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

Daria Pirozhkova

Advisor: Tatiana Batura

NSU,

December 2020

Content

- Introduction to the project
- Project goal and objective
- Task plan
- Papers review
- Data description
- Method description
- Future work

Introduction to the project

Taxonomy (general) is the practice and science of classification of things or concepts, including the principles that underlie such classification.

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.

Introduction to the project

- Antisymmetric relation:

$$\forall a,b \in S \text{ If } R(a,b) \wedge R(b,a) \text{ Then } a=b.$$

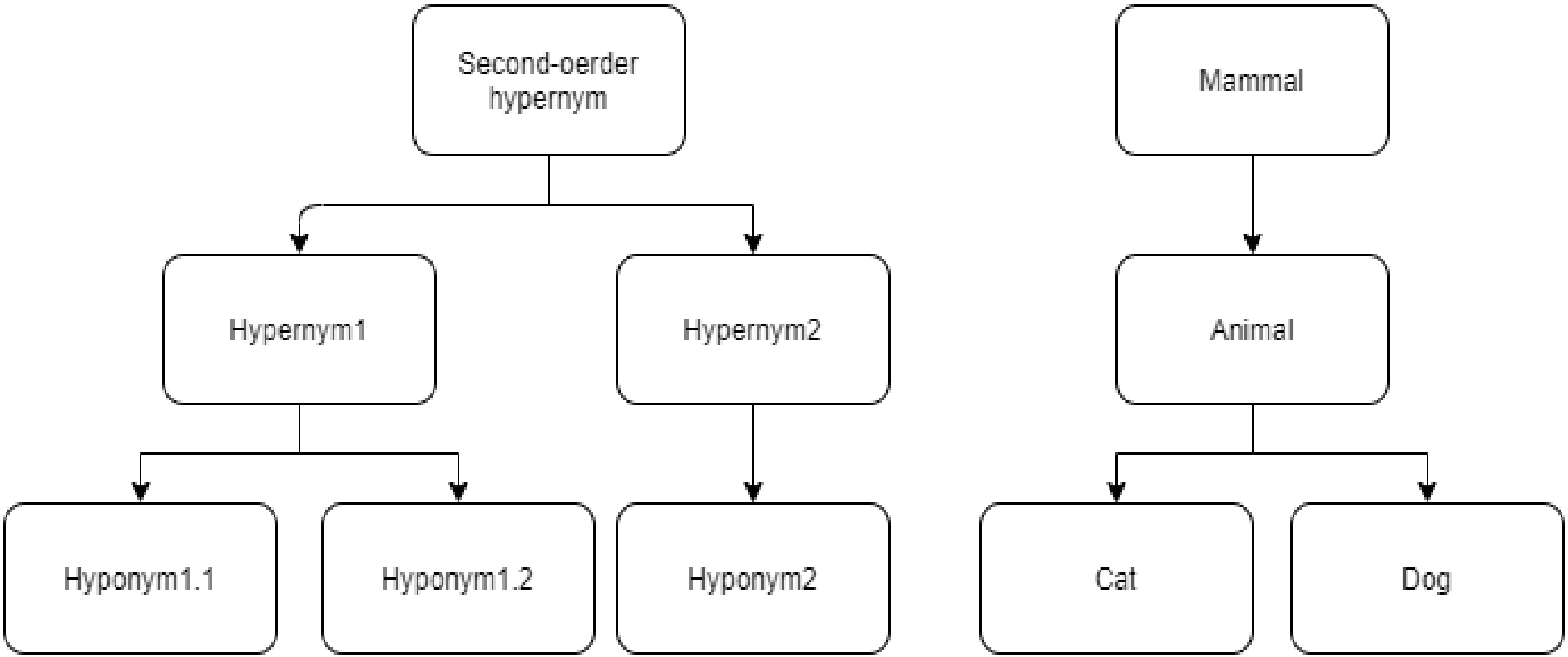
- Irreflexive Relation:

$$\forall s \text{ In } S: \neg R(s,s)$$

- Transitive Relation:

R is a **Transitive Relation** IF $\forall$ a,b,c $\in$ S: IF R(a,b) =>True $\wedge$ R(b,c) =>True THEN R(a,c) =>True.

Introduction to the project



Project goal and objective

The main 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 main task is to develop a method for prediction the hypernym relations between terms in texts of the Computer Science category.

