Application of neural network methods for automatic taxonomy enrichment for the Russian language

Daria A. Pirozhkova

Advisor: Tatiana V. Batura

NSU,

May 2021

Content

- Introduction to the project
- Tasks list
- Project goal and objective
- Existing approaches review
- Data preprocessing
- Task solving methods
- Results

Introduction to the project

A **Taxonomy** is a hierarchy whose tree nodes represent named entities related to other tree nodes with is-a-type-of relationships.

An **Is-A Relation** is a domain independent semantic relation that is a strict partial order relation (**antisymmetric**,**irreflexive**, **transitive**) between a subclass concept and a superclass concept.

Antisymmetric relation	Irreflexive relation	Transitive relation
$\forall\, a, b \in S,$ $if\ R(a,b) \rightarrow True\ and$ $R(b,a) \rightarrow True,$ $then\ a = b$	$\forall\, s \in S: \neg R(s,s)$	$R\ is\ transitive\ relation,$ $if\ \forall a, b, c \in S:$ $R(a,b) \rightarrow True\ and\ R(b,c) \rightarrow True,$ $then\ R(b,c) \rightarrow True$

Introduction to the project

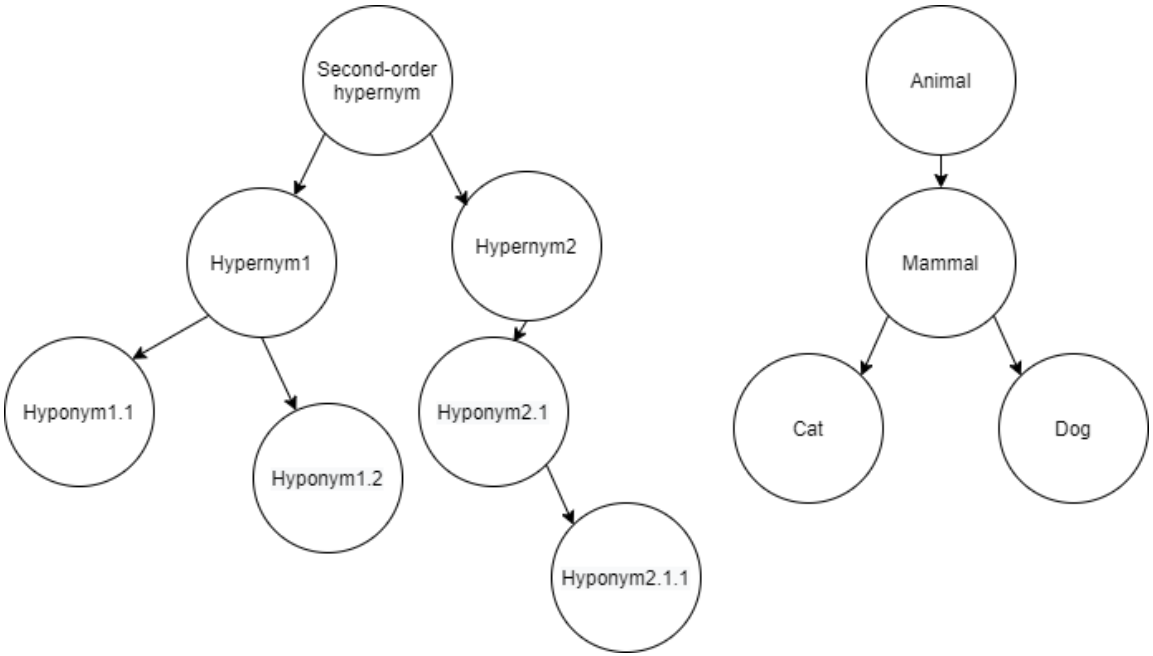

Figure 1. A diagram of Is-A relations between entities

Terms	**Type**	**Terms**	**Type**
Entity1	Hypernym	Entity2	Hyponym
Entity2	Hypernym	Entity3	Hyponym
Entity1	Second-order hypernym	Entity3	Hyponym
Entity3	Hypernym	Entity4	Hyponym

Table 1. A table of Is-A relations between entities

Tasks list

- The main goal and task formulation
- An existing approaches review
- Analysis of thesauri for taxonomy enrichment
- Data understanding and preprocessing
- Implementation the different methods for classification an ISA relation between entities
- Results analysis and summarizing

Project goal and objective

The **goal** is to create methods that automatically enrich any knowledge base with new terms, and at the same time connect them with existing words using Is-A relations.

