TextGraphs 2022 Shared Task on Natural Language Premise Selection

Marco Valentino[1,2], **Deborah Ferreira**[2], **Mokanarangan Thayaparan**[1,2], **André Freitas**[1,2], **Dmitry Ustalov**[3]
[1]Idiap Research Institute, Switzerland
[2]University of Manchester, United Kingdom
[3]Toloka, Serbia

Abstract

The Shared Task on *Natural Language Premise Selection (NLPS)* asks participants to retrieve the set of premises that are most likely to be useful for proving a given mathematical statement from a supporting knowledge base. While previous editions of the TextGraphs shared tasks series targeted multi-hop inference for explanation regeneration in the context of science questions (Thayaparan et al., 2021; Jansen and Ustalov, 2020, 2019), NLPS aims to assess the ability of state-of-the-art approaches to operate on a mixture of natural and mathematical language and model complex multi-hop reasoning dependencies between statements. To this end, this edition of the shared task makes use of a large set of approximately 21k mathematical statements extracted from the PS-ProofWiki dataset (Ferreira and Freitas, 2020a). In this summary paper, we present the results of the 1st edition of the NLPS task, providing a description of the evaluation data, and the participating systems. Additionally, we perform a detailed analysis of the results, evaluating various aspects involved in mathematical language processing and multi-hop inference. The best-performing system achieved a MAP of 15.39, improving the performance of a TF-IDF baseline by approximately 3.0 MAP.[1]

1 Introduction

The articulation of mathematical language represents a core feature of human intelligence, requiring complex reasoning capabilities and abstraction as well as a correct evaluation of the semantics of mathematical structures and its internal components (Greiner-Petter et al., 2019). Moreover, mathematical language consists in a combination of words and symbols, which act following different rules and alphabets, but preserving, at the same time, mutual dependencies that are necessary for the comprehension of mathematical discourse (Ganesalingam, 2013).

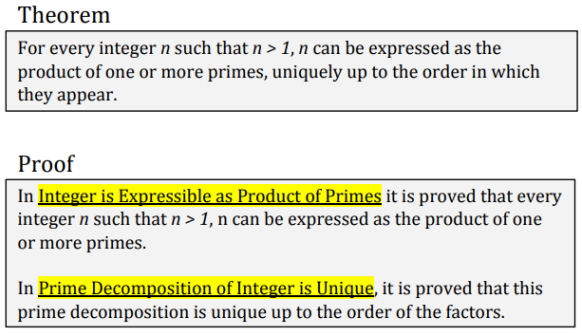

Figure 1: Given a mathematical statement s, that requires a mathematical proof, and a collection of premises P, the task of Natural Language Premise Selection (NLPS) consists in retrieving the premises in P that are most likely to be useful for proving s (Ferreira and Freitas, 2020a).

These features provide a unique set of opportunities for the evaluation of state-of-the-art models in Natural Language Processing (NLP) (Ferreira and Freitas, 2020a,b; Welleck et al., 2021). To encourage new lines of research at the intersection of natural language and mathematics, we propose the 1st Shared Task on *Natural Language Premise Selection (NLPS)*.

The NLPS task asks participants to retrieve the premises that are most likely to be useful for proving a given mathematical statement from a supporting knowledge base (see Figure 1). Specifically, NLPS is designed to assess the capabilities and behaviours of state-of-the-art approaches in dealing with a mixture of natural language and mathematical text along with the modelling of complex multi-hop dependencies between statements. To this end, this edition of the shared task makes use of a large set of approximately 21k mathematical statements extracted from the PS-ProofWiki dataset (Ferreira and Freitas, 2020a).

In this summary paper, we present the results of

[1]Data and code available online: https://github.com/ai-systems/tg2022task_premise_retrieval.

Proceedings of TextGraphs-16: Graph-based Methods for Natural Language Processing, pages 105–113
October 16, 2022.

the 1st edition of the Natural Language Premise Selection task, providing a detailed description of the evaluation data, and the participating systems. Moreover, we perform a detailed analysis of the behaviour of the participating systems, evaluating various aspects involved in mathematical language processing (i.e., the ability to deal with an increasing number of mathematical elements) and multi-hop inference. The best performing system achieved a MAP of 15.39, improving the performance of a TF-IDF baseline by approximately 3.0 MAP, while still leaving a large space for future improvements.

