Sentence Splitter
Multilingual Natural Language Processing
Homework 2

Ercoli Fabio Massimo
802397

Della Porta Nicolò
1920468

1 Introduction

Quoting [Redaelli and Sprugnoli, 2024] "Sentence splitting, that is the segmentation of the raw input text into sentences, is a fundamental step in text processing". According to [Frohmann et al., 2024] the main challenges are:

- robustness to missing punctuation
- effective adaptability to new domains
- high efficiency

According to [Redaelli and Sprugnoli, 2024] we can add to the list:

- multilinguality

Because a sentence splitting that works well for English may not work well to split another language.

In this project we implemented two models for sentence splitting, using an Italian corpus as train and validation set. The first one is based on an embedding model, the second one is based on generative LLM. We want to analyze compare the two approaches. We want also to test the models out of domains.

2 Methodology

2.1 Embedding-based

We fine tuned a pretrained embedding model (multilingual or trained on Italian language) using the test and validation datasets provided by the homework guide.

The original dataset provides two large texts (one to use as train and the other as validation) together with the golden labels to mark the end of sentence (1) and all the rest of the words (0).

In order to make the datasets suitable for the training we had to apply some transormations. We needed to group words into sequences of tokens aligning the golden labels consistently. The number of tokens of each sequence must fit the max length of the embedding models, for instance 512.

We generated different datasets using different number of words to generate each sequence: 64, 128, 192, 256. Notice that the number of tokens for each sequences will be strictly grater, since for each word the tokenizer of the model will produce one or more tokens.

Before to use the datasets we still need to align (as mentioned above) the labels to the tokens. The alignment strategy we applied consists in keep 1 as the first token generated from a word having label 1, use 0 for all the other cases.

One aspect that we had to address was the fact that the label distribution is very unbalanced for this use case. Most the labels are 0s, and few 1s. We applied 3 different ideas.

First of all, we set the *load_best_model_at_end* as training argument to *true*, using *metric_for_best_model* to F1. This to implement an early stopping based on F1, and not on the accuracy (that would be misleading for such a unbalanced dataset).

Second of all, we override the loss function to weight much less (1/30) a miss-labeling on 0s then the ones on 1s. We called it a weighted trainer.

2.2 LLM-based

3 Experiments

explain your experimental setup; here you give technical details on models, resources, etc. and how you implemented what you described before

Test the model(s) on the OOD. Building a sentence splitter that actually works Out-of-Domain (ODD) is a non-trivial task in NLP.

3.1 Results

present the results referencing tables and commenting them, i.e., why you accept or refuse your start-

ing hypothesis and how you explain the (most notable/unexpected) results.

Comparative analysis of the two approaches on the validation set – which one is better?

The LaTeX related items are [Frohmann et al., 2024].

The LaTeX related items are [Redaelli and Sprugnoli, 2024].

A Training results Appendix

References

Markus Frohmann, Igor Sterner, Ivan Vulić, Benjamin Minixhofer, and Markus Schedl. 2024. Segment any text: A universal approach for robust, efficient and adaptable sentence segmentation. *Preprint*, arXiv:2406.16678.

Arianna Redaelli and Rachele Sprugnoli. 2024. Is sentence splitting a solved task? experiments to the intersection between NLP and Italian linguistics. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 813–820, Pisa, Italy. CEUR Workshop Proceedings.