Kai Nylund
Alessio Tosolini
Elliott Zackrone

Bad Pair Encoding - Exploring the Effects of Novel Tokenization Techniques on Machine Translation

NLP Capstone
May 2024

Background

Byte-pair encoding

- **What's wrong with BPE?**
  - Byte-pair encoding (BPE) replaces frequently occurring byte pairs with a new byte to create token vocabulary.
  - Balance between information of word-level tokenization and flexibility of character-level tokenization.
  - Related work shows BPE is not perfect:
    - BPE causes language imbalance.
    - Tokenizing spaces separately improves accuracy.
- **How can we improve?**
  - Language-specific BPE: Train BPE on dataset of balanced languages.
  - Morpheme tokenizer: Segment words into tokens based on morphemes.
    - Known improvement for Korean NMT [1].

Methodology

Language_specific_tokenizer

"Persoonallinen"

↓ (tokenizer trained on Fi)

["_Persoonalli", "nen"]

↓ (default tokenizer on UNKs)

['__Perso', 'on', 'alli', 'nen']

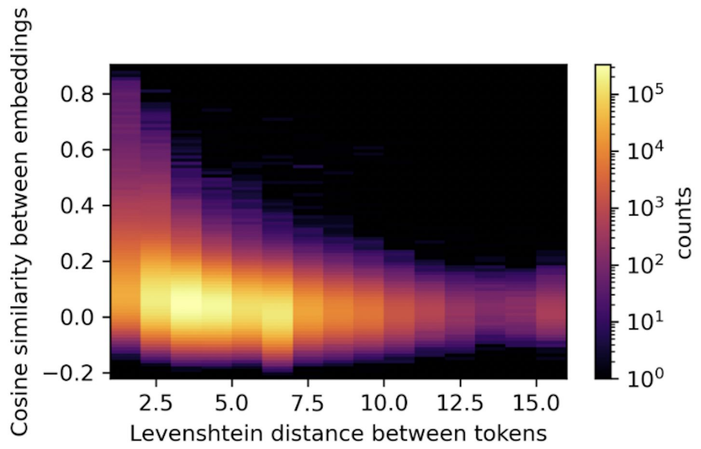
Vs. only using default tokenizer:
['__Perso', 'on', 'allinen']

Morpheme_tokenizer

- Greedy algorithm:
  - Tokenize the word according to tokenization provided in MorphyNet [2].
  - Split the word into the stem and the most frequent potential affix.
  - Use default tokenizer on UNKs.

Results

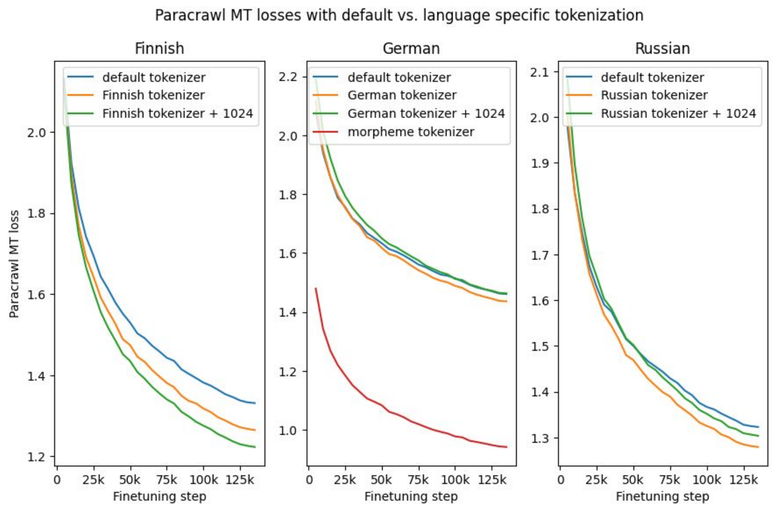
Token_similarity_vs._edit_distance



Substituting_new_tokenizers

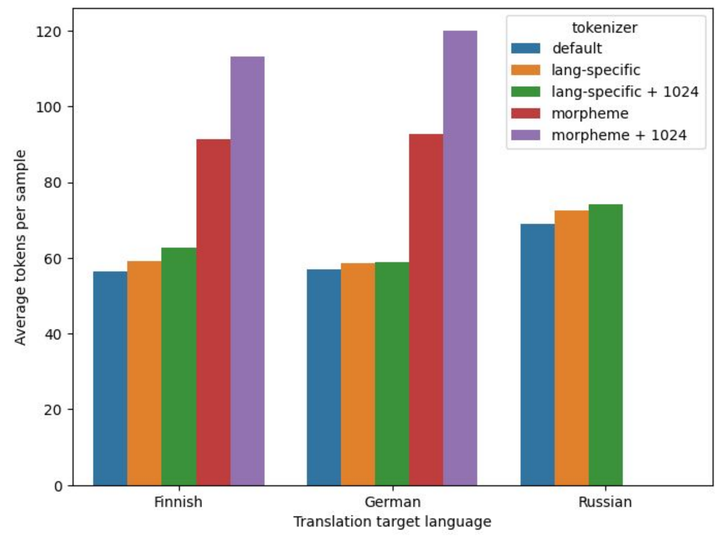
Finetuning Tokenizer	Finnish	German	Russian
Pretrained	20.80 (41.97)	18.77 (39.84)	13.37 (33.28)
Language-specific	20.18 (40.24)	18.68 (39.96)	12.97 (31.60)
Language-specific (Added tokens)	17.80 (38.39)		
Morpheme			N/A
Morpheme (Added tokens)			N/A

BLEU score and chrF++ for each strategy



Results Cont.

Token_segmentation_rates



Discussion and Future Work

Discussion

- Substituting linguistically-informed tokenizers hurt translation performance.
  - BLEU and chrF++ follow segmentation rates better than loss.

Limitations

- Unsure whether decreased accuracy from alternative tokenizers is due to (i) distribution shift or (ii) linguistically informed tokenization not helping

Future_work

- How to insert new tokens into vocabulary?
- Pre-train LLM from scratch using new tokenization strategies.
- Use a variety of more distantly-related languages.

Citations

[1] Kyubyong Park, Joohong Lee, Seongbo Jang, and Dawoon Jung. An empirical study of tokenization strategies for various korean nlp tasks. 2020
[2] Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. MorphyNet: a large multilingual database of derivational and inflectional morphology. In Garrett Nicolai, Kyle Gorman, and Ryan Cotterell, editors Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology,,and Morphology, pages 39–48, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.sigmorphon-1.5. URL https://aclanthology.org/2021.sigmorphon-1.5 .