UPPSALA
UNIVERSITET

Cross-lingual Word Embeddings beyond Zero-shot Machine Translation

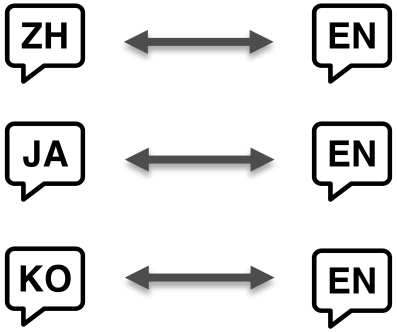
Shifei Chen & Ali Basirat

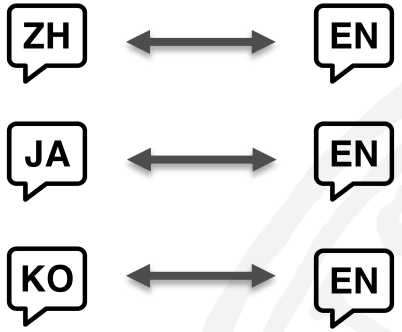
Department of Linguistics and Philology
Uppsala University

November 27, 2020

Introduction

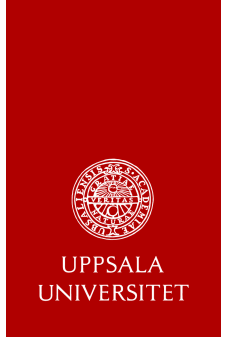
Training





Testing

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



Research Question

- Is there any *transferability* of a multilingual NMT system from seen languages to *completely unseen* languages?
- Follow up: How *language similarity* works in this scenario?

UPPSALA UNIVERSITET

NMT & Multilingual NMT

- NMT (Neural Machine Translation)
  - Uses neural networks to learn the translation relationship between a source and a target language.
  - Comes with an encoder, a decoder and an attention module

- Multilingual NMT (Johnson et al., 2017)
  - Same architecture as NMT
  - Additional token indicating the target language
  - Enables transfer learning between languages

  - Even zero-shot translation (Zoph et al., 2016; Nguyen and Chiang, 2017)!

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Vie´gas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: En- abling zero-shot translation. Transactions of the As- sociation for Computational Linguistics, 5:339–351.
Toan Q. Nguyen and David Chiang. 2017.
Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In Proceedings of the 2016 Conference on Empirical Methods in Natu- ral Language Processing, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.
Trans- fer learning across low-resource, related languages for neural machine translation. In Proceedings of the Eighth International Joint Conference on Natu- ral Language Processing (Volume 2: Short Papers), pages 296–301, Taipei, Taiwan. Asian Federation of Natural Language Processing.

UPPSALA
UNIVERSITET

Zero-shot Machine Translation

- Translation between unseen language *pairs*

- Benefit low-resource languages
  - By leverage knowledge from high-resource languages
  - No need for a pivot language, or *interlingua*
  - Requires much less data to train

- Our experiment differs
  - Test on *completely unseen* languages

UPPSALA UNIVERSITET

Cross-lingual Word Embeddings

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)

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.

UPPSALA UNIVERSITET

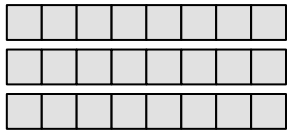
Methodology

Neural Network	Corpus	Languages	Word Embeddings
An Encoder-Decoder LSTM neural network with attention module	TED subtitle corpus (Qi et al., 2018) Training: varying from 490k to 1m sentences Test: varying from 9k to 28k	Training: EN+DE+FR Testing: SV/HE/HU	fastText(Joulin et al., 2018; Bojanowski et al., 2016) aligned 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.

