Joint task learning for relation extraction and named entity recognition

Authors

Institution

In this work we present our system for RuREBus challenge held together with Dialog 2020 conference. The task consisted of 3 subtasks: named entity recognition, relation extraction with provided named entity tags and end-to-end relation extraction. Our system took the first and the second place in the first and the second subtasks respectively. For the second subtask we submitted our solution only in the post-evaluation phase, however, it was among top 2 best performing systems. The systems for all tasks are based on Transformer models. Relation extraction was solved as a sequence labelling problem. We also used joint task named entity and relation extraction learning [1].

Key words: Relation Extraction, Named Entity Recognition, Transformer, BERT

Совместное обучение моделей для извлечения отношений и именованных сущностей

Авторы

Организация

В данной работе мы представляем нашу систему для соревнования RuREBus, проводящегося совместно с конференцией Dialog 2020. Задача состояла из 3 дорожек: распознавание именованных сущностей, классификация отношений между заранее аннотированными именованными сущностями и извлечение отношений из неаннотированного текста. Наша система заняла первое место на первой дорожке и второе место на третьей. Для второй задачи мы не успели своевременно представить решение, но оно бы оказалось в числе лучших систем. Системы для всех задач основаны на моделях Transformer. Извлечение отношений мы рассматривали как задачу разметки последовательностей. Также мы использовали совместное обучение для задач распознавания именованных сущностей и извлечения отношений.

Ключевые слова: извлечение отношений, распознавание именованных сущностей, Transformer, BERT

1 Introduction

This work is devoted to RuREBus challenge held together with the conference Dialog 2020. RuREBus competition was devoted to the problem of relation extraction and named entity recognition (NER) in a specialized business domain. The competition consisted of three subtasks: named entity recognition, relation extraction with provided named entity labels and end-to-end relation extraction. Our first subtask solution was a BERT-based [2] sequence labelling model. For the second one we applied joint named entity and relation extraction learning. We went with a similar approach for the third subtask. However, due to having no labelled named entities, they were inferred using the model trained for the first subtask.

Our NER model with the 0.561 F1-score at the test dataset took the first place in the competition. Our second subtask model took the second place with the F1-score equal to 0.394.

[1] https://github.com/AdisDavletov/DeftEval2020/tree/dev

Our work shows that the sequence labelling approach is viable for relation extraction. It also demonstrates that correct named entity labels are vital for relation extraction due to the difference in scores between the second and the third subtask models.

2 Related work

There are many ways to extract information from text. This task is often solved by extracting named entities and classifying relations between them. One of the most popular datasets for this task is TACRED [9] where semantic relations are understood as relations between two pairs of entities.

Nowadays, state-of-the-art results for this dataset are achieved with Transformer-based models [7]. The most advanced models (according to paperswithcode[2]) use extra training data or additional knowledge bases. For example, in [1] the authors use Wikipedia data. However, such data is useless for domain-specific relations.

Among the systems that do not use encyclopedias or other labeled data, the best results were achieved by Joshi et al. [4]. They pre-trained a BERT-like system, but instead of predicting individual masked tokens they trained the model to infer contiguous random spans. The model was also trained to predict each token in the masked span using output representations of only span boundary tokens. This significantly improved results of their model in comparison with the vanilla BERT. As in both described works we also incorporated information about named entity spans.

However, it is difficult to compare results for relation extraction systems for languages besides English because such annotated datasets are scarce for most languages including Russian. Some researchers have tried to solve this problem using unsupervised language-agnostic approaches and relying on knowledge databases such as Wikidata and various online encyclopedias such as Wikipedia [3]. Models trained this way tend to be not specialized because the original database does not contain relations from the required domain. The results are good only for the most popular relation types such as geographical or professional ones, which frequently appear in Wikipedia.

3 Shared task overview

The organizers of the competition have provided approximately 300 annotated texts in total. All texts were provided by the Ministry of Economic Development of the Russian Federation. The corpus consists of various regional and strategic plan reports. There are in total 8 named entity classes and 11 semantic relation classes (see Tables 1 and 2). The organizers have also provided a large unannotated dataset for language model fine-tuning. However, we did not use it. A named entity can consist of several words. All entities and relations do not span across sentences. There may be many-to-many, many-to-one and other types of relations (see Fig. 1).