2 Natural Language Premise Selection

Given a mathematical statement s that requires a mathematical proof, and a collection (or a knowledge base) of premises $P = \{p_1, p_2, \ldots, p_{N_p}\}$, with size N_p, the task of Natural Language Premise Selection (NLPS) consists in retrieving the premises in P that are most likely to be useful for proving s.

A mathematical statement can be a definition, an axiom, a theorem, a lemma, a corollary or a conjecture. Premises are composed of universal truths and accepted truths. Definitions and axioms are *universal truths* since the mathematical community accepts them without proof. *Accepted truths* include statements that need a proof before being adopted. Theorems, lemmas and corollaries are such types of statements. These statements were, at some point, framed as a conjecture before they were proven. As such, they can be grounded on past mathematical discoveries, referencing their own supporting premises (i.e., the background knowledge that was used to prove the conjecture). This network structure of available premises can be used as a foundation in order to predict new ones. The relationship between these statements can be leveraged to build models that can better perform inference for mathematical text (Ferreira and Freitas, 2020b,a).

The NLPS task can be particularly challenging for existing Information Retrieval systems since it requires the ability to process both natural language and mathematical text (Ferreira and Freitas, 2020a; Ferreira et al., 2022). Moreover, as shown in the example in Figure 2, the retrieval of certain premises necessitates complex multi-hop inference (Ferreira and Freitas, 2020b).

Statement Type	Data Split				
	KB	Train	Dev	Test	All (Unique)
Definitions	7,077	0	0	0	7,077
Lemmas	252	134	70	69	252
Corollaries	161	113	57	57	275
Theorems	8,715	5,272	2,652	2,636	14,003
Total	16,205	5,519	2,778	2,763	21,746

Table 1: Types of mathematical statements present in PS-ProofWiki. The table shows the number divided by the data split. The last columns shows the total unique entries for each mathematical type.

3 Training and Evaluation Data

PS-ProofWiki (Ferreira and Freitas, 2020a) has a total of 21,746 different entries, composed of definitions, lemmas, corollaries and theorems, as shown in Table 1. Note that only the Knowledge Base contains definitions since definitions do not contain proofs and, consequently, do not have premises. However, definitions are often used as premises playing a fundamental role in the NLPS task. There also exists an intersection between the KB and the training set. Accordingly, we include the last column to account for all unique entries in the dataset.

Figure 3 presents a histogram with the frequency of the different number of premises. We can observe that the statements usually have a small number of premises, with $9,640$ (Around 87% of the entries in the Train/Dev/Test set) statements containing between one and five premises. The highest number of premises for one theorem is 72.

Similarly, the histogram in Figure 4 shows the frequency of the dependencies between statements, reporting how many times each statement is used as a premise. A total of 4,236 statements is connected to between one and three dependants. On average, the statements contain a total of 289 symbols (characters and mathematical symbols).

The dataset provides a specific semantic modelling challenge for natural language processing as it requires specific tokenisation and the modelling of specific discourse structures tailored towards mathematical text, such as encoding mathematical elements along with natural language, and encoding the relationship between conjectures and premises.

4 System Descriptions and Performance

Following the previous editions of the TextGraphs Shared Tasks on Multi-Hop Inference for Explanation Regeneration (Thayaparan et al., 2021; Jansen

