Oops! I Translated Again

A project investigating iterative back-translation as a technique to improve machine translation for low-resource languages

Joline Hellström, Adam Samuelsson,
Cajsa Wargren, Ture Wramner

Motivation

Lack of bilingual data

Swedish-Sami

Synthetical data

Research questions

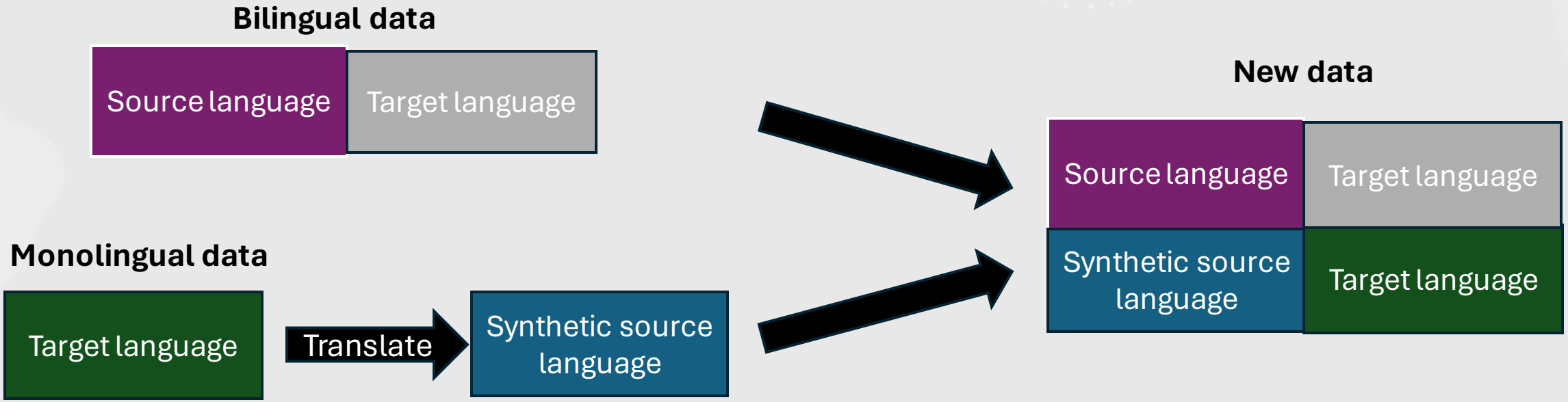
Can back-translation improve the performance of machine translation?

Swedish-Sami
Swedish-Norwegian
Swedish-Finnish

Do the effects of back-translation depend on whether they are from the same family of languages?

** Iterative Back-Translation for Neural Machine Translation (Hoang et. Al, 2018)*
**Revisiting Back-Translation for Low-Resource Machine Translation Between Chinese and Vietnamese (Li et. Al, 2020)*
** Neural Machine Translation Models with Back-Translation for the Extremely Low-Resource Indingenous Language Bribri (Feldman et. Al, 2020)*

