Relation Extraction from Scientific Texts in Russian with Limited Training Data

Olga Tikhobaeva
Novosibirsk State University / Russia
otikhobaeva10@gmail.com

Elena Bruches
A.P. Ershov Institute of Informatics Systems / Russia
Novosibirsk State University / Russia
bruches@bk.ru

Abstract

In this paper, we address the task of extracting semantic relations between entities in scientific articles in Russian, with a focus on scientific terms as entities. We present a dataset that includes annotated abstracts of scientific articles in Russian. This dataset was used to train and test models and develop an algorithm for the automatic extraction of semantic relations. We conducted experiments and compared one zero-shot and one few-shot approach for relation extraction: one based on the perplexity score and the other based on the use of prototype vectors of relations. Our results show that both methods can achieve reasonable performance, demonstrating the potential of zero-shot and few-shot approaches for relation extraction in scientific texts in Russian. The developed tool and annotated dataset are publicly available and could be valuable resources for other researchers [1].

1 Introduction

At the present time, the proliferation of electronic scientific publications has led to an increasing need for extracting various types of semantic information from scientific texts. One of the types of such information is semantic relations. By extracting these relations, machines can better understand the meaning of a text, and this can have a wide range of practical applications. For instance, relation extraction can be used in search and question-answering systems, as well as in ontology development and text classification.

However, currently, this problem is still difficult for any domain in any language. There are several factors that contribute to the difficulty of this task such as high variability in terms of syntax, grammar, and vocabulary and ambiguity of meanings in the texts. What's more, there is a problem of lack of labeled data, especially for the Russian language. Even though, there are some datasets with annotated relations such as (Zhang et al., 2017; Dunietz and Gillick, 2014; Li et al., 2016) in multi-domains and biomedical domain, it is still hard to find some publicly available datasets such as SciERC (Luan et al., 2018) for scientific fields other than biomedical, and in languages other than English.

Due to the problem of lack of data we decided to concentrate on some zero-shot and few-shot methods. Zero-shot relation extraction is a type of relation extraction that allows a model to identify and extract the types of relations that it has not been specifically trained on. In other words, the model can perform relation extraction in a "zero-shot" manner without any direct supervision for the relation types in question. Few-shot relation extraction assumes that the model is trained on a small set of labeled data. The purpose of this method is to allow the model to generalize to the new tasks based on a few examples.

Thus, we make the following contributions:

- Provide a new dataset for relation extraction tasks for Russian scientific texts.
- Compare one zero-shot and one few-shot approach for relation extraction (based on perplexity score and with the use of prototype vectors of relations).

2 Related Work

Relation extraction (RE) is one of the main tasks in the field of natural language processing (NLP). With the introduction of large language models (Devlin et al., 2019; Liu et al., 2019; Lewis et al., 2020) their use became one of the main methods of solving this problem. However, such methods require a lot of well-annotated data for training. Currently there are no datasets available for this task in a scientific field in Russian, and manual

[1] https://github.com/iis-research-team/terminator/tree/main/relation_extractor

Proceedings of the 2nd Workshop on Information Extraction from Scientific Publications, pages 25–30
Nov 1, 2023. ©2023 Association for Computational Linguistics

annotation takes a long time and requires the efforts of more than one person to objectively label the relations. Therefore, in this paper we decided to pay our special attention to zero-shot and few-shot approaches that do not require a lot of annotated data. There are some examples of them.

The first method is based on the scores of the probability of a sentence that the language model can give. (Henlein and Mehler, 2022) proposed to create a template for each relation type and then compute increased log probability of the sentences from these templates with the use of BERT as in (Kurita et al., 2019). For example, a template for the relation "LOCATED-IN" might look like this – *"the <e1> is in the <e2>"*. So if the first entity is *"toothbrush"* and the second is *"bathroom"*, the sentence from the template will be *"the toothbrush is in the bathroom"*. With the selected threshold of probability, it will be possible to separate the presence or absence of relation between two entities and also its type.

