Keyword-based Natural Language Premise Selection for an Automatic Mathematical Statement Proving

Doratossadat Dastgheib[†] and **Ehsaneddin Asgari**[*]
[†] Language Processing and Digital Humanities Lab, Tehran, Iran.
[†] Department of Computer and Data Science, Shahid Beheshti University, Tehran, Iran.
∗ NLP Expert Center, Data:Lab, Volkswagen AG, Munich, Germany
d_dastgheib@sbu.ac.ir, asgari@berkeley.edu

Abstract

Extraction of supportive premises for a mathematical problem can contribute to profound success in improving automatic reasoning systems. One bottleneck in automated theorem proving is the lack of a proper semantic information retrieval system for mathematical texts. In this paper, we show the effect of keyword extraction in the natural language premise selection (NLPS) shared task proposed in TextGraph-16 that seeks to select the most relevant sentences supporting a given mathematical statement.

1 Introduction

A mathematical statement requires a collection of appropriate definitions, previously proved statements, and inference rules to be proved. The automatic reasoning field deals with computing systems automating proof procedures and proof checking. One of the considerations in implementing automatic deduction and artificial intelligence approaches is restricting the proof search space and preventing the automatic prover from pursuing unfruitful reasoning paths. A dual aspect of search is looking for previous results that could be useful in proof completion (Portoraro, 2021).

Premise selection was initially introduced in (Blanchette et al., 2016) as a task to select a part of a formal library that improves the chance that an automatic prover can prove a mathematical conjecture. In (Irving et al., 2016), neural network-based premise selectors were applied for the first time, and (Ferreira and Freitas, 2021) reformulated the problem as a pairwise relevance classification problem.

Similar challenges in mathematical context have been proposed, such as ARQMATH (Zanibbi et al., 2020) seeking an answers retriever and ranker for a given mathematical question. An answer retriever system mainly needs to consider mathematical text similarities. However, the premise selector task also requires a mathematical concept understanding component.

In this study, we work on the shared-task introduced by the 16^{th} Workshop on Graph-Based Natural Language Processing (Valentino et al., 2022) on natural language premise selection. In this task, the teams are given a collection of mathematical statements in natural language and the goal is to retrieve supportive premises from a knowledge-base that can prove certain statements.

In this study we look into the effectiveness of keyword extraction in selecting premises for proving each statement outperforms the TF-IDF-based baseline.

2 Approach

2.1 Data Description

The dataset used in this task is a collection of mathematical statements and their premises extracted from ProofWiki, available in (Ferreira and Freitas, 2020). Each statement in the dataset is expressed in natural language, and the formulas are in LaTeX format. An overview of the dataset can be found in Table 1. The collection contains 21614, statements spanning 1227949, tokens in total.

2.2 Preprocessing

For data cleaning, we perform specific preprocessing steps, e.g., removing LaTeXcommands such as `begin` that describe a part of a formula in the sentence from the texts of statements. We perform this step to avoid their extractions as keywords in the next part of the pipeline. Then using an automatic keyword extractor (Campos et al., 2020), we generate up to 20 keywords for each sentence. Table 1 provides sample keywords for an example statement.

Embedding. To compare the semantic and context similarity of keywords, we also produce all keywords embeddings using fastText embedding pretrained on Wikipedia (Joulin et al., 2016).

Proceedings of TextGraphs-16: Graph-based Methods for Natural Language Processing, pages 124–126
October 16, 2022.

