UPPSALA
UNIVERSITET

Cross-lingual Word Embeddings beyond Zero-shot Machine Translation

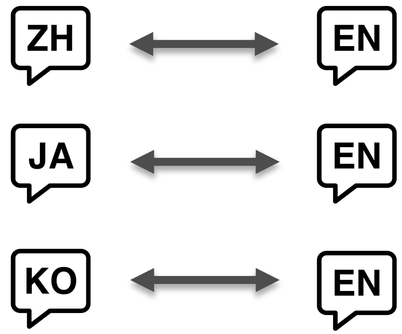
Shifei Chen

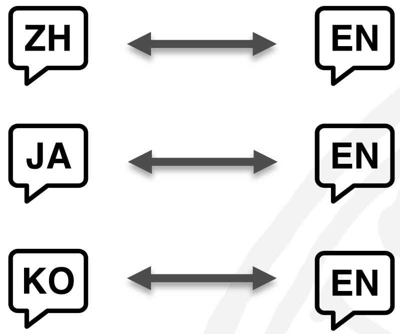
Department of Linguistics and Philology
Master Program in Language Technology
Master’s Thesis in Language Technology, 30 ECTS credits
October 28, 2020

Supervisor: Ali Basirat

Introduction

Training





SV is *never* in the training set
Will it work?

Testing

Research Question

- Is there any *transferability* of a multilingual NMT system from known languages to *completely unknown* languages?
- Follow up: How *language similarity* works in this scenario?
- Transfer source: Cross-lingual word embeddings

UPPSALA UNIVERSITET

Background

Relates to language similarity (Qi et al., 2018)

Who can transfer?

Similar vocabulary distribution exists across languages (Mikolov et al., 2013)

Which part in multilingual NMT is responsible for transferability?

Embedding layers are critical (Kim et al., 2019) or not (Aji et al., 2020)

Qi, Y., Sachan, D. S., Felix, M., Padmanabhan, S. J., and Neubig, G. When and why are pre-trained word embeddings useful for neural machine translation?
Mikolov, T., Le, Q. V., and Sutskever, I. Exploiting similarities among languages for machine translation.
Kim, Y., Gao, Y., and Ney, H. Effective cross-lingual transfer of neural machine translation models without shared vocabularies.
Aji, A. F., Bogoychev, N., Heafield, K., and Sennrich, R. In neural machine translation, what does transfer learning transfer? In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Online, July 2020), Association for Computational Linguistics, pp. 7701– 7710.

Methodology

An Encoder-Decoder LSTM neural network with attention module

Neural Network

TED subtitle corpus (Qi et al., 2018)