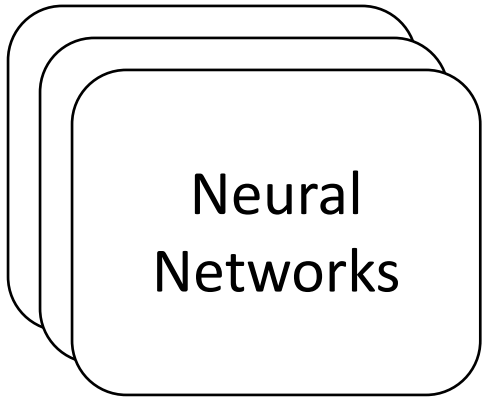
UPPSALA UNIVERSITET

Methodology



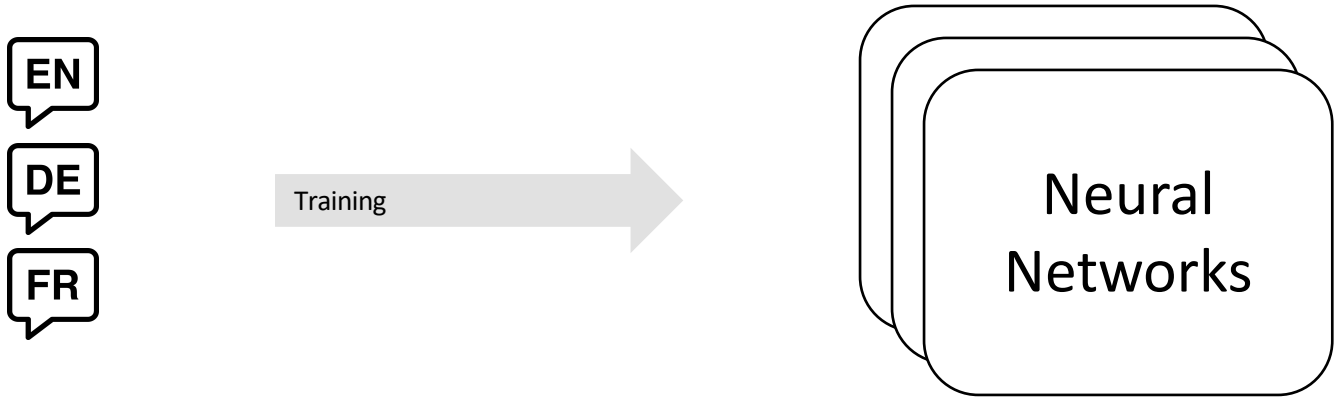
Embeddings in the training languages

Training