The **task** is to develop a method for prediction the Is-A relations between entities in texts of the Computer Science category.

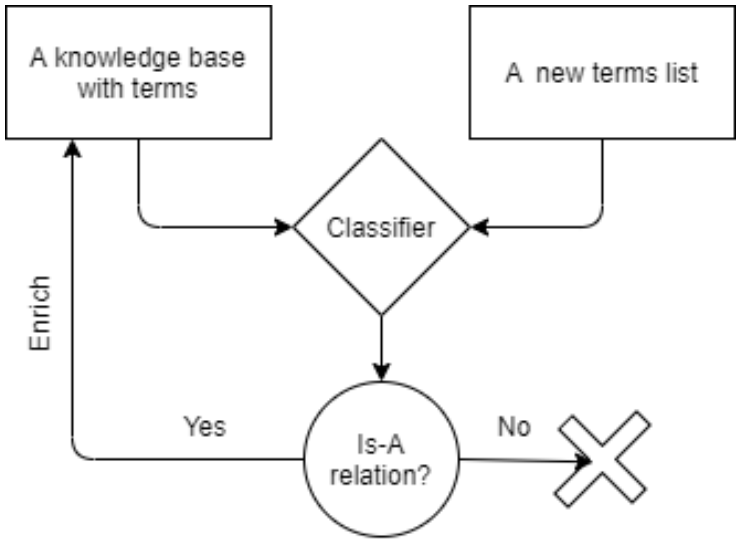

Figure 2. Task solving diagram.

Existing approaches review

English language	Russian language
The most approaches based on the vectorization, classification, and clustering.	The Russian language is complicated due to some properties of the language, namely, a highly inflectional morphology.
The hybrid approaches are better works than single method.	The graph-based method, using dictionaries and search engines, KNN algorithms, language models, and neural networks.
The solution for domain-specific data is better work that the one for language in general.	
Neural networks are used less than other methods. SVM+Word2Vec ELMo CNN + BiLSTM	
The results of proposed methods do not get good result for practical usage.	

Table 2. Summarizing the recaps of a review of articles.

Existing approaches review

The 70 entities from a manually marked text were used for analysis the following knowledge bases:

- a thesaurus on Information Technologies by Prof. A.M.Fedotov (NSU),
- a thesaurus for Russian language ruWordNet,
- an open knowledge base Wikidata.

One entity is a term that includes one or more words.

Data description

Train dataset: the 80 annotated texts: the 659 observations with the following types of relations: USAGE, PARTOF, SYNONYMS, ISA, COMPARE, CAUSE, NONE. It includes 90 Is-A relations.
Test dataset: the 74 observations with 11 Is-A relations.

Приводится сравнительный экспериментальный анализ трех методов коллаборативной фильтрации : н а основе документов, на основе пользователей и на основе <e1>гибридного метода</e1>, являющегося комбинацией первых двух <e2>методов</e2>.	ISA
Под расширенной моделью понимается тематическая модель, содержащая кроме однословных термин ов термины, состоящие из нескольких слов (также называемые <e1>многословные термины</e1> или <e2>ключевые фразы</e2>).	SYNONYMS

Table 3. Dataset examples.

An additional data is a dataset that includes 1000 scientific articles on Russian language in Computer Science