	Train	Dev	Test	Knowledge Base
Instance Number	5519	2778	2763	16205
Statement Example	Let $Q_n = \langle a_j \rangle_{0 \le j \le n}$ be a geometric sequence of length n consisting of positive integers only. Let a_1 and a_n be coprime. Then the jth term of Q_n is given by: $a_j = q^j p^{n-j}$			
Premise Example	Let $\langle x_n \rangle$ be a geometric sequence in $\mathbb{R}$ defined as $x_n = ar^n$ for $n = 0, 1, 2, 3, \dots$ The parameter: $r \in \mathbb{R} : r \neq 0$ is called the common ratio of $\langle x_n \rangle$.			
Statement Keywords				Premise Keywords
sequence, length, consisting, geometric, positive, integers, coprime, term				sequence, defined, geometric, parameter, called, common, ratio

Table 1: Overview of available dataset for retrieving supportive premises along with an example statement and one of its premises with their respective extracted keywords.

2.3 Retrieval Approach

The retrieval system should assign a score between the statements and their candidate premises. For sentences S_1, S_2 in dataset (coming from statement or premises) we extract the keyword sets KS_1, and KS_2 respectively. We define our suggested schemes for scoring as follows:

1. **Keyword Jaccardian Similarity.** The intersection cardinality over union cardinality of extracted keywords from the statement and the candidate premise:
$$Score(KS_1, KS_2) = \frac{|KS_1 \cap KS_2|}{|KS_1 \cup KS_2|}$$
2. **Keyword Affecting Relevance Score.** We measure the affecting relevance scores of keywords in the intersection keywords set:
$$Score(KS_1, KS_2) = \sum_{k_i \in KS_1 \cap KS_2} (1 - r_{i_1}) \times (1 - r_{i_2})$$
where r_{i_1} and r_{i_2} are keyword scores for keyword k_i in the sentences S_1 and S_2 respectively.
3. **Keyword Embedding Similarity.** Sum of cosine similarity of embeddings in two keyword sets:
$$Score(KS_1, KS_2) = \sum_{k_1 \in KS_1, k_2 \in KS_2} \textbf{cos-sim}(k_1, k_2)$$

We select the premises with maximum scores as the ultimate premise for each statement.

2.4 Evaluation

The systems are supposed to rank the sentences in the knowledge base premises for a given mathematical statement. We evaluate our NLPS system using Mean Average Precision (MAP) for 500 top premises retrieved from the knowledge base and introduced the term frequency (TF-IDF) model as a baseline.

3 Results

The results achieved using methods described in the previous section compared to the baseline score are presented in Table 2. Keyword-based approaches performed reasonably well in retrieving premises for given mathematical statements and outperformed the TF-IDF-based baseline. However, the embedding-based approach did not achieve competitive performance. One reason can be the ambiguity in the fixed embeddings as fastText.

4 Conclusions

In this paper, we checked the effectiveness of keyword extraction of mathematical statements for premise selection shared task NLPS and considered three keyword scoring schemas. Given statements,

Method	**Dev**	**Test**
Base line	0.1239	0.1228
Jaccardian Sim.	**0.1364**	**0.1414**
Affected Rel.	0.1256	0.129
Embedding Sim.	0.0539	0.05

Table 2: Mean Average Precision (MAP) socre for of our proposed methods in comparison with the tf-idf baseline.

we scored the keywords extracted for each statement and selected supportive sentences. Results show that keywords of statements can be effectively used in selecting relevant premises.

References

Jasmin C. Blanchette, Cezary Kaliszyk, Lawrence C. Paulson, and Josef Urban. 2016. Hammering towards qed. *Journal of Formalized Reasoning*, 9(1):101–148.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.

Deborah Ferreira and André Freitas. 2020. 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. 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.

Geoffrey Irving, Christian Szegedy, Alexander A Alemi, Niklas Een, Francois Chollet, and Josef Urban. 2016. Deepmath - deep sequence models for premise selection. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Frederic Portoraro. 2021. Automated Reasoning. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Fall 2021 edition. Metaphysics Research Lab, Stanford University.

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.

Richard Zanibbi, Douglas W. Oard, Anurag Agarwal, and Behrooz Mansouri. 2020. Overview of arqmath 2020: Clef lab on answer retrieval for questions on math. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings*, page 169–193, Berlin, Heidelberg. Springer-Verlag.