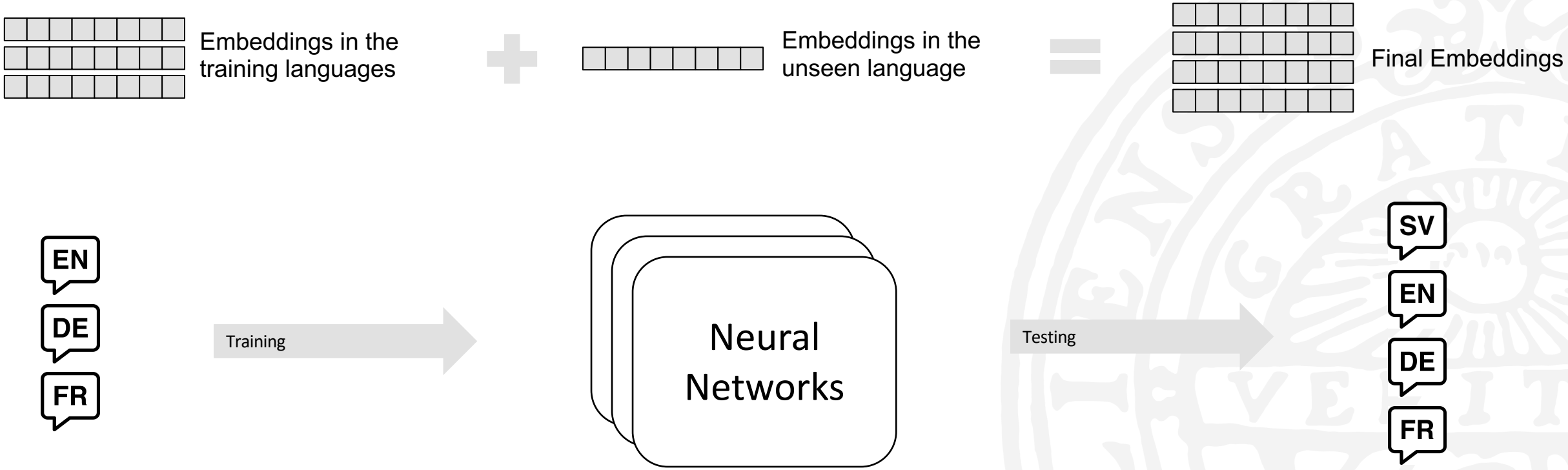
UPPSALA UNIVERSITET

Methodology



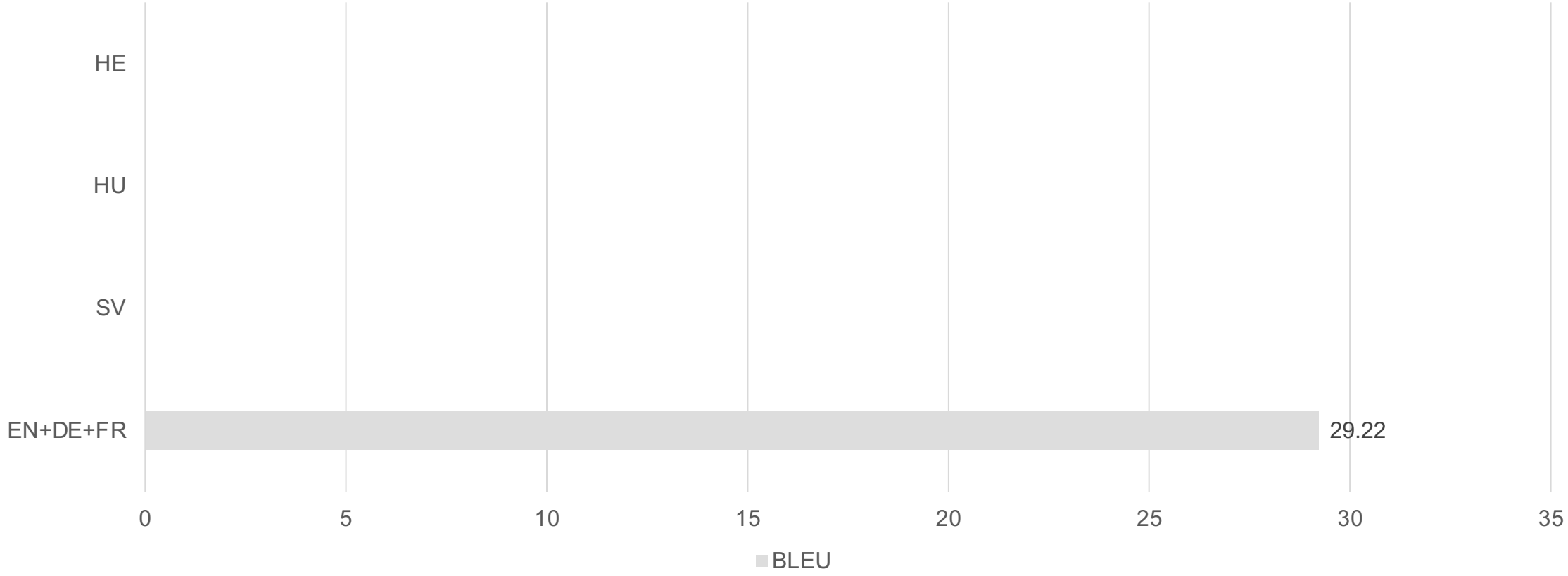
UPPSALA UNIVERSITET

Methodology



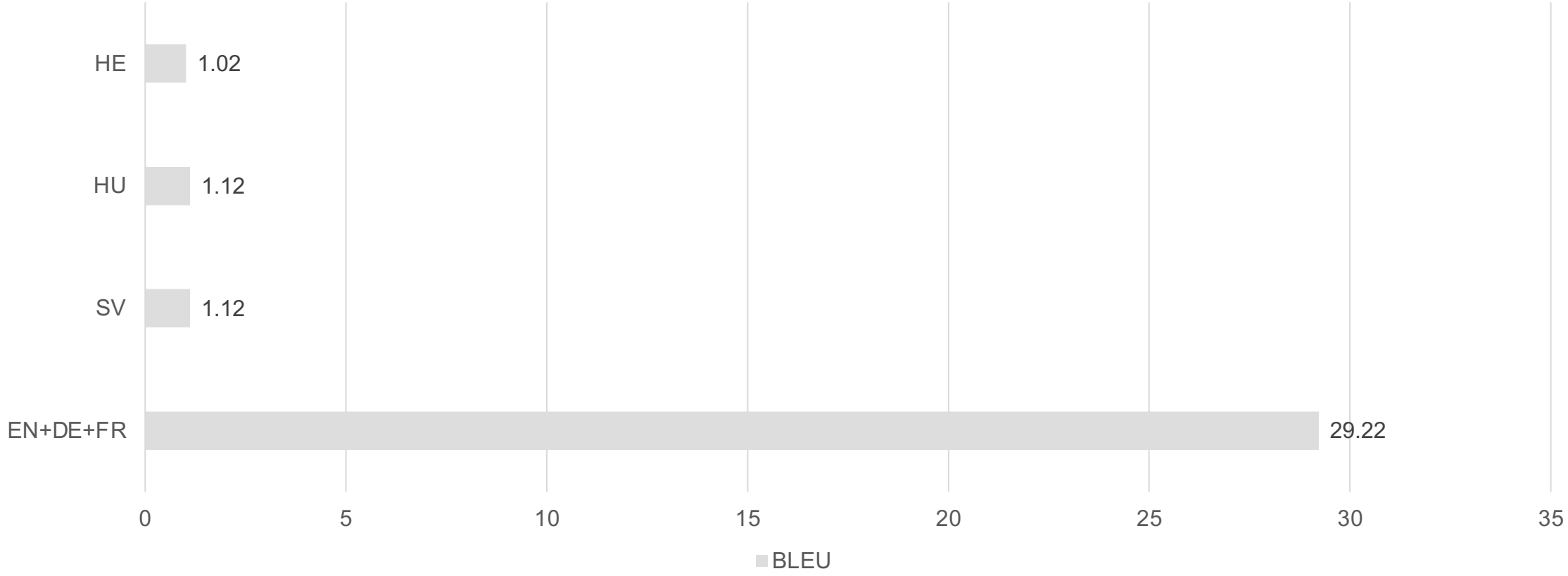
