UNLPS at TextGraphs-16 Natural Language Premise Selection Task: Unsupervised Natural Language Premise Selection in Mathematical Text using Sentence-MPNet

Paul Trust
University College Cork

Provia Kadusabe
Worldquant University

Haseeb Younis
University College Cork

Rosane Minghim
University College Cork

Ahmed Zahran
University College Cork

Evangelos Millos
Dalhousie University

Abstract

This paper describes our system for the submission to the TextGraphs 2022 shared task at COLING 2022: Natural Language Premise Selection (NLPS) from mathematical texts. The task of NLPS regards selecting mathematical statements called premises in a knowledge base written in natural language and mathematical formulae that are most likely to be used to achieve a particular mathematical proof. We formulated this solution as an unsupervised semantic similarity task by first obtaining contextualized embeddings of both the premises and mathematical proofs using sentence transformers. We then obtained the cosine similarity between these embeddings and then selected premises with the highest cosine scores as the most probable. Our system improves over the baseline system that uses bag of words models based on term frequency inverse document frequency in terms of mean average precision (MAP) by about 23.5% (0.1516 versus 0.1228).

1 Introduction

Deep learning methods have achieved state of the art performance across several natural language processing (NLP) tasks in a wide variety of applications in several fields. Despite the importance of the field of mathematics and its contribution to scientific discovery, the application of NLP to mathematical text is still under-explored (Ferreira and Freitas, 2020a).

The task of natural language premise selection in mathematical text is a novel application of NLP in the field of mathematics. It involves selecting mathematical statements (premises) which are written in natural language and mathematical formulae that are most likely to be useful in proving a given conjecture or mathematical proof from a knowledge base.

More formally, given a set of premises P and a new conjecture c, all written in a combination of free text and mathematical formulae, Natural Language Premise Selection (NLPS) aims to select premises from P that will be helpful in proving a conjecture or proof c (Ferreira and Freitas, 2021). This is not a trivial task since it involves comprehending mathematical text, which in turn requires understanding of distinctive structure, discourse, and dependencies within text.

Computational approaches have been proposed to solve this task. For example, Ferreira (2021) used a graph neural network trained in a supervised learning approach to extract the most relevant premises (Ferreira and Freitas, 2020b). In this work we formulate the Natural Language Premise Selection task as an unsupervised semantic similarity task by retrieving premises that have a higher cosine similarity score with the given conjecture or proof of interest. A straightforward way to solve the task would have been to encode the premises and conjectures using word embeddings, and then perform cosine similarity to obtain the most relevant premises as those with the highest cosine score. However, this naive approach requires that both sentences are fed into the neural network, which causes a massive computational overhead. Additionally, for better performance fine tuning or pretraining the models on downstream sentence pairs may be necessary (Devlin et al., 2019), and that is computationally expensive.

In this work, we propose to use SMPNet (Sentence Masked and Permuted Language Modeling), a computationally efficient and effective sentence transformer, which is a modification of the pre-trained MPNet (Song et al., 2020a) that uses Siamese and triplet network structure to derive semantically meaningful sentence embeddings that were used for this task. MPNet (Song et al., 2020a) is a variant of transformer models (Vaswani et al., 2017) that leverages the advantages of BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) and XLNet (Yang et al.,

Proceedings of TextGraphs-16: Graph-based Methods for Natural Language Processing, pages 119–123
October 16, 2022.

2019) (Generalized Autoregressive Pretraining for Language Understanding) while avoiding their limitations.

2 Related work

2.1 Natural Language Premise Selection

The applications of natural language processing to mathematical text is still an under-explored area despite its great potential. The following are some of the key previous work on natural language premise selection. (Ferreira and Freitas, 2020b) formulates the task of premise selection from mathematical text as a link prediction problem using a deep convolution neural network. (Ferreira and Freitas, 2021) proposes a cross-model attention to learn mathematical text for natural language premise selection.

Our work is different from the previous approaches in that we focus on unsupervised learning approach using sentence transformers based on MPNet. This is an attempt to circumvent the labeling issue, which is a hard one for mathematical text, as well as improve the performance of the baseline methods mentioned.

2.2 Text Representation

Text data in most cases need to be converted into numerical values to be able to perform any meaningful machine learning operations. The form in which text is encoded directly influences the performance of models on downstream tasks. The traditional way to represent text is count-based approaches (bag of words). Bag of words approaches (Ramos et al., 2003) represent text based on the frequency of occurrence of terms in a document. The challenges with these approaches is that they sometimes do not capture any notion of similarity among semantically related words.

