Prof. Dr. Goran Glavaš,
M.Sc. Fabian David Schmidt
M.Sc. Benedikt Ebing
Lecture Chair XII for Natural Language Processing, Universität Würzburg

2. Exercise for "Multilingual Natural Language Processing"

05.07.2024

1 Paper Readings

We segment the literature on token-level transfer with a focus on Named Entity Recognition as follows:

1. **Datasets**
   - MasakhaNER: Named Entity Recognition for African Languages
   - Toward More Meaningful Resources for Lower-resourced Languages
2. **Translate-train for Token-level Cross-Lingual Transfer**
   - Frustratingly Easy Label Projection for Cross-lingual Transfer

2 Discussions on Datasets and Dataset Quality

A lot of very recent work shows that, for instance, scaling laws of language modelling are significantly improved if the pre-training corpora are of "textbook quality" (cf. Textbooks Are All You Need), or 1000 hand-collected instances can suffice to learn strong alignment (i.e. ChatGPT-style models, cf. LIMA: Less is More Alignment.

Such findings motivate the below questions on how dataset generation affects the quality of the benchmark. With the shift towards ever larger models, an emphasis on dataset quality will only become more important.

1. Briefly explain how the WikiANN dataset has been sourced.

2. Can you think of reasons why the way WikiANN was generated negatively affects
   - The quality of the benchmark in general
   - Suitablity to evaluate cross-lingual transfer
3. How is MasakhaNER created? What are reasons that make MasakhaNER such a valuable benchmark?

3 Translate-Train for Token-Level Cross-Lingual Transfer

`translate-train` refers to approaches that lever state-of-the-art machine translation models to translate training data from English into the corresponding target languages to improve transfer. We will discuss neural machine translation in a later lecture and exercise more specifically and here focus on how translation is relevant for task-transfer.

1. What are approaches to machine-translate token-level annotations, for instance, for Named Entity Recognition?
2. Discuss the approaches. Highlight key considerations that determine whether approaches (beyond “good translation”) are successful!

   As a hint, scan for analyses on the quality of relevant approaches in Frustratingly Easy Label Projection for Cross-lingual Transfer. The authors offer valuable insights relating to both approaches in both the main body and Appendix.

3.1 Discourse (not required for bonus): Translate-Test for Sequence-Level Cross-Lingual Transfer

Thanks to ever improving machine translation, translation-based approaches (re-)gain a lot of popularity. Another paradigm is `translate-test`, in which test instances in the target language are translated to a high-resource language, typically English, in which you perform inference on models trained on high(er) quality annoations than in the target language.

As part of this exercise, we focus on a very recent paper that showcases important developments relevant for cross-lingual transfer.

Reading: Revisiting Machine Translation for Cross-lingual Classification

Elaborate on the key ingredients which the authors lever to materially improve `translate-test` over prior work!

4 Additional Exercises (not required for bonus)

1. Estimate all parameters of an IBM Model 2 (all q_p, q_w) on the given parallel corpus with MLE (using the alignments). Compute the probability your model assigns to the three translations given in the corpus (given the alignment).

Source	Target	Alignment (Target, Source)
the dog	Hund der	(1,2), (2,1)
the dog	der Hund	(1,1), (2,2)
the dog	die Hund	(1,1), (2,2)