UPPSALA UNIVERSITET

Initial Results



UPPSALA UNIVERSITET

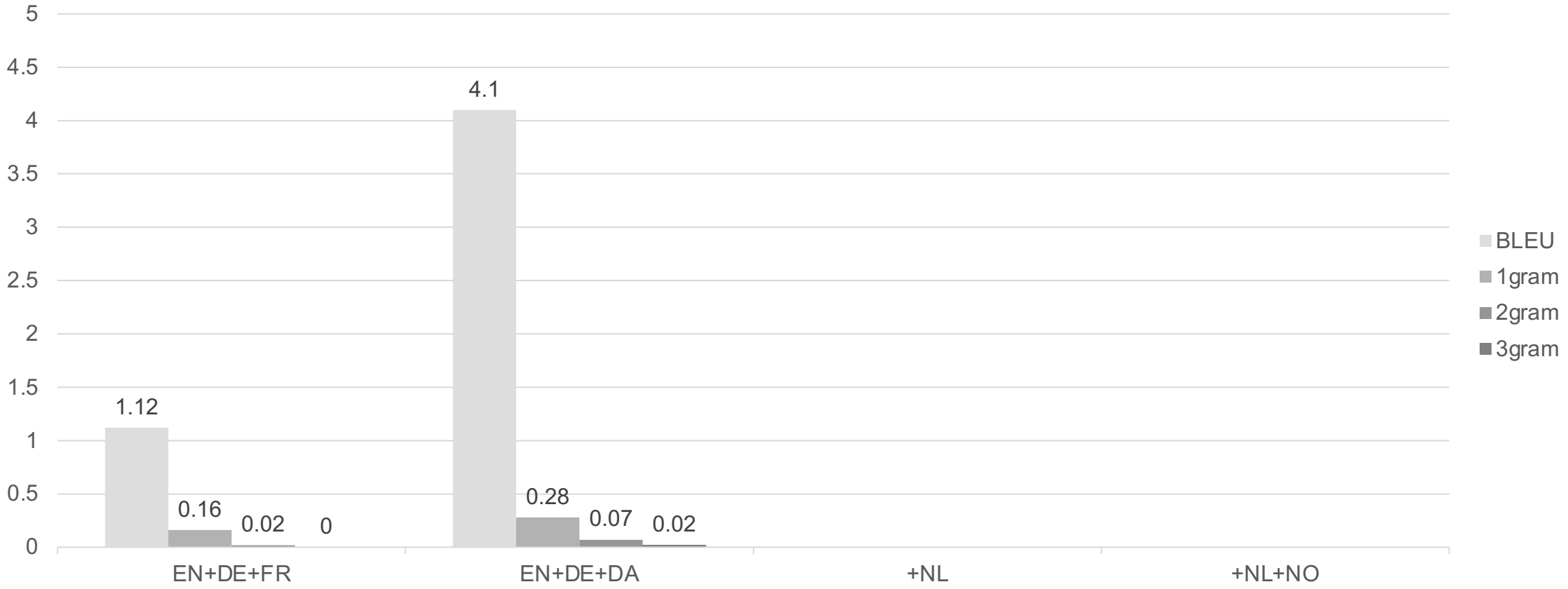
Initial Results



UPPSALA UNIVERSITET

Language Similarity

