# 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, ghazvini@fb.com

# Bitexts are not Always Parallel

| | |
|---|---|
| He was born in London. | त्याचा जन्म लंडनमध्ये झाला. |

GLOSS: He was born in London.

# Bitexts are not Always Parallel

| | |
|---|---|
| She visited her sister. | ते डॉक्टरांना भेट देत आहेत. |
| | GLOSS: They are visiting the doctor. |
| He was born in London. | त्याचा जन्म लंडनमध्ये झाला. |
| | GLOSS: He was born in London. |

# Bitexts are not Always Parallel

| | |
|---|---|
| She visited her sister. | ते डॉक्टरांना भेट देत आहेत. |
| | GLOSS: They are visiting the doctor. |
| He was born in London. | त्याचा जन्म लंडनमध्ये झाला. |
| | GLOSS: He was born in London. |
| I am not going back. | मांजर तिचे अन्न खात आहे. |
| | GLOSS: The cat is eating her food. |

# Audits of Mined Bitext Reveal Systematic Quality Issues

**Quality at a Glance:**
**An Audit of Web-Crawled Multilingual Datasets**

**Julia Kreutzer[a,b], Isaac Caswell[a], Lisa Wang[a], Ahsan Wahab[c], Daan van Esch[a], Nasanbayar Ulzii-Orshikh[d], Allahsera Tapo[b,e], Nishant Subramani[b,δ], Artem Sokolov[a], Claytone Sikasote[b,g], Monang Setyawan[h], Supheakmungkol Sarin[h], Sokhar Samb[b,i], Benoît Sagot[j], Clara Rivera[a], Annette Rios[k], Isabel Papadimitriou[l], Salomey Osei[b,m], Pedro Ortiz Suarez[j,n], Iroro Orife[b,o], Kelechi Ogueji[b,p], Andre Niyongabo Rubungo[b,q], Toan Q. Nguyen[r], Mathias Müller[k], André Müller[k], Shamsuddeen Hassan Muhammad[b,s], Nanda Muhammad[h], Ayanda Mnyakeni[h], Jamshidbek Mirzakhalov[c,t], Tapiwanashe Matangira[h], Colin Leong[b], Nze Lawson[h], Sneha Kudugunta[a], Yacine Jernite[b,u], Mathias Jenny[k], Orhan Firat[a,c], Bonaventure F. P. Dossou[b,v], Sakhile Dlamini[h], Nisansa de Silva[w], Sakine Çabuk Ballı[k], Stella Biderman[x], Alessia Battisti[k], Ahmed Baruwa[b,y], Ankur Bapna[a], Pallavi Baljekar[a], Israel Abebe Azime[b,i], Ayodele Awokoya[b,z], Duygu Ataman[c,k], Orevaoghene Ahia[b,α], Oghenefego Ahia[h], Sweta Agrawal[β], Mofetoluwa Adeyemi[b,γ],**

[a]Google Research, [b]Masakhane NLP, [c]Turkic Interlingua, [d]Haverford College, [e]RobotsMali, [f]Intel Labs, [g]University of Zambia, [h]Google, [i]AIMS-AMMI, [j]Inria, [k]University of Zurich, [l]Stanford University, [m]Kwame Nkrumah University of Science and Technology, [n]Sorbonne Université, [o]Niger-Volta LTI, [p]University of Waterloo [q]University of Electronic Science and Technology of China, [r]University of Notre Dame, [s]Bayero University Kano, [t]University of South Florida, [u]Hugging Face, [v]Jacobs University Bremen, [w]University of Moratuwa, [x]EleutherAI, [y]Obafemi Awolowo University, [z]University of Ibadan, [α]Instadeep, [β]University of Maryland, [γ]Defence Space Administration Abuja, [δ]Allen Institute for Artificial Intelligence

**Abstract**

With the success of large-scale pre-training and multilingual modeling in Natural Language Processing (NLP), recent years have seen a proliferation of large, web-mined text datasets covering hundreds of languages. We manually audit the quality of 205 language-specific corpora released with five major public datasets (CCAligned, ParaCrawl, WikiMatrix, OSCAR, mC4). Lower-resource corpora have systematic issues: At least 15 corpora have no usable text, and a significant fraction contains less than 50% sentences of acceptable quality. In addition, many are mislabeled or use non-standard/ambiguous language codes. We demonstrate that these issues are easy to detect even for non-proficient speakers, and supplement the human audit with automatic analyses. Finally, we recommend techniques to evaluate and improve multilingual corpora and discuss potential risks that come with low-quality data releases.

**1 Introduction**

Access to multilingual datasets for NLP research has vastly improved over the past years. A variety of web-derived collections for hundreds of languages is available for anyone to download, such as ParaCrawl (Esplà et al., 2019; Bañón et al., 2020), WikiMatrix (Schwenk et al., 2021) CCAligned (El-Kishky et al., 2020), OSCAR (Ortiz Suárez et al., 2019; Ortiz Suárez et al., 2020), and several others. These have in turn enabled a variety of highly multilingual models, like mT5 (Xue et al., 2021), M2M-100 (Fan et al., 2020), M4 (Arivazhagan et al., 2019).

Curating such datasets relies on the websites giving clues about the language of their contents (e.g. a language identifier in the URL) and on automatic language classification (LangID). It is commonly known that these automatically crawled and filtered datasets tend to have overall lower quality than hand-curated collec-

The vast majority of low-resource languages contain less than 50% valid translations.
[Kreutzer et al.]

