Can Synthetic Translations Improve Bitext Quality?

Eleftheria Briakou & Marine Carpuat
{ebriakou, marine}@cs.umd.edu

Synthetic Translations improve MT quality

ST: Translations generated by MT

Prior work primary uses synthetic translations to reliably improve MT translation quality in two ways:

Augmenting original translations:	Replacing original translations:
[1] Forward translation [2] Backward translation [3] Data Diversification	[4] Sequence Distillation [5] Data Rejuvenation

Yet, it remains unclear:
Where does this improvement come from?

WE HYPOTHESIZE:
Synthetic translations are of higher quality (i.e., preserve translation equivalence) better than naturally occurring bitext

WE CONTRIBUTE:
An extensive empirical evaluation of the quality of bitext revised with synthetic translations

Revising Bitext w/ Selective Replacement of Synthetic Translations

procedure REVISE($D = (S, T)$, $\mathscr{R}$) — Given semantic equivalence classifier*

$M_{S\to T} \leftarrow$ TRAIN_MT$(D = (S, T))$
$M_{T\to S} \leftarrow$ TRAIN_MT$(D = (T, S))$
— Train NMT models to translate in opposite directions

$\tilde{D} \leftarrow \emptyset$
for $i \in |D|$ **do**
  $(S_i, \hat{T}_i) \leftarrow (S_i, M_{S\to T}(S_i))$
  $(\hat{S}_i, T_i) \leftarrow (M_{T\to S}(T_i), T_i)$
  — Generate synthetic bitext by pairing original references w/ synthetic transl.
  $d_F \leftarrow \mathscr{R}(S_i, \hat{T}_i) - \mathscr{R}(S_i, T_i)$
  $d_B \leftarrow \mathscr{R}(\hat{S}_i, T_i) - \mathscr{R}(S_i, T_i)$
  — Compute equivalence scores for original & synthetic pairs
  if $\max(d_F, d_B) > t$ **then**
    if $\max(d_F, d_B) = d_F$ **then**
      $\tilde{D} \leftarrow \tilde{D} \cup \{(S_i, \hat{T}_i)\}$
    else
      $\tilde{D} \leftarrow \tilde{D} \cup \{(\hat{S}_i, T_i)\}$
    end if
    — Replace the original with a synthetic translation only if it yields a more equivalent translation
  else
    $\tilde{D} \leftarrow \tilde{D} \cup \{(S_i, T_i)\}$
    — otherwise keep the original
  end if
end for
return $\tilde{D}$
end procedure

*Semantic Equivalence Classifier:
Fine-tuned mBERT of synthetic divergences of varying granularity based on our previous work [7]

Original vs. Revised Bitext Evaluation

Intrinsic: Human Assessments of Equivalence

Extrinsic: Performance on downstream NLP tasks

- **Bilingual Lexicon Induction via word alignment**
  Why?
  Intuition: Better Bitext Quality yields more accurate cross-lingual lexical mappings
- **Machine Translation (from scratch & continued training ~ WMT Parallel Corpus Filtering evaluations)**
  Why?
  Intuition: Better Bitext Quality yields more reliable training signal

Experimental Settings

Medium Resource Focus:
(a) Sufficient MT Quality
(b) Bitext Improvement needed

DATA:
- ✓ Training bitexts: WikiMatrix
- ✓ Language-pairs: EL-EN [~750K], RO-EN [~600K]
- ✓ BLI Test Sets: MUSE lexicons
- ✓ MT Test Sets: TED data

MODELS:
- ✓ MT [from scratch]: Transformer
- ✓ MT [continued training]: mt5

Intrinsic Evaluation Results

"Which sentence (A vs.B) conveys the meaning of the source better?"

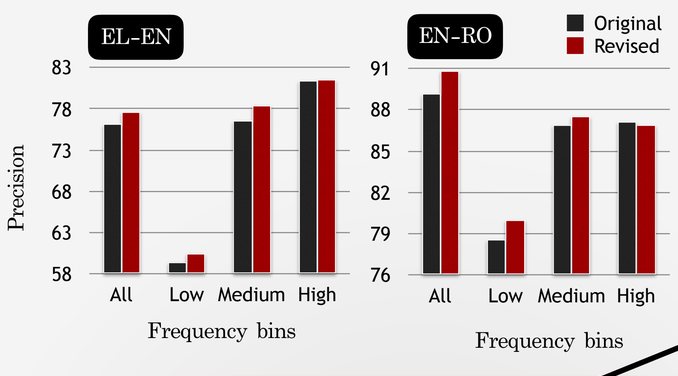
Source (original)
Ένας από τους οικισμούς που δημιούργησαν ήταν ο Καραβάς.
Target A (original)
One of the first towns to be created was Vila Barreto.
Target B (revised)
One of settlements to be created was Karavas.