Training: varying from 490k to 1m sentences

Test: varying from 9k to 28k

Corpus

Training: EN+DE+FR

Testing: SV/HE/HU

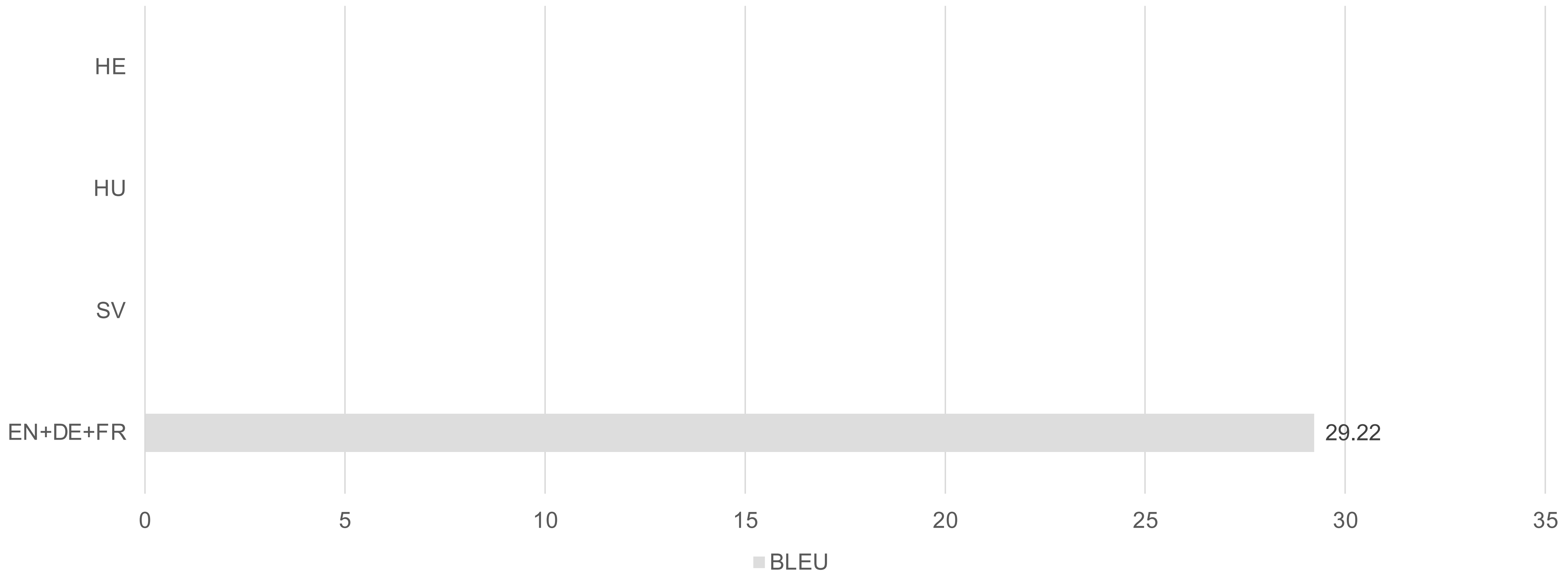
Languages

fastText(Joulin et al., 2018; Bojanowski et al., 2016) aligned word embeddings

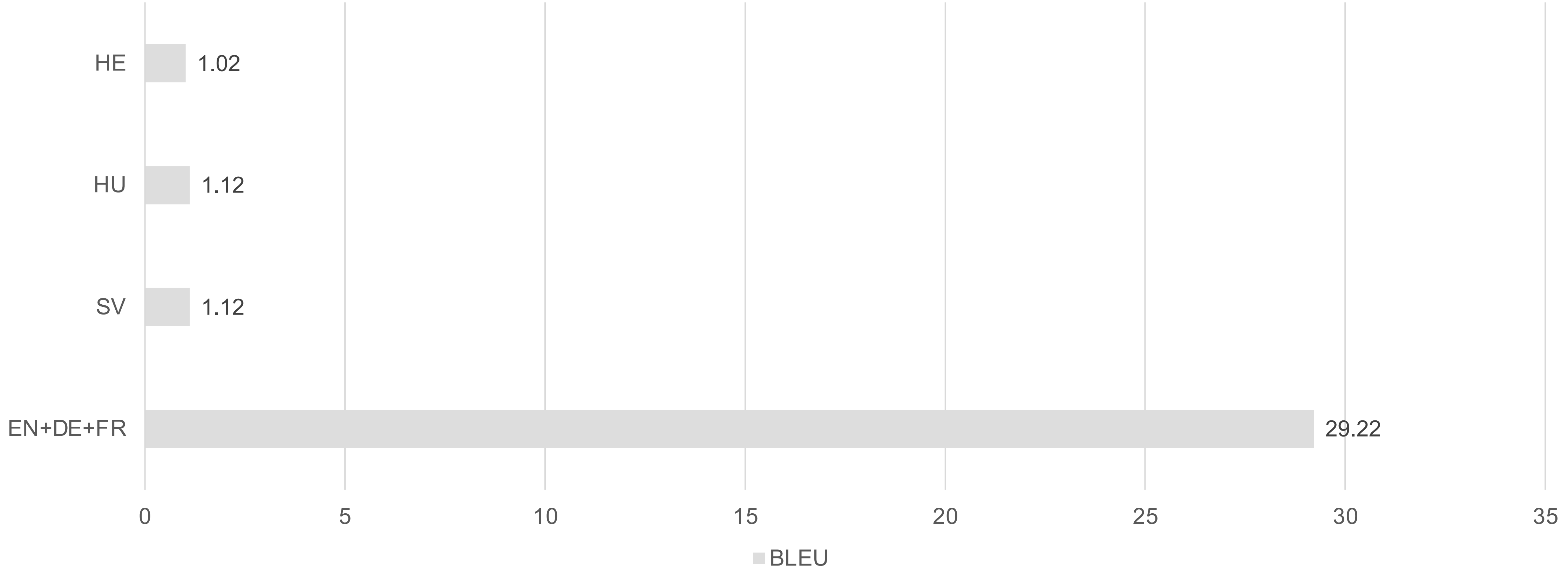
Word Embeddings

Qi, Y., Sachan, D. S., Felix, M., Padmanabhan, S. J., and Neubig, G. When and why are pre-trained word embeddings useful for neural machine translation?
Joulin, A., Bojanowski, P., Mikolov, T., Jegou, H., and Grave, E. Loss in translation: Learning bilingual word mapping with a retrieval criterion.
Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. Enriching word vectors with subword information.

