Acta Cybernetica **27** (2025) 83–91.

Hungarian Sentence Analysis Learning Application with Transformer Models

Noémi Evelin Tóth[a], Beatrix Oszkó[bcd], and Zijian Győző Yang[be]

Abstract

The purpose 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 would help students of the Hungarian education system to practise tasks related to native language lessons. In this way, we expect them to have a more accurate understanding of the grammar rules. The application allows them to learn in the comfort of their own homes by providing immediate and accurate feedback on the solutions to various tasks. Natural language processing has made spectacular progress with the application of neural network technology, especially the contextual transformer model. In our research, Hungarian transformer-based BERT models were trained for our sentence analyser task. The results showed that the transformer models were much more condescending than the previously trained convolutional neural network based SpaCy models. This allowed us to increase the reliability of our software.

Keywords: learning application, Hungarian grammar, sentence analysis, SpaCy, transformer models, BERT

1 Background

These days, there is a growing demand for self-studying. With the current state of technology, learning is increasingly accessible through mobile phones, tablets and laptops. Young people are familiar with this type of technology and use it daily. There are a lot of applications targeted at learning, such as Duolingo[1], Kahoot[2],

[a]Eszterházy Károly Catholic University, Eger, Hungary, E-mail: noemitth.10@gmail.com, ORCID: 0009-0006-4919-4338
[b]HUN-REN Hungarian Research Centre for Linguistics, Budapest, Hungary
[c]University of Novi Sad, Serbia
[d]E-mail: oszko.beatrix@nytud.hun-ren.hu, ORCID: 0000-0002-0169-4505
[e]E-mail: yang.zijian.gyozo@nytud.hun-ren.hu, ORCID: 0000-0001-9955-860X
[1]https://www.duolingo.com/
[2]https://kahoot.com/

DOI: 10.14232/actacyb.307878

Mateking[3] and many others. With this research, we wanted to find a good use of the results of computational linguistics and help Hungarian students learn Hungarian grammar, especially sentence analysis. Our main target group was primary and secondary school students. The rules of Hungarian sentence analysis are often not self-explanatory, and there is little time to practice them in class. Our idea was to provide a tool that would allow them to analyse any sentence in real time.

Thanks to the development of computational linguistics, the solution for many language and communication problems can now be automated. Therefore, to avoid having to constantly verify sentences and their associated handwritten analysis by hand, the central theme of the research - in addition to application development - is the teaching and testing of linguistic analytical models to analyse raw sentences.

In the early stage of development, we only used SpaCy to train two models and we compared the results [8]. Dependency analysis was used as the basis for the preparation of the source material, as it was most similar to the school analysis. 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. In the current research, we further developed our application with the new generation deep contextual transformer language models.

2 Rules of sentence analysis

In order to develop the application and teach several neural network models, we first must consider how sentence analysis is taught in school. The relationship between linguistics and the teaching of Hungarian grammar is not always clear and consistent [3]. There is a significant difference and gap between the methods of scientific linguistics and the material taught. Scientific linguistics is always slightly ahead of school grammar, as the latter always tries to teach theories and models that have already been proven. In today's modern syntax, generative grammar does not categorise linguistic structures and their functions, but promotes applicable knowledge and critical and analytical thinking. In its model, of course, regularities and ways of describing and defining sentences are present, but it focuses on linguistic skills and competence. In contrast, school grammar is built on the traditional levels of language, so that knowledge of sentences can be acquired through the knowledge of phonemes, morphemes, lexemes and synagems [10].

School grammar has a dependency approach, the aim is to establish the relationship between the syntagmas or words that make up the sentence. The order of the words is rather loose, given the characteristics of the Hungarian language, so it is not dealt with in school; the dependencies form the hierarchy of the words and thus ensure the meaning of the sentence. Sentences can be classified in many ways, for example according to their structure or logical quality. In this paper, we will look in detail at the structure and analysis of simple sentences. They consist of a single clause, i.e. a single statement. Sentences can be further broken down into

[3] https://www.mateking.hu/

word structures, which are formed by the grammatical combination of two words of a basic word type that are closely related. The subject and predicate form such a syntagm. They are the main parts of the sentence and form the grammatical, semantic and logical core of the sentence. In addition, the analysis usually takes into account other elements that are extensions of the sentence. Extensions are the subject, adverbs and adverbials.

To understand this, take a look at a real example. The sentence is the following in Hungarian: *A hatalmas jegesmedve az Északi-sarkon él.* Which means: *the giant polar bear lives in the Arctic.* The predicate of the sentence is '*lives*', and the subject is the '*polar bear*' itself. The word '*giant*' is an indicator of quality, and the word '*in the Arctic*' is a locative part of the sentence. It locates where the polar bear lives.