**Detecting Fine-Grained Cross-Lingual Semantic Divergences**
**without Supervision by Learning to Rank**

**Eleftheria Briakou** and **Marine Carpuat**
Department of Computer Science
University of Maryland
College Park, MD 20742, USA
ebriakou@cs.umd.edu, marine@cs.umd.edu

**Abstract**

Detecting fine-grained differences in content conveyed in different languages matters for cross-lingual NLP and multilingual corpora analysis, but it is a challenging machine learning problem since annotation is expensive and hard to scale. This work improves the prediction and annotation of fine-grained semantic divergences. We introduce a training strategy for multilingual BERT models by learning to rank synthetic divergent examples of varying granularity. We evaluate our models on the **R**ationalized **E**nglish-**Fr**ench **S**emantic **D**ivergences, a new dataset released with this work, consisting of English-French sentence-pairs annotated with semantic divergence classes and token-level rationales. Learning to rank helps detect fine-grained sentence-level divergences more accurately than a strong sentence-level similarity model, while token-level predictions have the potential of further distinguishing between coarse and fine-grained divergences.

**1 Introduction**

Comparing and contrasting the meaning of text conveyed in different languages is a fundamental NLP task. It can be used to curate clean parallel corpora for downstream tasks such as machine translation (Koehn et al., 2018), cross-lingual transfer learning, or semantic modeling (Ganitkevitch et al., 2013; Conneau and Lample, 2019), and it is also useful to directly analyze multilingual corpora. For instance, detecting the commonalities and divergences between sentences drawn from English and French Wikipedia articles about the same topic would help analyze language bias (Bao et al., 2012; Massa and Scrinzi, 2012), or mitigate differences in coverage and usage across languages (Yeung et al., 2011; Wulczyn et al., 2016; Lemmerich et al., 2019). This requires not only detecting coarse content mismatches, but also fine-grained differences in sentences that overlap in content. Consider the following English and French sentences, sampled from the WikiMatrix parallel corpus. While they share important content, highlighted words convey meaning missing from the other language:

EN *Alexander Muir's "The Maple Leaf Forever"* ***served for many years*** *as an* ***unofficial*** *Canadian* ***national anthem****.*
FR *Alexander Muir compose The Maple Leaf Forever (en) qui est un* ***chant patriotique*** *pro canadien* ***anglais****.*
GLOSS *Alexander Muir composes The Maple Leaf Forever which is an English Canadian patriotic song.*

We show that explicitly considering diverse types of semantic divergences in bilingual text benefits both the annotation and prediction of cross-lingual semantic divergences. We create and release the **R**ationalized **E**nglish-**Fr**ench **S**emantic **D**ivergences corpus (REFRESD), based on a novel divergence annotation protocol that exploits rationales to improve annotator agreement. We introduce Divergent mBERT, a BERT-based model that detects fine-grained semantic divergences without supervision by learning to rank synthetic divergences of varying granularity. Experiments on REFRESD show that our model distinguishes semantically equivalent from divergent examples much better than a strong sentence similarity baseline and that unsupervised token-level divergence tagging offers promise to refine distinctions among divergent instances. We make our code and data publicly available.[1]

[1]Implementations of Divergent mBERT can be found at: https://github.com/Elbria/xling-SemDiv; the REFRESD dataset is hosted at: https://github.com/Elbria/xling-SemDiv/tree/master/REFreSD.

Only 36% of En-Fr WikiMatrix sample are exact translations
[Briakou & Carpuat]

**BitextEdit: Automatic Bitext Editing for Improved Low-Resource Machine Translation.**
**Eleftheria Briakou, Sida I. Wang, Luke Zettlemoyer and Marjan Ghazvininejad. NAACL Findings 2022.**

# Bitext Filtering as the Standard Approach to Bitext Quality Improvement

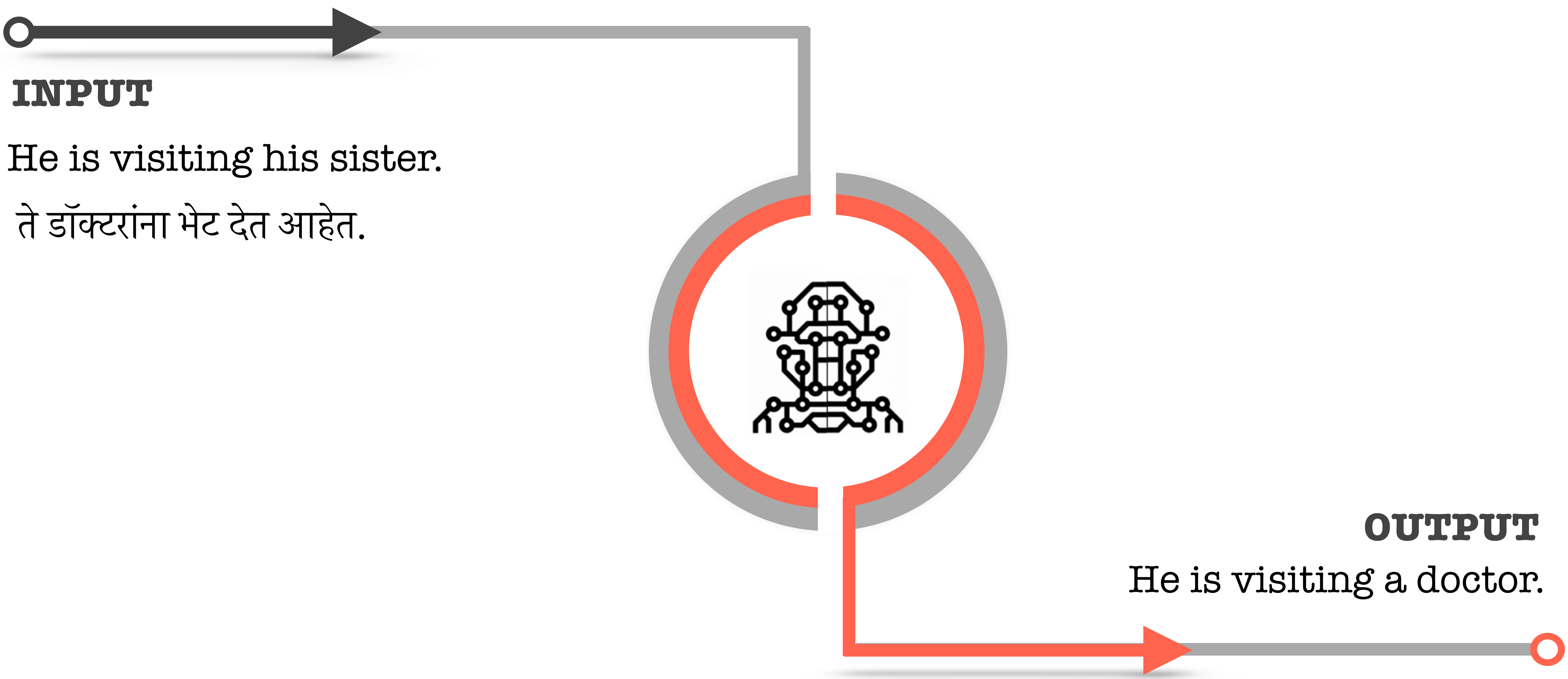
| | |
|---|---|
| She visited her sister. | ते डॉक्टरांना भेट देत आहेत. |
| He was born in London. | त्याचा जन्म लंडनमध्ये झाला. |
| I am not going back. | मांजर तिचे अन्न खात आहे. |

