A Neural Natural Language Processing System for Educational Resource Knowledge Domain Classification

Johannes Schrumpf [1], Felix Weber [2] , Tobias Thelen [3]

Abstract:
In higher education, educational resources are the vessel with which information get transferred to the learner. Information on the content discussed in the scope of the educational resources, however, is implicit and must be inferred by the user by reading the resource title or through contextual information. In this paper we present a state-of-the-art neural natural language processing system, based on Google-BERT, that maps educational resource titles into one of 905 classes from the Dewey Decimal Classification (DDC) system. We present model architecture, training procedure dataset properties and our performance analysis methodology. We show that aside from classification performance, our model implicitly learns the class hierarchy inherent to the DDC.

Keywords: Machine Learning, AI in Higher Education, Recommender Systems

1 Introduction

In German higher education institutions, a pivotal aspect of student learning success is the choice of courses: Aside from compulsory courses, students can choose from the vast offering of courses available at their local university as well as additional material such as books, courses on MOOCs or OER for self-study. However, finding these materials can pose a challenge for students unfamiliar with the terms frequently used in a particular field. A student unfamiliar with the field may use descriptions or single terms they know to look for an introductory course. When using traditional pattern-matching search, these search queries may yield no results or return courses within another domain that by chance contain the same words but are otherwise unrelated to the domain the student is looking for. This is a challenge from a technical perspective as students may not be able to give more information about their interest domains other than a description in natural language.

Conversely, categorizing educational resources into knowledge domains poses an equal challenge, different educational resource repositories use different, sometimes

[1] Osnabrück University, Institute of Cognitive Science, Wachsbleiche 27, Osnabrück, 49090, jschrumpf@uos.de, https://orcid.org/0000-0002-0068-273X

[2] Osnabrück University, virtUOS, Postfach 4469, Osnabrück, 49069, felix.weber@uos.de, https://orcid.org/0000-0002-7012-3378

[3] Osnabrück University, Institute of Cognitive Science, Wachsbleiche 27, Osnabrück, 49090, tthelen@uos.de, https://orcid.org/0000-0002-3337-6093

incompatible meta data. Additionally, the inclusion of knowledge domains a course or an OER covers is often only inferable from their title or description.

This study aims at creating a machine-learning based natural language processing system that solves these challenges by classifying any educational resource or interest description into one of 905 classes from the Dewey Decimal Classification (DDC). For this, the system solely relies on the subject information present in the title or description of an educational resource or conversely the input sentence of a student without the need of further metadata information.

1.1 The Dewey Decimal Classification System

The DDC is a knowledge domain representation system commonly used in libraries around the globe [Wi98]. It covers a large amount of knowledge domains that reflect current and past subjects of academic enquiry. We assume the difference between educational resource knowledge domain classes and DDC classes to be small enough that one can be translated into the other without the loss of essential information. The DDC is structured in a tree-like fashion: Every DDC class possess a unique identifier number. A DDC class possesses a parent class and up to 10 child classes. These child classes typically represent a sub-domain of their parent's knowledge domain. Entries within a child class can be transferred into their respective parent class by truncating their unique identifier number. The DDC has been subject of automatic book classification attempts since the 1970's [Li13]. Recent attempts [GHA20],[KK20] focussed on applying machine learning paradigms. However, to our knowledge, no classification system based on state-of-the-art deep natural language processing as proposed in this work has been attempted so far.

1.2 Bidirectional Encoder Representations from Transformers (BERT)

BERT (Bidirectional Encoder Representations from Transformers) is a deep neural network for natural language processing. First presented in 2018 by [De19], it is based on the transformer architecture as proposed by [Va17]. A neural network comprised of multiple Encoder stacks, a building block of the transformer architecture, BERT provides deep word embeddings, which then can be used for natural language processing tasks downstream. We utilize a multilingual version of BERT-base, called Bert-base multilingual cased, which was trained on 104 languages.