In computer linguistics, dependency analysis [1] has been used to represent sentence structure. The approach of traditional Hungarian grammar is vastly similar to this. The sentence structure is represented as a tree and the starting point is always the predicate. An important difference, however, is that while dependency analysis works with tokens, the grammar targeted in this research uses syntactic words. A syntactic word can sometimes consist of several tokens, but traditional analysis does not establish any further relationship between the individual elements, they simply appear as a node of several words in the tree. Another important difference is that the traditional analysis ignores certain words. These are typically function words: article words, conjunctions, participles, etc. School grammar does not take punctuation into account either, but in computer analysis these are also present as separate tokens. Later, these tokens will have their own label, different from the ones we discussed in this section.

3 Methods

Natural language processing is a branch of artificial intelligence based on linguistic research.The goal is to reduce the gap between the computer and the human as much as possible, so that the computer can read, interpret and process human language [4]. The first problem to be solved by using natural language processing tools was to translate a text into another language. To do this, the computer must be able to understand the rules, morphology and syntax. The latter requires knowledge of the semantics and vocabulary of the language. Today, machine translation is only a small part of computational linguistics and can be found in many different areas of life, with intelligent assistants and chatbots, it powers search engines and spell checkers, and there are now many people involved in computer processing of various textual data and its use in other disciplines.

We can say that there are different types of neural networks that are optimised for different data. As we have already mentioned, when we started developing LMEZZ, we only used HuSpaCy [7] to train two models and compared them. At that time SpaCy used convolutional neural networks to achieve its goals [6]. It was optimised for industrial use. However, as we will see later in the article, we reached certain limits with the convolutional neural networks. In this paper, our goal was

to try to train transformer models to get more reliable results with our previously created label set.

Before transformer models, recurrent neural networks were used to processing texts [9]. The problem with these types of models is that they work sequentially and can only analyse one word at a time, in order. But in human languages, word order is a big part of the meaning of the text. For this reason, recurrent neural networks are difficult to train. To solve this problem, the researchers developed the first transformer model in 2017 [11], which was originally designed for translation. It was a model that could scale up to a huge dataset. Transformer models are based on three main ideas: positional encoding, attention, and self-attention. With positional encoding, instead of looking at words sequentially, we store information about word order so that the model learns the meaning of word order directly from the data. Attention allows the model to look at each word in the original sentence when making a decision about how to translate a word in the output sentence. Self-attention allows a model to understand a word in the context of the words around it.

4 Models

At the start of this research, we collected and analysed sentences by hand for the training and testing. Most of the sentences were collected from a Hungarian grammar textbook, called *Magyar nyelv a középiskolások számára 9.* written by Adrienne Fráter. Other sentences were chosen from textbooks for secondary school students, written by Ágnes Szabó Antalné and Judit Raátz. We also used examples from the Grammarly Practice Book.[4] The corpus consisted of 268 sentences at the beginning. We used 82% of the dataset for training and the remaining part for testing. We defined a set of labels. These labels and their explanation are shown in Table 1. Then we trained two HuSpaCy models. The models had some problems identifying labels, which were not as common in the corpus as predicates or subjects. In addition to collecting more sentences for our corpus, we wanted to improve our application with this research in order to find a better model with more reliable results for each label.

In recent years, natural language processing tasks can be solved with high performance by fine-tuning a pre-trained transformer language model. One of the most popular transformer based language model is BERT. BERT (Bidirectional Encoder Representations from Transformer) is defined as a multi-level, bidirectional transformer encoder architecture [2]. The BERT model is pre-trained on two language modeling tasks: word masking and next sentence prediction. In the recent years, two state of art BERT models have been trained for Hungarian: huBERT [5] (BERT base model – 110 million parameter) and PULI BERT-Large [12] (BERT large model – 345 million parameter).

[4] https://gepeskonyv.btk.elte.hu/adatok/Magyar/31Lakatos/Digi_TK_v2/Gyakorlokonyv.html

Table 1: Label set we used to train the models

Label	Meaning of the label
ROOT	Predicate
A	Subject
T	Object
H	Adverbial
J	Indicator
P	Element of a multi-word group of the sentence
X	Not analysed part of the sentence

We solved this sentence analysis problem as a token classification task. To fine-tune the Hungarian BERT models, we used the code provided by Hugging Face.[5]

5 Results

In Table 2, you can see the results of the fine-tuned transformer models compared with the results of one of the HuSpaCy models.

Table 2: 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 the case of the models we have previously produced, we found that the label set was too large for the size for the corpus we used, resulting in too many rare labels. Therefore, in this comparison, we only use the results of the small model, which did not take into account the different types of adverbs and indicators. Finally, this model was implemented for the first time in the application. The code of the application itself can be found here[6].

The results of the contextual transformer models turned out far more condescending than the previously trained models. The new models predicted with 100%

[5] https://github.com/huggingface/transformers
[6] https://github.com/noemitth10/Learning-App

accuracy the labels of the predicate, the object and the X. We also achieved results above 90% for the subject and adverbs. However, we still need to achieve some improvement with the indicators. The PULI BERT-Large model only achieved a two percent improvement over the previous model. This result is due to the fact that only a very few test sentences contained any indicator type. The same applies to words marked with a P label.

