Report for assignment 3 on NLP course

Student name: Vladislav Urzhumov

Student e-mail: v.urzhumov@innopolis.university

Nickname in CodaLab: v_like

Github: link

Solution 1:

Simple yet reliable approach of dictionary construction. Dictionary consists of entities with corresponding labels counted. That means the counter is created for each entity in the train set and then the appearance of labels is counted, then, with the help of constructed vocabulary, dev and test sets are processed by taking the most common label for each ngram present in the vocabulary. So, this is ngram-based approach: we look at ngrams to extract nested entities.

Solution 2:

Solution is based on the pre-trained large-language model with proper prompt-engineering, embracing one-shot or few-shot prompting techniques. LLM is not re-trained nor fine-tuned, instead it is given the sentence, the set of labels (they are named in the way that is quite easy to understand), and one to few examples of sentence plus extracted entities.

Since LLMs are usually quite bad in terms of working with numbers, we only ask to extract the token sequences and label, and then we convert it to the format required for the submission (start and end indices of the entity, label).

Results:

Solution 1 happened to be better in terms of F1-score due to non-specificity of LLM model. Yet, both solutions are a subject to upgrade with heuristics in the future, if needed.

Thank you for the attention to this solution report!