The second method was used in (Zhang and Lu, 2022; Zhang et al., 2022). The primary idea behind this approach is that one can get prototype vectors for each type of relation and then use them to define the relations between pairs of entities. To create a prototype vectors the authors used sentences from the train part of the dataset, as well as the name and the description of the relations. A prototype vector of each relation can be compared with actual sentences that contain the pair of entities. The closest prototype in vector space will reflect the relation in the sentence. In (Zhang et al., 2022) the authors employed BERT (Devlin et al., 2019) as the encoder to map the sentences into a low-dimensional vector space.

Last but not least, (Lan et al., 2022) proposed a third method that trains the model to extract relations from unstructured text, while the train and test sets of relations do not intersect. At first, the model was trained to find the probability for different sets of potential relations from the train dataset and then to find the boundaries of two entities. After that it can process any new texts and does not need to know the types of relations. To find the probability for some relations in the sentence the authors offer to encode semantics of the relation types by given the combined sentence like *"[CLS] text-of-the-sentence [SEP] text-of-the-relation [SEP]"* to BERT. If the model has these sentences for each relation type, it is possible to get the probability distributions over candidate relations.

3 Data Preparation

To conduct the experiments with different approaches we created an annotated dataset which is composed of abstracts of scientific papers on 10 domains in Russian. The list of domains includes the following: Biology and Medicine, History and Philology, Journalism, Law, Linguistics, Math, Pedagogy, Physics, Psychology and Information Technology.

To test the approaches we used 20% of the texts on each of the subject areas.

Statistics for our dataset is presented in Table 1.

Unit	number
texts	400
tokens	17 481
terms	5 834
relations	976

Table 1: Dataset statistics

Each abstract was annotated by two annotators. The task was to classify the relations between each possible pair of terms in each sentence in the abstract. The terms in the texts were already extracted. During the annotation, we followed the instructions proposed in (Bruches et al., 2020).

For our experiments we chose 3 following oriented semantic relations: USAGE, ISA, PART-OF. Those relation were selected because they are common to all considered domains. The types of relations in the corpus, along with their meanings and distribution across the dataset, are provided in Table 2.

Relation type	Meaning	number
USAGE	x is used for/in y	544
ISA	x is y	270
PART_OF	x is part of y	162

Table 2: Types of relations

In Table 3 sample sentences of all three relation types in the dataset are presented. In each sample two terms and the relation between them are highlighted.

The dataset is available for other researchers[2].

[2] https://github.com/iis-research-team/ruserrc-dataset

Relation type	Example	Translation
USAGE	*В статье рассматривается способ <e1>формирования тектовых сообщений</e1> на основе <e2>метода движения губ</e2>, сооветствующего определенной фонеме.*	The article considers a method of <e1>formation of text messages</e1> based on <e2>the method of movements of lips</e2> corresponding to a certain phoneme.
ISA	*Одним из самых точных и эффективных <e1>способов управления жестами</e1> является <e2>управление активностью мышц</e2>.*	One of the most accurate and effective <e1>ways to control gestures</e1> is to <e2>control muscle activity</e2>.
PART_OF	*Метод обработки и определения форм слов позволяет в отличие от аналогов обрабатывать формы слов <e1>естественных языков</e1> различных групп и <e1>семейств</e1>.*	Unlike analogies, the method of processing and definiting forms of words allows to process the forms of words from <e1>languages</e1> of different groups and <e2>families</e2>.

Table 3: Examples of relations

4 Zero-shot and few-shot approaches for relation extraction

4.1 Using perplexity scores

In the first place, we tried an approach for relation extraction based on perplexity scores. It can be traced to zero-shot approaches. It consists in the following: for each type of relation we had made 3 patterns of the sentences. The patterns and their meaning are provided in Table 4.

Then the terms were added to these templates to make sentences. For example, the pattern for USAGE is "{term1} are used in {term2}". So if the first term is "multimedia technologies" and the second is "the educational process", the sentence from the template will be *"multimedia technologies are used in the educational process"*.

