Name: Mostafa Kira

Email: m.kira@innopolis.university

Nickname in CodaLab: MostafaKira

Github Repo: https://github.com/MostafaKhaled2017/NLP-Assignments

First Solution

The initial approach I tried involved using a SpaCy-based model tailored for the Russian language. I started by sorting the entities based on their length, which prioritizes more significant entities that potentially contain others. Next, I adjusted entity boundaries to align correctly with token boundaries essential for model training. To manage the challenge of overlapping entities, a filtering method was implemented to exclude any entity that overlapped with spans already assigned to another entity. This preprocessing ensured that the training data was free from conflicts, allowing the model to train on clear, distinct entity boundaries.

Second Solution

Alternatively, I explored the capabilities of a BERT-based model, leveraging its state-of-the-art architecture for NER tasks. BERT's deep learning framework is highly adept at understanding context in text, making it a strong candidate for recognizing nested and overlapping entities. I employed a pre-trained BERT model fine-tuned on our specific dataset. This method focused on leveraging contextual embeddings to discern and classify entities at multiple levels without the need for explicit preprocessing to handle overlaps.

Best Solution

Comparatively, the BERT model emerged as the more effective approach, as it achieved a higher F1 score. Its ability to integrate deep contextual understanding allowed it to outperform the SpaCy model, particularly in scenarios involving complex nested entities.