University of Tehran

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

Introduction

- Translation quality estimation (QE)
- Applications
- WMT Shared Task
- QE datasets
- Direct assessment
- EPOQUE → DA labels for an English–Persian test dataset of 1000 sentences

Related Work

- English to Spanish translation (Specia et al. (2010))
- Japanese to English, Chinese and Korean translation (Fujita and Sumita (2017))
- QE dataset for 6 language pairs including high-, medium-, and low-resource NMT training (Fomicheva et al. (2020))
- A multilingual QE dataset of 11 language pairs (Fomicheva et al. (2022))
- English to Persian translation (Azadi et al. (2023))

Data Collection

- We Use English sentences and their Persian MT outputs from the dataset presented in Azadi et al. (2023) → 1000 English sentences
- Each sentence pair was given to three human annotators
- 3 DA scores from 0-100 obtained for each sentence pair

Guideline

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 meaning 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 meaning is very different from the English sentence.	10-29
Completely wrong translation	The translated sentence does not convey the meaning of any part of the English sentence, and it is completely irrelevant or impossible to understand.	0-9

Example

		Text	DA	HTER
Sample 1	Source	However, in reality, different users of the network have different incentive demands.		
	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.		
Sample 2	Source	It is commonly used for its easy of interpretation and low calculation time.		
	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.		

Statistics and Analysis

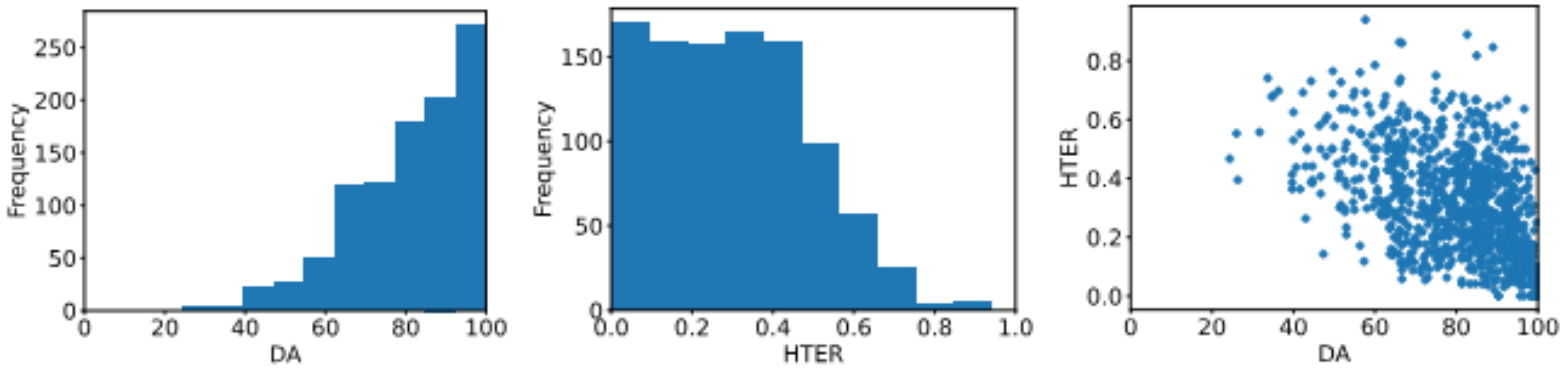

Figure 1: Distribution of DA scores, HTER scores and their scatter plot for the English–Persian dataset