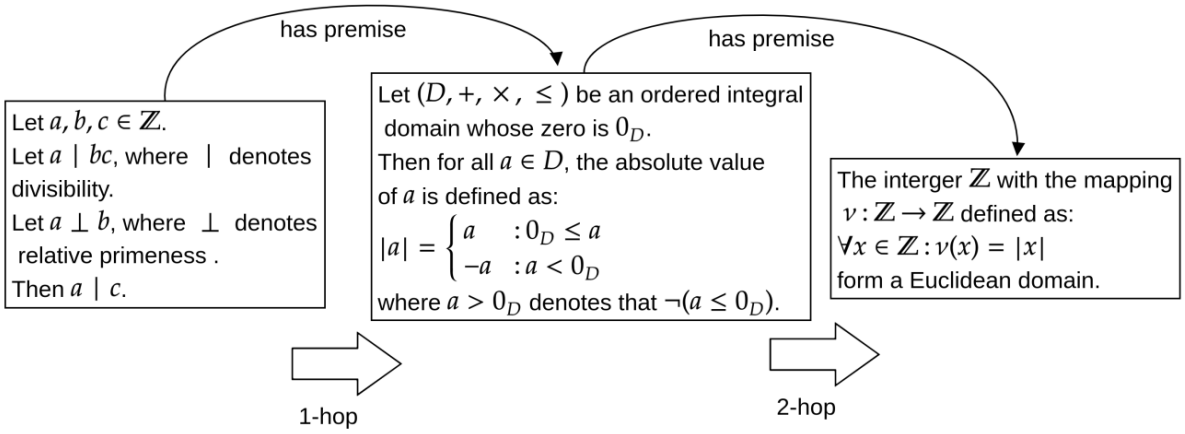

Figure 2: Example of premises requiring multi-hop inference.

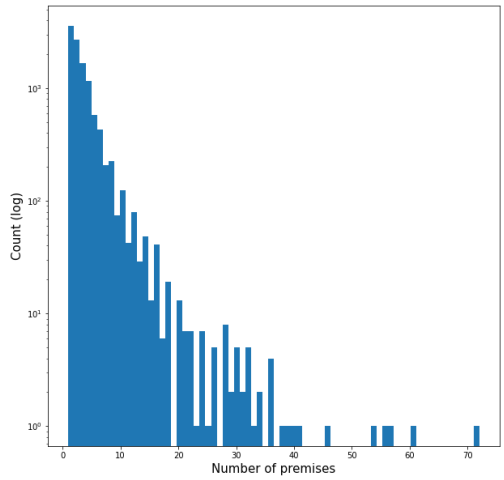

Figure 3: Distribution of the number of premises in the ProofWiki corpus. Log transformation is applied to facilitate visualisation for the y axis.

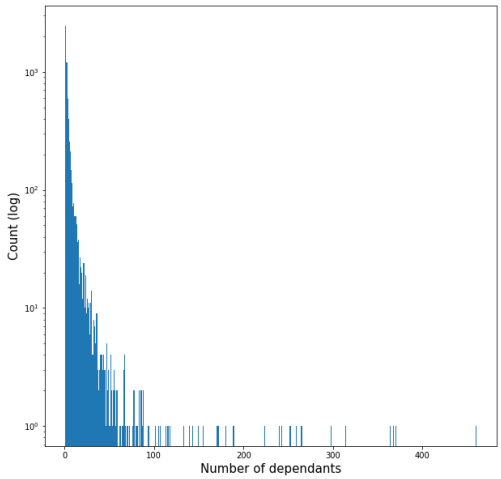

Figure 4: Number of times a statement is referred as a premise. Log transformation is applied to facilitate visualisation for the y axis.

and Ustalov, 2020, 2019), we frame Natural Language Premise Selection (NLPS) as a ranking problem. To this end, the participating systems have been evaluated using Mean Average Precision (MAP) at K, with $K = 500$. Specifically, the top 500 premises retrieved for supporting a given mathematical statement are compared against the gold premises in the corpus via MAP.

The competition has been organised on CodaLab (Pavao et al., 2022),[2] with a total of four teams submitting their solutions to the leaderboard (Tran et al., 2022; Trust et al., 2022; Kovriguina et al., 2022; Dastgheib and Asgari, 2022). Table 2 presents the overall results of the evaluation phase (test-set). In general, the shared task attracted a diverse set of submissions adopting methods spanning from state-of-the-art Transformers (Vaswani et al., 2017) to lexical-based approaches. All the participating systems improved the performance of a TF-IDF baseline, with the best performing system (IJS) achieving a MAP score of 15.39. However, the relatively low performances of the systems demonstrate that the task is still challenging for existing models, leaving large space for future improvements.