Then we got an estimate of the probability of each sentence using the model GPT2 (Radford et al., 2019). After choosing the most probable pattern for each relation, we again compared the probability of sentences from these best templates. The most likely sentence would reflect the true relation between the terms. The schematic work of the method is presented in the Figure 1.

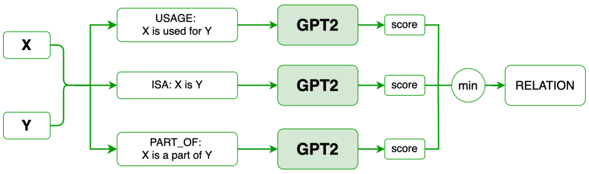

Figure 1: Schema for the perplexity scores approach

To measure the probability we used the perplexity score. In general, this value can be described as the model uncertainty measure when predicting each of the next token, hence the lower the perplexity, the more certain the model in predicting this sequence.

The obtained metrics for this approach are shown in Table 5.

4.2 Using prototype vectors of relations

The second approach for relation identification that we tried is based on the usage of the prototype vectors of relations. It can be attributed to few-shot approaches. First of all, we manually chose 138 best examples from the train part of the dataset to create a prototype vectors for each type of relations. In selecting the best examples we were guided by the following criterion: the example shows only one type of relations and has short context which includes only two terms of interest. Then we got the vectors of these of sentences. Vectors of sentences are the embeddings of CLS token from BERT(Devlin et al., 2019). Each prototype vector is an average of the vectors of sentences reflecting each relation. Once these prototype vectors are obtained, they can be used to classify test examples. By computing the value of the cosine similarity of the example and the prototypes, we can determine which relation is most similar to this example. Schematic graphics that reflect the work of this method can be seen in Figure 2.

The obtained metrics for this approach are shown in Table 6.

However, this method falls short in defining the "ISA" relation type and generally performs most effectively in identifying the "USAGE" relation. There are several reasons for this. First of all, quite often the relations are not expressed explicitly by some specific words or phrases, but with semantics, which are difficult to automatically find and understand in the text. The second reason is the fact

Relation type	Patterns	Meaning
USAGE	x используется для y x применяется для y y выполняется при помощи x	x is used for y x is used for y y is done with x
ISA	x является y x представляет собой y x – это y	x is y x represents y x is y
PART-OF	x является частью y y состоит из x y включает в себя x	x is a part of y y consists of x y includes x

Table 4: Patterns of relations

Relation type	Precision	Recall	F1
USAGE	0.69	0.37	0.48
ISA	0.46	0.38	0.42
PART_OF	0.15	0.41	0.22
macro-average	0.43	0.39	0.37

Table 5: Metrics for perplexiry score approach

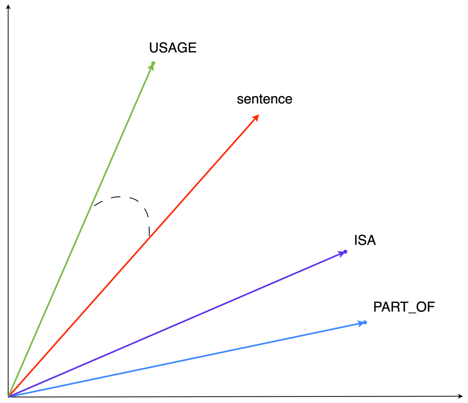

Figure 2: Plot for the prototype vectors approach

Relation type	Precision	Recall	F1
USAGE	0.59	0.81	0.68
ISA	0.00	0.00	0.00
PART_OF	0.22	0.24	0.23
macro-average	0.38	0.51	0.30

Table 6: Metrics for the prototype vectors approach

that all of these relations are expressed in similar contexts. For example, parentheses or colons can associate terms with both "ISA" and "PART-OF" relations. At the same time, the preposition *"в"* (*in*) depending on the terms it links, can express the relation "PART-OF" as well as "USAGE".

