STAR: Cross-modal STAtement Representation for selecting relevant mathematical premises

Deborah Ferreira[1], **André Freitas**[1,2]
[1]The University of Manchester, Department of Computer Science, Manchester, UK
[2]Idiap Research Institute, Switzerland
{deborah.ferreira, andre.freitas}@manchester.ac.uk

Abstract

Mathematical statements written in natural language are usually composed of two different modalities: mathematical elements and natural language. These two modalities have several distinct linguistic and semantic properties. State-of-the-art representation techniques have demonstrated an inability in capturing such an entangled style of discourse. In this work, we propose STAR, a model that uses cross-modal attention to learn how to represent mathematical text for the task of Natural Language Premise Selection. This task uses conjectures written in both natural and mathematical language to recommend premises that most likely will be relevant to prove a particular statement. We found that STAR not only outperforms baselines that do not distinguish between natural language and mathematical elements, but it also achieves better performance than state-of-the-art models.

1 Introduction

Natural language understanding has been applied to several different tasks and areas, from question answering to visual grounding. Even though Mathematics is a well-established field with immense importance for most areas of science, applications of NLP in this field are still limited.

Natural language premise selection (NLPS) (Ferreira and Freitas, 2020a) is a task that requires the combination of natural language reasoning and mathematical reasoning. Given a certain conjecture (a mathematical statement written in natural language) that needs to be proven, we attempt to recommend useful premises that can be relevant for developing that mathematical argument.

Mathematical statements have a particular discourse structure that makes it challenging to use traditional NLP techniques. Some of its distinctive features are: (1) Entangled dual lexical spaces for the mathematical elements (ME) and natural language (NL); (2) Distinct syntactic phenomena between ME and NL.

Given this entangled nature of the discourse, where two very different linguistic modalities coexist in the same text, traditional information retrieval approaches are not able to capture the different semantics for each modality (Greiner-Petter et al., 2019). For example, in the mathematical domain, variables are represented using generic symbols; this lexical layer does not necessarily ground the semantics of the variables. The context surrounding the variables is more important than the symbol itself. When interpreting mathematical discourse, such particulars need to be taken into account.

In this work, we propose STAR, a cross-modal representation for mathematical statements for addressing the task of premise selection. In order to interpret the different modalities in the mathematical discourse (natural language and equational), STAR uses two different self-attention layers, one focused on the mathematical elements, such as expressions and variables, while the other attends to natural language features. STAR is taught to see these tokens as parts of different languages, the mathematical language and the English Language, similar to what our human brain does (Butterworth, 2002). Even though the brain interprets mathematics as a language, it requires different parts for processing it (Amalric and Dehaene, 2016). Using different attention layers, STAR can learn that understanding mathematics requires a different type of reasoning than natural language, approximating the behaviour of the brain when faced with mathematical tokens.

The approach presented in this work is based on the hypothesis that the use of cross-modal attention-based mechanisms provides a better encoding of the semantic content of mathematical statements

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 3234–3243
April 19 - 23, 2021.

for the task of premise selection.

The contributions of this work can be summarised as follows:

- Proposal of a novel cross-modal embedding that captures the different modalities inside mathematical text: mathematical elements (expressions) and words.
- A systematic analysis of the transferability of this representation across different mathematical domains.
- An empirical evaluation, comparing our approach with state-of-the-art models and the performance of supporting ablation studies.
- We demonstrate an improvement of up to 70.34% in F1-Score, compared to a baseline that does not distinguish between mathematical elements and natural language. We also obtain competitive results with state-of-the-art approaches, using a smaller model and no pre-training.

2 Background: Natural Language Premise Selection

In this work, we address the problem of *Natural Language Premise Selection* (Ferreira and Freitas, 2020a) (premise selection or NLPS). 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 requiring proof.

On the other hand, accepted truths include statements that need 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. This network structure of known premises can be used as a foundation in order to predict new ones.

Given a new conjecture c, that requires a mathematical proof, and a collection of premises $P = \{p_1, p_2, \ldots, p_{N_p}\}$, with size N_p, the NLPS task aims to retrieve the premises that are most likely to be useful for proving c. Premises of accepted truth statements can also have a subset of premises $\tilde{P} \subseteq P$.

Figure 1 presents an example of a conjecture containing two premises. Both *Premise 1* and *Premise 2* can be used as part of the proof for this conjecture.

