Character-based Neural Machine Translation

Encuentro estudiantes de doctorado

Antonio Manuel Larriba Flor

June 10, 2018

Universitat Politècnica de València

Index

Why characters?

Lack of perfect word segmentation algorithms (Chung et al., 2016). Similar words treated as completely different entities.

- Difficulties when generalizing into new words.
- Problems modeling morphological variants.

Why characters? 2

Pros

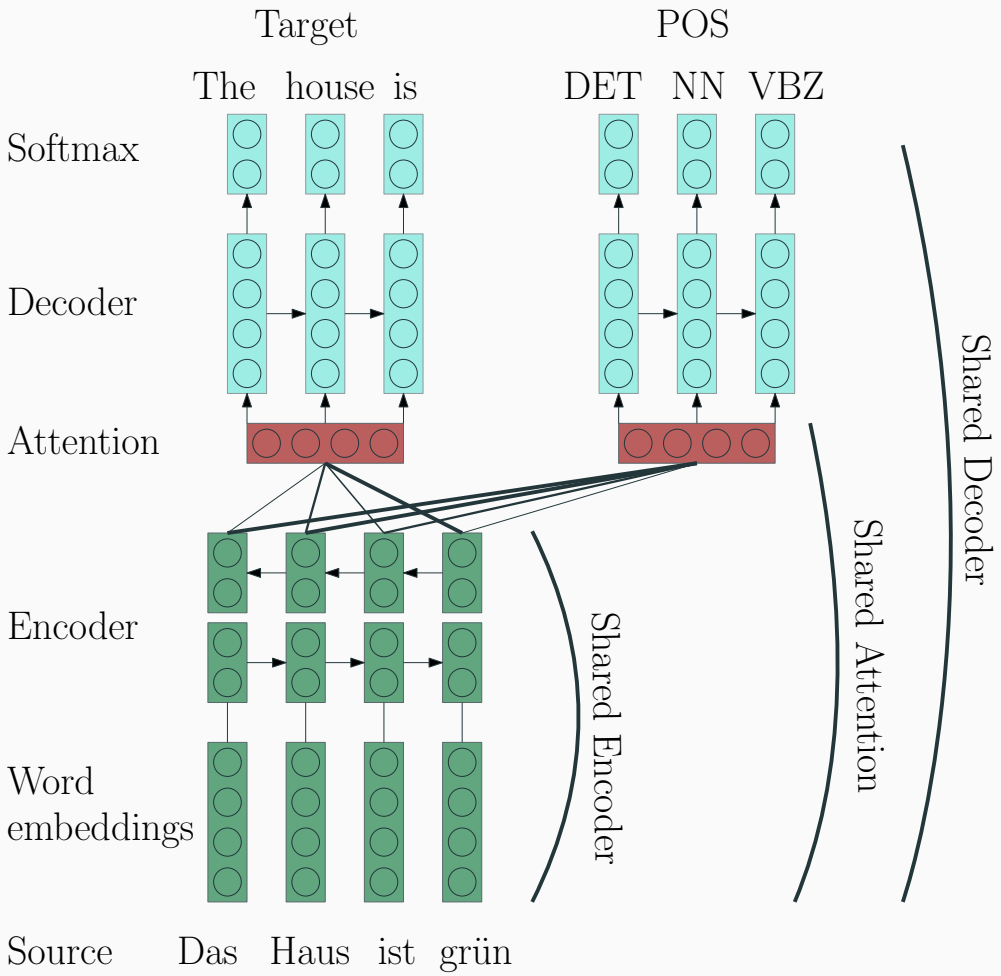
- Smaller vocabularies.
- Reduce out-of-vocabulary words.
- Profit from in-word lexical information.

Cons

- Longer input sequences.
- References at word level.

How?

- Based on Niehues and Cho (2017) work. (POS Tagging)
- Shared information may benefit both sides.
- Different granularities.



Architecture overview (Niehues and Cho, 2017)

Conclusions

- Characters act as an important piece of information in the translation process.
- Less out-of-vocabulary words, better results and more fluent translations can be obtained.

References

Chung, J., Cho, K., and Bengio, Y. (2016). A character-level decoder without explicit segmentation for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1.

Niehues, J. and Cho, E. (2017). Exploiting linguistic resources for neural machine translation using multi-task learning. In *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*, pages 80–89.