BERT-BASED NAMED ENTITY RECOGNITION AND RELATION EXTRACTION FOR BUSINESS DOCUMENTS IN RUSSIAN

Pauls A.E. (aleksey.pauls@mail.ru), NSU
Berezin S.A. (sergeyberezin123@gmail.com), NSTU

Abstract

Named Entity-Recognition (NER) and Relation Extraction (RE) are one of the most demanded by business tasks of natural language processing, the basis for many solutions. Today there are many researches devoted to solving this problem on academic corpora of texts, which are often far from the typical business setting. The aim of the work is to compare the NER and RE methods in texts in Russian using a close to reality task.

Key words: NLP, BERT, NER, RE

1. Introduction

Named Entity-Recognition (NER) is a task of highlighting words or phrases in the text that indicate an object or phenomenon of a certain category, for example, names of organizations, names of people, etc. Selected entities often have semantic relationships (for example, “demand has grown”) — the detection of such relationships is the task of extracting relationships between them (Relation Extraction, RE).

Being intuitive to people these tasks have been beyond the power of automated systems for long. For years, the best solutions were based on a set of rules made by hand or in automatic way [1]. The using of recurrent neural networks (Lample et al., Ma & Hovy) [19, 20] and models based on vector representations of words like «word2vec» [2] and «GloVe» [3] was a significant breakthrough. However, a quantum leap in this area occurred only with the advent of language models based on deep neural networks with the attention mechanism [4].

For the Russian language, the task of solving this problem is significantly complicated by the small number of prepared text corpora. Moreover, the existing data corpora are quite far from a typical business setting for the following reasons:

- Firstly, relationships are highlighted in the text very tightly. On the contrary, in business tasks there are often only 1-2 occurrences of relations in sufficiently voluminous texts.
- Secondly, in standard corpora relations of domestic and daily nature are determined (employee-company relationship, owner-property relationship, family relations, facts of birth and death, etc.), whereas in business tasks it is usually required to single out relations that have a specific nature related to the subject area.

The aim of this work is to analyze the application of various methods of NER and RE to Russian texts in a conditions close to practice. The corpus of reports of the Ministry of Economic Development of the Russian Federation with a volume of about 280 million tokens was used for this [5].

2. Related Work

The starting point for the study of the problem of NER in Russian texts is the research of Rinat Gareev et al., “Introducing Baselines for Russian Named Entity Recognition 2013” [6], where the authors present a standardized dataset for training of NER algorithms and describe several basic approaches that have become starting point for further researches.

In 2016, the participants of FactRuEval-2016 [14] competition, devoted to the NER task, proposed many solutions to this problem. For example, Sysoev A. A. and Andrianov I. A. had proposed an approach based on word2vec language model [2].

At the beginning of 2019 the research named “A Deep Neural Network Model for the Task of Named Entity Recognition” by The Anh Le and Mikhail S. Burtsev was published. In this research authors describe an approach based on CharCNN-BLSTM-CRF, which showed notable results [7].

During BSNLP-2019 [13] significant progress was achieved by Tatiana Tsygankova et al. using the approach based on BiLSTM-CRF with embeddings obtained by the BERT multilingual model and by Arkhipov et al. with a modified BERT model, which uses CRF as the top-most layer [16].

3. Model

The proposed models for NER and RE are based on the BERT [8] architecture. BERT is a language model, it's training is carried out on two tasks simultaneously: the restoration of masked words in a sentence and the classification of sentences based on logical connectivity (whether the second sentence is a continuation of the first or not). During training in solving these two problems, the language model studies the syntax and semantics of the language, which allows it to be used to create informative vector representations of words (embeddings).

Having received vector representations for words and / or text, we can apply them to the input of another model that solves a specific problem (for example, classifications), continue to teach the entire model as a whole or to train only the added model by freezing the weight of BERT itself.

3.2 BERT NER

The BERT "head" for solving the NER task is a seq2seq classifier that receives words embeddings and returns BIO (Begin-inside-outside) entities tags corresponding to these words.

