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.

Pan et al. generate "silver-standard" named entity annotations "by transferring annotations from English to other languages through cross-lingual links and KB [knowledge-base] properties, refining annotations through self-training and topic selection, deriving language-specific morphology features from anchor links, and mining word translation pairs from cross-lingual links."

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

   - General issues, WikiANN –
     a) is very dense in mentions, i.e., few entities with lot of references
     b) contains many "sentences" not ending in periods (which are likely not actually sentences at all)
     c) has a high number of "sentences" that consist of only a single mention
   - Misalignment in labels muddies their definition, e.g., {*Independently released*, *If I were a boy*, *List of books written by teenagers*} are annotated as **organization**" and span issues (i.e., named entities exceeding the original span such that verbs might get attributed to persons
   - Shallow cross-lingual alignment from knowledge-base links and word-translation mining

Pekin	Beijing
Pekin metrosu	Beijing Subway
Pekin Ulusal Stadyumu	Beijing National Stadium

     Bootstrapping from word translations substantially narrows coverage to specifities of each language
   - Stratified sampling to a standardized number of sequences (i.e., 100, 1000, 10000) discards a lot of valuable data

   The quality issues of WikiANN, for instance, materialize in the fact that there generally is very little performance improvement associated with larger models: When fine-tuning on the English training portion, XLM-R-Large hardly better transfers zero-shot to other languages than XLM-R-base.

   The key take-away is that for any type of machine learning task, the quality of

data matters. This extents to both pre-training and fine-tuning.

3. How is MasakhaNER created? What are reasons that make MasakhaNER such a valuable benchmark?

The data was obtained from local news sources to ensure relevance of the dataset for native speakers from those regions. The dataset was annotated using the ELISA tool (Lin et al., 2018) by native speakers who come from the same regions as the news sources and volunteered through the Masakhane community.

- Covers language vastly underrepresented in NLP for both pre-training and fine-tuning
- Sizable number of high quality human annotated instances
- High densitiy of entities in every single sequence

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?

a) Alignment-based projection: the source-language sentence is first translated, and then words are aligned to map labels from the original to the translated sentence

b) Mark-Then-Translate: add (e.g. HTML) tags around each labelled token and hope machine translation preserves the tags to circumvent post-hoc word alignment

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.

- The quality of “Mark-then-translate” highly depends on the quality of the neural machine translation model, e.g., the authors of "Frustratingly Easy Label Projection for Cross-lingual Transfer" found "fine-tuning [...] improve the projection rate on TyDiQA dataset from 70% to 96.4% maintaining the translation quality"

  In other words, off-the-shelf open source neural machine translation models may not retain tags without appropriate adaptation.

  Professional translation services like Google Translate are very robust to added tags and yield translations that are highly useful for `translate-train`.

- Alignment-based projections rise or fall with the quality of the alignment. The primary issues that *good* word alignments between English-to-many languages to this day remains very difficult. Translating without tags might yield translations that are cannot easily be mapped word-by-word back to the source-language.

  This is particularly true for low-resource languages, where it is difficult to source word-alignments as large-scale dictionaries are not broadly available. Today’s focus shifted more towards collecting parallel data (i.e., sentence pairs that are mutual translations of one another).

  Modern word-alignment thus frequently bases on sentence transformers (to be discussed) that implictily align a pair of translations well on the word-level for alignment.

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!

- **Better translation models**: the original translations by today's standards are subpar and state-of-the-art models produce much better performance to translate to better inference
- **Account for distribution shifts**: Using the original human-generated data does align well with what models see at (translate-)test time. Machine translation models introduce what is referred to as "translationese" (e.g., copying certain words from the input, etc.). Hence, models need to be trained on both original and translated data.
- **Monolingual models**: rather than evaluating `translate-test` on multilingual models, evaluate the approach on stronger, focused monolingual models (RoBerta & DeBerta). The model's tokenizers are tailored to English which typically improves performance vis-a-vis multilingual models by slightly less than 2 points.

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)

$q_w(der|_)) = 0$
$q_w(die|_)) = 0$
$q_w(Hund|_)) = 1/3$
$q_w(der|the) = 2/3$
$q_w(die|the) = 1/3$
$q_w(Hund|the) = 0$
$q_w(Hund|dog) = 3/3$
$q_w(der|dog) = 0$
$q_w(die|dog) = 0$

$q_p(2|1) = 1/3$
$q_p(1|2) = 1/3$
$q_p(1|1) = 2/3$
$q_p(2|2) = 2/3$
$q_p(0|1) = 0$
$q_p(0|2) = 0$

$p(t_1|s_1, a_1) = q_p(2|1) * q_w(Hund|dog) * q_p(1|2) * q_w(der|the) = 1/3 * 1 * 1/3 * 2/3 = 2/27$