Initial Results

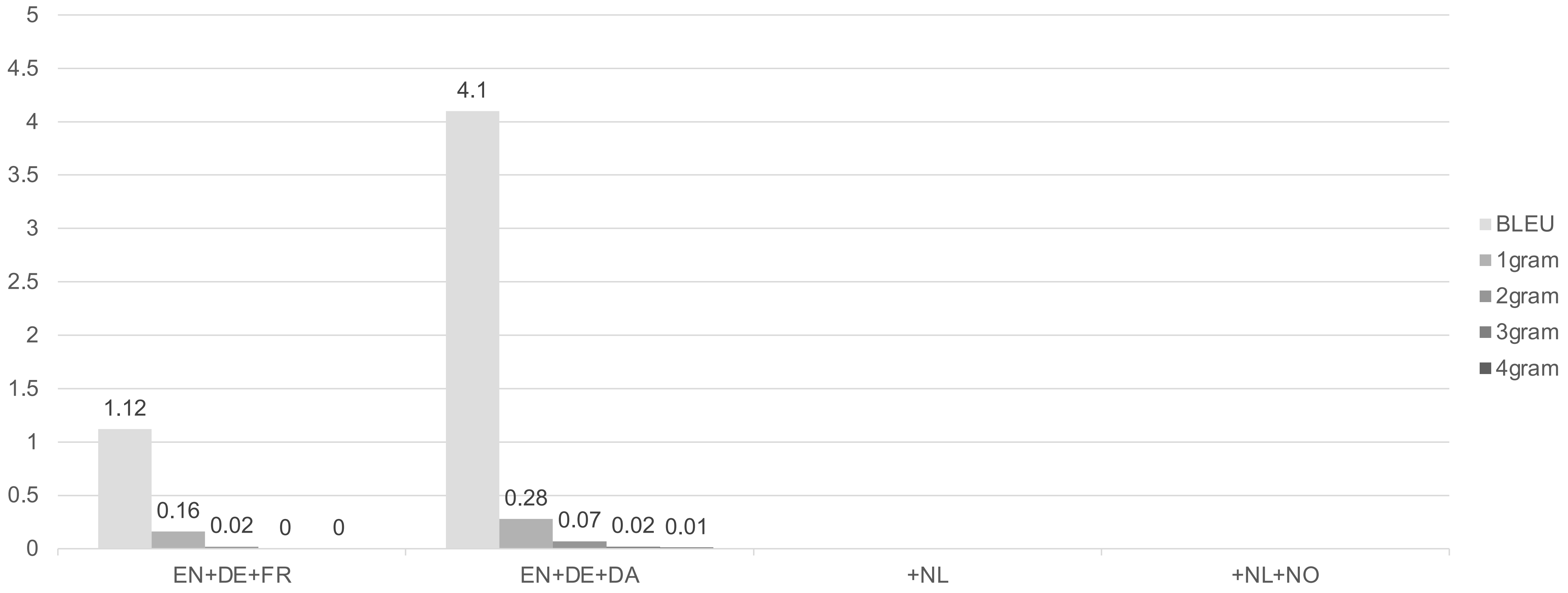


Initial Results



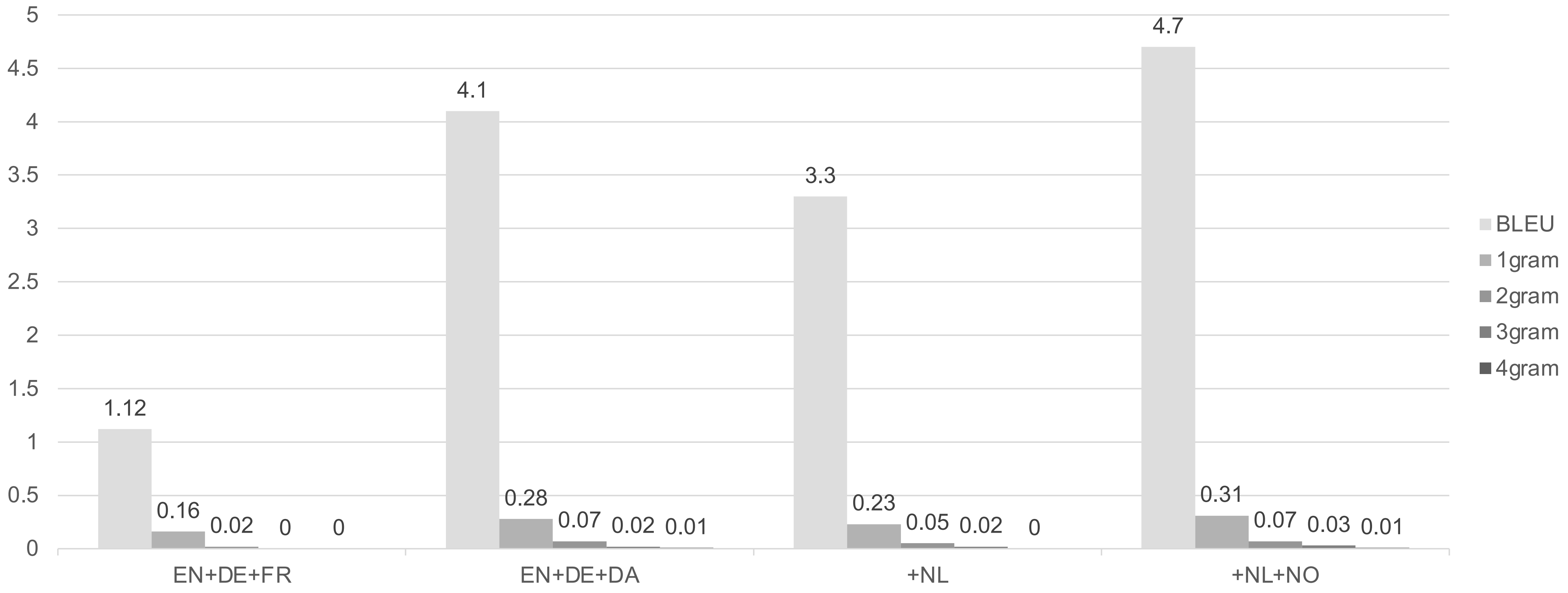
Language Similarity

- Remove FR and replace it with DA = better training set homogenization
- Later add NL and NO one by one to the training set



Language Similarity

- Remove FR and replace it with DA = better training set homogenization
- Later add NL and NO one by one to the training set



UPPSALA UNIVERSITET

Source of Language Similarity

Differentiate every token by its origin

```
__de__ <<sv>>och <<sv>>vi <<sv>>kämpar <<sv>>med <<sv>>dem .
```