Data Preprocessing

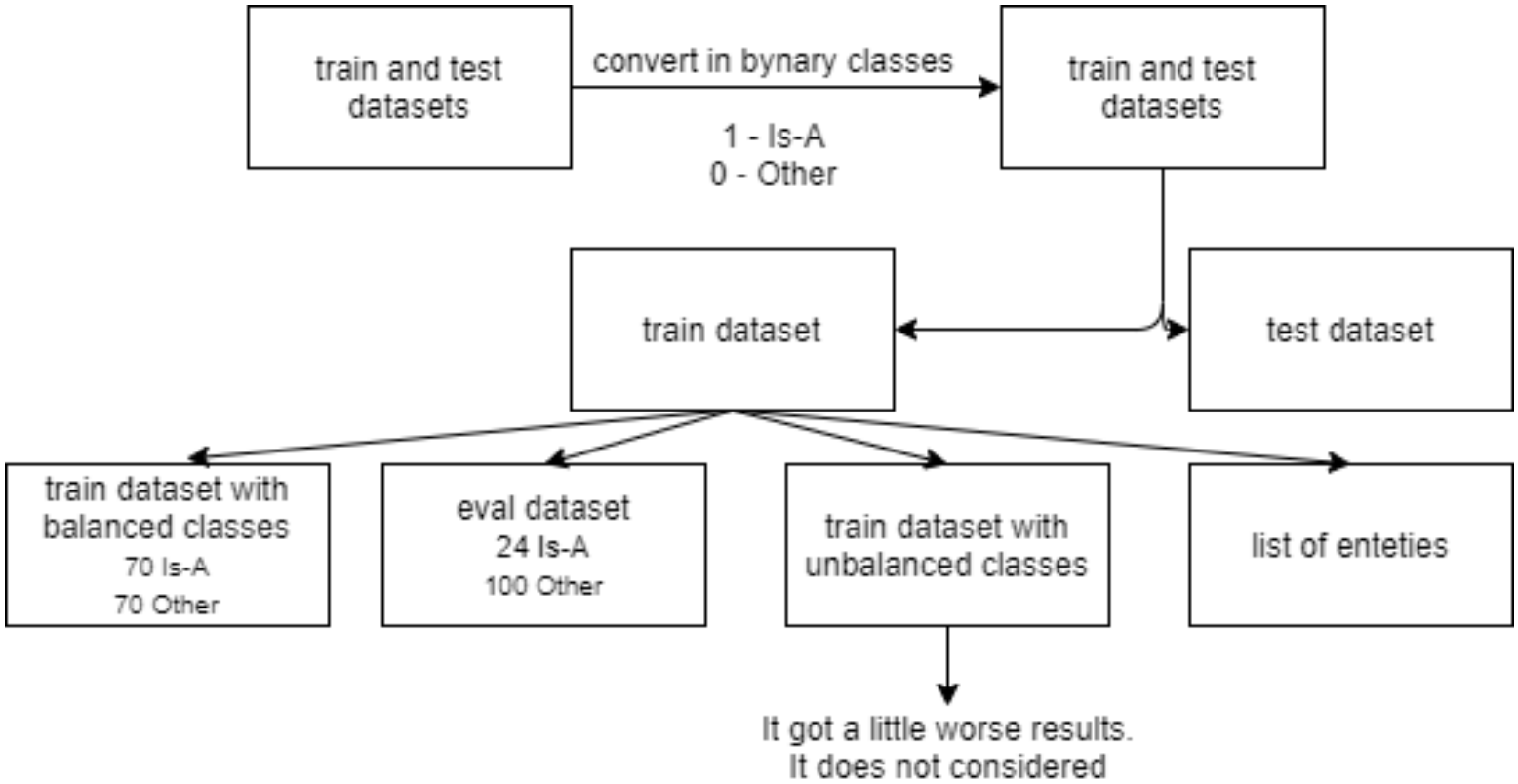

Figure 3. Data preprocessing workflow.

Task solving methods

- R-BERT (model 'bert-base-uncased')
- R-BERT (model 'bert-base-multilingual-cased')
- Universal Sentense Encoder + CNN
- SentencePiece + Neural Networks (different combinations)

R-BERT [1]