# Bitext Filtering as the Standard Approach to Bitext Quality Improvement

| | |
|---|---|
| She visited her sister. | ते डॉक्टरांना भेट देत आहेत. |
| He was born in London. | त्याचा जन्म लंडनमध्ये झाला. |
| I am not going back. | मांजर तिचे अन्न खात आहे. |

- What if we cannot afford filtering (e.g., low-resource)?
- How do we handle imperfect translations beyond noise?

# BitextEdit: Automatic Editing for Improved Bitext Quality




**BitextEdit: Automatic Bitext Editing for Improved Low-Resource Machine Translation.**
**Eleftheria Briakou, Sida I. Wang, Luke Zettlemoyer and Marjan Ghazvininejad. NAACL Findings 2022.**

# BitextEdit: Automatic Editing for Improved Bitext Quality

**INPUT**

He is visiting **his sister.**

ते डॉक्टरांना भेट देत आहेत.

Gloss He is visiting a doctor

**OUTPUT**

He is visiting **a doctor.**

**Edit as necessary…**

✓ Light editing
✓ Translate from scratch
✓ No editing (copy)

**BitextEdit: Automatic Bitext Editing for Improved Low-Resource Machine Translation.**
**Eleftheria Briakou, Sida I. Wang, Luke Zettlemoyer and Marjan Ghazvininejad. NAACL Findings 2022.**

# BitextEdit: Automatic Editing for Improved Bitext Quality

**INPUT**

I am not going back.

मांजर तिचे अन्न खात आहे.

Gloss The cat is eating her food.

**Edit as necessary...**

✓ Light editing
✓ Translate from scratch
✓ No editing (copy)

**OUTPUT**

मी परत जात नाही.

Gloss I am not going back.

**BitextEdit: Automatic Bitext Editing for Improved Low-Resource Machine Translation.**
**Eleftheria Briakou, Sida I. Wang, Luke Zettlemoyer and Marjan Ghazvininejad. NAACL Findings 2022.**

# BitextEdit: Automatic Editing for Improved Bitext Quality

**INPUT**

He was born in London.

त्याचा जन्म लंडनमध्ये झाला.

Gloss He was born in London.

**OUTPUT**

He has born in London.

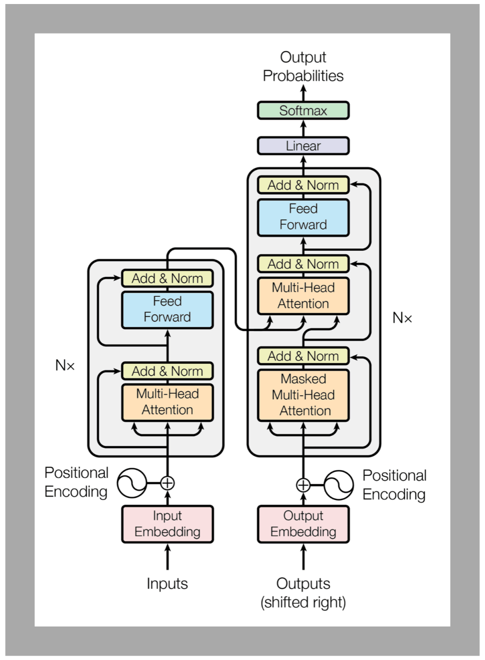
**Edit as necessary…**

- ✓ Light editing
- ✓ Translate from scratch
- ✓ No editing (copy)

**BitextEdit: Automatic Bitext Editing for Improved Low-Resource Machine Translation.**
**Eleftheria Briakou, Sida I. Wang, Luke Zettlemoyer and Marjan Ghazvininejad. NAACL Findings 2022.**

# BitextEdit: Learn to Reconstruct Original from Noisy References & Translate

Sequence-to-Sequence Transformer

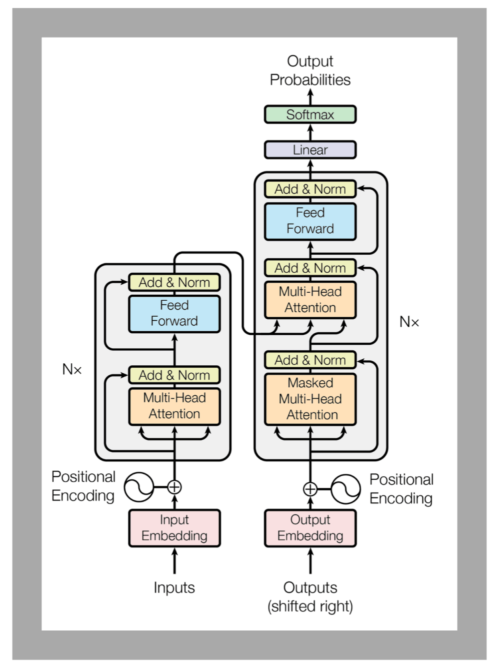


# BitextEdit: Learn to Reconstruct Original from Noisy References & Translate

Original reference given in parallel texts

Sequence-to-Sequence Transformer

$x^{(n)}$ ते डॉक्टरांना भेट देत आहेत.