BLEU drops: from 4.1 to 1.7

The system mainly learns lexicon translation

Improvements came from shared vocabularies

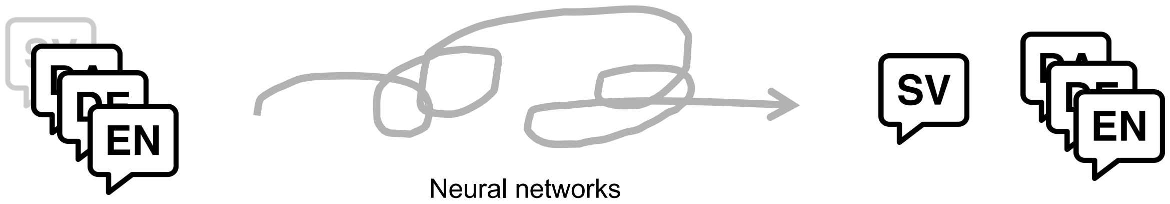
UPPSALA UNIVERSITET

Transformed Vector Space

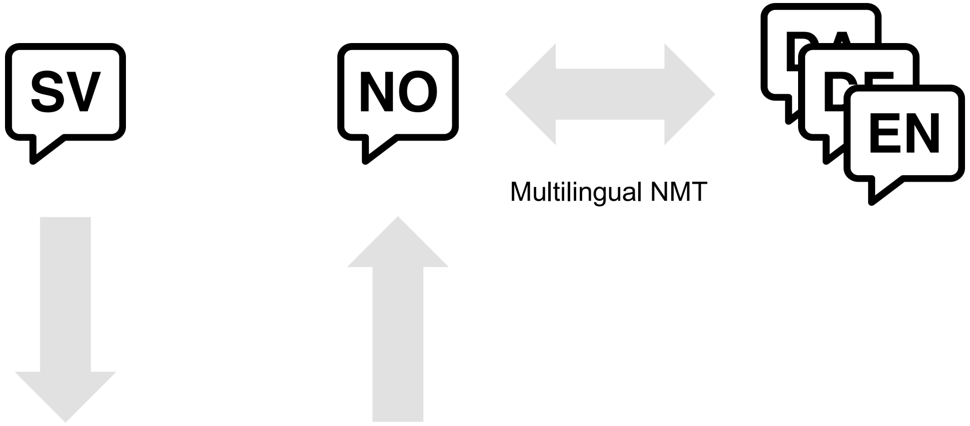
Translation quality differs in different direction:

SV to EN+DE+DA = 6 BLEU scores
EN+DE+DA to SV = 0.65 BLEU scores

The system *never* sees positive examples
Decoder's V_{out} may *no longer* align with the V_{in}(word embeddings)



Lexicon Replacement



$$d(w_s, w_t) = \sqrt{\sum_{i=1}^{n} (w_{s_i} - w_{t_i})^2}$$

If d < *threshold*, replace w_s with w_t

d value	BLEU	1gram	2gram	3gram	4gram
SV ↔ EN+DE+DA	4.1	0.28	0.07	0.02	0.01
No replacement	2.99	0.27	0.05	0.02	0.00
0.25	2.99	0.27	0.05	0.02	0.00
0.5	2.99	0.27	0.05	0.02	0.00
1	6.18	0.34	0.10	0.04	0.01
2	6.17	0.34	0.10	0.04	0.01
3	6.17	0.34	0.10	0.04	0.01
4	6.17	0.34	0.10	0.04	0.01
0	6.00	0.33	0.08	0.03	0.01

Conclusion

- Weak transferability exists
- Language similarity is related to transferability because of the shared vocabularies
- No positive examples caused the transformed output vector space
- Embedding layer alone is not enough for knowledge transfer (Aji et al., 2020)

Aji, A. F., Bogoychev, N., Heafield, K., and Sennrich, R. In neural machine translation, what does transfer learning transfer? In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Online, July 2020), Association for Computational Linguistics, pp. 7701– 7710.

Future Work

- Add a regularization layer to the loss function
- Explore other NMT architecture, e.g., Transformer (Vaswani et al., 2017)
- Explore other embeddings, e.g.,
    - language embeddings (Littell et al., 2017; Malaviya et al., 2017)
    - contextual word embeddings (Devlin et al., 2019)
    - sub-word embeddings (Heinzerling and Strube, 2017)

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need.
Littell, P., Mortensen, D. R., Lin, K., Kairis, K., Turner, C., and Levin, L. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers (Valencia, Spain, Apr. 2017), Association for Computational Linguistics, pp. 8–14.
Malaviya, C., Neubig, G., and Littell, P. Learning language representations for typology pre- diction. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (Copenhagen, Denmark, Sept. 2017), Association for Computational Linguistics, pp. 2529–2535.
Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirec- tional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (Minneapolis, Minnesota, June 2019), Association for Computational Linguistics, pp. 4171–4186.
Heinzerling, B., and Strube, M. Bpemb: Tokenization-free pre-trained subword embeddings in 275 languages.