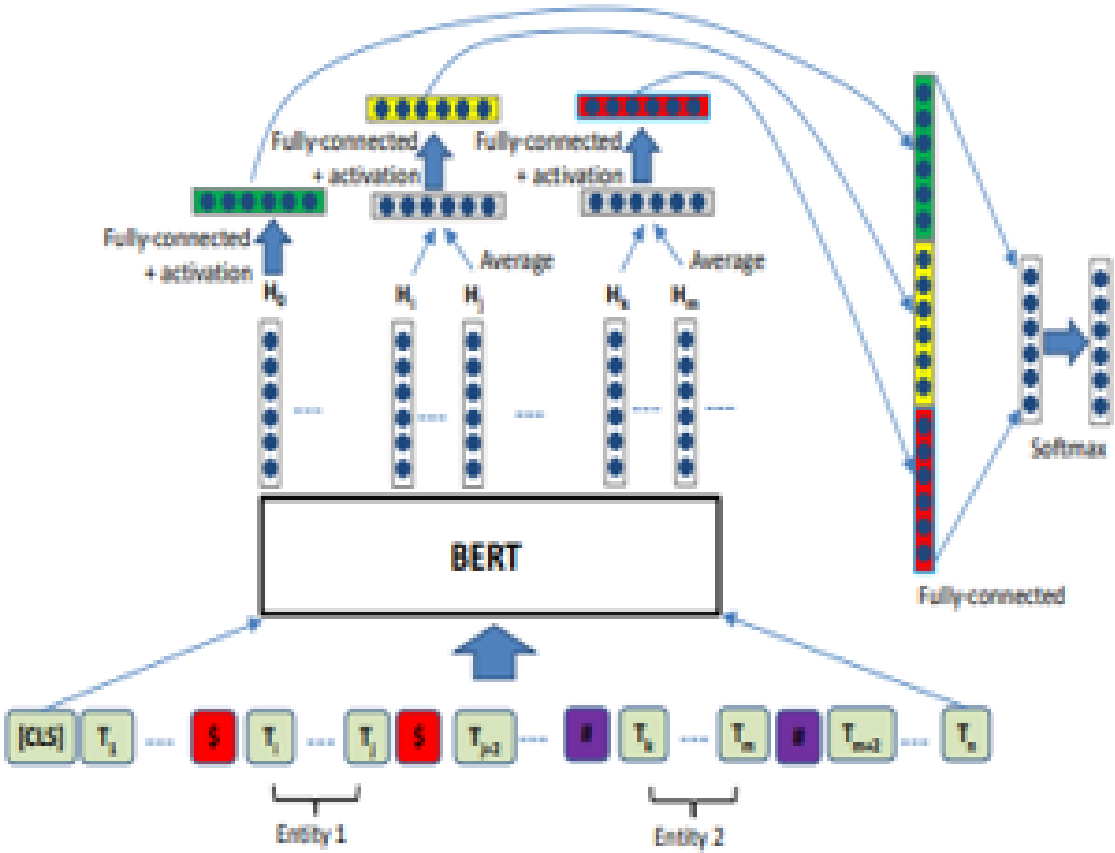

Figure 4. R-BERT architecture.

[1] Wu S., He Y. (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. pp. 2361-2364.

USE+NN

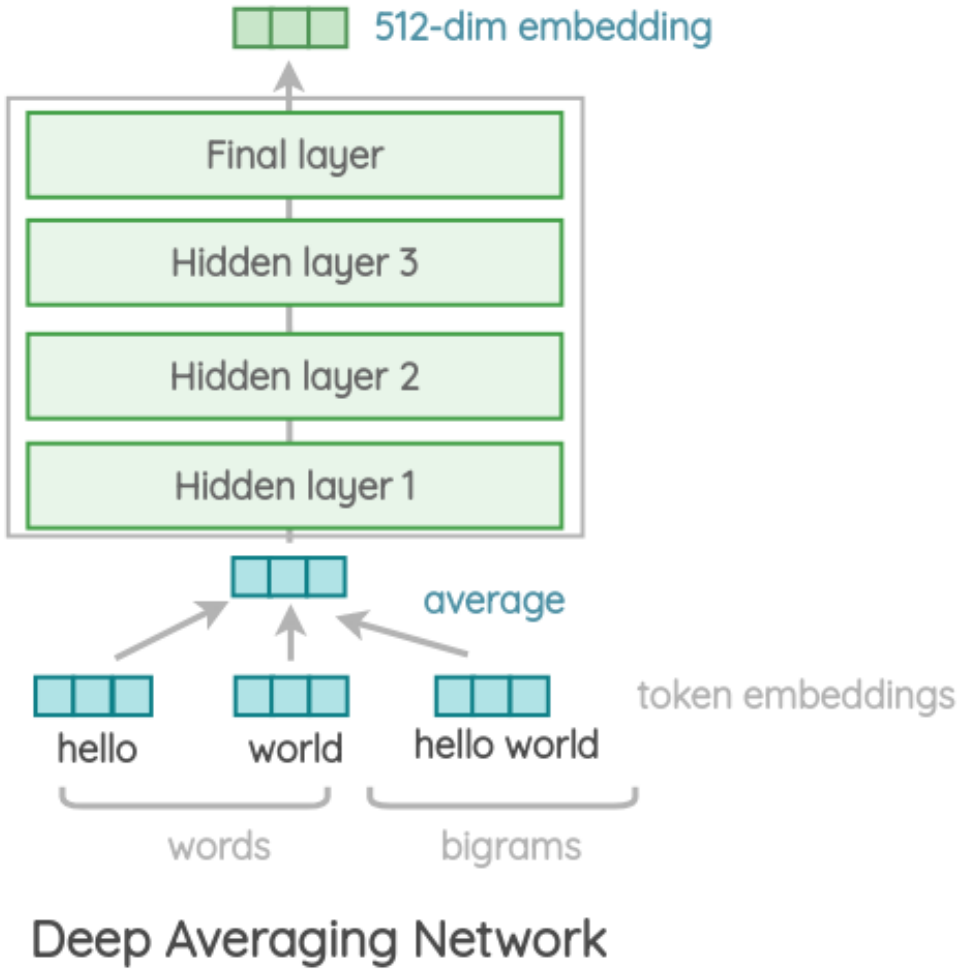

Figure 5. USE architecture.