2 Methods

This section describes our model, SidBERT, its architecture and the training dataset we used for training.

2.1 Architecture & Training

SidBERT is a deep neural network classifier model. It is comprised of a BERT-base multilingual cased model with a custom classification head. SidBERT is trained to classify an input text into one of 905 DDC classes. Input flows through our model via the first layer of the BERT-base layers. The input gets passed through a stack of pre-trained Encoder layers and is extracted via the output of the final Encoder. This information then gets passed to the custom classification head, consisting of a pooling operation, three fully connected layers with dropout layers in between. The fully connected layers of our model possess ReLU activations except the last layer which possesses a softmax activation function. To implement SidBERT we utilize the Huggingface-transformers library in conjunction with the Keras Tensorflow interface. Derived from Lee et al. [LTL19], we use a training strategy that segments training into three phases: In phase one and phase three, only the weights of the classification head are being trained while the weights of the original BERT network are frozen. We chose the learning rate for phase one and three to be is 3e-5. In the first phase, we train for two epochs. In phase three, we train for six epochs. In phase two, we train for four epochs with all weights being trainable. The learning rate is reduced to 1e-5.

2.2 Dataset & Class selection

Our dataset was accumulated by collecting metadata from libraries of three German University libraries as well as the repository of the German National Library. It is comprised of 1.315.962 book titles and descriptions of books from multiple languages but predominantly contains books in German. We selected 905 DDC classes for our model to be trained on. DDC classes from level one and two were excluded. Classes with at least 200 training samples are selected for training. This results in our training dataset containing classes from DDC level 3 and level 4 only. We limited the maximum number of samples per class to 1600. The statistical properties of the training dataset are listed in Table 1.

max # samples	min # samples	mean	median	SD
1600	200	659.22	444	476.28

Table 1: Statistical properties of the training dataset

3 Results

To assess model's performance, we evaluated it on a test dataset. We use two metrics for evaluation: The mean classification accuracy on the test dataset, and a structural analysis of our model's misclassification behavior. The statistical properties of the test dataset can be found in Table 3.

max #	min #	mean	median	SD
250	50	136,65	110	75.28

Table 2: Statistical properties of the test dataset

3.1 Accuracy

During training, our model achieves a 62.2% mean recall accuracy on the training dataset as well as a 45.2% mean accuracy on the test dataset.

3.2 Class relationship analysis

When it comes to machine learning classifiers, orthogonality between classes is assumed, meaning that they are independent from one another. Any classifier that does not achieve a classification accuracy of 100% possesses properties which lead to misclassification cases. By analyzing the misclassification behavior, we can draw conclusions about relationship between classes within the classifier's learned representations. By creating and analyzing a misclassification matrix, we derive a hierarchical clustering of classes utilizing Ward's minimum variance method [Wa63]. We compare the structure of this hierarchical clustering to the ground-truth structure for our 905 classes. This way, classes that have a common ancestor and that are clustered together to form a new cluster with a respective label denoting the identity of their common ancestor class. Clusters that group together classes with no common ancestor get assigned no label and are excluded from the count. Finally, we count the number of clusters that retained a label for DDC level one to three and present the results in table 4.

Label level	# of labels retained	Difference to total # of classes	Total Percentage Matching
1	109	796	12,04
2	431	474	47,62
3	674	231	74,47

Table 3: Difference between the number of clusters with label and ground truth clusters per DDC level. Difference is computed by subtracting the number of clusters with labels from 905.