6 The Application

LMEZZ is a web application built with React and Javascript. Based on our target audience it has a user-friendly interface and a colourful design. For this phase of the development, our main goal was to make it responsible, besides teaching the new transformer models. This way, students can easily use the application from their phones and tablets. Most of the services on the site can be used by a registered user. If someone does not have an account, they can create one using the Register option by entering their email address, password, name and other details. After logging in, the *Elemezz!* option becomes available. By clicking on it, the site redirects the user to the model that can analyse any given sentence. After entering the sentence, the user has to click on the *Kész* button. The application will return the analysed sentence in a minute. Figure 1 shows an example of how this works. The example is the following: *Peti könyvet olvas a verandán.* Which means *Peti reads a book on the porch.* Peti is the subject of the sentence, the könyv is the object. He reads the book, so olvas is the predicate. The porch, is the adverbial of the sentence, which gives us information about the location, where Peti reads the book. Other parts of the sentences are labeled with X, the app will simply not analyze those parts. The newly implemented transformer model is running on a different server, the application just calls the API when a certain sentence is given.

7 Conclusion

In summary, we can say that transformer models are much better suited to the problem of analysing sentences. They have reached a more reliable state when it comes to analysing new sentences. This is a really important factor in teaching. We need to avoid the possibility of giving students incorrect information. However, they still have shortcomings due to the low number of test data. In the future, we need to increase the size of the corpus. In parallel with this research, we have already collected 1000 new raw sentences that need to be annotated. After that we can retrain the transformer models to further improve the results. In the next step we can test again the extended label set with the newly collected data to see where the model needs further improvements.

Figure 1: Screenshot of the application in mobile view

References

[1] De Marneffe, M.-C. and Nivre, J. Dependency grammar. *Annual Review of Linguistics*, 5:197–218, 2019. DOI: 10.1146/annurev-linguistics-011718-011842.

[2] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 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)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. DOI: 10.18653/v1%2FN19-1423.

[3] Gábor, T. N. A magyar nyelv leírása és iskolai oktatása (the description and education of the hungarian language), 2002. URL: https://mta.hu/data/dokumentumok/i_osztaly/1_Eloadasok_tara/Magyar_nyelv_es_kutatasa_20020502/TolcsvaiN_leiras_oktatas_20020502.pdf.

[4] Nadkarni, P. M., Ohno-Machado, L., and Chapman, W. W. Natural language processing: An introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551, 2011. DOI: 10.1136/amiajnl-2011-000464.

[5] Nemeskey, D. M. Introducing huBERT. In *XVII. Hungarian Computational Linguistics Conferences*, pages 3–14, Szeged, Hungary, 2021. Institute of Informatics, University of Szeged.

[6] Orosz, G., Szabó, G., Berkecz, P., Szántó, Z., and Farkas, R. Advancing Hungarian text processing with HuSpaCy: Efficient and accurate NLP pipelines. In Ekštein, K., Pártl, F., and Konopík, M., editors, *Text, Speech, and Dialogue*, pages 58–69, Cham, 2023. Springer Nature Switzerland. DOI: 10.1007/978-3-031-40498-6_6.

[7] Orosz, G., Szántó, Z., Berkecz, P., Szabó, G., and Farkas, R. HuSpaCy: an industrial-strength Hungarian natural language processing toolkit. In *XVIII. Hungarian Computational Linguistics Conferences*, pages 59–73, 2022. URL: https://arxiv.org/abs/2201.01956.

[8] Oszkó, B., Tóth, N. E., and Yang, Z. G. Az általános és középiskolai magyar nyelvtan tananyag elsajátítását segítő alkalmazás (Supporting application for learning Hungarian grammar in elementary and secondary schools). In *A digitális oktatás nyelvi dimenziói: Válogatás a PeLiKon2020 oktatásnyelvészeti konferencia kerekasztal-beszélgetéseiből és előadásaiból (Linguistic dimensions of digital education. Selected papers of lectures and roundtable discussions of the PeLiKOn 2020 conference)*, pages 145–157, Eger, Hungary, 2022. Eszterházy Károly Catholic University. URL: http://publikacio.uni-eszterhazy.hu/id/eprint/7566.

[9] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, 2018. Association for Computational Linguistics. DOI: 10.18653/v1/N18-1202.

[10] Szabó, V. A magyar mondat modelljei a nyelvtanoktatásban (The models of the Hungarian sentence in grammar education), 2010. Manuscript, University of Pécs.

[11] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017. URL: https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[12] Yang, Z. G., Dodé, R., Ferenczi, G., Héja, E., Jelencsik-Mátyus, K., Kőrös, d., Laki, L. J., Ligeti-Nagy, N., Vadász, N., and Váradi, T. Jönnek a nagyok! BERT-Large, GPT-2 és GPT-3 nyelvmodellek magyar nyelvre (The Heavy

Guys are Coming! BERT-Large, GPT-2 and GPT-3 Language Models for Hungarian). In *XIX. Hungarian Computational Linguistics Conferences (MSZNY 2023)*, pages 247–262, Szeged, Hungary, 2023. Institute of Informatics, University of Szeged. URL: https://acta.bibl.u-szeged.hu/78417/.