IJS at TextGraphs-16 Natural Language Premise Selection Task: Will Contextual Information Improve Natural Language Premise Selection?

Hanh Thi Hong TRAN[1,2,3], **Matej MARTINC**[1], **Antoine DOUCET**[3], **Senja POLLAK**[1]
[1]Jožef Stefan Institute, Slovenia
[2]Jozef Stefan International Postgraduate School, Slovenia
[3]University of La Rochelle, France

Abstract

Natural Language Premise Selection (NLPS) is a mathematical Natural Language Processing (NLP) task that retrieves a set of useful relevant premises to support the end-user finding the proof for a particular statement. In this paper, we evaluate the impact of Transformer-based contextual information and different fundamental similarity scores towards NLPS. The results demonstrate that the contextual representation is better at capturing meaningful information despite not being pretrained on mathematical background in comparison with the statistical approach (e.g., the TF-IDF) with a boost of around 3.00% MAP@500. Our code is publicly available at https://github.com/honghanhh/premise-selection.

Keywords: *Premise selection, NLPS, contextual information, Transformers.*

1 Introduction

Natural Language Premise Selection (NLPS) (Ferreira and Freitas, 2020a), inspired by the field of Automated Theorem Proving, is a mathematical NLP task that retrieves a set of useful relevant premises. Given a mathematical statement written in natural language as the input, NLPS systems predict the relevant premises that could support an end-user finding a proof for that mathematical statement.

Mathematically, NLPS task can be defined as:

Definition 1.1. *Given a new mathematical statement s, that requires a mathematical proof, and a collection (or a knowledge base) of premises $P = p_1, p_2, \ldots, p_{Np}$, with size Np, retrieve the premises in P that are most likely to be useful for proving s.*

The premises often include supporting definitions and propositions, which can act as explanations for the proof process. Figure 1 presents examples of 2 premises that support a given mathematical statement or theorem.

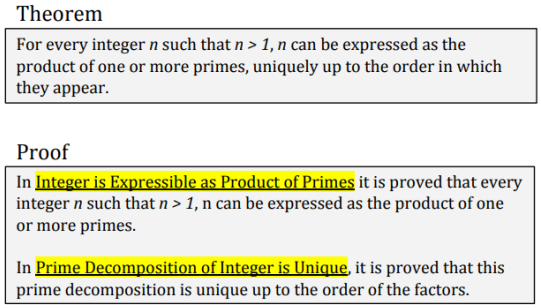

Figure 1: Example premises supporting a given theorem (Ferreira and Freitas, 2020a).

Most of the existing systems focus on manual feature engineering or statistical approaches to extract meaningful mathematical knowledge, with one exception being the study by Ferreira and Freitas (2020b), where they tackle the task by employing Deep Convolutional Graph Neural Networks (DCGNN) on graph representations. The state of the art models for NLP such as BERT (Devlin et al., 2016) are not fully explored under the assumption that they do not encode the intricate mathematical background knowledge needed to reason over mathematical discourse.

The 1^{st} Shared Task on Natural Language Premise Selection (Valentino et al., 2022), organized as part of the TextGraphs 2022 workshop, presented one of the first opportunities to systematically compare different approaches towards a NLPS task in an Information Retrieval setting, by adopting PS-ProofWiki (Premise Selection-ProofWiki) dataset (Ferreira and Freitas, 2020a). This dataset can be considered as the baseline corpus for our specific shared task.

The contributions of this paper can be summarised as follows:

- An empirical evaluation of several contextual representations relying on Transformer-based language models;

Proceedings of TextGraphs-16: Graph-based Methods for Natural Language Processing, pages 114–118
October 16, 2022.

- Evaluation of the performance of different similarity scores, including Cosine, Euclidean, and Manhattan score on the NLPS task.

This paper is organised as follows: Section 2 presents the related work in premise selection. Next, we introduce our methodology, experimental setup and evaluation metrics in Section 3. The corresponding results are presented in Section 4. Finally, we conclude our work and suggest future directions in Section 5.

2 Related work

In this section, we present the related research in NLP applied to the NLPS task in the domain of Automated Theorem Proving.

The research was first introduced by Alama et al. (2014), who employed corpus analysis and kernel-based methods, in order to showcase the usefulness of automatic premise selection systems for proving the conjectures in the field of Automated Theorem Proving (ATP). Few years later, Irving et al. (2016) proposed a neural deepmath-deep sequence architecture for premise selection using formal statements from the Mizar corpus, which solved 67.90% of the conjectures present in the Mathematical Mizar Library. Other machine learning based approaches have also been investigated for the task at hand (e.g. KNN (Gauthier and Kaliszyk, 2015), Random Forest (Färber and Kaliszyk, 2015), to mention a few).

Similar to the previous research, (Ferreira and Freitas, 2021) formulate this problem as a pairwise relevance classification problem and present STAR, a cross-modal representation for mathematical statements with two layers of self-attention, one for each language modality present in the mathematical text.