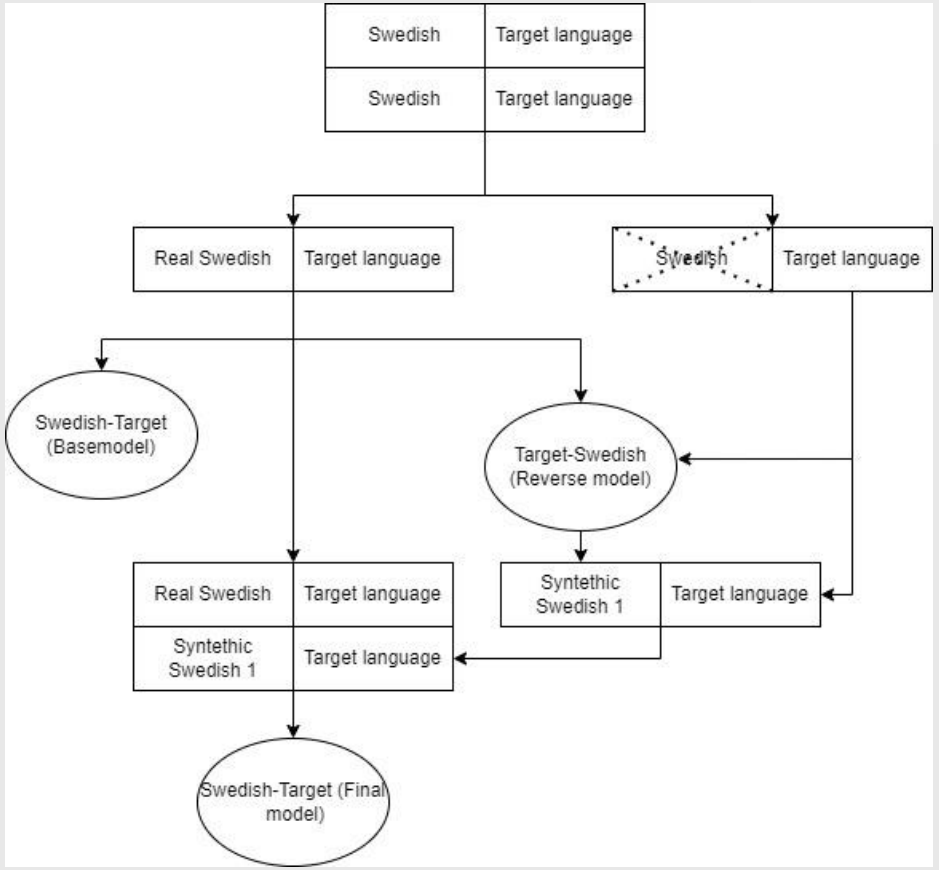
Back-translation



- Different methods of back-translation is suggested in the literature. The illustration describes our chosen method (Neural Machine Translation Models with Back-Translation for the Extremely Low-Resource Indigenous Language Bribri, Feldman & Coto-Solano, 2020)

Implementation

- The bilingual corpus is first split. 2000 sentences are saved for testing.
- The Basemodel and Reverse model is trained on one half.
- We pretend that the second half of the data is monolingual target data and create synthetic Swedish data.
- The back-translation process is repeated to implement an iterative back-translation.
- Transformer based neural machine translation models are used. (Attention is all you need, Vaswani et al. 2017)
- Preprocessing is done on all sets of training data.



Evaluation and Results

Translation	Bleu baseline	Bleu backtranslation	Δ Bleu
Swe-Nor	35.86	36.41	0.55
Swe-Fin	22.68	30.95	8.27
Swe-Sami	4.64	24.35	19.71

+2 Bleu estimated

Sometimes negative (Feldman)

No indications that linguistic genetic relation affects back-translation effectiveness

Indications that closely related languages gives better baseline model, which gives lower delta Bleu (Hoang)

Insufficient data to draw conclusive results about language families and Bleu

Dataset

Importance of data!

Our Data:
Swedish-Finnish: Opus
Swedish-Norwegian: Opus
Swedish-Sami: Masters-thesis

Other Research:
Bribri-Spanish: textbooks
Vietnamese-Chinese: native speakers fact-checking
English-German: Commonly used datasets

Data splitting:
Splitting after preprocessing
No guarantee that half the data is synthetic

Model

Feldman et al. trained for 4000 steps

“The longer trained systems have much better translation quality, and their synthetic parallel corpora prove to be beneficial.”
- Hoang et al.

“As the datasets are extremely small to train better models with more complicated architectures, we decide not to use the popular Transformer architecture [26], which will be one part of our future work.”
- Li et al.

Conclusion

- Why is our work important?
- Results and litterature shows that back-translation improves NMT for low resource languages
- Not enough evidence that back-translation is affected by language families