Hungarian Named Entity Recognition with BERT

Dávid Bárdos
Budapest University of Technology and Economics
davidbardos@edu.bme.hu

Dávid Pristyák
Budapest University of Technology and Economics
davidpristyak@edu.bme.hu

Mátyás Tarnay
Budapest University of Technology and Economics
matyas.tarnay@edu.bme.hu

Abstract

Using Machine Learning [6] methods to understand and process natural language is getting more and more important in business. In this paper, we research a state-of-the-art Transformer model for natural language processing, called BERT [3], which stands for Bidirectional Encoder Representations from Transformers. BERT is designed to pretrain deep bidirectional representations from unlabelled text by jointly conditioning on both left and right context in all layers. Using the baseline BERT model, it is easy to create state-of-the-art models for a wide range of tasks, by adding one additional output layer. On top of being simple and empirically powerful, it obtains state-of-the-art results on eleven natural language processing tasks. We set out the goal to use a basic BERT model and create a version, which understands Hungarian input. Using the hunNERwiki[10] database, we attempt to train, evaluate and test the ability of the BERT algorithm to use the Hungarian language.

Magyar nyelvű entitás felismerés a BERT-tel

Abstract

A gépi tanulási módszerek alkalmazása a természetes nyelv megértésére és feldolgozására egyre fontosabbá válik az üzleti életben. Ebben a cikkben a természetes nyelvi feldolgozás legkorszerűbb Transformer modelljét kutatjuk, az úgynevezett BERT-t[3], amely a Bidirectional Encoder Representations from Transformers jelenti. A BERT-et úgy tervezték, hogy előtanítsa a mély kétirányú reprezentációkat a címkézetlen szövegből úgy, hogy a bal és a jobb kontextusban együttesen kondicionálja az összes réteget. A BERT alapmodell használatával egy további kimeneti réteg hozzáadásával egyszerűen hozhatóak létre a legkorszerűbb modellek számos feladathoz. Amellett, hogy a BERT egyszerű és empirikusan erős, a legjobb eredményeket éri el az élő természetes nyelvi feldolgozási feladatokban. Célul tűztük ki egy alap BERT-modell használatát, és egy olyan változat létrehozását, amely megérti a magyar nyelvet. A hunNERwiki[10] adatbázis segítségével megkíséreljük betanítani, értékelni és tesztelni a BERT algoritmus magyar nyelvhasználati képességét.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

1 Introduction

Natural language processing is becoming more and more important in today's world. One of the biggest challenges in natural language processing (NLP) is the shortage of training data. Because NLP is a diversified field with many distinct tasks, most task-specific datasets contain only a few thousand or a few hundred thousand human-labelled training examples. However, modern deep learning-based NLP models see benefits from much larger amounts of data, improving when trained on millions, or billions, of annotated training examples. To help close this gap in data, researchers have developed a variety of techniques for training general purpose language representation models using the enormous amount of unannotated text on the web (known as pre-training). BERT, which stands for Bidirectional Encoder Representations from Transformers, is based on Transformers, a deep learning model in which every output element is connected to every input element, and the weightings between them are dynamically calculated based upon their connection. Using this bidirectional capability of BERT, we attempt to use it for Named Entity Recognition (NER)[8] in Hungarian text. Name Entity recognition build knowledge from unstructured text data. It parses important information from the text like email address, phone number, etc.

In this paper, we will introduce the background of our work and what kind of architecture was used. We present the challenges and difficulties of the Implementation and summarize our semesters work.

2 Background

We started researching the state-of-the-art NLP architectures and narrowed down our possible models to three solutions. Decoding-enhanced BERT with Disentangled Attention (DeBERTa)[5] improves the BERT and RoBERTa models using two novel techniques. The first is the disentangled attention [9] mechanism, where each word is represented using two vectors that encode its content and position, respectively, and the attention weights among words are computed using disentangled matrices on their contents and relative positions, respectively. Second, an enhanced mask decoder is used to incorporate absolute positions in the decoding layer to predict the masked tokens in model pre-training. CamemBERT [7] sets out a goal to use a monolingual Transformer-based language model and training it with other languages, in this case French. The authors show that a relatively small web crawled dataset (4 GB) leads to results that are as good as those obtained using larger datasets (130+GB). Lastly, Deep Bidirectional Transformers for Language Understanding (BERT) [3]. Due to the lack of successful operation of the first two models, we used the simple BERT architecture in our work.