Recently, Ferreira and Freitas (2020a) introduced a new systematic formulation of the task under the name Natural Language Premise Selection (NLPS) and published a new evaluation corpus called NL-PS. They propose two baseline approaches, using TF-IDF and PV-DBOW (Le and Mikolov, 2014). Additionally, they also suggested to model the task as a pairwise relevance classification problem and tackled it by employing neural contextual representations, namely BERT and SciBERT (Beltagy et al., 2019).

While the previous work focused on capturing either content (local) or structural dependencies (global) across natural language mathematical statements, Ferreira and Freitas (2020b) were the first to consider NLPS as a link prediction problem using Deep Convolutional Graph Neural Networks (DCGNN), with the aim of capturing both local and global information. Their study demonstrates the capability of graph embeddings to capture structural and content elements of mathematical statements.

3 Methodology

3.1 Data

The experiments are conducted on PS-ProofWiki (so-called Premise Selection-ProofWiki) dataset (Ferreira and Freitas, 2020a), which contains 3 subsets: training set, development set, and test set. Each mentioned subset includes a list of mathematical statements and their relevant premises. The number of instances in each subset are presented in Table 1. Besides, there is a knowledge base supporting these statements, which contains approximately 16,205 premises.

Subsets	Amount
Training set	5,519
Development set	2,778
Test set	2,763

Table 1: The number of examples in PS-ProofWiki's subsets.

Initially, the dataset was used for evaluating semantic representations (e.g., textual entailment and inference for mathematics (Ferreira and Freitas, 2020a), embeddings (Ferreira and Freitas, 2021), or mathematical discourse (Ferreira et al., 2022)). Regarding our research, we adopt the dataset for NLPS task with the aim to retrieve the set of relevant premises for a given statement in the test set by ranking the sentences contained in the supporting knowledge base.

3.2 Methods

Our research focuses on the impact of contextual information from Transformer-based language models compared with the statistical approaches (baselines) towards NLPS task. For simplification and better comparison, we extract contextual representations from different Transformer-based language models and compute several similarity scores to rank how likely the sentences in the knowledge

base are a part of the set of premises for a given mathematical statement. The overall workflow is presented in Figure 2.

We employ several Transformer-based models, including PatentSBERTa (Bekamiri et al., 2022) (*PatentSBERTa*), T5-Large (Raffel et al., 2020) (*gtr-t5-large* and *sentence-t5-large*), RoBERTA-Large (Liu et al., 2019) (*all_datasets_v3_roberta-large*), Mpnet-Base (Song et al., 2020) (*all-mpnet-base-v2* and *all-mpnet-base-v_outcome_sim*), MiniLM (Wang et al., 2020) (*all-MiniLM-L6-v2* and *ll-MiniLM-L12-v2*). The models were obtained from the Hugging Face library[1] and were chosen according to the number of downloads and likes criteria.

Note that all the chosen models share the same pretraining purpose: they aim to train sentence embedding models on very large textual datasets using a self-supervised learning objective. As sentence Transformer models, they map the sentences and paragraphs to a dense vector space. Thus, we encoded the statements and premises into vector representations and then used different similarity metrics to calculate the similarity between a specific premise and the corresponding statement. The obtained similarity scores are afterwards used for ranking the premises in a descending order. We keep top 500 most relevant premises for each statement. We compare three similarity metrics, namely Cosine, Euclidean, and Manhattan similarity. All the experiments have been ran on a A100-PCIE-40GB GPU.

3.3 Evaluation metrics

For each model, we retrieve the top 500 premises from the knowledge base that support a given statement. We use Mean Average Precision at K (MAP@K) with K = 500 for the evaluation. This evaluation metric has also been used in the related work (Ferreira and Freitas, 2020a), thus our results are directly comparable to the state of the art methods.

4 Results

In this Section, we evaluate the suitability of different contextual representations of premises from the knowledge base for retrieving the top relevant premises for a given statement in the test set. We also compare the obtained results with the results of the shared task baseline (Valentino et al., 2022).

[1] https://huggingface.co/

Table 2 presents the performance of contextual representations extracted from different Transformer-based pretrained language models using Cosine similarity as the similarity metric. The shared task baseline to which we compare our approaches uses a simple term frequency model (TF-IDF) to rank how likely the sentences (premises) in the knowledge base are a part of the set of premises for a given mathematical statement.

Representation	MAP@500
sentence-t5-large	0.134110
gtr-t5-large	0.139367
all-mpnet-base-v_outcome_sim	0.144706
PatentSBERTa	0.146141
all-MiniLM-L6-v2	0.146995
all-mpnet-base-v2	0.151724
all-MiniLM-L12-v2	0.152427
all_datasets_v3_roberta-large	**0.153897**
Baseline	0.122800

Table 2: Performance of different representations on the test data using Cosine similarity score.