First, we have word embeddings which are the main words features in sequence and which are obtained using BERT. Further, we use BiLSTM block that is used to highlight additional dependencies in the sequence. The next step is a Linear layer designed to increase the expressiveness of the network. Finally, we use CRF layer for labeling. The architecture of the model is illustrated in Figure 1.

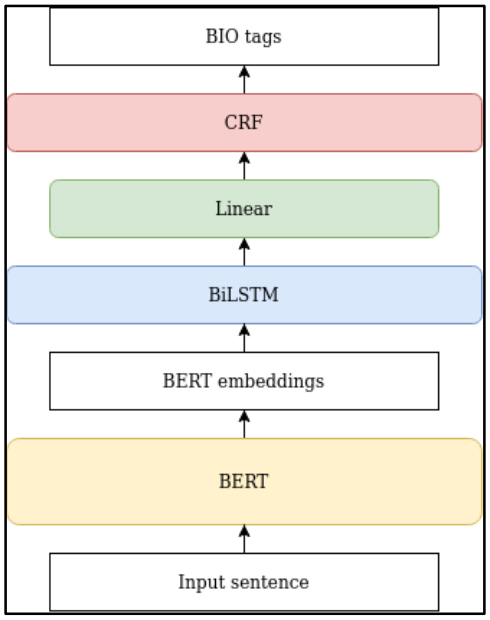

Figure 1 — BERT NER architecture

We also considered many other classifier variations, for example, we replaced CRF with Time-distributed layer and BiLSTM with CNN — such models were also tested during our experiments.

3.3 BERT RE

To solve the RE problem, a classifier model has been used. At the input it receives a sequence of words and bit masks that highlight specific entities in this sequence. These masks are used to obtain embeddings of entities by averaging the corresponding token embeddings. Also, as for a sentence embedding, a special token [CLS] embedding is taken, which accumulates information about the sentence as a whole. Embeddings of entities and sentences are used as features to classify the semantic relationship between entities. First, they are passed through the linear layers separately, then they are concatenated into one tensor, which passes through the another linear layer. At the output of the model, the Softmax layer returns the probability that the relationship belongs to each class. The architecture is shown in Figure 2.

