ICAI 2023
12th International Conference on Applied Informatics
Eger, Hungary, 2023

Hungarian Sentence Analysis Learning Application with Transformer Models

Noémi Evelin Tóth[a], **Beatrix Oszkó**[b], **Zijian Győző Yang**[b]

[a]Eszterházy Károly Catholic University
noemitth.10@gmail.com

[b]Hungarian Research Centre for Linguistics
{oszko.beatrix,yang.zijian.gyozo}@nytud.hu

Abstract

The intention of our research is to present a project in which we started to develop an educational support tool that helps primary and high school students to use the correct techniques of sentence analysis, based on the rules of Hungarian grammar taught in school. The aim was to create an application, called LMEZZ, that helps students of the Hungarian education system to practise tasks related to native language lessons. This way, we expect that they will have a more accurate understanding of grammar rules. With the help of the application, they can learn from the comfort of their home by receiving immediate and accurate feedback on solutions to different tasks. Thanks to the development of computational linguistics, many language and communication problems can now be automated. Therefore, to avoid having to constantly update and verify thousands of sentences and their associated handwritten analysis, the central theme of the research, in addition to application development, is the teaching and testing of linguistic analytical models capable of correctly learning to analyse raw sentences. In an early stage of development [4], we only used HuSpaCy [3] to train two models and compare them. Dependency analysis was used as the basis for the preparation of the source material, as it was most similar to the school analysis and then the annotations of the teaching sentences were made. We distinguished the two models based on the label set we defined. In the case of the smaller model, we were only interested in the most important and basic labels, while in the case of the extended model, we also covered the analysis of different types of adjectives and indicators.

The models developed with version 2 HuSpaCy had some problems identifying

labels, that were not as frequent in the corpus as the predicates or the subjects. Besides collecting more sentences to our corpus, with this research we wanted to improve our application to find a better model with more reliable outcomes for each label.

Natural language processing has seen spectacular progress with the application of neural network technology, in particular, the contextual transformer model [5]. In recent years, natural language processing tasks can be solved with high performance, if a pre-trained transformer language model is fine-tuned. One of the most popular transformer based language model is the BERT. BERT (Bidirectional Encoder Representations from Transformer) is defined as a multi-level, bidirectional transformer encoder architecture [1]. The BERT model is pre-trained on two language modeling tasks: word masking and next sentence prediction. In the recent years, there are two state of the art (in different natural language processing tasks) BERT model were trained for Hungarian: huBERT [2] (BERT base model – 110 million parameter) and PULI BERT-Large [6] (BERT large model – 345 million parameter).

We solved this sentence analysis as a token classification task. To fine-tune the Hungarian BERT models, we used the code provided by Hugging Face[1]

Table 1. F-Score results

	HuSpaCy	huBERT	PULI BERT-Large
Predicate (R)	94.12%	100%	100%
Subject (A)	73.91%	93.02%	90.48%
Object (T)	86.75%	100%	100%
Adverbial (H)	78.87%	96.15%	96.15 %
Indicator (J)	76.92%	96.97%	78.57%
P	58.33%	86.49%	94.44%
X	95.96%	100%	100%

In Table 1, you can see the results of the fine-tuned transformer models. The results of the contextual transformer models turned out far more condescending than the previously trained HuSpaCy models.

References

[1] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova: *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), Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186.

[2] D. M. Nemeskey: *Introducing huBERT*, in: XVII. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Magyarország: Szegedi Tudományegyetem, Informatikai Intézet, 2021, pp. 3–14.

[1] https://github.com/huggingface/transformers

[3] G. Orosz, Z. Szántó, P. Berkecz, G. Szabó, R. Farkas: *HuSpaCy: an industrial-strength Hungarian natural language processing toolkit*, Szeged, 2022.

[4] B. Oszkó, N. E. Tóth, Z. G. Yang: *Az általános és középiskolai magyar nyelvtan tananyag elsajátıtását segıtő alkalmazás* (2022).

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin: *Attention is All you Need*, in: Advances in Neural Information Processing Systems 30, ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett, Curran Associates, Inc., 2017, pp. 5998–6008.

[6] Z. G. Yang, R. Dodé, G. Ferenczi, E. Héja, K. Jelencsik-Mátyus, Á. Kőrös, L. J. Laki, N. Ligeti-Nagy, N. Vadász, T. Váradi: *Jönnek a nagyok! BERT-Large, GPT-2 és GPT-3 nyelvmodellek magyar nyelvre*, in: XIX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2023), Szeged, Hungary: Szegedi Tudományegyetem, Informatikai Intézet, 2023, pp. 247–262.