The results demonstrate that by employing Transformer-based models we can outperform the statistical baseline by a relatively large margin in terms of the MAP@500 evaluation metric. The best contextual representation for the task at hand was obtained by employing the large version of RoBERTa. Using this model, we can improve on the baseline performance by 3.11 percentage points. All tested contextual representations manage to outperform the baseline, with the performance improvement ranging from about 1.00 to 3.00 percentage points in terms of MAP@500. This indicates that contextual representations from Transformer-based language models are capable of encoding meaningful information from intricate mathematical background knowledge despite not being pretrained on domain-specific mathematical texts.

Similarity score	MAP@500
Cosine	0.153897
Euclidean	0.153896
Manhattan	**0.153902**

Table 3: Similarity score performance on the test data using RoBERTa embeddings

.

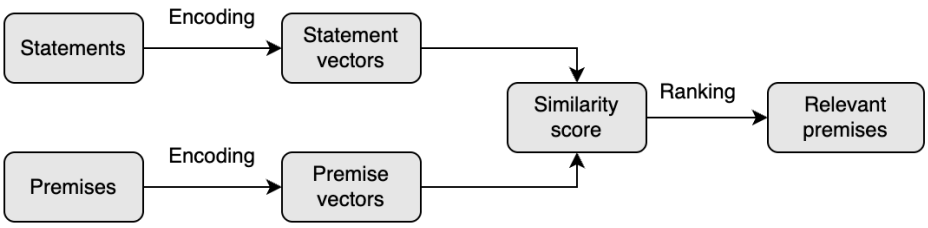

Figure 2: Our general workflow.

Using the contextual representations obtained from our best model, i.e. the large version of RoBERTa, we also evaluate three different similarity scores used for measuring similarity between premise and statement representations, namely Cosine, Euclidean, and Manhattan similarities. The results presented in Table 3 show that Manhattan similarity works slightly better than the other two similarity measures, although the difference is marginal in terms of MAP@500.

Teams	MAP@500	Ranking
IJS	0.1539	1
PaulTrust	0.1516	2
kamivao	0.1460	3
langml	0.1414	4
Organizers	0.1228	5

Table 4: Ranking on the shared task leaderboard.

Table 4 presents comparison between our proposed approach and the approaches proposed by other teams participating in the shared task in terms of rank and MAP@500. As can be seen, our system outperforms all others. Regarding the reproducibility and complexity, our approach uses a simple paradigm that is easy to reproduce and scale to large knowledge bases, but nevertheless offers a relatively efficient retrieval of premises.

5 Conclusion

In this paper, we have investigated the performance of contextual representations towards the task of Natural Language Premise Selection. We also evaluated the impact of different similarity scores. By using the contextual information obtained from the pretrained Transformer-based models in order to obtain premise and statement representations, we manage to outperform the baseline statistical approach using TF-IDF (the baseline) by a decent margin of around 3 percentage points in terms of MAP@500. These findings serve as a good initiative to explore the potential of using language models' for the NLPS task further. We also showed that by using the Manhattan distance for measuring similarity between representations, we can improve the performance by a small margin.

There remains a lot of room for improvement. In the future, we would like to investigate the effect of different mathematical representations on the performance of the model, e.g., by feeding the model graph representations. Combinations of contextual and graph representations will also be explored.

6 Acknowledgements

The work was partially supported by the Slovenian Research Agency (ARRS) core research programme Knowledge Technologies (P2-0103) and the project CANDAS (J6-2581). The first author was partly funded by Region Nouvelle Acquitaine. This work has also been supported by the TERMITRAD (2020-2019-8510010) project funded by the Nouvelle-Aquitaine Region, France.

References

Jesse Alama, Tom Heskes, Daniel Kühlwein, Evgeni Tsivtsivadze, and Josef Urban. 2014. Premise selection for mathematics by corpus analysis and kernel methods. *Journal of automated reasoning*, 52(2):191–213.

Hamid Bekamiri, Daniel S Hain, and Roman Jurowetzki. 2022. Patentsberta: A deep nlp based hybrid model for patent distance and classification using augmented sbert. *Technological Forecasting and Social Change*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2016. Bert: Bidirectional encoder representations from transformers.

Michael Färber and Cezary Kaliszyk. 2015. Random forests for premise selection. In *International Symposium on Frontiers of Combining Systems*, pages 325–340. Springer.

Deborah Ferreira and André Freitas. 2020a. Natural language premise selection: Finding supporting statements for mathematical text. *arXiv preprint arXiv:2004.14959*.

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.

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.

Deborah Ferreira, Mokanarangan Thayaparan, Marco Valentino, Julia Rozanova, and André Freitas. 2022. To be or not to be an integer? encoding variables for mathematical text. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 938–948.

Thibault Gauthier and Cezary Kaliszyk. 2015. Premise selection and external provers for hol4. In *Proceedings of the 2015 Conference on Certified Programs and Proofs*, pages 49–57.

Geoffrey Irving, Christian Szegedy, Alexander A Alemi, Niklas Eén, François Chollet, and Josef Urban. 2016. Deepmath-deep sequence models for premise selection. *Advances in neural information processing systems*, 29.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

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

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.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.