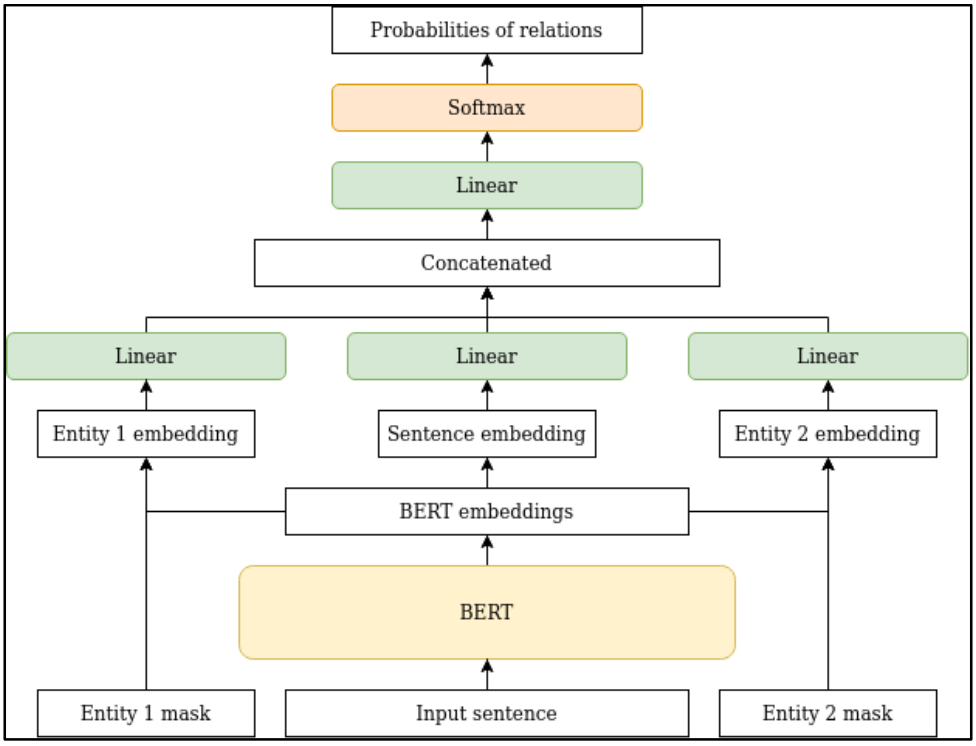

Figure 2 — BERT RE architecture

4. Experiments

4.1 BERT pre-training

We took the RuBERT — fine-tuned on the Russian Wikipedia and Russian news articles datasets BERT models as s basis [9]. We took the RuBERT weights as initial and fine-tuned model on the text corpus of RuREBus competition.

According to the results of training at 72 epochs, which took about 20 hours, the loss function of the model fell from 4.2823567, to 0.60947275.

In this way, we got a model adapted not for the Russian language as a whole, but specifically for the formal clerical language of state reports.

4.2 Vocabulary generation

A dictionary of 119547 tokens was generated using the SentencePiece tokenizer [17] in its BPE version. After this dictionary has been converted to the WordPiece format — it is expected by the BERT model for input. It was decided to keep the case of words and not to transfer all characters to lower case — this is motivated by the specifics of the task: named objects often start with a capital letter, and this can serve as an important feature.

4.3 Training data

For training the corpus of reports of the Ministry of Economic Development of the Russian Federation with a volume of about 280 million tokens was used. This corpus represents various reports of regional authorities about work done and planned activities, as well as forecasts and plans for the future. A certain subset of the corpus is marked with special named entities (8 classes) and semantic relations between them (11 classes).

The markup competed in the brat standoff format [18] and includes the number of the entity or relationship, the type, the position of first and last characters of the entity or an enumeration of the arguments of the relationship. Data example is shown in Figure 3.

```
T127    SOC 6730 6742   правопорядка
T128    BIN 6955 6965   реализации
T129    BIN 7009 7019   достижение
T130    INST 7198 7214  Правительства РК
T131    INST 7387 7403  Правительства РК
R1      GOL Arg1:T3 Arg2:T4
R2      GOL Arg1:T3 Arg2:T5
R3      GOL Arg1:T70 Arg2:T6
R4      GOL Arg1:T70 Arg2:T7
R5      TSK Arg1:T10 Arg2:T11
```

Figure 3 — data example

4.4 Data augmentations for RE

An important stage in the preparation of data is its augmentation. We introduce inverse relations for asymmetric relations (in our case, all relations are asymmetric) and take the corresponding sentences with the reverse word order. Such sentences often have the wrong syntax, but still contain useful information. The model is trained on a unified set of relations, and during testing, the direct and inverse classes are considered one class.

4.5 Model parameters

In all models size of linear and other layers set to 128, the maximum length of the input sequence is 128 tokens. Each model was trained for 50 epochs using the AdamW optimizer with learning rate 5e-5 and linear scheduler.

5. Results

The model's results on the test data are presented in Table 1. We compare our results with the results of the RuREBus competition participants. Despite the fact that the result was obtained on the same test data set, our result was not obtained during the official evaluation.

	NER					RC
	Linear	LSTM-Linear	LSTM-CRF	CNN-Linear	CNN-CRF	Base
Our results	0.3	**0.306**	0.307	0.284	0.296	**0.39**
davletov-aa	**0.561**					0.394
Sdernal	0.464					**0.441**
ksmith	0.464					0.152
viby	0.417					0.218
dimosolo	0.4					—
bond005	0.338					0.045
Student2020	0.253					—

Table 1 — results comparison

6. Future work

In the process of working on the solution, we identified the following ways to improve:

1. Using more complex classifiers — for example, Bayesian Neural Network for the NER task [10]
2. Using more advanced multitask learning based language models, e.g. ERNIE [11]
3. The use of data augmentation using a language model to evaluate learning examples generated automatically from existing ones [12]

7. Conclusion

We considered the modern models for solving NER and RE problems within the problem that is close to real. The models showed a good result and can be improved and used to solve other problems.