Static word embedding approaches that represent words as outputs of a neural network, such as word2vec (Mikolov et al., 2013) improved word representations since it was very easy to retrieve the most semantically related words for a given target word. The key weakness of these approaches is that the context in which the word is used is not captured. That means that a word has the same vector representation regardless the context in which it is used.

Contextualized language representations (Peters et al., 2018; Devlin et al., 2019) captures the context in which words are used improving performance on downstream tasks. The challenges with naive contextualized representations is that they are are not adapted for semantic similarity tasks.

Sentence embeddings (Reimers and Gurevych, 2019) modify contextualized embeddings by combining word embeddings in a sentence through a pooling strategy. They are additionally pre-trained and fine-tuned on a large corpus of sentence pairs making them ideal for semantic similarity tasks.

3 Methodology

Consider a knowledge base K from which we retrieve a collection of N mathematical premises $P = \{p_1, ..., p_N\}$ written in natural language. We would like to retrieve the premises P in K that are most likely to be useful in proving a mathematical statement or conjecture $c \in \{c_1, ...c_M\}$. We formulate this task as a semantic similarity task by retrieving premises P in K that were semantically close to a given statement or conjecture c.

3.1 Embedding Construction

The organizers of the shared task on Natural Language Premise Selection released a baseline method alongside the data, which uses a term-frequency inverse document frequency (TF-IDF) model to find the semantically related premises from a knowledge base given a mathematical conjecture, which is used to compare with our method.

3.1.1 Bag of words Baseline (TF-IDF)

TF-IDF (Term Frequency Inverse Document frequency) is a combination of two word statistics: term frequency, which is a measure of how many times a word appears in a document and Inverse Document frequency (IDF), which is a measure of whether a term is common in a given document (Ramos et al., 2003).

3.1.2 BERT

BERT (Devlin et al., 2019) is a transformer model(Vaswani et al., 2017) that was pre-trained in a bi-directional context with two objectives: masked language modeling and next sentence prediction using the bookcorpus (800 million words) and English wikipedia (2,500 million words). Additionally, the trained model can be fine-tuned for downstream tasks. Word embeddings are extracted from the last layers of the network.

3.1.3 SBERT

SBERT (Sentence-BERT) (Reimers and Gurevych, 2019) is a modification of the pre-trained BERT networks using Siamese and triplet networks, which make it able to derive semantically meaningful sentence embeddings. This model was trained using Stanford Natural Language Inference(SNLI) and Multi-Genre Natural Language Inference (MNLI) datasets. SNLI contained $570,000$ annotated sentence pairs and MNLI contained 430000 annotated sentence pairs.

3.1.4 SMPNet

SPMNet is a sentence transformer that uses a pre-trained MPNet model and fine-tuned on a large and diverse dataset of 1 billion sentence pairs using a contrastive learning objective. In our experiments, we particularly used the version named "all-mpnet-base-v2" which maps sentences and paragraphs to a 768 dimensional dense vector space (Reimers and Gurevych, 2019).

The contextualized representations of the premises and mathematical statements were obtained using sentence embeddings with SPMNet (Reimers and Gurevych, 2019). Let the obtained sentence embeddings for mathematical premises be P_E and those for conjectures be C_E.

3.2 Premise Selection

To identify the most important premises given a mathematical conjecture, we calculate the cosine similarity between the embeddings of the mathematical conjectures C_E and those of the mathematical premises P_E as follows:

$$CosineSimilarity(P_E, C_E) = \frac{P_E * C_E}{||P_{E]}|| * ||C_E||} \tag{1}$$

To retrieve the most important premises from knowledge base K for proving a conjecture C, we rank the premises according to the cosine similarity scores and select those premises that had the highest cosine similarity scores with the given mathematical conjectures

4 Experiments and Results

4.1 Datasets

The dataset used for experiments in this paper was provided by the organizers of the shared task on Natural Language Premise Selection organized at TextGraphs-16, a workshop on Graph theory and natural language processing at EMNLP 2022 (Valentino et al., 2022).

The dataset is composed of a training set ($5,519$ instances), a development set ($2,778$ instances), and a test set ($2,763$ instances), each including a list of mathematical statements and their relevant premises. The knowledge base supporting these statements contains approximately $16,205$ premises (Valentino et al., 2022).