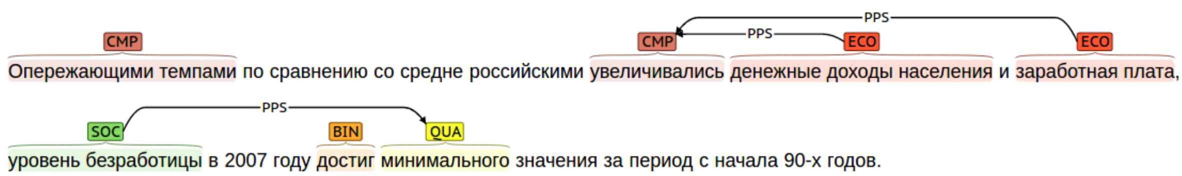

Figure 1: RuREBus annotation example.

[2]https://paperswithcode.com/sota/relation-extraction-on-tacred

Named entity groups could contain rather broad types of entities, for example "SOC" entities contained social groups as well as various social attributes - phrases like 'blue collar workers' and 'housing accessibility' corresponded to this group.

Type	Description	Examples
MET	Some quantitative metric	доля сельского населения (rural population ratio); положение в округе (ranking in the neighbourhood)
ECO	An economy entity or facility	обрабатывающим сектором промышленности (processing industry); экономического кризиса (economic crisis)
BIN	A binary attribute	входит в состав (is part of)
CMP	Comparative attribute	рост (growth); увеличился (increased); в наибольшей степени (to the greatest extent)
QUA	Qualitative attribute	лидирующее (leading)
ACT	Activity, actions, implemented policies	восстановление экономики региона region economy reconstruction
INST	Institutions and organizations	Алтайского края (Altai region); Сибири (Siberia)
SOC	Social groups and characteristics	населения края (region population); здравоохранение (health care)

Table 1: Named entity types

Group	Type	Description
Current state of affairs	NNG	now negative
Current state of affairs	NNT	now neutral
Current state of affairs	NPS	now positive
Results	PNG	past negative
Results	PNT	past neutral
Results	PNS	past positive
Forecasts	FNG	future negative
Forecasts	FNT	future neutral
Forecasts	FNS	future positive
Goals	GOL	some abstract goals
Tasks	TSK	tasks and actions performed to achieve goals

Table 2: Semantic relation types

The organizers first held tracks 1 and 3 and after that track 2 was also run. We describe our solutions in the same order (first tracks 1 and 3, then track 2).

4 Named entity recognition and relation extraction as sequence labelling

The data for the competition was presented in brat format [6] where texts were given as plain text files and annotations were provided in another file with mixed labels for named entities and relations between them. Thus, we first had to separate the labels and transform the data into special formats used by our models.

We used Razdel library to split plain texts into sentences and tokens [3]. It is a rule-

[3]https://github.com/natasha/razdel

based system that along with splitting sentences can also provide sentence and token offsets in the source text. Offset ranges provided by Razdel were used during preprocessing and postprocessing to map tags and relations to text spans which are required by the brat format (see Table 3).

Dataset	Number of			
	Sentences	Tokens	NER tags	Original NER tags
train	10460	336023	54377	54388
test	20483	643668	89006	89879

Table 3: Named entity types

4.1 Subtask 1: Named Entity Recognition

The first task was to annotate named entities. First we transformed the data into format where in each example we have pairs of tokens and their corresponding tags. Sentences were separated with newlines. We randomly split the data into training and validation datasets in 0.7 to 0.3 ratio. After hyperparameter tuning we did not retrain the model using the left validation data. We used a BERT-based system [2] with PyTorch model code and pretrained weights provided by Hugging Face [8]. Due to competition being in Russian, we used the multilingual uncased base BERT model.

BERT is a Transformer based model [7]. On top of BERT outputs we added a linear layer with softmax activation function and dropout regularization. The cross entropy loss function was used to train the model. For each word token in the sentence we took a BERT embedding from its first BPE-token and fed it to the dropout layer followed by the linear layer. All non entity tokens were ignored (i.e. padding tokens and tokens describing borders between sentences and various spans).