There is no ability to add special tokens in USE, the model encodes the text without it.

SentencePiece+Keras+NN

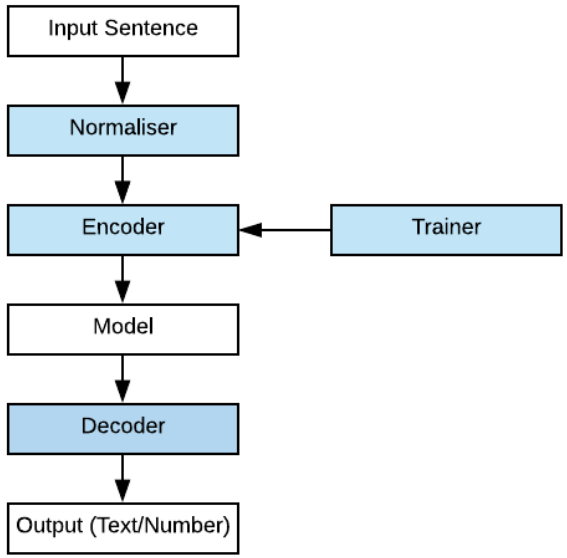

Figure 6. SentencePiece architecture.

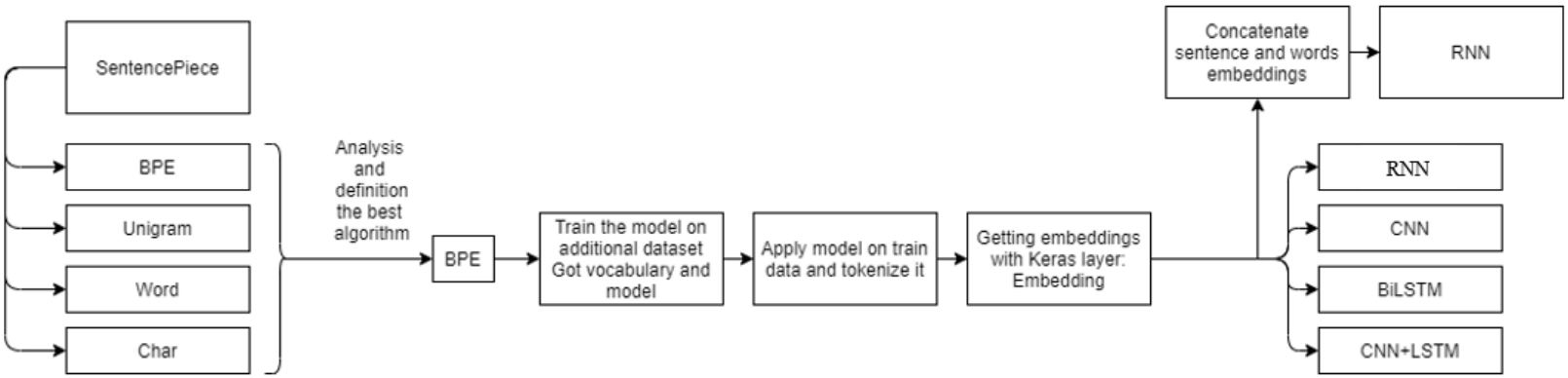

Figure 7. Workflow.

Visualization with TSNE

Figure 8. Visualization entities using embeddings received with BERT tokenizer and model.

