BitextEdit: Automatic Bitext Editing for Improved Low-Resource Machine Translation

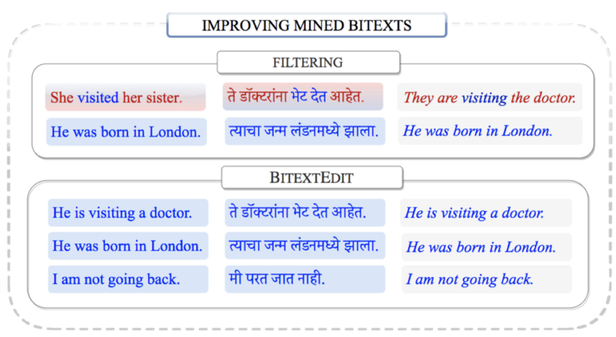
Eleftheria Briakou, Sida I. Wang, Luke Zettlemoyer, Marjan Ghazvininejad

ebriakou@umd.edu, sida@fb.com, lsz@fb.com, ghazvi@fb.com

Problem Definition

- ➡ Mined bitexts contain imperfect or noisy translations [1] [2]
- ➡ Bitext Quality Matters for Neural Machine Translation [3] [4]
- ➡ Bitext Filtering improves final model quality [5] *but it is suboptimal in low-resource conditions where data are limited*
- ➡ What can we do? **We propose to refine the mined bitexts via automatic editing!**

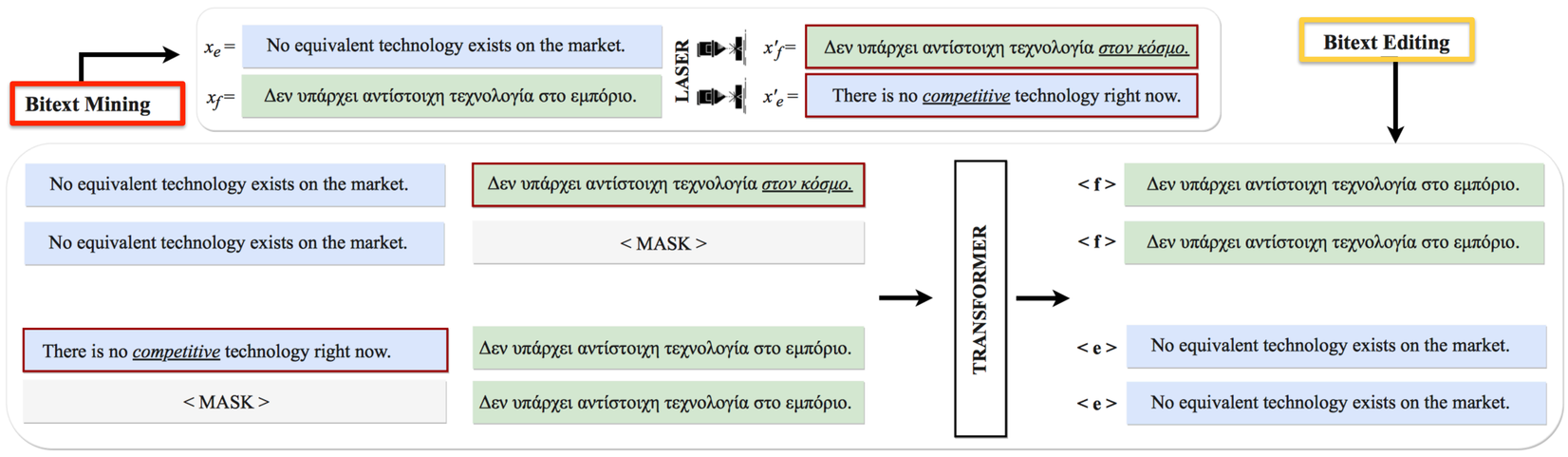
Noisy bitexts consist of a mixture of good-quality, imperfect, and poor-quality translations



Filtering decreases the size of training samples which is crucial for low-resource NMT.

BitextEdit revises noisy bitexts via utilizing imperfect translations in a more effective way, while keeps the size of training data untouched.

BitextEdit Training Strategy



Our **multi-task** model is trained using synthetic supervision from mined bitexts.

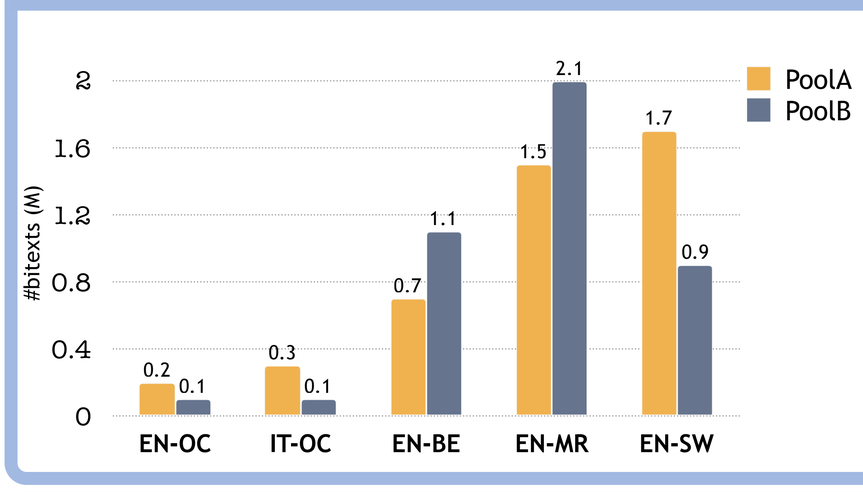
Bitext Mining

Starting from an original bitext (x_e, x_f), we mine imperfect translations x'_f and x'_e for each reference using LASER [6].