$y^{(n)}$ He is visiting a doctor.


**BitextEdit: Automatic Bitext Editing for Improved Low-Resource Machine Translation.**
**Eleftheria Briakou, Sida I. Wang, Luke Zettlemoyer and Marjan Ghazvininejad. NAACL Findings 2022.**

# BitextEdit: Learn to Reconstruct Original from Noisy References & Translate

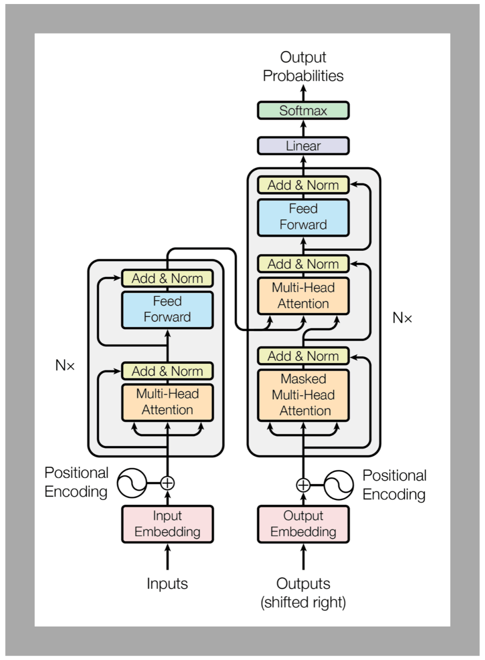
Sequence-to-Sequence Transformer

Original reference given in parallel texts

$x^{(n)}$ ते डॉक्टरांना भेट देत आहेत.

$\hat{y}^{(n)}$ He is visiting his sister.

Mined Reference extracted using LASER

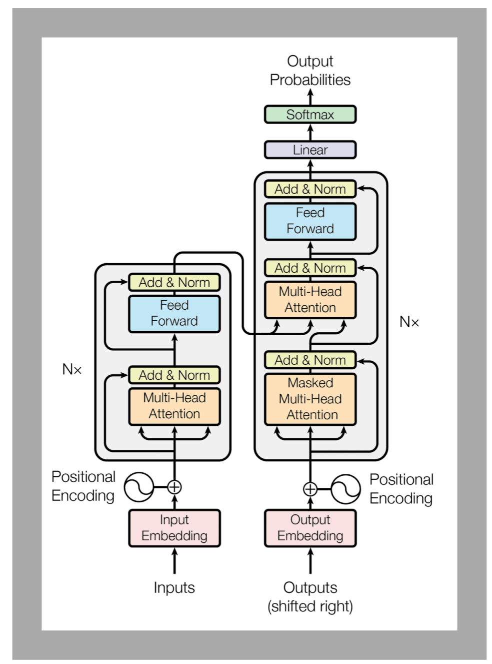



$y^{(n)}$ He is visiting a doctor.

BitextEdit: Automatic Bitext Editing for Improved Low-Resource Machine Translation.
Eleftheria Briakou, Sida I. Wang, Luke Zettlemoyer and Marjan Ghazvininejad. NAACL Findings 2022.

# BitextEdit: Learn to Reconstruct Original from Noisy References & Translate

Sequence-to-Sequence Transformer




Original reference given in parallel texts

$x^{(n)}$ ते डॉक्टरांना भेट देत आहेत.

$\hat{y}^{(n)}$ He is visiting his sister.

$y^{(n)}$ He is visiting a doctor.

Mined Reference extracted using LASER

$$\log \mathrm{p}\left( [ < e > y^{(n)} ] \mid \left( x^{(n)}, \hat{y}^{(n)} \right) \right) + \log \mathrm{p}\left( [ < e > y^{(n)} ] \mid \left( x^{(n)}, < \mathrm{MASK} > \right) \right)$$


**BitextEdit: Automatic Bitext Editing for Improved Low-Resource Machine Translation.**
**Eleftheria Briakou, Sida I. Wang, Luke Zettlemoyer and Marjan Ghazvininejad. NAACL Findings 2022.**

# BitextEdit: Learn to Reconstruct Original from Noisy References & Translate

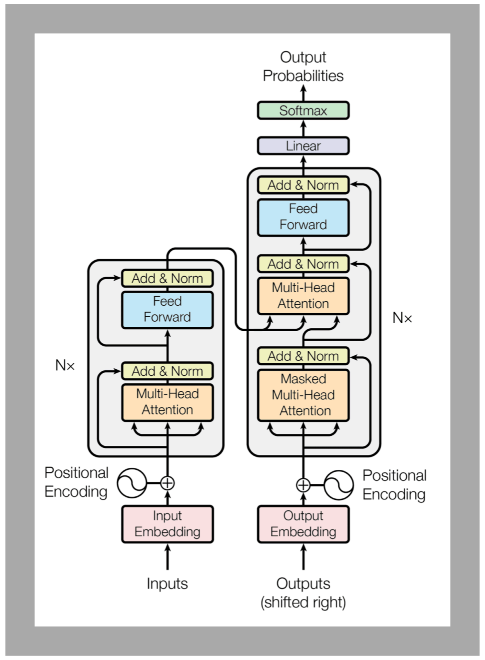
Sequence-to-Sequence Transformer

$x^{(n)}$ ते डॉक्टरांना भेट देत आहेत.

$\hat{y}^{(n)}$ He is visiting his sister.

$x^{(n)}$ ते डॉक्टरांना भेट देत आहेत.