Method	test	dev
test_xlm_r_tag_0.1_wd_0.2_relation_run_predictions	0.497	0.465
test_xlm_r_tag_0.01_relation_run_predictions	0.33	0.294
test_xlm_r_tag_0.1_relation_drop_0.2_run_predictions	0.489	0.456
⋆ whole_train_tag_0_xlm_r_run_predictions	0.002	0.003
test_xlm_r_tag_relation_run_predictions	0.465	0.463
test_xlm_r_tag_0.1_relation_run_predictions	0.494	0.467
test_xlm_r_tag_0_relation_run_predictions	0.0398	0.0495
test_xlm_r_tag_0.05_relation_run_predictions	0.482	0.440
test_tag_0.1_and_relation_run_predictions	0.189	0.172
test_tag_0_and_relation_run_predictions	0.002	0.001
test_xlm_r_tag_0.2_relation_run_predictions	0.503	0.468

Table 4: Subtask 1: Results on test and development sets on multitask system. ⋆ denotes out of competition result

Our system with 0.561 micro F1-score on the public leaderboard outperformed solutions presented by other contestants.

4.2 Subtask 3: End-to-end Relation Extraction

The second and the third subtasks were relation classification. In the second subtask the organizers provided named entity tags while in the third they did not. For both tracks we used the equivalent approach.

Akin to BERT-multitask learning, in this competition we wanted to experiment with simultaneous finetuning for separate tracks. RuREBus competition provided an excellent framework for this idea because we had separate tracks with different target values but

Method	test	dev
test_tag_0.1_and_relation_run_predictions	0.262	0.757
test_tag_0_and_relation_run_predictions	0.249	0.784
⋆ whole_train_tag_0_xlm_r_run_predictions	**0.391**	0.729
test_xlm_r_tag_0.1_relation_run_predictions	0.378	0.667
test_xlm_r_tag_0.1_wd_0.2_relation_run_predictions	0.356	0.668
test_xlm_r_tag_0.1_relation_drop_0.2_run_predictions	0.354	0.668
test_xlm_r_tag_0.01_relation_run_predictions	0.38	0.677
test_xlm_r_tag_relation_run_predictions	0.27	0.679
test_xlm_r_tag_0_relation_run_predictions	**0.389**	0.678
test_xlm_r_tag_0.05_relation_run_predictions	0.389	0.685
test_xlm_r_tag_0.2_relation_run_predictions	0.339	0.662

Table 5: Subtask 2: Results on test and dev sets. ⋆ denotes out of competition result

the same input data. Thus, we tried a multitask architecture to jointly predict tags and relations. To do so, we consider relation extraction as a sequence labeling problem (similar to how named entity recognition is usually solved). In each example we have one marked main entity and we predict all named entity tags and all relations between the main token and all other tokens in the sentence (see Fig. 2). We put an empty relation label ('0') if a token does not have relation to the marked entity and the relation tag otherwise. Special tokens marking the beginning and the ending of the main entity are added to input to tell the system which entity it should predict relations with. Thus, for each sentence we had to make n predictions where n is the number of named entities in the sentence. We did not relabel previously inferred named entity tags with new predictions.

Sequence labelling might be a preferable solution if we are interested in limiting the number of model calls and our model run time does not depend on the sequence length (unlike recurrent neural networks).

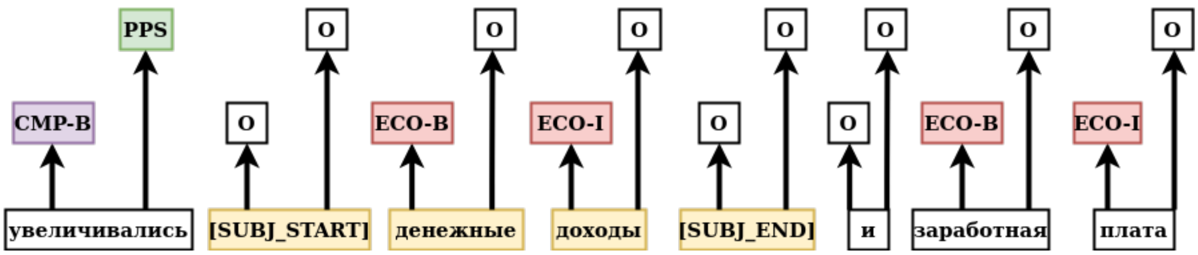

Figure 2: Joint relation extraction and named entity recognition training.