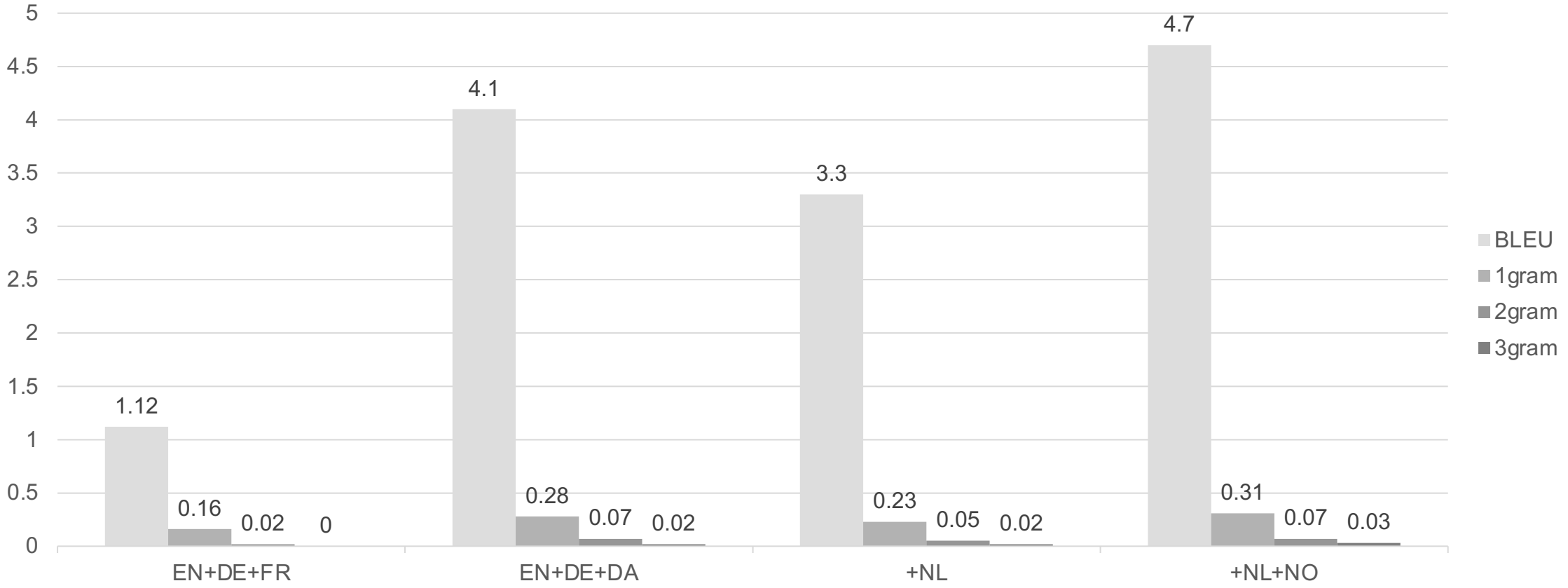
- Remove FR and replace it with DA = better training set homogenization
- Later add NL and NO one by one to the training set