Statistics and Analysis

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

Experiments

Model	Pearson	Spearman
Transquest	0.49	0.53
CometKiwi	0.66	0.69

Table 3: Pearson and Spearman correlations of baseline models

Experiments

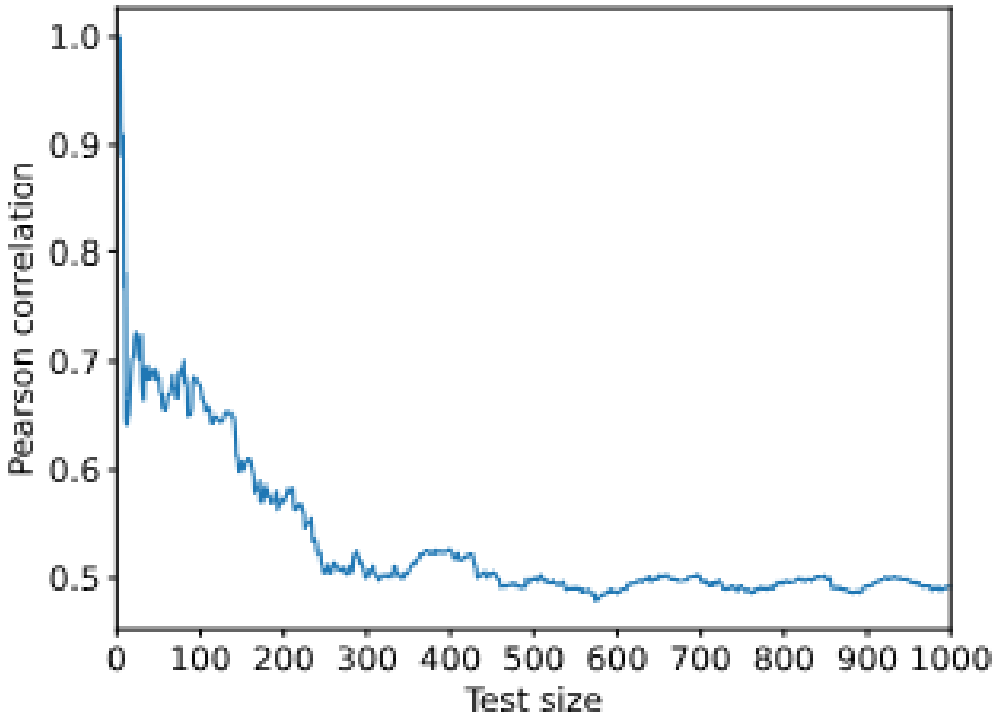

Figure 2: Pearson correlation of TransQuest model for various test set sizes

Experiments

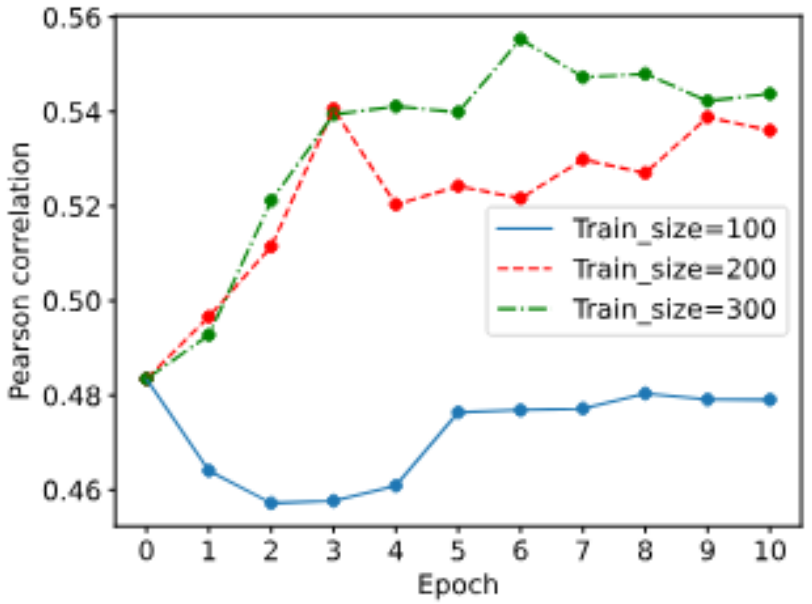

Figure 3: Pearson correlation of the fine-tuned model for various training data sizes and epochs

Conclusion

- EPOQUE: An English-Persian Quality Estimation Dataset
- DA annotations
- Test set for zero-shot QE
- Publicly available
- Improving current QE models by adding a very small-scale training data

Thanks for your Attention