Influence of training corpora in neural machine translation output quality

Proposer(s) / Proposatzailea(k): Nora Aranberri

Contact / Kontaktua: nora.aranberri@ehu.eus

Description / Deskribapena

Most research on neural machine translation (NMT) systems focuses on the different techniques and approaches to process the training data that will allow us to obtain the best possible output quality. However, despite the acknowledged importance of the training data itself, little research has been carried out on how it influences NMT output. This project aims to explore this aspect. What impact do the different features of a training corpus have on an NMT output? How does the output vary when we use corpora of different sizes, topics, lexical variation, density, word and structural distribution, word and structural frequency, etc.?

Goals / Helburuak

To study the influence of training corpora in the output of neural machine translation systems

Requirements / Betebeharrak

Linguistic background, at least basic programming skills, proficiency in at least two languages

Framework / Esparrua

Machine translation evaluation

Tasks and plan / Atazak eta plana

- Analyse the literature on NMT training processes, corpus characteristics and MT evaluation
- Identify/build an baseline NMT system
- Identify potentially relevant characteristics of training corpora
- Identify/select/modify training corpora according to the selected characteristics
- Train the NMT system with the different (versions of the) corpora
- Analyse the output
- Write up the report

References

Khadivi, S., & Ney, H. (2005, June). Automatic filtering of bilingual corpora for statistical machine translation. In *International Conference on Application of Natural Language to Information Systems* (pp. 263-274). Springer, Berlin, Heidelberg.

Muischnek, K., & Müürisep, K. (2018, September). Impact of corpora quality on neural machine translation. In *Human Language Technologies–The Baltic Perspective: Proceedings of the Eighth International Conference Baltic HLT 2018* (Vol. 307, p. 126). IOS Press.

Srivastava, J., Sanyal, S., & Srivastava, A. K. (2019). An Automatic and a Machine-assisted Method to Clean Bilingual Corpus. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, *19*(1), 1-19.

Xu, G., Ko, Y., & Seo, J. (2019). Improving Neural Machine Translation by Filtering Synthetic Parallel Data. *Entropy*, *21*(12), 1213.