criterion. Fomicheva et al. (2020) also presented a QE dataset for 6 language pairs including high-, medium-, and lowresource NMT training.³ It consists of 10,000 segments per language pairs, annotated with the DA methodology (Graham et al., 2017) on a 0 to 100 scale. In a more recent work Fomicheva et al. (2022) provided a multilingual QE dataset of 11 language pairs, adding adding 5 other new language pairs to the previous dataset.⁴ Each language pair has at least 1000 sentences for testing QE models, with both DA and HTER tags. Apart from these, there are also many other publicly available datasets, prepared and used for the WMT shared task on QE (Bojar et al., 2014; Fonseca et al., 2019; Specia et al., 2018a, 2020, 2021).

Recently, Azadi et al. (2023) have created a QE test dataset for English to Persian translation containing 1000 sentences with their HTER tags, which was used for the word level subtask of WMT23 QE shared task. As the current practice in QE is using the DA labels, which is shown to be more reliable than HTER ones (Graham et al., 2016), in this paper we provide DA annotations for this dataset, which leads to the first dataset with DA labels for English–Persian language pair.

3. Data Collection and Statistics

In this section, we present our English–Persian dataset with DA annotations and describe its preparation process, while discussing some statistics and analysis about it.

3.1. Data Collection

To prepare our dataset, we use English sentences and their Persian MT outputs from the dataset presented in Azadi et al. (2023)⁵, and provide DA annotations for them. As mentioned in their paper, this dataset consists of 1000 English sentences derived from a collection of scientific papers, which are translated with an RNN based encoder-decoder commercial MT system, named Faraazin⁶. More details about this MT system are described in Azadi et al. (2023). They also provided a human post-edited translation, and computed the HTER score as QE label for each translated sentence. Instead, we have given each sentence pair to three human translators to annotate DA scores for each translation.

For DA annotations, we follow the guidelines from Fomicheva et al. (2022) and Guzmán et al. (2019), inspired by the work of Graham et al. (2013). Thus, we get 3 scores from 0-100 for each sentence pair, according to its perceived translation quality. Our detailed annotation guidelines, given to the human annotators, are presented in Table 4 in the Appendix. We finally take an average across the scores from individual annotators, to produce a single QE label for each translation. We also convert raw scores into z-scores, that is, standardizing according to each individual annotator's overall mean and standard deviation. All these scores are included in our published dataset.

3.2. Statistics and Analysis

Statistics of the prepared dataset including the number of annotated sentences, as well as the number of the source and target tokens in the test set are shown in Table 1. To assure the consistency among three collected annotations for each translated sentence, we measured the agreement among annotators using Cohen's kappa coefficient (Cohen, 1960), following Specia et al. (2010). The kappa score obtained was 0.61, which is substantial according to Landis and Koch (1977).

Figure 1 shows the distribution of our DA scores in comparison to existing HTER ones, as well as the scatter plot of DA against HTER scores. The DA distribution shows that most of the translations

³The language pairs include En-De, En-Zh, Ro-En, Et-En, Si-En and Ne-En

⁴The new language pairs include Ru-En, Ps-En, Km-En, En-Ja, and En-Cs

⁵https://github.com/fatemeh-azadi/ Unsupervised-QE



Figure 1: Distribution of DA scores, HTER scores and their scatter plot for the English–Persian dataset

Sentences English Tokens		Persian Tokens		
1000	26,739	26,470		

Table 1: Statistics of the English-Persian dataset

received a high score, i.e., they have high quality, while there are some representatives of low-quality translations, which make this dataset inclusive.

To better compare DA scores with HTERs, the average values for each score, in addition to their correlations, are shown in Table 2. The DA vs. HTER scatter plot, as well as Table 2, shows that although these scores are weakly correlated and may consistently assess the overall translation quality of the whole dataset, they evaluate many of the translated sentences quite differently. As pointed in Fomicheva et al. (2022), and we acknowledged by observing such examples in our dataset, this is because HTER is less sensitive to some serious errors like mistranslated words, which may affect the overall meaning of the sentence, and more sensitive to the translation fluency, while DA is the opposite. Some examples from our dataset where DA and HTER indicated different translation gualities are shown in Table 5 in the appendices

4. Experiments

Using the created dataset, we conduct our experiments to evaluate the effectiveness of some existing state-of-the-art well-known QE models, as baselines for English to Persian translation quality estimation. For this purpose, we consider two open-source QE toolkits, namely Transquest⁷ (Ranasinghe et al., 2020) and CometKiwi⁸ (Rei et al., 2022), and evaluate them on our dataset. Furthermore, by considering a small portion of our dataset for fine-tuning Transquest, we discuss how adding English–Persian training data could benefit this model.

