www.inesc-id.pt

SIGdial 2023

Towards Multilingual Automatic Open-Domain Dialogue Evaluation

John Mendonça, Alon Lavie and Isabel Trancoso

How can we make Dialogue Evaluation metrics multilingual?

A) Finetune multilingual pretrained model → Strong baseline!

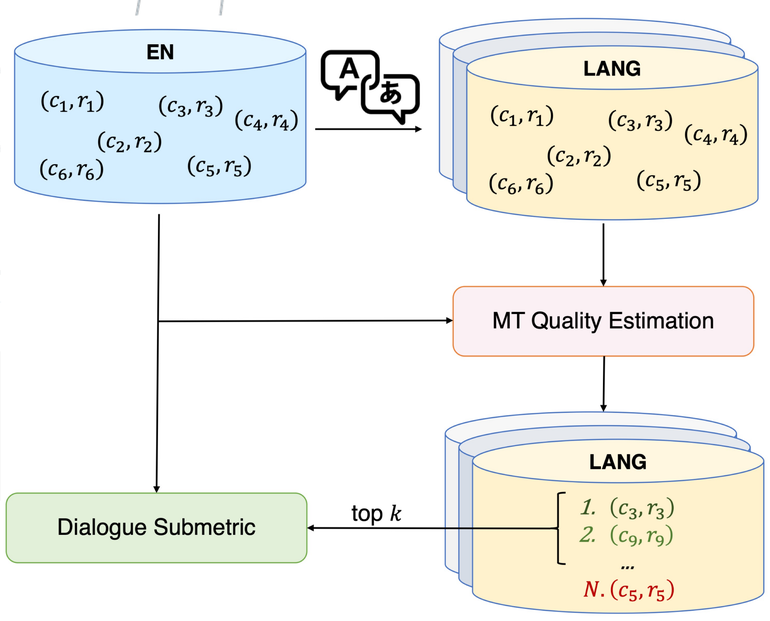
B) Translate everything and finetune → Performs worse

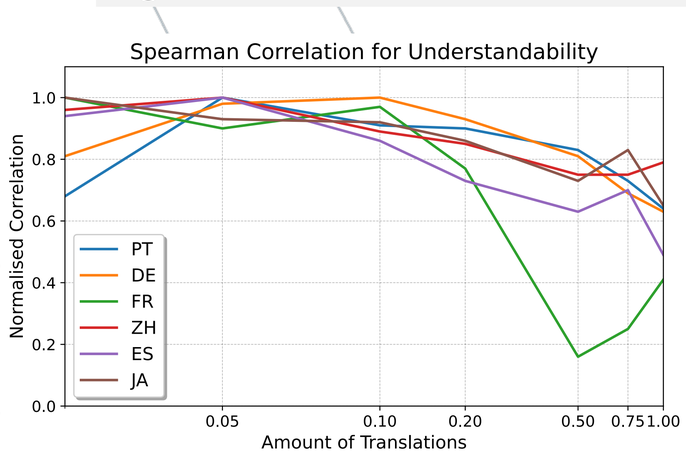
C) Prompt SoTA LLMs → Expensive

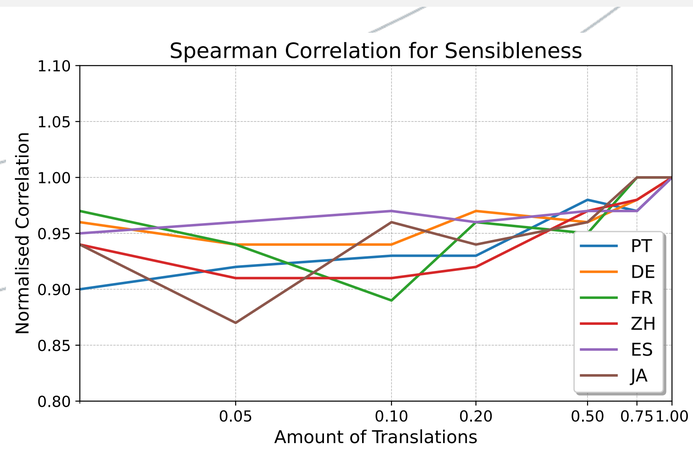
D) Finetune using only the best translations →