Findings: Synthetic translations in the revised bitexts are more equivalent to the source than the original references **88%** of the times.

- 100 samples
- 3 annotators
- Lang: EL-EN

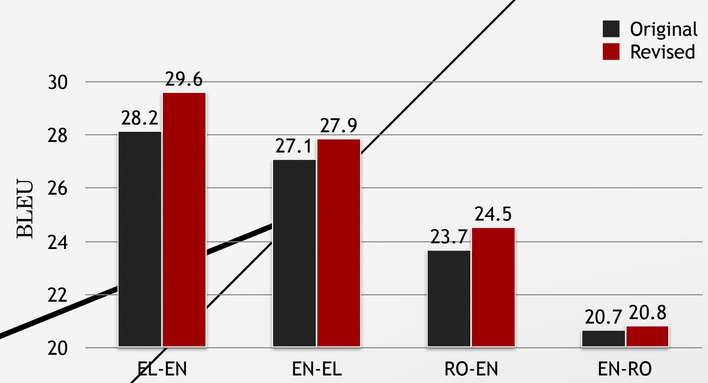
Extrinsic Evaluation Results

Unsupervised Bilingual Lexicon Induction

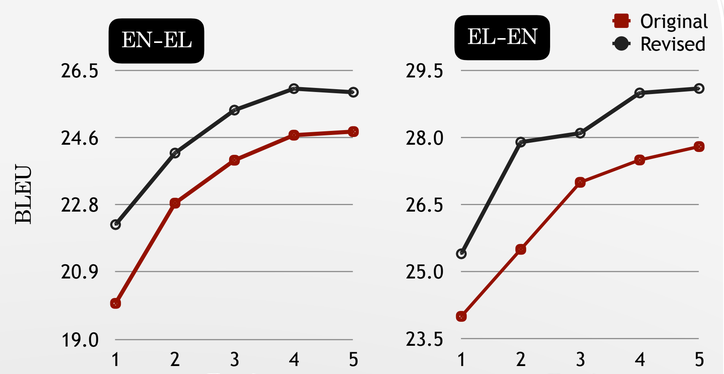


Findings: Revised bitexts yields better induction of low- & medium-frequency words, which we are more **sensitive to noisy misalignments** that result from poor quality bitext.

Machine Translation [Training from scratch]



Machine Translations [Continued Training]



Findings Revised bitexts yields better translation quality than training on the original for both MT settings (training from scratch & continued training), which further confirms that it yields more reliable training signal due to the **reduced noise** in the synthetic samples.

REFERENCES

[1] Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In EMNLP
[2] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In ACL
[3] Xuan-Phi Nguyen, Shafiq Joty, Kui Wu, and Ai Ti Aw. 2020. Data diversification: A simple strategy for neural machine translation. In. NEURIPS
[4] Jiatao Gu, James Bradbury, Caiming Xiong, Victor O.K. Li, and Richard Socher. 2018. Non-autoregressive neural machine translation. In *ICLR*
[5] Wenxiang Jiao, Xing Wang, Shilin He, Irwin King, Michael Lyu, and Zhaopeng Tu. 2020. Data Rejuvenation: Exploiting Inactive Training Examples for Neural Machine Translation. In *EMNLP*
[6] Haoyue Shi, Luke Zettlemoyer, and Sida Wang. 2021 Bilingual lexicon induction via unsupervised bitext construction and word alignment. In ACL
[7] Eleftheria Briakou, Marine Carpuat. 2020. Detecting Fine-Grained Cross-Lingual Semantic Divergences without Supervision by Learning to Rank. In *EMNLP*

Can Synthetic Translations Improve Bitext Quality?

Yes, when… they selectively replacing imperfect translations in naturally occurring bitexts under a semantic equivalence condition

According to… intrinsic evaluations of semantic equivalence and extrinsic evaluations on BLI and MT tasks

Data: https: //github.com/Elbria/xling-SemDiv-Equivalize.