4.1. Performance of Baseline Models

In this section, we assess the performance of two of the most well-known QE toolkits on our dataset, and report them as baseline results for the task of direct assessment of the English–Persian translation. Here, we use all 1000 sentences in our dataset for testing. We use the publicly available models in the Hugging Face platform (Wolf et al., 2020), which are trained multilingually for the DA task, both for Transquest⁹ and CometKiwi¹⁰. The Pearson and Spearman correlations of these models with our DA labels are shown in Table 3, which indicates the more recent CometKiwi model performs much better than Transquest.

4.2. Fine-tuning Transquest with a Tiny Dataset

The baseline models used in the previous section, as well as none of the other existing QE models, have not used any English–Persian QE dataset during their training, as no training dataset is available for this language pair. In this section, we try to consider a small portion of our dataset for training, and investigate the effect of adding a small-scale English–Persian training data to Transquest on its performance.

As we need a sufficient amount of test data to obtain reliable results in our experiments, we first evaluate the baseline Transquest model on different sizes of test data in terms of the Pearson correlation with DA labels. The results are shown in Figure 2, which indicates that the Pearson correlation becomes relatively smooth and converges for test sizes above 650 sentences. Consequently, it can be said that holding 700 sentences for testing gives us reliable results close to the results of the whole dataset. Thus, we randomly sample 700

⁷https://github.com/TharinduDR/ TransQuest

⁸https://github.com/Unbabel/COMET

⁹https://huggingface.co/TransQuest/ monotransquest-da-multilingual ¹⁰https://huggingface.co/Unbabel/ wmt22-cometkiwi-da

Average DA	Average HTER	Correlation between DA and HTER		
		Pearson	Spearman	
81.09	0.29	-0.53	-0.56	

Table 2: Average DA scores and HTER scores, along with the Pearson and Spearman correlations between DA and HTER scores

Model	Pearson	Spearman
Transquest	0.49	0.53
CometKiwi	0.66	0.69

Table 3: Pearson and Spearman correlations of baseline models

sentences from our dataset, hold them as a test set and consider the remaining 300 sentences for training in the rest of our experiments.



Figure 2: Pearson correlation of TransQuest model for various test set sizes

We perform three experiments on fine-tuning Tranquest with different training set sizes, including 100, 200, and 300 samples from the part of the dataset reserved for training. We fine-tune the Transquest model on each of these training sets with the batch size of 10, the learning rate of 2×10^{-5} and its default values for other hyperparameters, for 10 epochs. The Pearson correlations are illustrated in Figure 3, for each training size on each epoch of the fine-tuning, using the 700-sentence test set. The model at the 0th epoch is the base Transquest model, without any finetuning. Figure 3 shows that although using 100 sentences for training cannot improve the baseline's performance; utilizing a tiny training set of 200 and 300 sentences can improve it significantly.

In another experiment, we fine-tune Transquest with 300 sentences for 5 different random seeds to reassure the reliability of the previous experiments.



Figure 3: Pearson correlation of the fine-tuned model for various training data sizes and epochs

We then compute the average and standard deviation of Pearson correlations for these 5 runs at each epoch, and plot them in Figure 4. This shows that using a tiny set of 300 English–Persian sentences to fine-tune Transquest, can consistently enhance its performance. The average improvement in Pearson correlation with DA labels is about 8%, which is a remarkable improvement.



Figure 4: Pearson Correlation of the fine-tuned model for various random seeds and epochs

5. Conclusion

In this paper, we introduced EPOQUE, the first English–Persian dataset with direct assessment annotations for translation quality estimation. EPOQUE can be a complementary to the previous publicly available dataset with HTER tags published in (Azadi et al., 2023), while providing DA tags as more reliable QE labels. This dataset contains 1000 sentences, and can be used as a test set for zero-shot QE or other scenarios. EPOQUE is publicly available, and can encourage further work on QE for this low-resource language pair. We also assess two of the recent state-of-the-art QE models on our dataset as baselines, while showing that adding a very small-scale English–Persian training data to them can make improvements.

References

- Fatemeh Azadi, Heshaam Faili, and Mohammad Javad Dousti. 2023. Mismatchingaware unsupervised translation quality estimation for low-resource languages. *arXiv preprint arXiv:2208.00463, to appear in Language Resources and Evaluation.*
- Frederic Blain, Chrysoula Zerva, Ricardo Rei, Nuno M. Guerreiro, Diptesh Kanojia, José G. C. de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, Constantin Orasan, and André Martins. 2023. Findings of the WMT 2023 shared task on quality estimation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 629–653, Singapore. Association for Computational Linguistics.
- 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*, pages 315–321.
- Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, et al. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the ninth workshop on statistical machine translation*, pages 12–58.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.
- Marina Fomicheva, Shuo Sun, Erick Fonseca, Chrysoula Zerva, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia

Specia, and André F. T. Martins. 2022. MLQE-PE: A multilingual quality estimation and postediting dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4963–4974, Marseille, France. European Language Resources Association.