Here, we summarize the key features of the models proposed by the participating teams:

[2] https://codalab.lisn.upsaclay.fr/competitions/5692

Team Name	MAP
IJS (Tran et al., 2022)	15.39
UNLPS (Trust et al., 2022)	15.16
Kamivao (Kovriguina et al., 2022)	14.60
langml (Dastgheib and Asgari, 2022)	14.14
TF-IDF baseline	12.28

Table 2: Overall results of the 1st Shared Task on Natural Language Premise Selection (NLPS).

TF-IDF baseline. The shared task data distribution included a baseline that employs a term frequency model (TF-IDF) (see, e.g. Manning et al., 2008, Ch. 6). Specifically, the TF-IDF baseline employs sparse vector representations in combination with cosine similarity to estimate how likely a given premise in the knowledge base supports the mathematical statements provided as input. This baseline achieves a MAP score of 12.28.

IJS (Tran et al., 2022). The team investigates the task of NLPS evaluating the impact of Transformer-based contextual representations along with several similarity metrics for retrieval. Specifically, the authors propose a systematic evaluation of different pre-trained Sentence-Transformers (Reimers and Gurevych, 2019) using a bi-encoder architecture. In order to rank the premises, the authors extract the contextual representation from different Transformers, computing the similarity scores to rank how likely the sentences in the supporting knowledge base are to be a part of the set of premises for a given mathematical statement. The authors observe that the best performance are obtained via RoBERTa large (Liu et al., 2019) and Manhattan distance achieving a MAP score of 15.39.

UNLPS (Trust et al., 2022). Similar to IIJS, the team explore the usage of Sentence-Transformers (Reimers and Gurevych, 2019), employing a bi-encoder architecture for addressing the NLPS task. The team does not rely on fine-tuning techniques but, instead, adopts pre-trained Transformers to retrieve the most relevant premises via a cosine similarity score. The team demonstrated that employing the Sentence-Transformer SMPNet model, which internally adopts a pre-trained MPNet (Song et al., 2020), yields a MAP score of 15.16.

Kamivao (Kovriguina et al., 2022). The team proposes an approach based on a mixture of dense retrieval and prompt-based methodology. Specifically, the proposed model combines a bi-encoder based on a pre-trained Sentence-Transformer (Reimers and Gurevych, 2019) (BERT (Devlin et al., 2019) and MathBERT (Peng et al., 2021)) with a GPT3 model (Brown et al., 2020) which is instructed to re-rank a set of candidate premises. In the first stage, the model uses bi-encoders and cosine similarity to retrieve a list of potentially relevant premises, while in the re-ranking stage, the authors adopt a prompt-based methodology to construct specific instructions for GPT-3. This approach achieves a MAP score of 14.60.

Iangml (Dastgheib and Asgari, 2022). The team proposes a method that relies on keywords extraction and matching to select relevant premises. The proposed approach employs a keyword extractor (Campos et al., 2020) to generate up to 20 keywords for each sentence. The team proposes and evaluates a range of similarity functions based on the extracted keyworkds through the generation of sparse embeddings. The embeddings are generated using the fastText model (Joulin et al., 2017). The scoring functions are then applied to re-rank the top 500 premises retrieved by the TF-IDF baseline. Their experiments show that the Jacardian similarity scoring function yields the best MAP performance of 14.14.

5 Detailed Analysis

In order to better evaluate and characterise the behaviour of the proposed systems beyond the aggregated MAP score, we carried out an additional analysis by partitioning the set of mathematical statements according to different categories.

Specifically, we categorise the statements in the test-set according to the total number of occurring mathematical elements (e.g., equations, variables, etc.) and the total number of gold premises. In particular, these categories allow for the evaluation of the behaviour of the systems when (a) dealing with a mixture of natural language and mathematical text and (b) retrieving premises that require multi-hop inference. The larger the number of premises supporting a given mathematical statement, in fact, the higher the number of inference steps that are likely to be required in the NLPS task.