<MASK>



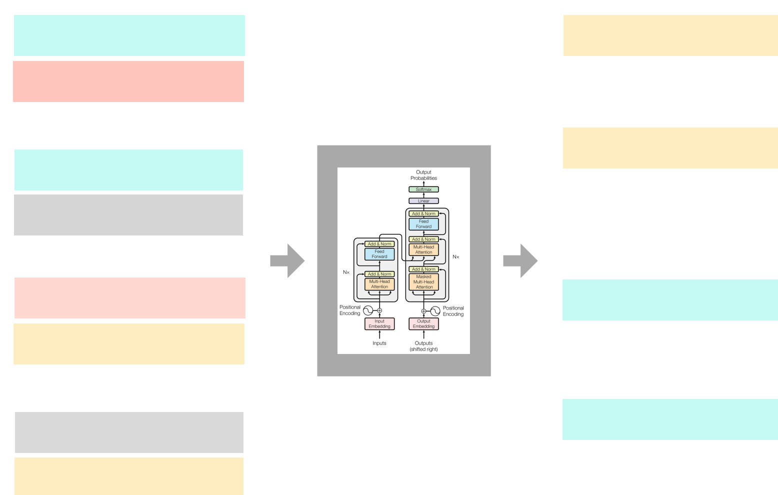

$y^{(n)}$ He is visiting a doctor.

$$\log \mathrm{p}\left(\left[<e>\, y^{(n)}\right] \mid \left(x^{(n)}, \hat{y}^{(n)}\right)\right) + \log \mathrm{p}\left(\left[<e>\, y^{(n)}\right] \mid \left(x^{(n)}, <\mathrm{MASK}>\right)\right)$$


**BitextEdit: Automatic Bitext Editing for Improved Low-Resource Machine Translation.**
**Eleftheria Briakou, Sida I. Wang, Luke Zettlemoyer and Marjan Ghazvininejad. NAACL Findings 2022.**

# BitextEdit: Learn to Reconstruct Original from Noisy References & Translate

Bi-directional Training



$$\log \mathrm{p}\left(\left[ < f > x^{(n)}\right] \mid \left(y^{(n)}, \hat{x}^{(n)}\right)\right) + \log \mathrm{p}\left(\left[ < f > x^{(n)}\right] \mid \left(y^{(n)}, < \mathrm{MASK} > \right)\right) +$$

$$\log \mathrm{p}\left(\left[ < e > y^{(n)}\right] \mid \left(x^{(n)}, \hat{y}^{(n)}\right)\right) + \log \mathrm{p}\left(\left[ < e > y^{(n)}\right] \mid \left(x^{(n)}, < \mathrm{MASK} > \right)\right)$$

BitextEdit: Automatic Bitext Editing for Improved Low-Resource Machine Translation.
Eleftheria Briakou, Sida I. Wang, Luke Zettlemoyer and Marjan Ghazvininejad. NAACL Findings 2022.

# BitextEdit: Learn to Reconstruct Original from Noisy References & Translate

Where does this supervision $(x^{(n)}, y^{(n)}, \hat{x}^{(n)}, \hat{y}^{(n)})$ come from?

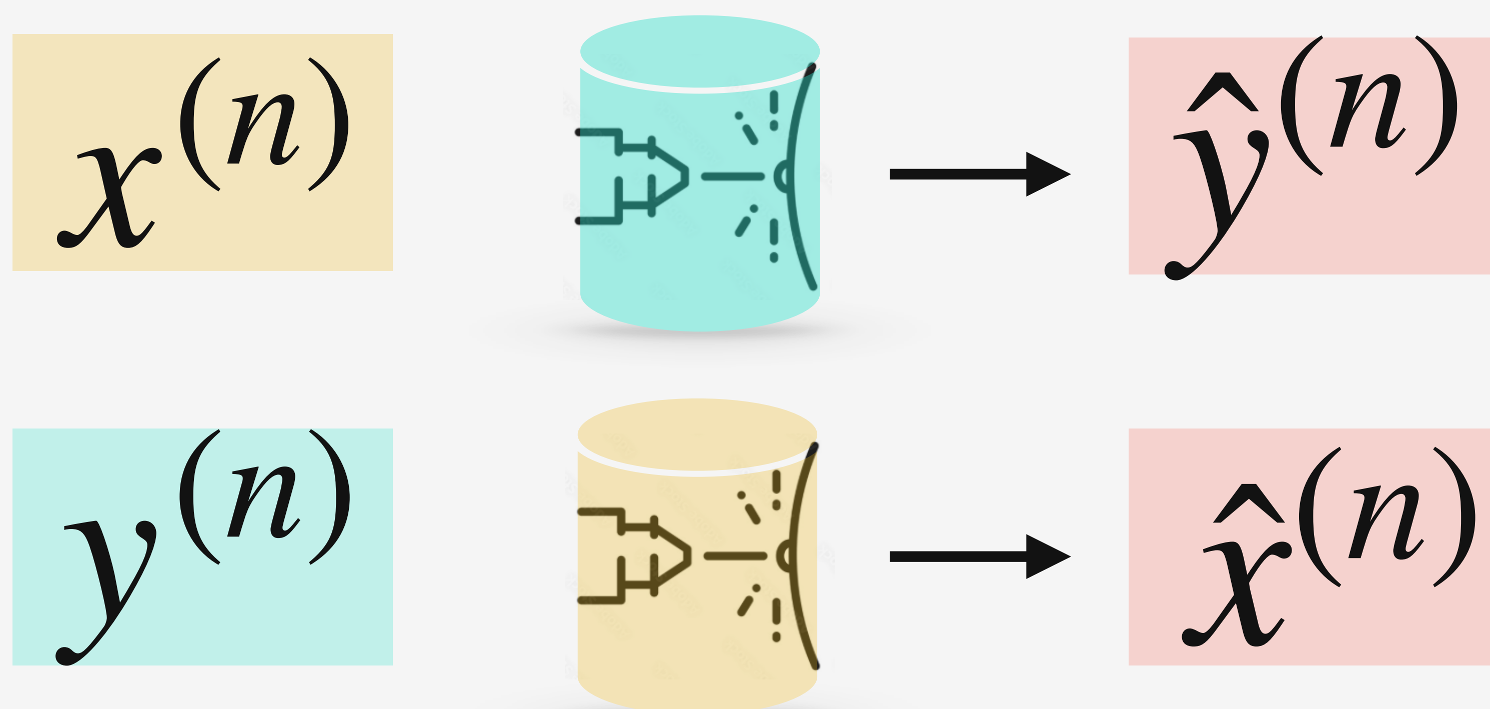
$$\log \mathrm{p}\Big(\,[<f>\; x^{(n)}] \mid \big(y^{(n)}, \hat{x}^{(n)}\big)\Big) + \log \mathrm{p}\Big(\,[<f>\; x^{(n)}] \mid \big(y^{(n)}, <\mathrm{MASK}>\big)\Big) +$$

