Your Title

First FAMILY

INSTITUTE

Jul. 5th, 2024

九州大学 | 大学院システム情報科学府
大学院システム情報科学研究院
Graduate School and Faculty of Information Science and Electrical Engineering

Background & Related Works

Data

German to Upper Sorbian NMT

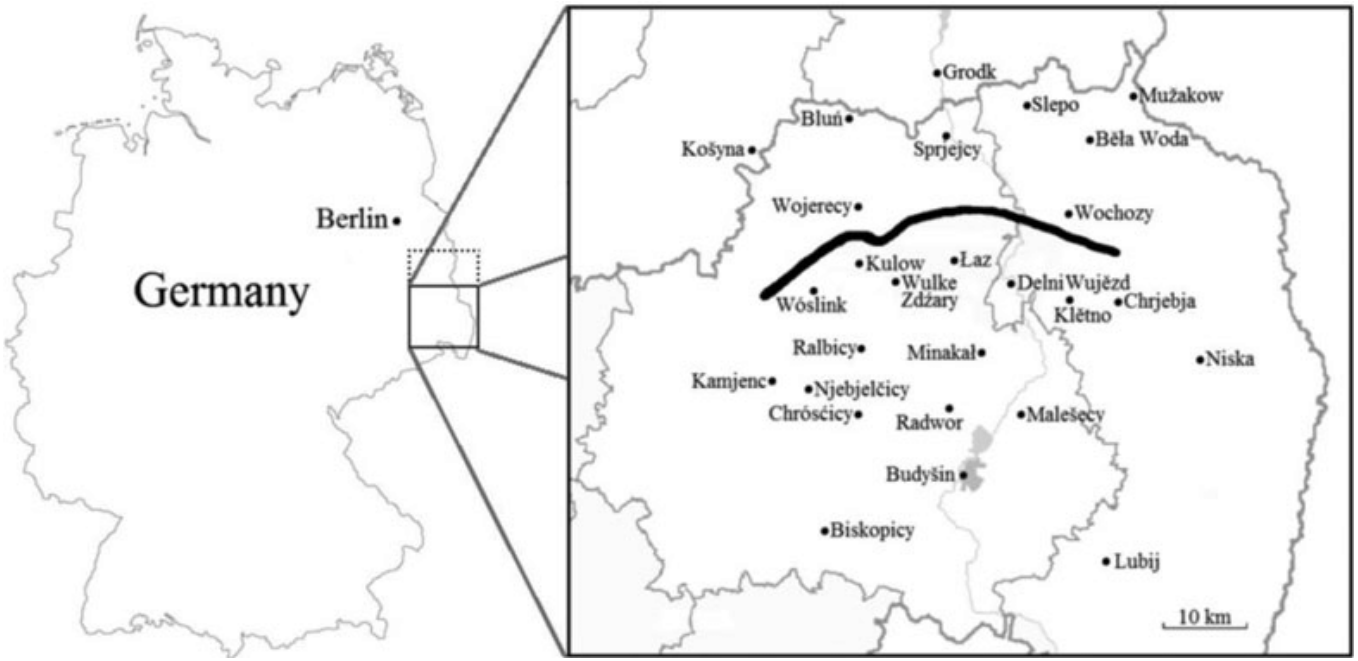

Figure: Upper Sorbian is spoken south of the solid black line. Image taken from (Howson, 2017).

Related works: Back Translation

Back-translation is a popular data augmentation technique to expand parallel corpora to enhance model performance.

Related works: Back Translation

Back-translation is a popular data augmentation technique to expand parallel corpora to enhance model performance.

- A proposed method utilizes a monolingual corpus on the target side for back-translation (Sennrich et al., 2016).

Related works: Back Translation

Back-translation is a popular data augmentation technique to expand parallel corpora to enhance model performance.

- A proposed method utilizes a monolingual corpus on the target side for back-translation (Sennrich et al., 2016).
- Fadaee and Monz (2018) investigated various aspects of back-translation and proposed several sampling strategy variations.

Related works: Back Translation

Back-translation is a popular data augmentation technique to expand parallel corpora to enhance model performance.

- A proposed method utilizes a monolingual corpus on the target side for back-translation (Sennrich et al., 2016).
- Fadaee and Monz (2018) investigated various aspects of back-translation and proposed several sampling strategy variations.
- Additionally, iterative back-translation to expand low-resource corpora was explored by (Hoang et al., 2018).