Visualization with TSNE

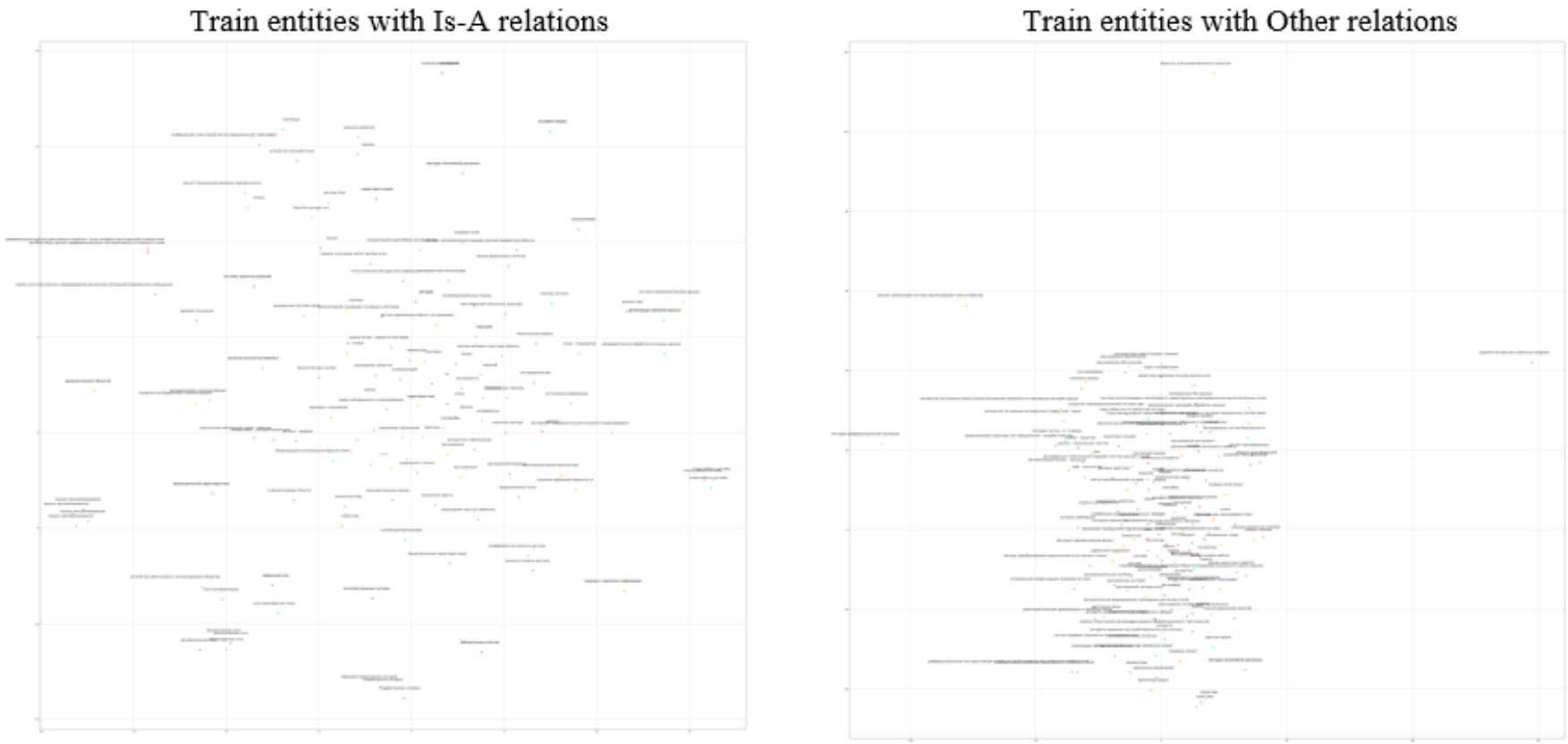

Figure 9. Visualization entities using embeddings received with SentencePiece tokenizer and Keras layer.

Results

Results	F1_score	ROC_AUC_score
RBERT ('bert-base-uncased')	66,6%	85,35%
RBERT ('bert-base-multilingual-cased')	72%	86,94%

Table 4. Results

The methods using USE, SentencePiece and Neural Networks recieved worse results: f1 score less than 50%.

Entity1	**Entity2**	**True label**	**Predicted label**
шард	сервер	Is-A	Other
предобработка	процесс реферирования	Other	Is-A
ООСУБД	программный инструментарий	Is-A	Other
многословный термин	ключевая фраза	Other	Is-A

Table 5. Prediction mistakes of the type of relationships between entities

Publications

1. Pirozhkova D.A., Goncharova I.V. Automatic prediction of hyperonyms for the Russian language. International Scientific Student Conference. Novosibirsk. April, 2020. p. 105. (in Russian, RSCI)
2. Pirozhkova D.A. Application of neural network methods for automatic taxonomy enrichment for the Russian language. International Scientific Student Conference. Novosibirsk. April, 2021. (in English, RSCI)

Thank you for your attention!