$$\log \mathrm{p}\Big(\,[<e>\; y^{(n)}] \mid \big(x^{(n)}, \hat{y}^{(n)}\big)\Big) + \log \mathrm{p}\Big(\,[<e>\; y^{(n)}] \mid \big(x^{(n)}, <\mathrm{MASK}>\big)\Big)$$

BitextEdit: Automatic Bitext Editing for Improved Low-Resource Machine Translation.
Eleftheria Briakou, Sida I. Wang, Luke Zettlemoyer and Marjan Ghazvininejad. NAACL Findings 2022.

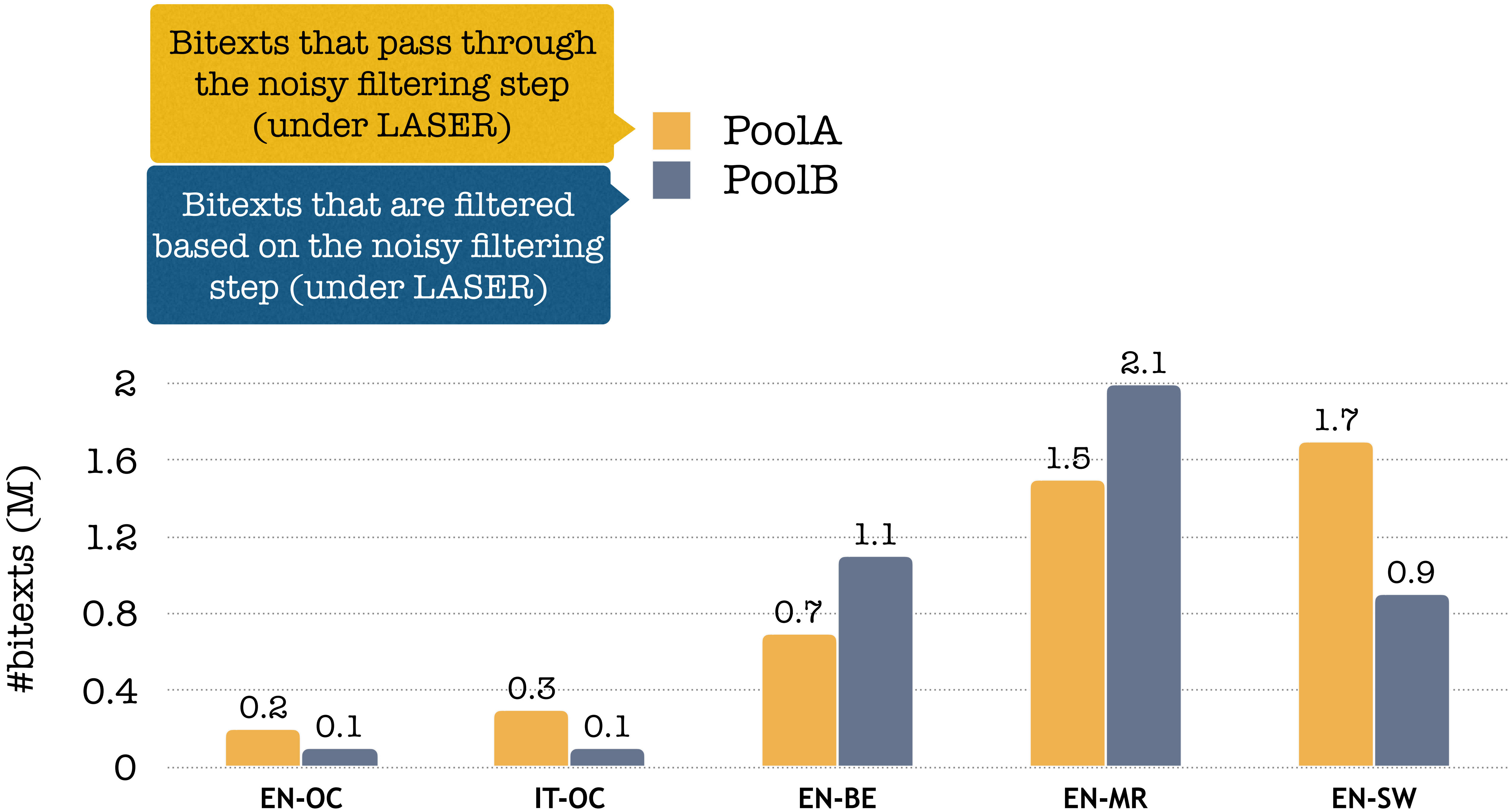
# BitextEdit: Mine Potentially Imperfect Translations for each Text in Given Bitext

CCMatrix
Parallel Text




**Noisy Reference Extraction:**
Mine Imperfect Translation from unlabeled texts
Similarity based multilingual sentence representations