4 Discussion

With an average accuracy of 45,2% on the test dataset, our model performs above a random chance accuracy. This signals that SidBERT can learn to classify input sequences into their corresponding DDC classes solely based on sample title and sample description. Comparatively, Golub et al. [GHA20] investigated the applicability of two machine learning models to classify Swedish works into one of 802 DDC classes of level 3, using titles. The first model, which uses a multinomial naïve bayes classifier, achieves a 34.89% accuracy while the second approach, a Support Vector Machine classifier, achieves a 40.91% accuracy. Our model achieves a 10.31% higher mean accuracy when compared to the naïve bayes classifier and a 4,29% higher accuracy compared to the support vector machine classifier. This performance increase is achieved while covering 905 DDC classes compared to 802, leading our model to cover a more granular distribution of the DDC. Our model's classification granularity is further elevated through its coverage of DDC classes of level 3 and 4 compared to only level 3.

When analyzing the misclassification behavior of our model, we observe a cluster retention rate of 74,47% for clusters at DDC level 3. This means that classes that are closer within the original DDC tree structure are also closer in the learned representation of SidBERT. This behavior could result from two different sources: Firstly, a sufficiently high heterogeneity between samples of classes within our training dataset. Secondly, an implicitly learned hierarchical representation of DDC classes solely based on a learned hierarchical structure from the word embeddings. Unfortunately, no clear answer can be drawn as the word embeddings are a direct result of the input samples from our training dataset. However, by looking at the retention rate of DDC level 2, we hypothesize that at least some representation of DDC structure is preserved within the representation of our model. This is because even though our model has not explicitly learned DDC classes of level 2, a label retention rate of 47,62% remains, significantly higher than what is to be expected if clusters were to be grouped together randomly. At DDC level 1, a retention rate of 12,04% remains, a rate higher than what is expected from a random clustering of classes at this level.

5 Conclusion & Outlook

In this work, we presented SidBERT, a BERT based deep neural network natural language processing model for the classification of input text into one of 905 DDC classes. Our work shows that firstly, BERT based NLP classifiers are indeed able to generalize into useable models for DDC classification. Secondly, we showed that classes with high proximity within the DDC are also close within the representations of SidBERT.

Bibliography

[DCLT19] Devlin, Jacob ; Chang, Ming Wei ; Lee, Kenton ; Toutanova, Kristina: BERT: Pre-training of deep bidirectional transformers for language understanding. In: NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference Bd. 1 (2019), Nr. Mlm, S. 4171–4186 — ISBN 9781950737130

[GoHA20] Golub, Koraljka ; Hagelbäck, Johan ; Ardö, Anders: Automatic Classification of Swedish Metadata Using Dewey Decimal Classification: A Comparison of Approaches. In: Journal of Data and Information Science Bd. 5 (2020), Nr. 1, S. 18–38

[KrKl20] Kragelj, Matjaž ; Kljajić Borštnar, Mirjana: Automatic classification of older electronic texts into the Universal Decimal Classification–UDC. In: Journal of Documentation Bd. 77 (2020), Nr. 3, S. 755–776

[LeTL19] Lee, Jaejun ; Tang, Raphael ; Lin, Jimmy: What Would Elsa Do? Freezing Layers During Transformer Fine-Tuning (2019)

[Liu13] Liu, Xiaozhong: Full-Text Citation Analysis : A New Method to Enhance. In: Journal of the American Society for Information Science and Technology Bd. 64 (2013), Nr. July, S. 1852–1863

[VSPU17] Vaswani, Ashish ; Shazeer, Noam ; Parmar, Niki ; Uszkoreit, Jakob ; Jones, Llion ; Gomez, Aidan N. ; Kaiser, Łukasz ; Polosukhin, Illia: Attention is all you need. In: Advances in Neural Information Processing Systems Bd. 2017-Decem (2017), Nr. Nips, S. 5999–6009

[Ward63] Ward, Joe H.: Hierarchical Grouping to Optimize an Objective Function. In: Journal of the American Statistical Association Bd. 58 (1963), Nr. 301

[Wieg98] Wiegand, Wayne A: The „Amherst Method": The Origins of the Dewey Decimal Classification Scheme. In: Libraries & Culture Bd. 33, University of Texas Press (1998), Nr. 2, S. 175–194