3 Architecture

BERT's model architecture is a multi-layer bidirectional Transformer encoder. The architecture is based on a previous work from Vaswani et al. (2017)[11] The basics can be seen on Figure 3.

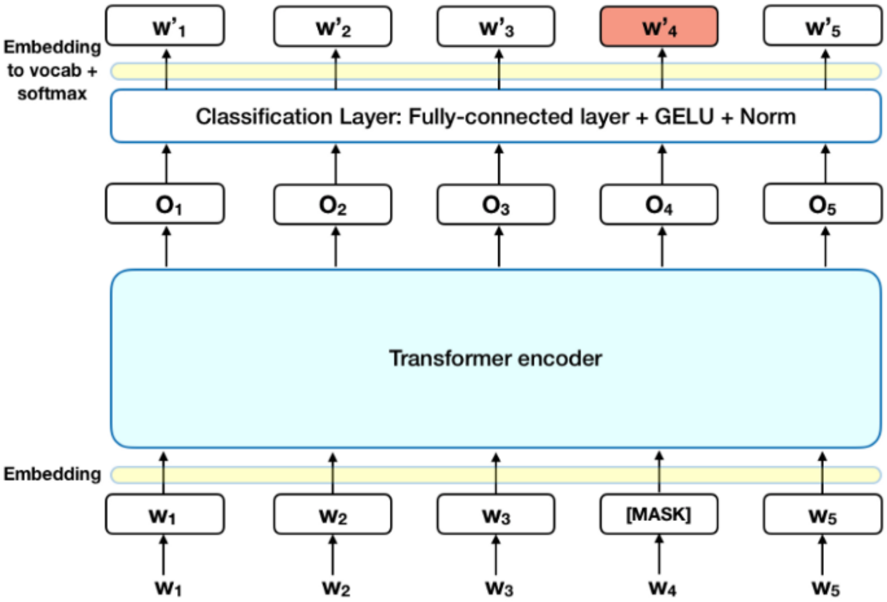
w'1
w'2
w'3
w'4
w'5
Embedding to vocab + softmax
Classification Layer: Fully-connected layer + GELU + Norm
O1
O2
O3
O4
O5
Transformer encoder
Embedding
w1
w2
w3
[MASK]
w5
w1
w2
w3
w4
w5

Figure 1: Architecture of BERT [1]

4 Implementation

During the implementation phase of the project, we used Google Colab as our development platform. We attempted to use Google Cloud, but it was hard to work with and did not solve our problems. Later in the semester we changed to Kaggle, where the development environment was much faster and easy to use and also the hardware, that was provided free, was more powerful.

4.1 Data

For training the baseline BERT model, we need to find a Hungarian database optimized for Named Entitiy Recognition. Our choice is called hunNERwiki [10]. It is a silver standard corpus for Hungarian Named Entity Recognition. The corpus has been automatically generated from the

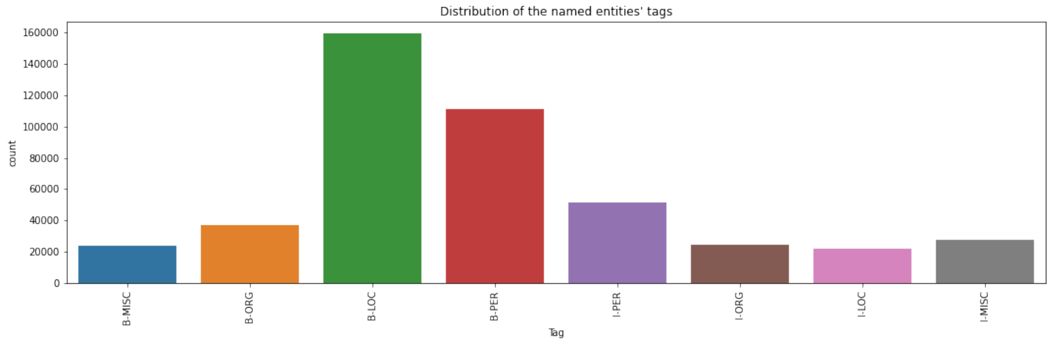

Figure 2:

Hungarian Wikipedia, using the entity categorization of DBpedia. At 19 108 597 tokens, it is the largest Hungarian NER training corpus by far; by comparison, the largest gold standard corpus for Hungarian, the Criminal NE corpus, consists of 562 822 tokens. The first step of the processing of the data is to clear it from unusable rows and transform it to a usable format. We kept the Sentence, the Word and the Tag information from the hunNERwiki database. The next step is to tokenize the transformed data. For tokenizing, we used the BertTokenizer[2]. BertTokenizer uses a WordPiece tokenizer. It works by splitting words either into the full forms (e.g., one word becomes one token) or into word pieces — where one word can be broken into multiple tokens. Using word pieces allows BERT to easily identify related words as they will usually share some of the same input tokens, which are then fed into the first layers of BERT. After tokenizing, we split our data into three parts: training, validating and testing data.

4.2 Training

Since the amount of data and the limited GPU access, we decided to use fewer data, than we originally have. The model was trained with 20 percentage of the data. The optimizer of the model is Adam[12], the loss function of the model is Sparse Categorical Cross entropy [4] and the metric is Accuracy. After the training with 3 epochs, the model produced 99.72% accuracy.

4.3 Testing

For testing, we use 10% of our data. We created an input pipeline which uses the model and the tokenizer to give us the results. The tests first gave us a low accuracy of 7%. This trained model did not contain the words with 'O' tag.

5 Summary and future plans

In this work, we researched the Named Entity Recognition field and chose the most optimal architecture for us. We used a baseline BERT pretrained model and trained it with a Hungarian corpus.

Using the available resources and time, we ran into several bottlenecks and had to adjust our goals accordingly. Hyperparameter optimization could have been explored more widely, and our evaluation technique could also be refined. We managed to train a model and produce decent accuracy.

To sum up, we achieved the set goals and learned a lot during the process. In the future, we are planning to refine our training algorithm to achieve better accuracy, and we would like to attempt to create an interface where the algorithm can be used in real applications.

References

[1] *Architecture of Bert*. https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270. Accessed: 2021-12-10.

[2] *BertTokenizer*. https://huggingface.co/docs/transformers/model_doc/bert. Accessed: 2021-10-05.

[3] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805 (2018). arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

[4] Elliott Gordon-Rodriguez et al. *Uses and Abuses of the Cross-Entropy Loss: Case Studies in Modern Deep Learning*. 2020. arXiv: 2011.05231 [stat.ML].

[5] Pengcheng He et al. "DeBERTa: Decoding-enhanced BERT with Disentangled Attention". In: *CoRR* abs/2006.03654 (2020). arXiv: 2006.03654. URL: https://arxiv.org/abs/2006.03654.

[6] Christian Janiesch, Patrick Zschech, and Kai Heinrich. "Machine learning and deep learning". In: *CoRR* abs/2104.05314 (2021). arXiv: 2104.05314. URL: https://arxiv.org/abs/2104.05314.

[7] Louis Martin et al. "CamemBERT: a Tasty French Language Model". In: *CoRR* abs/1911.03894 (2019). arXiv: 1911.03894. URL: http://arxiv.org/abs/1911.03894.

[8] Arya Roy. "Recent Trends in Named Entity Recognition (NER)". In: *CoRR* abs/2101.11420 (2021). arXiv: 2101.11420. URL: https://arxiv.org/abs/2101.11420.

[9] Shriraj P. Sawant and Shruti Singh. "Understanding Attention: In Minds and Machines". In: *CoRR* abs/2012.02659 (2020). arXiv: 2012.02659. URL: https://arxiv.org/abs/2012.02659.

[10] Eszter Simon and Dávid Márk Nemeskey. "Automatically generated NE tagged corpora for English and Hungarian". In: *Proceedings of the 4th Named Entity Workshop (NEWS) 2012*. Jeju, Korea: Association for Computational Linguistics, July 2012, pp. 38–46. URL: http://www.aclweb.org/anthology/W12-4405.

[11] Ashish Vaswani et al. "Attention Is All You Need". In: *CoRR* abs/1706.03762 (2017). arXiv: 1706.03762. URL: http://arxiv.org/abs/1706.03762.

[12] Huaqing Xiong et al. *Non-asymptotic Convergence of Adam-type Reinforcement Learning Algorithms under Markovian Sampling*. 2020. arXiv: 2002.06286 [cs.LG].