Background & Related Works

Data

Data

Corpus	Language	# sentences	Sentence length	
			in words	in characters
Bilingual	German (de) Upper Sorbian (hsb)	60,000 × 2	12.1 ± 6.9 10.7 ± 6.3	83.4 ± 51.7 71.6 ± 45.4
Monolingual	German (de)	1,879,765	15.4 ± 9.2	108.5 ± 64.8

Table: Statistics of the corpora.

We use the German-Upper Sorbian (de-hsb) parallel corpus from WMT20[1] as the original bilingual corpus, while the German corpus from WMT14[2] is chose as the original monolingual corpus.

[1] https://www.statmt.org/wmt20/unsup_and_very_low_res/

[2] `https://www.statmt.org/wmt14/training-monolingual-news-crawl/`

Thank you

Thanks for your attention!

Related works: Back Translation

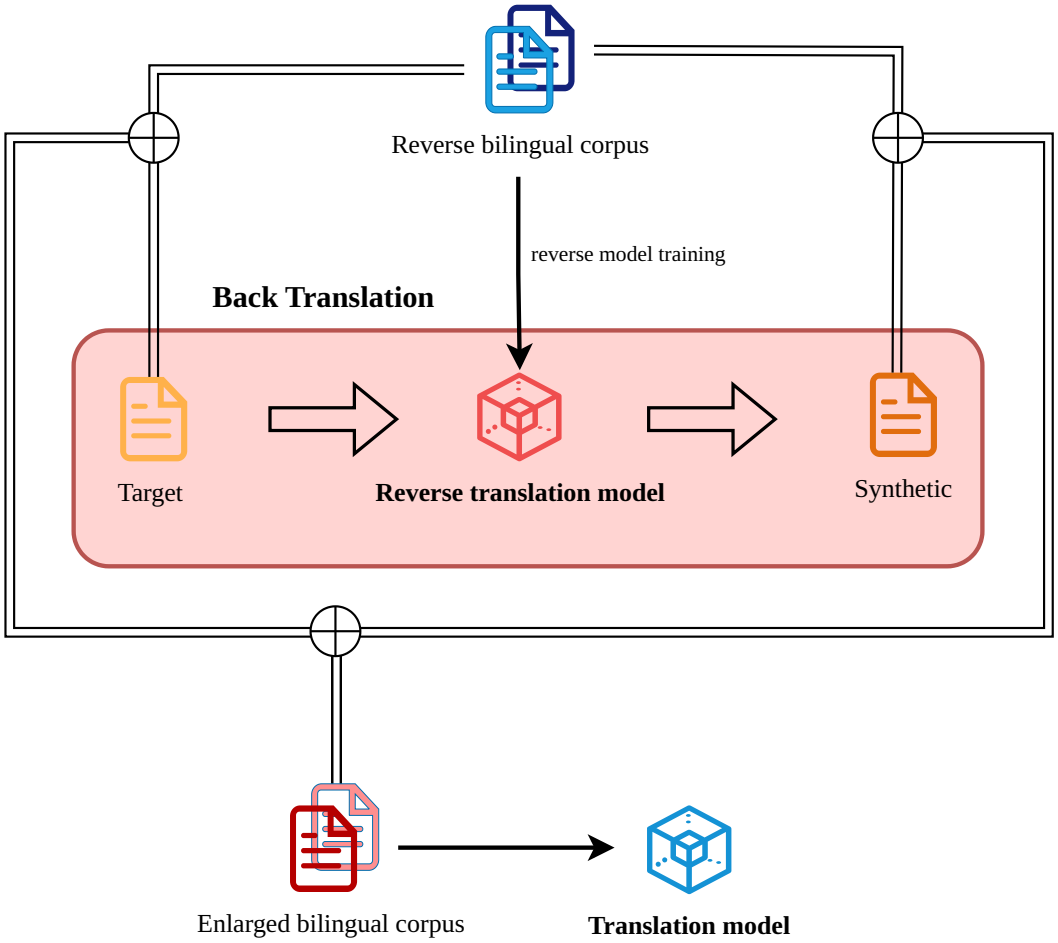

Figure: (Sennrich et al., 2016)

References I

Marzieh Fadaee and Christof Monz. Back-translation sampling by targeting difficult words in neural machine translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 436–446, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1040. URL https://aclanthology.org/D18-1040.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. Iterative back-translation for neural machine translation. In Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, pages 18–24, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-2703. URL https://aclanthology.org/W18-2703.

Phil Howson. Upper sorbian. Journal of the International Phonetic Association, 47(3):359–367, 2017.

References II

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 86–96, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1009. URL https://aclanthology.org/P16-1009.