For end-to-end relation extraction we went with a two-stage approach. At first we used the model from the first track to label named entities. After that using the provided named entity predictions we trained our model to infer semantic relations.

In this task we used the same multilingual uncased BERT model as in subtask 1. However, to get simultaneous relation and named entity predictions on top of the model we added another dropout layer followed by tag and relation linear layers. We use weighted sum of cross entropy losses for tag and relation labeling as our final loss for optimization. Padding tokens do not contribute to our loss calculation.

However, joint task learning only worsened our results and the best result at the validation set was

The system showed 0.132 micro F1-score using public test data and it would have taken the first place among the provided systems, if we had managed to submit our solution before the deadline.

4.3 Subtask 2: Relation Extraction for given Named Entities

The model for this track is equivalent to the system used for end-to-end relation extraction. This track was very similar to end-to-end relation extraction. However, instead of using named entity labels predicted by our model, we could use the manual annotation provided by the organizers of the competition.

We also attempted at using the multi-task learning procedure described in previous section. However, as in the previous case the quality deteriorated when the model was trained to predict named entity tags. Thus, the loss coefficient for named entity recognition was also set to zero.

For subtask 2 we also tried a base XLM-RoBERTa [5] model also provided by Hugging Face. RoBERTa is BERT inspired model which optimized many hyper-parameter choices in the underlying model. RoBERTa authors have replaced static masking with random masking during language training. They also removed additional sentence prediction loss, increased the batch size, trained on longer sequences and enhanced the original Wikipedia dataset with various Common Crawl datasets. All these adjustments helped RoBERTa to outperform BERT in many benchmarks such as GLUE or SQuAD 2.0.

5 Results

All in all, our named entity recognition model with micro F1-score equal to 0.561 took the first place in the competition. However, the results are lower than for other named entity recognition datasets (e.g. for the Ontonotes dataset Transformer-based models usually get > 0.85 in F1-score [4]). It can be attributed to the small number of training examples and complexity of the domain.

Our end-to-end relation extraction model despite being one of the best solutions at the competition was much worse than the model trained with manual annotations provided by the organizers. In future we will try to use approaches similar to pseudo labelling where we include only those named entity predictions that have high logit scores instead of all predictions. The difference in results also demonstrates that correct named entity labels are vital for relation extraction.

Multi-task learning did not improve our results for this task as well.

6 Conclusion

In this work we present our system for RuREBus challenge held together with Dialog 2020 conference. The task consisted of 3 tracks: named entity recognition, relation extraction with provided named entity tags and end-to-end relation extraction. All tracks we considered as sequence labelling problems. We show that sequence labelling might be a decent approach for the relation extraction problem. We also attempted to use joint-task learning. However, it did not improve our results on the validation dataset. The system took the first place in the named entity recognition track and the second place in the third track. For the second task we failed to submit the solution till the deadline but it was among the best systems. The systems for all tasks are based on Transformer models.

[4] see http://docs.deeppavlov.ai/en/master/features/models/ner.html

7 Acknowledgments

We would like to thank the organizers of the competition. We believe that their work will be very helpful for the development of natural language processing for the Russian language.

References

[1] Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. Matching the Blanks: Distributional Similarity for Relation Learning. In *arxiv.org*, pages 2895–2905, 2019.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. oct 2018.

[3] Nicolas Heist and Heiko Paulheim. Language-agnostic relation extraction from wikipedia abstracts. In *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, volume 10587 LNCS, pages 383–399, 2017.

[4] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, Omer Levy, and † Allen. SpanBERT: Improving Pre-training by Representing and Predicting Spans. Technical report.

[5] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arxiv.org*, 2019.

[6] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proc. Demonstr. Sess. EACL 2012*, Avignon, France, 2012. Association for Computational Linguistics.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Adv. Neural Inf. Process. Syst.*, volume 2017-Decem, pages 5999–6009, 2017.

[8] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*, abs/1910.0, 2019.

[9] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. Position-aware attention and supervised data improve slot filling. In *EMNLP 2017 - Conf. Empir. Methods Nat. Lang. Process. Proc.*, pages 35–45, 2017.