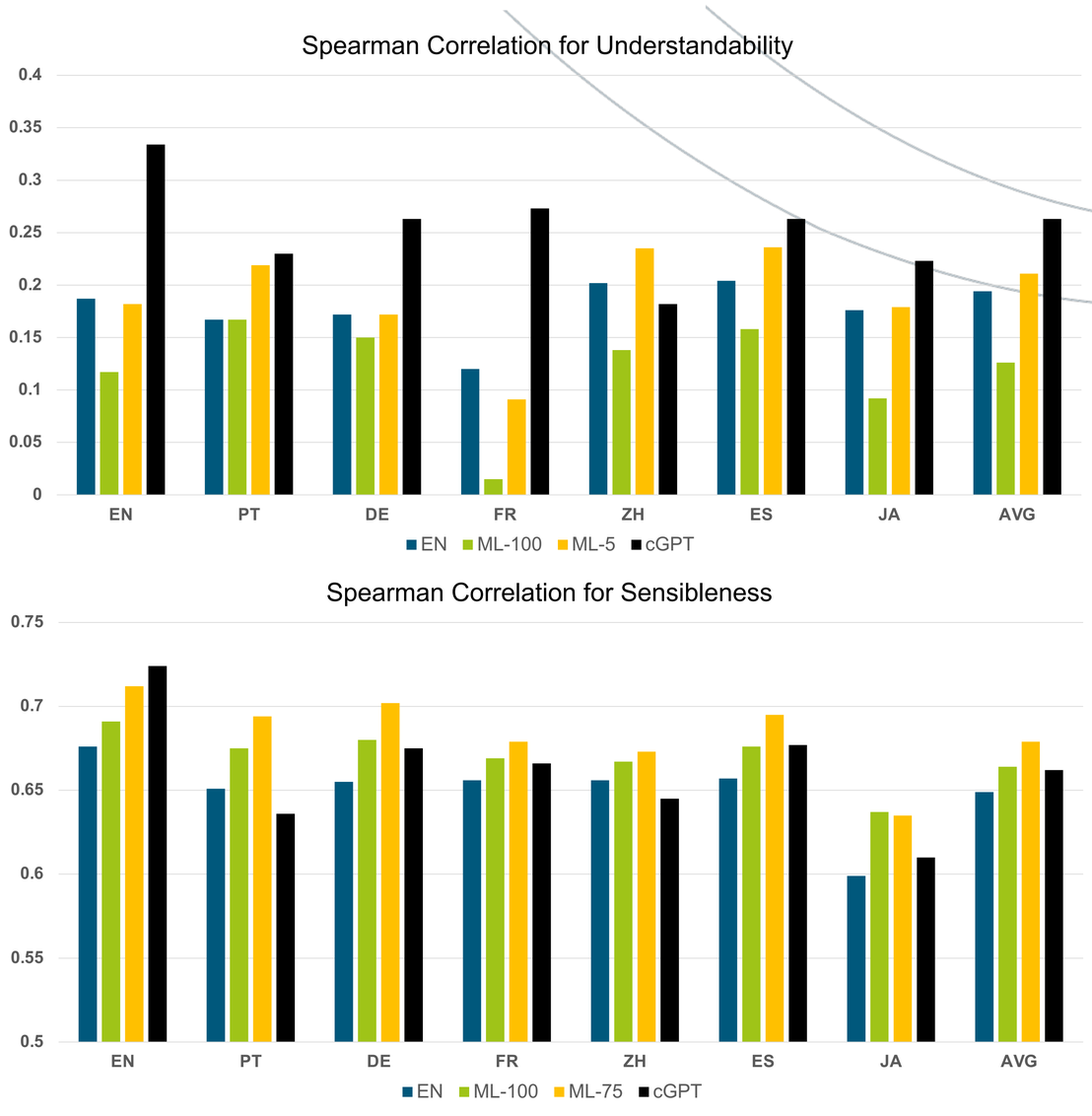
*We found that, depending on the subquality and target language, the optimal amount of translated data for multilingual models can be as **low as 5% and as high as 75%.***

Our solution:

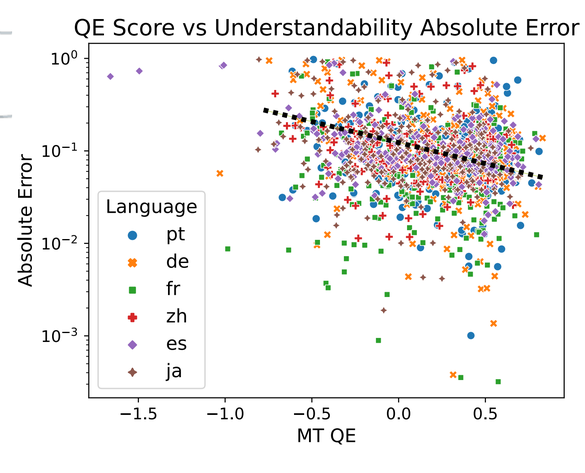






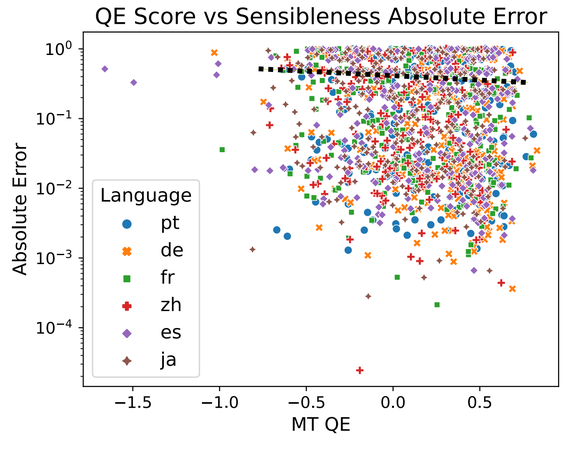


Does translation quality affect quality predictions?



Negatively correlated

*This model is highly sensitive to low quality translations, since **MT can affect the fluency of the response**.*



Weakly correlated

*This model showed robustness when including more translations during training as **MT typically translates context dependent keywords correctly.***

Acknowledgments:
This research was supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Responsible.AI), and by national funds through Fundação para a Ciência e a Tecnologia (FCT) with references PRT/BD/152198/2021 and UIDB/50021/2020, and by the P2020 program MAIA (LISBOA-01-0247-FEDER-045909).

Co-funded by the European Union