UPPSALA UNIVERSITET

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

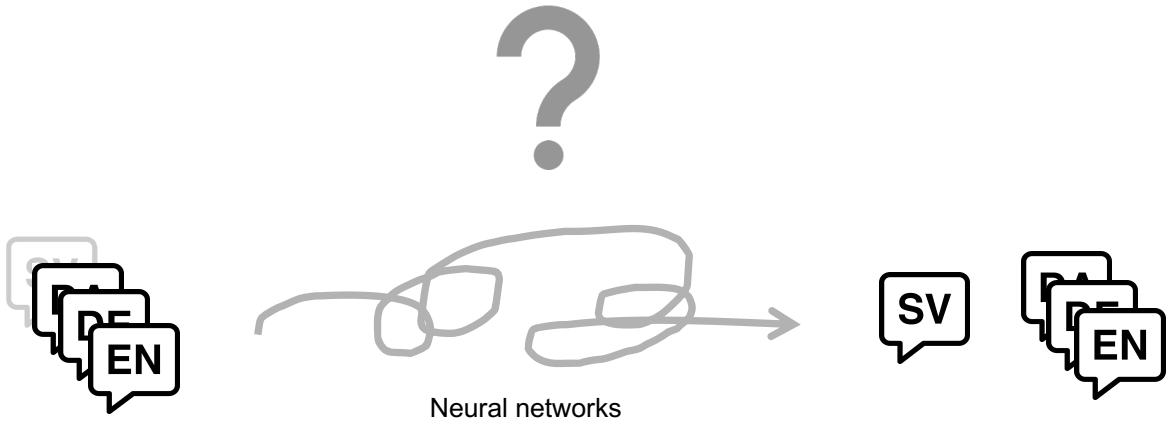
The system mainly learns lexicon translation

Improvements came from shared vocabularies

Gap in Translation Directions

Translation quality differs in different direction:

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



UPPSALA UNIVERSITET

Transformed Vector Space

The system *never* sees positive examples

Output vector weight gets *continuously* deducted

Decoder's V_{out} may *no longer* align with the V_{in}(word embeddings)

UPPSALA UNIVERSITET

Conclusion

- Transferability from unseen language to seen language exists
- Language similarity is related to the transferability because of similar vocabularies

We have also observed:

- No positive examples caused the transformed output vector space
- Cross-lingual word embedding along is not enough for transfer learning (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.

UPPSALA UNIVERSITET

Future Work

- Add a regularization layer to the loss function
- Add language information to the model, e.g.,
  - Language embeddings (Littell et al., 2017; Malaviya et al., 2017)
- Explore other NMT architecture, e.g., Transformer (Vaswani et al., 2017)
- Explore other embeddings, e.g.,
  - 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.