Automatic Extraction of Lithuanian Cybersecurity Terms Using Deep Learning Approaches

Aivaras ROKAS
Sigita RACKEVIČIENĖ
Andrius UTKA

Aim of the research:

- To determine the most effective deep learning method for automatic extraction of Lithuanian terms from specialised (cybersecurity) domain.

Objectives:

- Compilation of cybersecurity (CS) corpus
- Development of the gold standard (GS) corpus with manually labelled terminology
- Performing experiments to test various deep learning models

The Cybersecurity Corpus

The compiled corpus is composed of 6 main categories of texts covering the period 1999-2019:

- legal acts of the Republic of Lithuania and related documents,
- reports of the National Cybersecurity Centre,
- translated EU legislation,
- translated international conventions,
- academic papers,
- information publications for the general public.

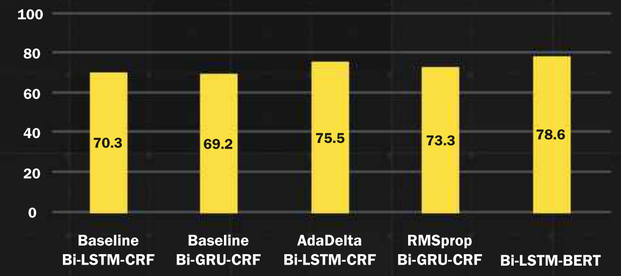
Gold standard (GS):

- The GS small-scale corpus (66,706 words) contains 1,258 manually annotated CS terms in BIESO annotation format.
- GS dataset was divided into 3 parts: 70% for training, 20% for validation and 10% for testing.
- GS dataset was supplemented by the entire Lithuanian Wikipedia database (27,907,392 words) in order to make embeddings more effective.

Experiments:

- Testing various baseline LSTM and GRU networks using Adam optimizer and FastText embeddings
- Testing each of the best baseline LSTM and GRU networks with various optimizers
- Comparing the best model with a model that has been trained using BERT contextual embeddings and testing if contextual embeddings can further improve our model

F1 Results of Top Performing Models



Conclusions:

- The results of our experiments suggest that for Lithuanian term extraction the semi-supervised deep learning approach is a way to go: although deep neural networks were trained on a very small amount of annotated data in these experiments, our top performing model managed to reach F1 78.6%.
- In order to achieve an even higher score, the quality and quantity of annotated data have to be increased.

Acknowledgements

The research is carried out under the project “Bilingual automatic terminology extraction” funded by the Research Council of Lithuania (LMTLT, agreement No. P-MIP-20-282). The project is also included as a use case in COST action “European network for Web-centred linguistic data science” (CA18209).

MYKOLAS ROMERIS UNIVERSITY
VYTAUTAS MAGNUS UNIVERSITY
Research Council of Lithuania