Task plan

- Literature review
- Building a classifier for an ISA relation
- Assessment of the quality of the classifier and its results
- Implementation of a method for selecting candidates for adding to a taxonomy
- Method quality assessment
- Analysis of the results, possible ways to improve the method and its results
- Preparation of a publication for the journal

Papers review

- Different solutions to this task for the English language includes approaches based on the vectorization, classification, and clustering of the words that are hypernym relations.
- The less articles describe the solution with neural networks.
- The solution for domain-specific data is better work that the one for language in general.
- The solution for Russian language [Karaeva et al. 2018] includes approach using the word embedding and calculating the distance between them.The obtained precision value is less than 65 percent.

Data description

80 annotated texts with 90 Is-A relations

```
Для восстановления трехмерной геометрии применяется комбинация двух основных <e2>методов</e2> :
трехмерного алгоритма роста области из семени и <e1>ячеечного метода</e1>, использующего разбиение
пространства на тетраэдры.  ISA
```

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

Method description

RBERT

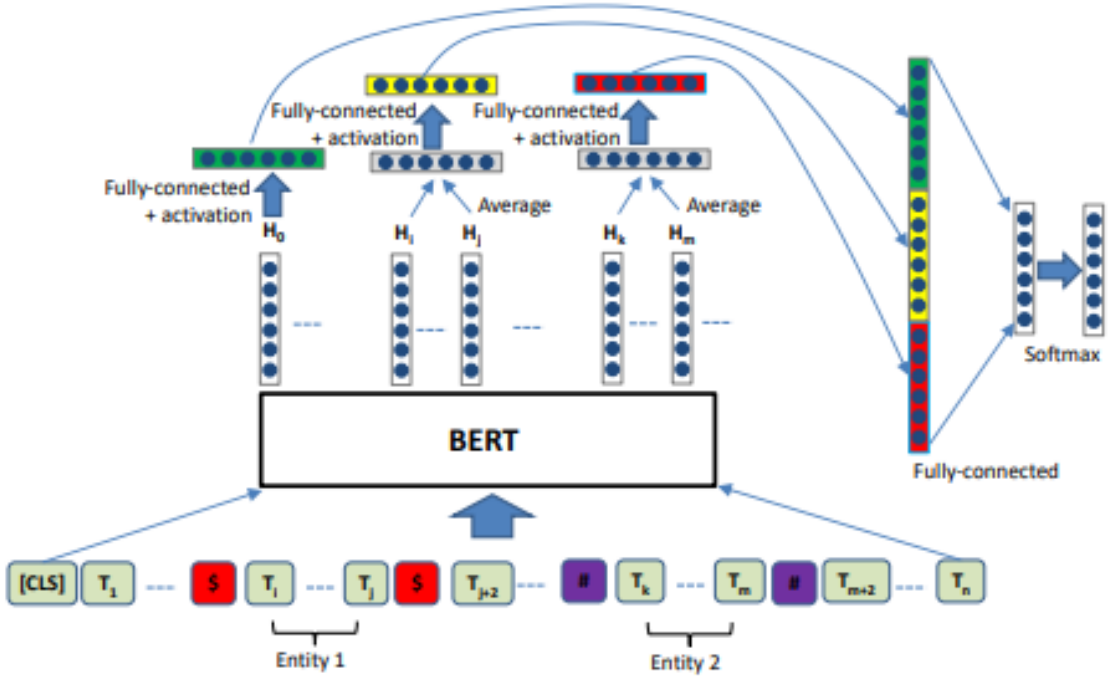


Figure 1: The model architecture.

More details: https://arxiv.org/pdf/1905.08284.pdf

A method for selecting candidates for adding to a taxonomy

Argument1	LinkToArg1	Argument2	LinkToArg2	Relation	
Entity1	Link1	Entity2	Link2	None	Not add
Entity1	Link1	Entity3	--	ISA	Add
Entity2	Link2	Entity4	--	None	Not add
Entity5	--	Entity6	Link6	None	Not add
Entity5	--	Entity7	Link7	ISA	Add
Entity8	--	Entity9	--	ISA	Add
Entity8	--	Entity10	--	None	Not add
Entity11	Link11	Entity12	Link12	ISA	Can be added

Future work

- Literature review (done)
- Building a classifier for an ISA relation (December, 2020)
- Assessment of the quality of the classifier and its results (December,2020 -January,2021)
- Implementation of a method for selecting candidates for adding to a taxonomy (January - February,2021)
- Method quality assessment (February,2021)
- Analysis of the results, possible ways to improve the method and its results (February,2021)
- Preparation of a publication for the journal (March,2021)

Thank you for your attention!