Report on NMT + Label Projection

Method description

One of the approaches to the multilingual NER is to train the NER model only on one language, having sufficient amount of labelled data for training, and then translate the texts from the other languages to the model's language, obtain NER tags for the translation, and finally project the labels to the original language via some language alignment model.

This experiment was intended to make a comparison between the method described above and the direct application of a multilingual model.

Thus, we have two different methods.

The first method directly performs NER on the Russian sentence, the `deeppavlov` NER model based on `ner_ontonotes_bert_mult_torch` is used.

In the second model, the Russian sentence was translated to English using the `EasyNMT` model based on `mBART50`. Next, the `deeppavlov` NER model based on `ner_ontonotes_bert_torch` runs on the translated sentence, and the returned labels are then aligned with the original words by using the token-level awesome-align.

Obtained results

For both methods, the precision, recall, and f1-score were measured for each NER tag (except for `GRP`, as there is no equivalent tag in the pretrained `deeppavlov` NER model). Here is the comparative table with the results:

	CORP		**LOC**		**PER**		**CW**		**PROD**		**O**	
Model	**1**	**2**	**1**	**2**	**1**	**2**	**1**	**2**	**1**	**2**	**1**	**2**
Precision	0.42	0.16	0.38	0.23	0.56	0.37	0.44	0.32	0.52	0.27	0.91	0.91
Recall	0.30	0.24	0.39	0.39	0.45	0.5	0.20	0.32	0.13	0.04	0.96	0.89
F1	0.35	0.19	0.39	0.29	0.50	0.43	0.28	0.32	0.20	0.07	0.94	0.9

Furthermore, I have conducted a brief mistakes analysis for the second model. 200 randomly selected sentences from the dev set were run through the model. There are three possible sources for mistakes: inaccuracies introduced by translation model, NER model, or

alignment model. Here is the table representing the ratio of mistakes coming from each source:

	Translation	**NER model**	**Alignment**
Ratio	0.25	0.32	0.1

Also, the labelling of the dev set is weird: sometimes there are entities that should be tagged, however, they are not. As an example, in the sentence *“Марта вышла замуж за наследника трона Швеции Биргера в 1298 году” Биргер* is tagged as PER, however, *Марта* is tagged as O. Such inaccuracies were detected in 26% of the dev data.