# Experimental Settings



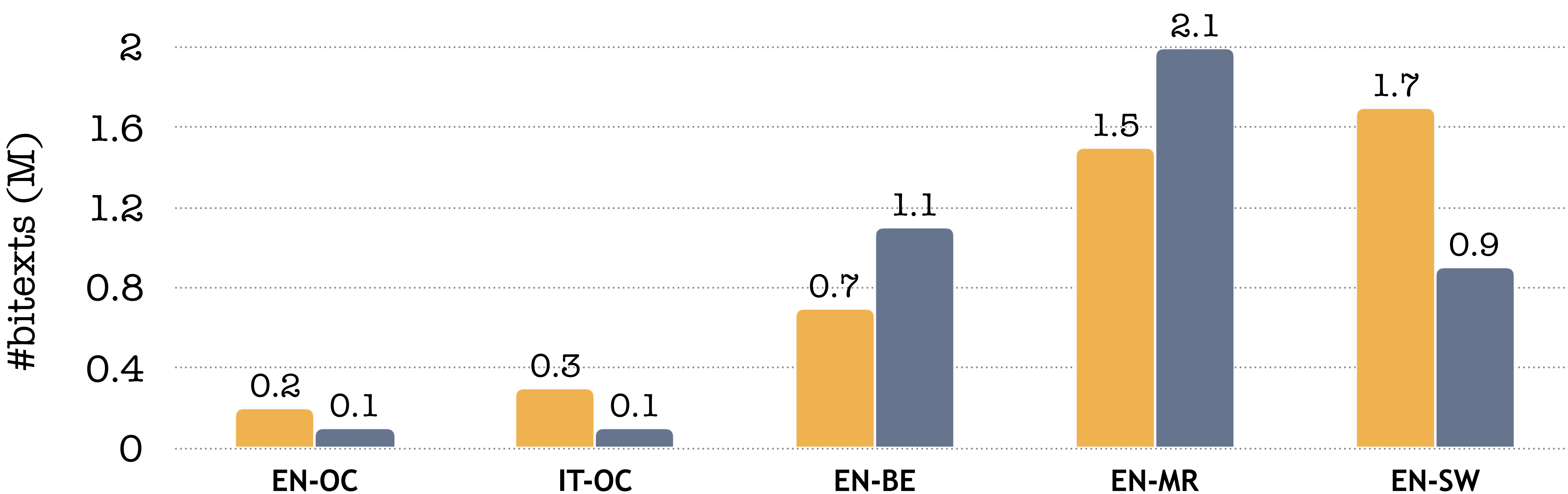
# Experimental Settings

Bitexts that pass through the noisy filtering step (under LASER)

Bitexts that are filtered based on the noisy filtering step (under LASER)

