Language Models for Morphologically Rich Languages

Proposer(s) / Proposatzailea(k): Aitor Soroa, Izaskun Aldezabal

Contact / Kontaktua: a.soroa@ehu.eus

Description / Deskribapena

Pre-trained language models such as BERT have been successful at addressing many natural language processing tasks, specially in languages with many resources. In languages where resources are scarce (e.g. smaller or lower quality corpora), however, the situation is very different. Many of those languages do not have enough resources to properly build a language model. At best, languages are included in multilingual models, where each language shares the quota of substrings and parameters with the rest. Besides, unsupervised sub-word tokenization methods commonly used in these models such as byte-pair encoding are sub-optimal at handling morphologically rich languages.

Goals / Helburuak

In this project we propose to work on language models for Basque. Following recent work on language modeling in morphological rich languages, we propose to try different alternatives and methods to infuse morphological information into the language model architecture. In particular, we propose to:

- Replace unsupervised sub-word tokenization method with sub-tokens derived from a morphological analyzer for Basque;
- Implement a variant of the method described in (Nzeyimana and Niyongabo, 2022; https://arxiv.org/pdf/2203.08459.pdf), tailored for the Basque language.

Requirements / Betebeharrak

The project needs to build language models from scratch, which is usually a complex task that requires a good understanding on deep learning and neural language model techniques. In any case, we plan to start building (comparativelly) small models, which are typically easier to build. We will work towards building a Basque language model, but knowledge of Basque is not strictly required (although recommended).