The results of this analysis are reported in Table 3 and Table 4.

Team Name	Overall	0–5	5–10	10–20	20+
IJS	15.39	**13.89**	17.37	13.95	9.36
UNLPS	15.16	13.53	**17.43**	**14.03**	**10.08**
Kamivao	14.60	13.58	16.07	13.73	7.46
langml	14.14	12.24	16.20	13.86	7.62
TF-IDF baseline	12.28	11.27	13.29	11.70	7.15

Table 3: MAP score by number of *mathematical elements* in a mathematical statement.

Team Name	Overall	0–5	5–10	10–20	20+
IJS	15.39	**15.96**	**13.37**	**10.92**	5.89
UNLPS	15.16	15.67	13.33	10.57	5.40
Kamivao	14.60	15.05	12.84	10.03	6.76
langml	14.14	14.45	12.95	10.86	**8.02**
TF-IDF baseline	12.28	12.64	11.47	8.93	7.84

Table 4: MAP score by number of *gold premises* supporting a mathematical statement.

5.1 Number of Mathematical Elements

In order to count the number of mathematical elements in a given statement, we create apposite regular expressions leveraging the special characters used to write equations in LaTeX (e.g., "$"). Subsequently, we recompute the performance of the systems, grouping the statements in the test-set by the number of occurring mathematical elements (see Table 3).

Overall, the analysis reveals that the performances significantly decrease for all the participating systems, including the TF-IDF baseline. In addition, we observe that the second system in the overall ranking (UNLPS) is actually the most robust when dealing with an increasing number of mathematical elements. Since IJS and UNLPS employ a similar architecture based on pre-trained Sentence-Transformers (Reimers and Gurevych, 2019), the difference in results might be attributed to the specific model adopted in the experiments. UNLPS, in fact, adopts a pre-trained MPNet (Song et al., 2020) while IJS uses RoBERTa-large (Liu et al., 2019). At the same time, the overall decrease in performance confirms that additional work is still required to make Transformer-based representations able to deal with a mixture of natural language and mathematical text (Ferreira et al., 2022).

5.2 Number of Gold Premises

We perform a similar analysis by grouping the mathematical statements in the test-set according to the number of gold supporting premises. In this case, we assume that the larger the number of premises, the higher the probability of systems required to perform multi-hop inference for addressing the NLPS task (see Table 4).

Overall, a similar trend can be observed when investigating the behaviours of the systems on statements requiring an increasing number of supporting premises. The results in Table 4, in fact, show that the performances substantially decrease as the number of gold premises increases, with comparable MAP scores across different systems when considering a number of premises varying from 5 to 20. Surprisingly, when considering statements with more than 20 premises, we observe an almost entirely inverse ranking in the leaderbord, with Iangml becoming the best performing system, outperforming more complex models based on Transformers. Moreover, we observe that with 20+ premises the top 3 participating systems achieve worse performance than the TF-IDF baseline. These results indicate that pre-trained Transformers are still not robust on multi-hop inference in this context, and might suffer from a phenomenon of semantic drift similar to what previously observed in scientific explanation regeneration tasks (Jansen and Ustalov, 2019; Valentino et al., 2022, 2021).

6 Related Work

Mathematical Language Processing. Several areas of research apply Natural Language Processing for domain-specific tasks, Mathematics being one of these areas. One crucial task in this field is solving mathematical word problems, where the goal is to provide the answer to a mathematical problem written in natural language (Zhang et al., 2020; Kushman et al., 2014; Ran et al., 2019). These problems are usually self-contained and are structured in a didactic and straightforward manner, not containing complex mathematical expressions.