4.2 Evaluation and Experimental setup

Our proposed model was compared with the bag of words baseline and other models using Mean Average Precision (MAP). MAP is computed as follows:

$$MAP = \frac{\sum_{i=1}^{N} AvgP(S_i)}{N} \tag{2}$$

where N is the total number of statements, S_i is the $i-$th mathematical statements and $AvgP(S_i)$ is the average precision. The test set was hosted on codalab (Valentino et al., 2022) by the organizers of the shared task.

We used Sentence-Transformers library [1] (Reimers and Gurevych, 2019) for computing sentence embedding for SBERT and SMPNet, bag of words baseline was implemented using sklearn package [2] (Pedregosa et al., 2011) and BERT word embeddings were obtained using the huggingface [3] library (Wolf et al., 2019). Our code used for the experiments can be found on https://github.com/TrustPaul/Premise-selection-coling.git

4.3 Discussion

Table 1 shows the results of our proposed approach (SPMNet) and the baseline models. Our approach (SPMNet) achieves mean average precision (MAP) of 0.151638 which is about 23% above the baseline comparison method of bag of words, which achieved MAP of 0.1228.

Additionally, we performed experimental comparison with SBERT, which is also a sentence transformer but with BERT as an underlying transformer. Experiment results from the Table 1 reveal that SBERT also outperforms the baseline bag-of-words model but is outperformed by SMPNet. We hypothesize that this is due to the impact of the underlying

[1] https://www.sbert.net/

[2] https://scikit-learn.org/

[3] https://huggingface.co/

Model	Mean Average Precision (MAP)
TF-IDF (Baseline)	0.1228
BERT Word Embedding	0.1109
Sentence BERT	0.1465
Sentence MPNet (Ours)	**0.1516**

Table 1: Mean average Precision (MAP) for models used in our experiments. TF-IDF stands for Term Frequency-Inverse Document Frequency which is a bag of words baseline, BERT stands for Bidirectional Encoder Representations from Transformers and MPNet represents Masked and Permuted Pre-training for Language Understanding

transformer model used to generate sentence embeddings, since MPNet is often a better performing model compared to BERT and also because SPMNet was fine-tuned on a larger dataset compared SBERT (1 billion sentence pairs versus $570,000$ sentence pairs)(Song et al., 2020b).

Contrary to our expectations, naive word embeddings obtained by BERT are outperformed even by the bag-of-words baseline model. This re-enforces the role played by the pre-training procedure and domain specific data employed in sentence transformers for semantic similarity tasks (Reimers and Gurevych, 2019).

5 Conclusion

In this work, we introduce an approach (SPMnet) for natural language premise selection, which is a task that involves finding the relevant theorems, axioms and definitions in natural language mathematical texts. Our proposed approach uses sentence embeddings based on the state-of-the-art transformer MPNet (Masked and Permuted Language Modeling) generating high quality embeddings that we used for retrieving the most important premises for a given mathematical conjecture. The results from our experiment show that the proposed approach (SPMNet) outperforms the baseline method (TF-IDF) by 0.028838 in mean average precision (MAP) which is a 23.5% improvement.

6 Acknowledgements

We thank Science Foundation Ireland (SFI) Center for Research Training in Advanced Networks and Future communications at University College Cork for funding this research and Irish Centre for High-End Computing (ICHEC) for providing access to computing power for running some of our experiments.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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. Association for Computational Linguistics.

Deborah Ferreira and André Freitas. 2020a. Natural language premise selection: Finding supporting statements for mathematical text. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2175–2182, Marseille, France. European Language Resources Association.

Deborah Ferreira and André Freitas. 2020b. Premise selection in natural language mathematical texts. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7365–7374, Online. Association for Computational Linguistics.

Deborah Ferreira and André Freitas. 2021. STAR: Cross-modal [STA]tement [R]epresentation for selecting relevant mathematical premises. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3234–3243, Online. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. 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. Association for Computational Linguistics.

Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020a. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020b. Mpnet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*, volume 33, pages 16857–16867. Curran Associates, Inc.

Marco Valentino, Deborah Ferreira, Mokanarangan Thayaparan, André Freitas, and Dmitry Ustalov. 2022. Textgraphs 2022 shared task on natural language premise selection. In *Proceedings of the Sixteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-16)*. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.