Readings for 17. 3. talk (Petrovič)

Kudo: Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates
unigram-DBLP:journals/corr/abs-1804-10959

Taku Kudo. "Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates". In: *CoRR* abs/1804.10959 (2018). arXiv: 1804.10959. URL: http://arxiv.org/abs/1804.10959.

Objective: Better understanding of inner workings of SentencePiece tokenizer.

Ideas: SubWord regularization - makes NMT models more robust, probabilistic approach to vocabulary building.

Relations: Used in SentencePiece tokenizer. on-the-fly subword sampling, subword segmentation based on unigram lang. model.

Kudo et al.: SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing
kudo-richardson-2018-sentencepiece

Taku Kudo and John Richardson. "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 66–71. DOI: 10.18653/v1/D18-2012. URL: https://aclanthology.org/D18-2012.

Abstract: This paper describes SentencePiece, a language-independent subword tokenizer and detokenizer designed for Neural-based text processing, including Neural Machine Translation. It provides open-source C++ and Python implementations for subword units. While existing subword segmentation tools assume that the input is pre-tokenized into word sequences, SentencePiece can train subword models directly from raw sentences, which allows us to make a purely end-to-end and language independent system. We perform a validation experiment of NMT on English-Japanese machine translation, and find that it is possible to achieve comparable accuracy to direct subword training from raw sentences. We also compare the performance of subword training and segmentation with various configurations. SentencePiece is available under the Apache 2 license at https://github.com/google/sentencepiece.

Objective: SentencePiece de/tokenizer - language independent, used by MarianMT from huggingface.

Ideas: Better performance on unknown words, interesting vocabulary building, really language independent, solves rare words better.

Relations: Used by MarianMT model in from huggingface.

Sennrich et al.: Neural Machine Translation of Rare Words with Subword Units **BPE-DBLP:journals/corr/SennrichHB15**

Rico Sennrich, Barry Haddow, and Alexandra Birch. "Neural Machine Translation of Rare Words with Subword Units". In: *CoRR* abs/1508.07909 (2015). arXiv: 1508.07909. URL: http://arxiv.org/abs/1508.07909.

Objective: Better translations for rare words and open vocabulary. Using BPE for subword translations.

Ideas: NMT model is able to generalize relationship between transparent words and create new (never seen) words.

Relations: Used in SentencePiece tokenizer. Improvement of vanilla BPE.