Some contributions focus on the representation of mathematical text and mathematical elements. Zinn (2004) proposes a representation for mathematical proofs using Discourse Representation Theory. Similarly, Ganesalingam (2013) introduces a grammar for representing informal mathematical text, while Pease et al. (2017) presents this style of text using Argumentation Theory. Such explicit representations are relevant for representing the reasoning process behind mathematical thinking. However, it is still not possible to accurately extract these representations at scale. Representations of mathematical elements are often used in the context of Mathematical Information Retrieval, used, for example, for obtaining a particular equation or expression, given a specific query. Tangent-CFT (Mansouri et al., 2019) is an embedding model that uses the subparts an expression or equation, to represent its meaning. This type of representation (Fraser et al., 2018; Zanibbi et al., 2016) often removes the expression for its original discourse, losing the textual context that can help to find a semantic representation. In this work, we focus on creating a representation that can integrate both of these aspects, natural language and mathematical elements. Similar to our work, Yuan et al. (2020) uses self-attention for mathematical elements in order to generate headlines for mathematical questions. Other relevant tasks for NLP applied to Mathematics include typing variables according to its surrounding text (Stathopoulos et al., 2018), obtaining the units of mathematical elements (Schubotz et al., 2016) and generating equations on a given topic (Yasunaga and Lafferty, 2019).

Premise Selection. Premise selection is a well-defined task in the field of Automated Theorem Proving (ATP), where proofs are encoded using a formal logical representation. Given a set of premises P, and a new conjecture c, premise selection aims to predict those premises from P that will most likely lead to an automatically constructed proof of c, where P and c are both written using a formal language. (Alemi et al., 2016) is one of the first models to use Deep Learning for premise selection in ATPs. Ferreira and Freitas (2020a) proposed an adaptation of this task, focusing on mathematical text written in natural language. A model based on Graph Neural Networks has been previously introduced for this task (Ferreira and Freitas, 2020b), however, the authors do not take into account the differences between mathematical and natural language terms, representing all statements homogeneously. The premise selection task can also be seen as an explanation reconstruction task, where premises are considered explanations for mathematical proofs.

Multi-Hop Natural Language Inference. The proposed NLPS task is related to previous work on Multi-Hop Inference and Explanation Regeneration as the set of premises retrieved by a given model can be interpreted as an explanation supporting the mathematical statement provided as input (Thayaparan et al., 2020; Xie et al., 2020; Valentino et al., 2022). Previous editions of the shared tasks series have focused on evaluating multi-hop inference in the context of science question answering (Thayaparan et al., 2021; Jansen and Ustalov, 2020, 2019). In this work, instead, we aim to assess the multi-hop inference capabilities of NLP models in a context requiring the articulation of both natural language and mathematical expressions.

7 Conclusion

Our shared task on Natural Language Premise Selection (NLPS) attracted a total of four participating teams, allowing for the evaluation of a diverse set of solutions ranging from Transformers to lexical-based approaches. The participating systems have all contributed to improving the performance of a TF-IDF baseline. The best-performing team, IJS, presented an approach based on pre-trained Sentence-Transformers, which has been shown to achieve a MAP score of 15.39. Given the challenges involved in the task, supported by the relatively low performance of state-of-the-art approaches, we hope this work will encourage future research in the field, exploring NLPS as a benchmark for testing complex inference capabilities and exploring the limit of AI and NLP models.

Acknowledgements

This work is partially funded by the SNSF project NeuMath (200021_204617).

References

Alexander A. Alemi, François Chollet, Niklas Een, Geoffrey Irving, Christian Szegedy, and Josef Urban. 2016. DeepMath - Deep Sequence Models for Premise Selection. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS 2016, pages 2243–2251, Barcelona, Spain. Curran Associates Inc.

Tom Brown et al. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33*, NeurIPS 2020, pages 1877–1901, Montréal, QC, Canada. Curran Associates, Inc.

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.

Doratossadat Dastgheib and Ehsaneddin Asgari. 2022. Keyword-based Natural Language Premise Selection for an Automatic Mathematical Statement Proving. In *Proceedings of TextGraphs-16: Graph-based Methods for Natural Language Processing*, Gyeongju, Republic of Korea. Association for Computational Linguistics.

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)*, NAACL-HLT 2019, pages 4171–4186, Minneapolis, MN, USA. 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*, LREC 2020, 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*, ACL 2020, pages 7365–7374, Online. Association for Computational Linguistics.