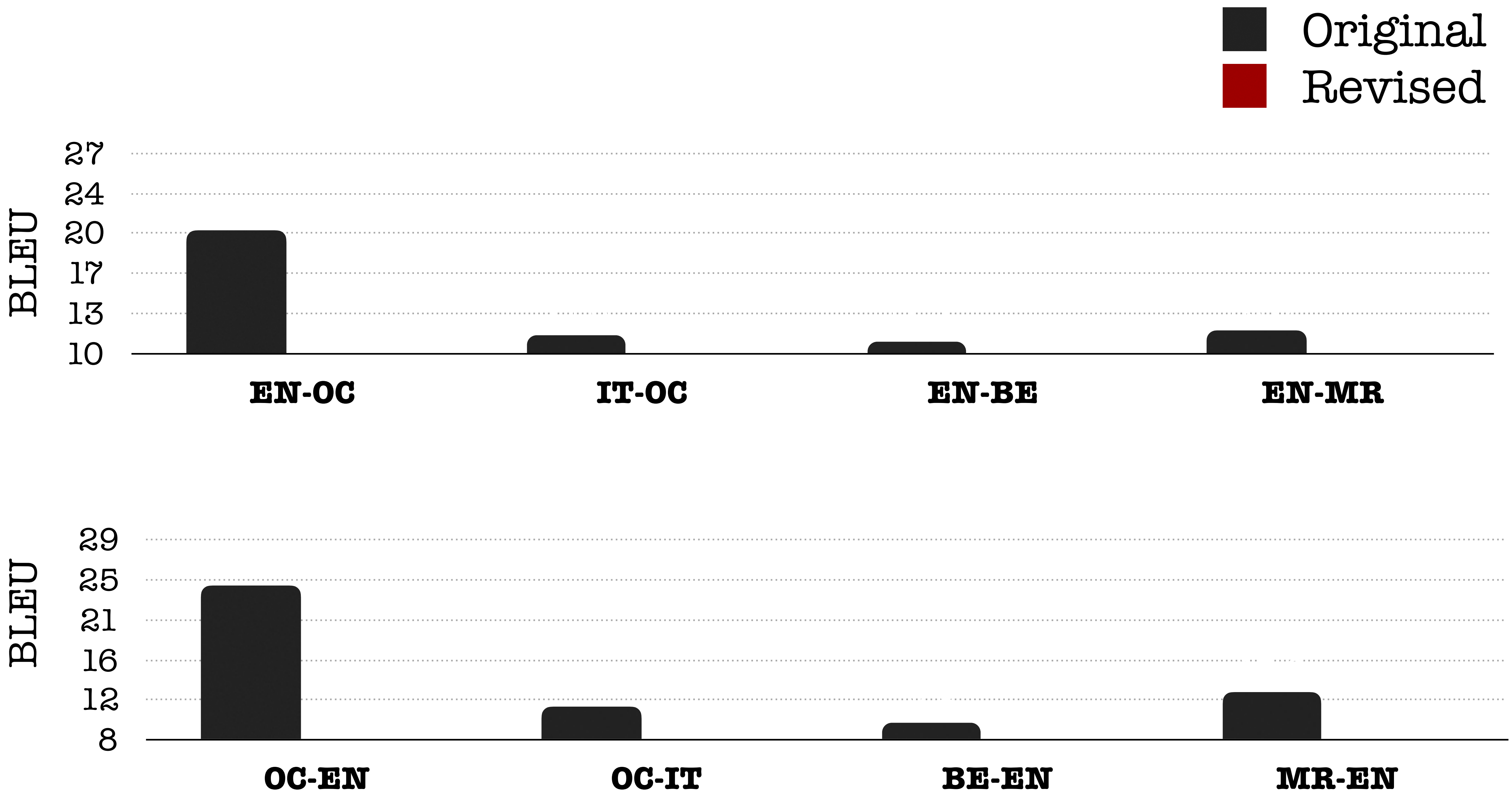
PoolA
PoolB

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



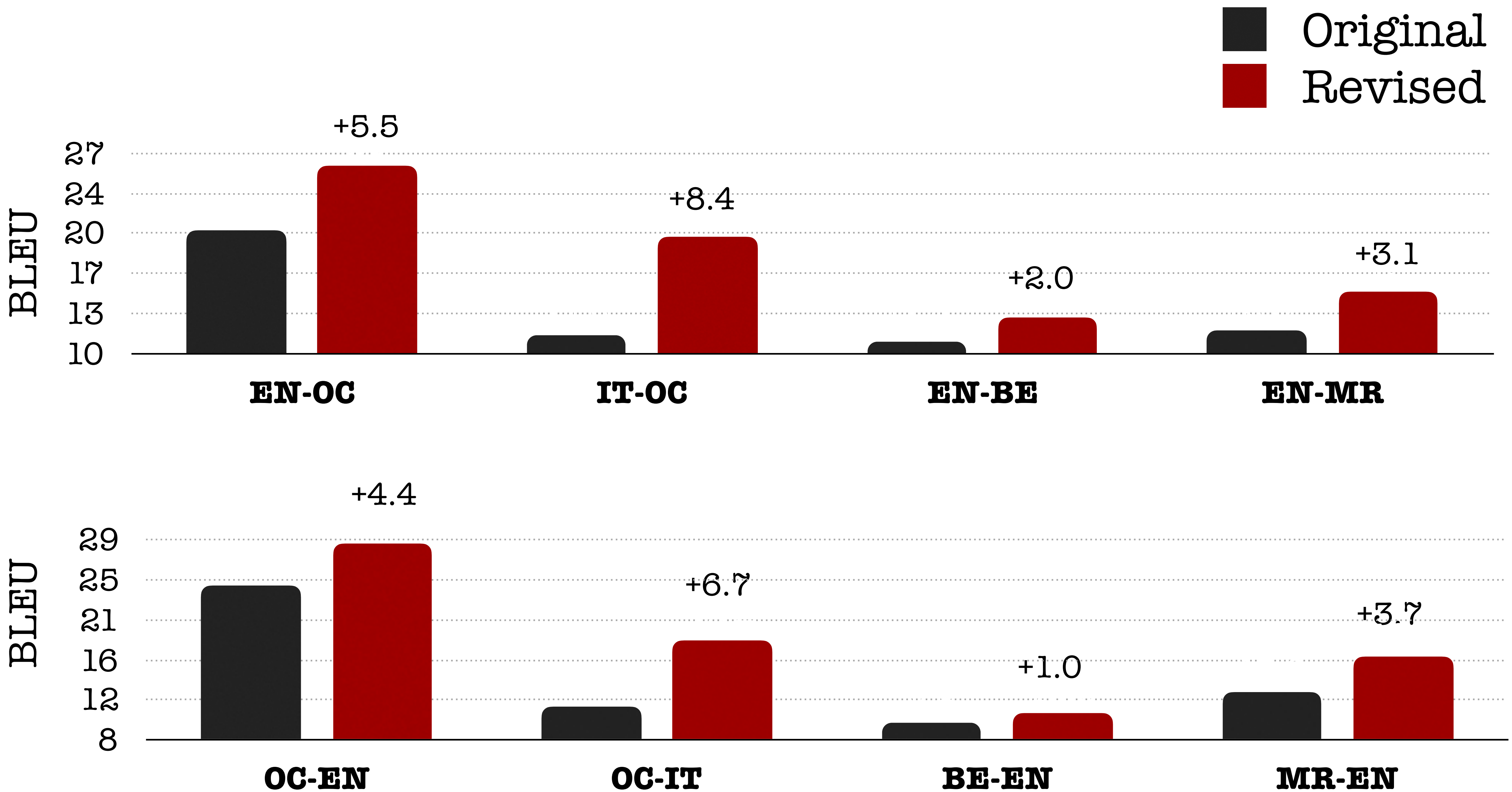

BitextEdit: Automatic Bitext Editing for Improved Low-Resource Machine Translation.
Eleftheria Briakou, Sida I. Wang, Luke Zettlemoyer and Marjan Ghazvininejad. NAACL Findings 2022.

# Does the Revised Bitext Provide More Reliable Training Signal than the Original?




BitextEdit: Automatic Bitext Editing for Improved Low-Resource Machine Translation.
Eleftheria Briakou, Sida I. Wang, Luke Zettlemoyer and Marjan Ghazvininejad. NAACL Findings 2022.

# BitextEdit: Revised Bitext yields better Translation Quality than the Original




BitextEdit: Automatic Bitext Editing for Improved Low-Resource Machine Translation.
Eleftheria Briakou, Sida I. Wang, Luke Zettlemoyer and Marjan Ghazvininejad. NAACL Findings 2022.

alternative approach for bitext quality improvement

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

synthetic supervision

alternative approach for bitext quality improvement

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

synthetic supervision

alternative approach for bitext quality improvement

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

translation quality improvements on 10 tasks

synthetic supervision

alternative approach for
bitext quality improvement

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

translation quality
improvements on 10 tasks

QUESTIONS?