Skoltech
Skolkovo Institute of Science and Technology

INDUSTRIAL IMMERSION 2019

MTS

Entity Linking

Prof. Maxim Fedorov
Skoltech supervisor

Student
Alisa Alenicheva [1]
Program name: Information Science and Technology

Vladimir Volkov
Company supervisor

Background

Entity linking (EL) is a task of finding **mentions** of **entities** inside the text and connecting them to the unique entry inside the **knowledge base** (for example, Wikipedia). EL models are applicable to tasks connected with extraction semantic information from texts, as in **chat-bots** – an application which is significantly important for **MTS** as a company that keeps in touch with its users on the daily basis.

For example: "What is the Worst TV-series finale in 2019?"

The eighth and final season of the fantasy drama television series *Game of Thrones*, produced by HBO, premiered on April 14, 2019, and concluded on May 19, 2019.

Objectives

1) To investigate existing approaches of solving Entity Linking task for the English language.
2) To develop a strategy of applying these algorithms to the Russian language.
3) To create marked datasets for Entity Linking tasks in the Russian language.

Process

The pipeline of the work was as follows:
I've spent the first month investigating existing solutions of the EL problem and looking for suitable github-repositories. At the same time I've trained chosen models trying to obtain the results mentioned in the articles, compared these models and analyzed which approach could possibly be transferred from the English Language to the Russian. During the second month I was creating the datasets that were needed for running the chosen End2End approach (the scheme of the model is presented in Fig.1).
During the Immersion every two weeks we had meetings with the whole NLP-team, where we were discussing our results and issues.

Academic cheating is a significantly common occurrence in high schools and colleges in the United States. Statistically, 64% of public high school students admit to serious test cheating. 58% say they have plagiarized. 95% of students admit to some form of cheating.

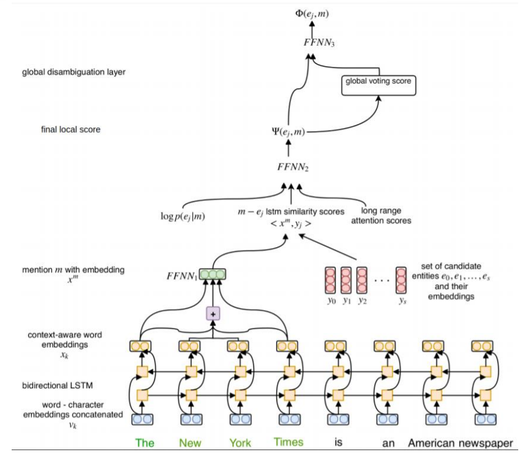

Fig. 1 End2End EL model

	DeepType	Deep-ed	End2End
Programming language	Python	Lua	Python
Type of the model	Neural Networks	Vectors	Vectors + Neural Networks
Type of the task	Dis-ambiguation	Dis-ambiguation	End2End
Dataset	Wikipedia	AIDA CoNLL-YAGO	AIDA CoNLL-YAGO
Knowledge base	Wikipedia	Wikipedia	Wikipedia
Micro-Precision	94.88	92.22	-
Micro-F1-strong	-	-	89.4
Micro-F1-strong	-	-	86.6

Table 1 Comparison of EL models

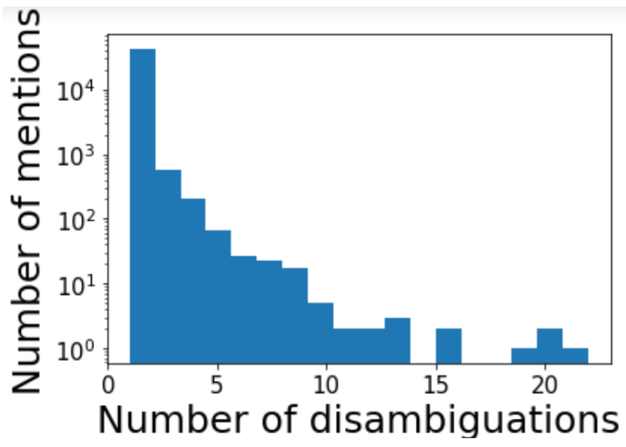

Fig. 2 Distribution of disambiguation in mention-entity dictionary

Results

1) I've made an analysis of existing and most effective SOTA approaches. These said approaches were as follows: **DeepType**, **Deep Joint Entity Disambiguation with Local Neural Attention (Deep-ed)** and **End-to-End Neural Entity Linking (End2End)**. The results of comparison can be seen in Table 1.

2) I've found the only one approach able to work with the Russian language from quick start and it was **DBpedia Spotlight** system. Although its results weren't satisfying, it could still be used as a baseline.

3) I've investigated the datasets in the Russian language connected to Entity Recognition tasks and found out that there was no dataset for the specific area of entity linking and disambiguation task, so I've created a script producing **marked dataset for Entity Linking** from Wikipedia dump in the Russian language.

4) Also, I've made a script which calculates **LinkCounts** p(e|m) dictionary - the number of times mention **m** is pointing to exact entity **e**. The example of the mention pointed on two different entities is presented in Fig. 3. The algorithm has worked with 865 Wikipedia articles and obtained **~43000** mentions, the distribution of the multiplicity of mentions is presented in Fig. 2.

```
Mention:  нанести поражение
Entities:
Невская битва
Битва под Аустерлицем
```

Fig. 3 Entry from mention: entities dictionary

Conclusions

The internship gave me an opportunity to put the knowledge I obtained earlier into practice. Obviously, real tasks are much more difficult to work with than academical ones but at the same time they provide additional challenges and valuable skills for the rosy future.

Unemployment, or **joblessness**, is a situation in which able-bodied people who are looking for a job cannot find a job.

[1] alisa.alenicheva@skoltech.ru, tel: +7 915 086 52 15