Deborah Ferreira, Mokanarangan Thayaparan, Marco Valentino, Julia Rozanova, and Andre 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, Dublin, Ireland. Association for Computational Linguistics.

Dallas Fraser, Andrew Kane, and Frank Wm. Tompa. 2018. Choosing Math Features for BM25 Ranking with Tangent-L. In *Proceedings of the ACM Symposium on Document Engineering 2018*, DocEng '18, pages 17.1–10.10, Halifax, NS, Canada. Association for Computing Machinery.

Mohan Ganesalingam. 2013. *The Language of Mathematics*, pages 17–38. Springer Berlin Heidelberg, Berlin, Heidelberg.

André Greiner-Petter, Moritz Schubotz, Fabian Müller, Corinna Breitinger, Howard Cohl, Akiko Aizawa, and Bela Gipp. 2019. Why Machines Cannot Learn Mathematics, Yet. In *Proceedings of the 4th Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2019)*, number 2414 in CEUR Workshop Proceedings, pages 130–137, Paris, France.

Peter Jansen and Dmitry Ustalov. 2019. TextGraphs 2019 Shared Task on Multi-Hop Inference for Explanation Regeneration. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, pages 63–77, Hong Kong. Association for Computational Linguistics.

Peter Jansen and Dmitry Ustalov. 2020. TextGraphs 2020 Shared Task on Multi-Hop Inference for Explanation Regeneration. In *Proceedings of the Graph-based Methods for Natural Language Processing (TextGraphs)*, pages 85–97, Barcelona, Spain (Online). Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, EACL 2017, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Liubov Kovriguina, Roman Teucher, and Robert Wardenga. 2022. TextGraphs-16 Natural Language Premise Selection Task: Zero-Shot Premise Selection with Prompting Generative Language Models. In *Proceedings of TextGraphs-16: Graph-based Methods for Natural Language Processing*, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Nate Kushman, Yoav Artzi, Luke Zettlemoyer, and Regina Barzilay. 2014. Learning to Automatically Solve Algebra Word Problems. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2014, pages 271–281, Baltimore, MD, USA. Association for Computational Linguistics.

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.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Behrooz Mansouri, Shaurya Rohatgi, Douglas W. Oard, Jian Wu, C. Lee Giles, and Richard Zanibbi. 2019. Tangent-CFT: An Embedding Model for Mathematical Formulas. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '19, pages 11–18, Santa Clara, CA, USA. Association for Computing Machinery.

Adrien Pavao, Isabelle Guyon, Anne-Catherine Letournel, Xavier Baró, Hugo Escalante, Sergio Escalera, Tyler Thomas, and Zhen Xu. 2022. CodaLab Competitions: An open source platform to organize scientific challenges. Technical report, LISN, CNRS, Université Paris-Saclay.

Alison Pease, John Lawrence, Katarzyna Budzynska, Joseph Corneli, and Chris Reed. 2017. Lakatos-style collaborative mathematics through dialectical, structured and abstract argumentation. *Artificial Intelligence*, 246:181–219.

Shuai Peng, Ke Yuan, Liangcai Gao, and Zhi Tang. 2021. MathBERT: A Pre-Trained Model for Mathematical Formula Understanding.

Qiu Ran, Yankai Lin, Peng Li, Jie Zhou, and Zhiyuan Liu. 2019. NumNet: Machine Reading Comprehension with Numerical Reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, EMNLP-IJCNLP 2019, pages 2474–2484, Hong Kong, China. Association for Computational Linguistics.

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 and the 9th International Joint Conference on Natural Language Processing*, EMNLP-IJCNLP 2019, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Moritz Schubotz, David Veenhuis, and Howard S. Cohl. 2016. Getting the Units Right. In *Joint Proceedings of the FM4M, MathUI, and ThEdu Workshops, Doctoral Program, and Work in Progress at the Conference on Intelligent Computer Mathematics 2016 (CICM-WS-WIP 2016)*, number 1785 in CEUR Workshop Proceedings, pages 146–156, Bialystok, Poland.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and Permuted Pre-training for Language Understanding. In *Advances in Neural Information Processing Systems 33*, NeurIPS 2020, pages 16857–16867, Montréal, QC, Canada. Curran Associates, Inc.