- 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. *Transactions of the Association for Computational Linguistics*, 8:539–555.
- Erick Fonseca, Lisa Yankovskaya, André FT Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the wmt 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10.
- Atsushi Fujita and Eiichiro Sumita. 2017. Japanese to English/Chinese/Korean datasets for translation quality estimation and automatic post-editing. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 79–88, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Yvette Graham, Timothy Baldwin, Meghan Dowling, Maria Eskevich, Teresa Lynn, and Lamia Tounsi. 2016. Is all that glitters in machine translation quality estimation really gold? In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3124–3134, Osaka, Japan. The COLING 2016 Organizing Committee.
- Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.
- 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):3–30.
- Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali– English and Sinhala–English. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural

Language Processing (EMNLP-IJCNLP), pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

- J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.
- Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020. TransQuest: Translation quality estimation with cross-lingual transformers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5070–5081, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: ISTunbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings 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.
- Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. 2020. Findings of the WMT 2020 shared task on quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 743–764, Online. Association for Computational Linguistics.
- Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André FT Martins. 2021. Findings of the wmt 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725.
- Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón F. Astudillo, and André F. T. Martins. 2018a. Findings of the WMT 2018 shared task on quality estimation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 689–709, Belgium, Brussels. Association for Computational Linguistics.

- Lucia Specia, Nicola Cancedda, and Marc Dymetman. 2010. A dataset for assessing machine translation evaluation metrics. In *Proceedings* of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta. European Language Resources Association (ELRA).
- Lucia Specia, Carolina Scarton, Gustavo Henrique Paetzold, and Graeme Hirst. 2018b. *Quality estimation for machine translation*, volume 11. Springer.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José GC De Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orašan, Marina Fomicheva, et al. 2022. Findings of the wmt 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation* (*WMT*), pages 69–99.

A. Annotation Guideline

We present the annotation guideline for DA labels in Table 4. This guideline was given to three annotators to provide DA labels, which is a score from 0 to 100 for each translation pair.

B. Samples for DA vs. HTER Scores

Table 5 shows some examples from our dataset, where DA and HTER indicated different translation qualities. The first example was assigned a high DA score as the MT output preserve the meaning of the source sentence, but its HTER score is also high because the MT outputs is not fluent enough and needs substantial changes during post-editing. Conversely, in the second example, DA score is low as the MT output does not contain the translation of the word "its", which made a serious change in the meaning of the sentence. However, the sentence is easy to post-edit by only adding two words, resulting in a low HTER score.

Range type Definition		Score range
Prefect translation	The translation is completely correct and in terms of meaning, it fully expresses the meaning of the English sentence and has no errors.	90-100
Good translation	The translation of the English sentence is clear and has no grammatical problems and reads like a normal text. The translation is very good and very close to the perfect translation of the English sentence. There is no error, but better words can be used in Persian translation.	70-89
Medium translation	The translation is understandable and conveys the general meaning of the English sentence, but there are some grammatical or lexical translation errors in it.	50-69
Bad translation	The translated sentence conveys parts of the mean- ing of the English sentence, but it is difficult to get the general meaning of the sentence due to the big errors in the translation.	30-49
Very bad translation	The translation contains a few correctly translated words, but it is impossible to understand or its mean- ing is very different from the English sentence.	10-29
Completely wrong translation	The translated sentence does not convey the mean- ing of any part of the English sentence, and it is com- pletely irrelevant or impossible to understand.	0-9

Table 4: Direct assessment guideline

		Text	DA	HTER
	Source	However, in reality, different users of the network have dif- ferent incentive demands.		
Sample 1	MT	با این حال, در واقعیت, کاربران مختلف این شبکه نیاز به درخواستهای تشویقی مختلف دارند.	97	0.5
	Post-Edit	با این حال، در واقعیت، کاربران مختلف شبکه خواستههای انگیزشی متفاوتی دارند.		
	Source	It is commonly used for its easy of interpretation and low calculation time.		
Sample 2	MT	معمولاً برای تفسیر آسان و زمان محاسبه پایین استفاده می شود.	64.66	0.15
	Post-Edit	معمولاً برای تفسیر آسان و زمان محاسبه کم از آن استفاده می شود.		

Table 5: Two samples from EPOQUE dataset where DA and HTER indicated different translation qualities