References

[1] Georgios Petasis, Frantz Vichot, Francis Wolinski, Georgios Paliouras (2018) Using Machine Learning to Maintain Rule-based Named-Entity Recognition and Classification Systems, available at: https://www.researchgate.net/publication/220875078_Using_Machine_Learning_to_Maintain_Rule-based_Named-Entity_Recognition_and_Classification_Systems
[2] Sysoev A. A., Andrianov I. A. (2016) Named Entity Recognition in Russian: the Power of Wiki-Based Approach, available at: http://www.dialog-21.ru/media/3433/sysoevaaandrianovia.pdf
[3] Pennington Jeffrey, Socher Richard, Manning Christopher (2014) Glove: Global Vectors for Word Representation, available at: https://www.aclweb.org/anthology/D14-1162.pdf
[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin (2017) Attention is all you need, available at: https://arxiv.org/pdf/1706.03762.pdf
[5] Ivanin, V., Artemova, E., Batura, T., Ivanov, V., Sarkisyan, V., Tutubalina, E., & Smurov, I. (2020). RuREBus-2020 Shared Task: Russian Relation Extraction for Business. In Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog"].
[6] Rinat Gareev, Maksim Tkachenko, Valery D Solovyev, Andrey Simanovsky, Vladimir Ivanov (2013) Introducing Baselines for Russian Named Entity Recognition 2013, available at: https://www.researchgate.net/publication/262203599_Introducing_Baselines_for_Russian_Named_Entity_Recognition
[7] Anh Le, Mikhail Burtsev (2019) A Deep Neural Network Model for the Task of Named Entity Recognition, available at http://www.ijmlc.org/vol9/758-ML0025.pdf
[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, available at: https://arxiv.org/pdf/1810.04805.pdf
[9] Kuratov, Y., Arkhipov, M. (2019) Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language, available at: https://arxiv.org/pdf/1905.07213.pdf
[10] Vikram Mullachery, Aniruddh Khera, Amir Husain (2018) Bayesian Neural Networks, available at: https://arxiv.org/ftp/arxiv/papers/1801/1801.07710.pdf
[11] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, Qun Liu (2019) ERNIE: Enhanced Language Representation with Informative Entities, available at: https://arxiv.org/pdf/1905.07129.pdf
[12] Jonathan Rotsztejn, Nora Hollenstein, Ce Zhang (2018) ETH-DS3Lab at SemEval-2018 Task 7: Effectively Combining Recurrent and Convolutional Neural Networks for Relation Classification and Extraction, available at: https://arxiv.org/pdf/1804.02042.pdf
[13] Jakub Piskorski, Laska Laskova, Michał Marcinczuk, Lidia Pivovarova, Pavel Priban, Josef Steinberger, Roman Yangarber (2019) The Second Cross-Lingual Challenge on Recognition, Normalization,Classification, and Linking of Named Entities across Slavic Languages, available at: http://bsnlp.cs.helsinki.fi/shared_task_BNSLP_2019.pdf

[14] Starostin A. S., Bocharov V. V., Alexeeva S. V., Bodrova A. A., Chuchunkov A. S., Dzhumaev S. S., Efimenko I. V., Granovsky D. V., Khoroshevsky V. F., Krylova I. V., Nikolaeva M. A., Smurov I. M., Toldova S. Y. (2016) FactRuEval 2016: Evaluation of Named Entity Recognition and Fact Extraction Systems for Russian, available at: http://www.dialog-21.ru/media/3430/starostinaetal.pdf
[15] Tatiana Tsygankova, Stephen Mayhew, and Dan Roth (2019) BSNLP2019 shared task submission: Multisource neural NER transfer. In Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing. Association for Computational Linguistics.
[16] Mikhail Arkhipov, Maria Trofimova, Yuri Kuratov, and Alexey Sorokin. 2019. Tuning multilingual transformers for language-specific named entity recognition. In Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing. Association for Computational Linguistics.
[17] Taku Kudo, John Richardson (2018) SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing, available at: https://arxiv.org/pdf/1808.06226.pdf
[18] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou and Jun'ichi Tsujii (2012). brat: A Web-based Tool for NLP-Assisted Text Annotation. In Proceedings of the Demonstrations Session at EACL 2012.