5 Classification task with a CLS-vector

To compare the approaches that were specified above with the classic supervised learning method we used the neural network architecture described by the authors in (Wu and He, 2019).

The algorithm of this model is as follows: We use the vector of a special token CLS (which is regarded as the input text vector) and the vector of two terms connected by the relation. These three vectors are concatenated and the resulting vector is fed to the classifier. We used 80% of our annotated dataset to train the model.

The results that we were able to achieve are described in the Table 7.

It is clear that "PART-OF" relation type has the lowest F1-score of all relations. The reason for this is likely to be the lack of examples of this relation in the training data.

Relation type	Precision	Recall	F1
USAGE	0.84	0.95	0.89
ISA	0.83	0.76	0.79
PART_OF	0.58	0.41	0.48
macro-average	0.75	0.71	0.72

Table 7: Metrics for supervised learning

6 Discussions

The results of our experiments show that zero-shot and few-shot approaches are generally able to distinguish semantic relations. But these methods still lose in quality in comparison with the supervised learning. It gives us the understanding that metrics obtained in the experiments are not a limit and there is a space for the research to grow.

For example, we assume that if we add more patterns for the model to choose from in the perplexity score method or put more appropriate examples in the set for prototype vectors this will greatly improve the results.

Of course, still there are some aspects of relation extraction that are extremely difficult to solve. For instance, the extraction of the terms that are not connected by any relation.

7 Future Work

We are definitely going to further develop relation extraction area for the Russian language since it is still low-resource language. Due to the lack of data the Russian language requires adaptation of existing solutions for English or development of brand new ones.

One of the ideas that we are about to thy in the foreseeable future is to translate the sentences from Russian to English and use some good quality method for relation extraction from English text.

It would also be interesting to conduct cross-domain experiments for each of the methods as the annotated dataset has been prepared for a number of disciplines. We are not entirely sure that the results will be representative in all domains because the texts of some of the disciplines have a limited amount of the examples of some relations. But it is still worth to try.

8 Conclusion

This study aimed to address the problem of lack of labeled data for relation extraction in Russian scientific texts by constructing a new dataset. One zero-shot and one few-shot approach for relation extraction were then evaluated, one based on perplexity score and the other utilizing prototype vectors of relations. The experimental results indicated that both methods can achieve reasonable performance, highlighting the potential of zero-shot and few-shot approaches for relation extraction in Russian scientific texts across different domains. These findings suggest that zero-shot and few-shot approaches could be a promising direction for relation extraction research, especially in low-resource languages such as Russian.

References

Elena Bruches, Alexey Pauls, Tatiana Batura, and Vladimir Isachenko. 2020. Entity recognition and relation extraction from scientific and technical texts in russian. In *2020 Science and Artificial Intelligence conference (S.A.I.ence)*, pages 41–45.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jesse Dunietz and Daniel Gillick. 2014. A new entity salience task with millions of training examples. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 205–209, Gothenburg, Sweden. Association for Computational Linguistics.

Alexander Henlein and Alexander Mehler. 2022. What do toothbrushes do in the kitchen? how transformers think our world is structured. *arXiv preprint arXiv:2204.05673*.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Yuquan Lan, Dongxu Li, Hui Zhao, and Gang Zhao. 2022. PCRED: Zero-shot relation triplet extraction with potential candidate relation selection and entity boundary detection. *arXiv preprint arXiv:2211.14477*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016. Baw068.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Shanchan Wu and Yifan He. 2019. Enriching pre-trained language model with entity information for relation classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2361–2364. ACM.

Peiyuan Zhang and Wei Lu. 2022. Better few-shot relation extraction with label prompt dropout. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6996–7006, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.

Yuzhe Zhang, Min Cen, Tongzhou Wu, and Hong Zhang. 2022. RAPS: A novel few-shot relation extraction pipeline with query-information guided attention and adaptive prototype fusion. *arXiv preprint arXiv:2210.08242*.