Bitext Editing

A sequence-to-sequence Transformer model is trained to *translate* and *reconstruct* the original references given synthetically extracted bitexts representing imperfect translations.

Experimental Settings



Bitexts that pass through the noisy filtering step

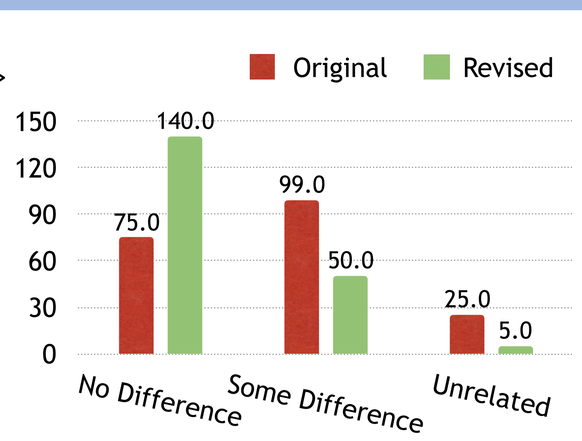
Bitexts that are filtered based on the filtering step

- ✓ **Training data:**
  - CCMatrix Bitexts [7]
  - Revise Pool B
- ✓ **Model:**
  - Transformer NMT
- ✓ **Evaluation:**
  - BLEU (spm-BLEU)
  - FlORES test set

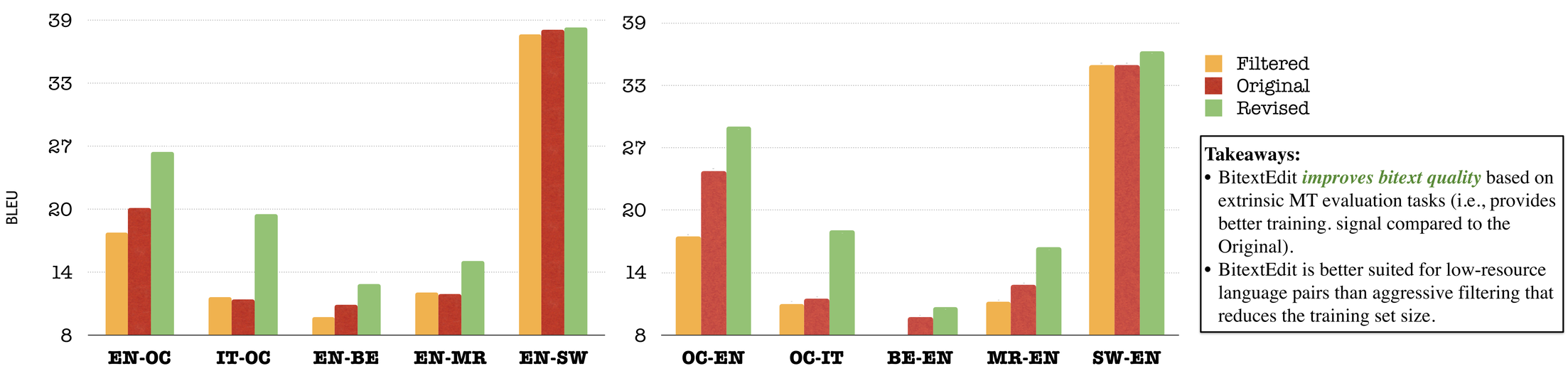
Intrinsic Evaluation Results

Number of bitexts manually rated as perfect translations (i.e., No difference), partial translations (i.e., some meaning difference), and wrong translations (i.e., unrelated) for a random sample of original vs. refined CCMatrix EN-EL data.

Takeaway: BitextEdit *improves bitext quality* based on intrinsic evaluation results.



Extrinsic Evaluation Results



Takeaways:

- BitextEdit *improves bitext quality* based on extrinsic MT evaluation tasks (i.e., provides better training. signal compared to the Original).
- BitextEdit is better suited for low-resource language pairs than aggressive filtering that reduces the training set size.

References

[1] Julia Kreutzer et al. 2022. Quality at a Glance:An Audit of Web-Crawled Multilingual Datasets. In ACL

[2] Eleftheria Briakou, Marine Carpuat. 2020. Detecting Fine-Grained Cross-Lingual Semantic Divergences without Supervision by Learning to Rank.In EMNLP

[3] Huda Khayrallah, Philipp Koehn. 2018 On the Impact of Various Types of Noise on Neural Machine Translation. In WNLG

[4] Eleftheria Briakou, Marine Carpuat. 2021. Beyond Noise: Mitigating the Impact of Fine-grained Semantic Divergences on Neural Machine Translation. In ACL

[5] Philipp Koehn, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco Guzmán. 2020. Findings of the WMT 2020 shared task on parallel corpus filtering and alignment. In WMT

[6] Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero- shot cross-lingual transfer and beyond. IN TACL

[7] Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. CCMatrix: Mining billions of high-quality parallel sentences on the web. In ACL

Conclusion

We introduce **BitextEdit**—an ***editing approach for bitext*** quality improvement that we show is better suited for low-resource language pairs. Those findings highlight the importance of the good *quality* bitexts in scenarios where large *quantities* cannot be guaranteed and motivate future research on improving low-resource NMT further.