Yiannos Stathopoulos, Simon Baker, Marek Rei, and Simone Teufel. 2018. Variable Typing: Assigning Meaning to Variables in Mathematical Text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, NAACL-HLT 2018, pages 303–312, New Orleans, LA, USA. Association for Computational Linguistics.

Mokanarangan Thayaparan, Marco Valentino, and André Freitas. 2020. A Survey on Explainability in Machine Reading Comprehension.

Mokanarangan Thayaparan, Marco Valentino, Peter Jansen, and Dmitry Ustalov. 2021. TextGraphs 2021 Shared Task on Multi-Hop Inference for Explanation Regeneration. In *Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15)*, pages 156–165, Mexico City, Mexico. Association for Computational Linguistics.

Thi Hong Hanh Tran, Matej Martinc, Antoine Doucet, and Senja Pollak. 2022. IJS at TextGraphs-16 Natural Language Premise Selection Task: Will Contextual Information Improve Natural Language Premise Selection? In *Proceedings of TextGraphs-16: Graph-based Methods for Natural Language Processing*, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Paul Trust, Provia Kadusabe, Haseeb Younis, Rosane Minghim, Evangelos Milios, and Ahmed Zahran. 2022. SNLP at TextGraphs 2022 Shared Task: Unsupervised Natural Language Premise Selection in Mathematical Texts Using Sentence-MPNet. In *Proceedings of TextGraphs-16: Graph-based Methods for Natural Language Processing*, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Marco Valentino, Mokanarangan Thayaparan, Deborah Ferreira, and André Freitas. 2022. Hybrid Autoregressive Inference for Scalable Multi-Hop Explanation Regeneration. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11403–11411.

Marco Valentino, Mokanarangan Thayaparan, and André Freitas. 2021. Unification-based Reconstruction of Multi-hop Explanations for Science Questions. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, EACL 2021, pages 200–211, Online. 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. In *Advances in Neural Information Processing Systems 30*, NIPS 2017, pages 6000–6010, Vancouver, BC, Canada. Curran Associates, Inc.

Sean Welleck, Jiacheng Liu, Ronan Le Bras, Hannaneh Hajishirzi, Yejin Choi, and Kyunghyun Cho. 2021. NaturalProofs: Mathematical Theorem Proving in Natural Language. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

Zhengnan Xie, Sebastian Thiem, Jaycie Martin, Elizabeth Wainwright, Steven Marmorstein, and Peter Jansen. 2020. WorldTree V2: A Corpus of Science-Domain Structured Explanations and Inference Patterns supporting Multi-Hop Inference. In *Proceedings of the 12th Conference on Language Resources and Evaluation*, LREC 2020, pages 5456–5473, Marseille, France. European Language Resources Association (ELRA).

Michihiro Yasunaga and John D. Lafferty. 2019. TopicEq: A Joint Topic and Mathematical Equation Model for Scientific Texts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7394–7401.

Ke Yuan, Dafang He, Zhuoren Jiang, Liangcai Gao, Zhi Tang, and C. Lee Giles. 2020. Automatic Generation of Headlines for Online Math Questions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9490–9497.

Richard Zanibbi, Kenny Davila, Andrew Kane, and Frank Wm. Tompa. 2016. Multi-Stage Math Formula Search: Using Appearance-Based Similarity Metrics at Scale. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 145–154, Pisa, Italy. Association for Computing Machinery.

Jipeng Zhang, Lei Wang, Roy Ka-Wei Lee, Yi Bin, Yan Wang, Jie Shao, and Ee-Peng Lim. 2020. Graph-to-Tree Learning for Solving Math Word Problems. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL 2020, pages 3928–3937, Online. Association for Computational Linguistics.

Claus Zinn. 2004. *Understanding Informal Mathematical Discourse*. Ph.D. thesis, Universität Erlangen-Nürnberg, Institut für Informatik.