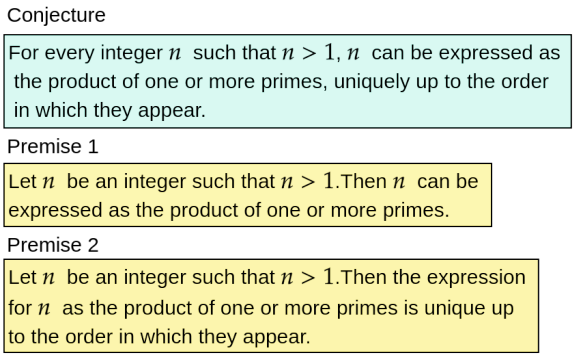


Figure 1: Example of a conjecture and its premises.

Similar to previous approaches (Irving et al., 2016; Ferreira and Freitas, 2020a), we formulate this problem as a pairwise relevance classification problem. Given a pair (c, p_i), we classify if p_i can be used for proving c. Our approach is built on top of a cross-modal representation for mathematical statements, as the following section presents.

3 Our Approach: Cross-modal STAtement Representation (STAR)

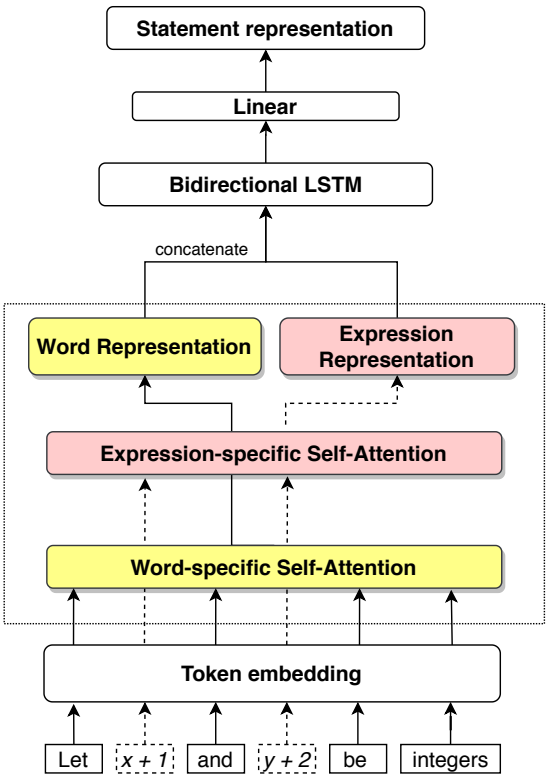
Mathematical language follows a regular pattern (in contrast to natural language) (Ganesalingam, 2013), regardless of representing a conjecture, universal truth or accepted truth. In this work, we consider mathematics written in natural language, instead of mathematics expressed in logical formal languages. The target corpus is composed of a combination of mathematical symbols and natural language words.

Given the set of mathematical statements $\mathcal{M}$ and a statement $m \in \mathcal{M}$, m is defined as a sequence of elements $m = \{s_1, s_2, \ldots, s_n\}$, where $s_i \in \mathcal{W}$, the set of words, or $s_i \in \mathcal{E}$, the set of mathematical elements present in $\mathcal{M}$. These components are situated in different lexical spaces; therefore, a function to generate a representation for m should take this into account.

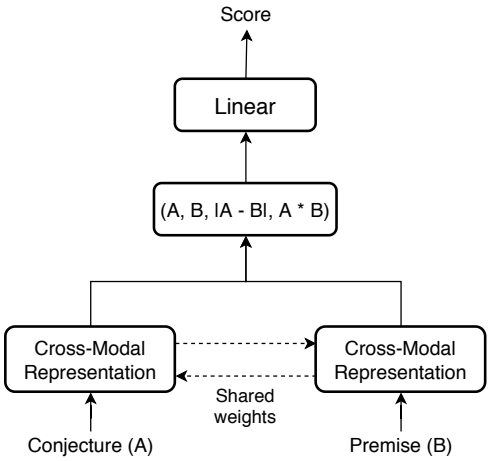
We define an embedding model $\gamma : \mathcal{M} \mapsto \mathbb{R}^d$, where d is the dimension of the output vector. The complete architecture is presented in Figure 2a, where part of a statement is shown as an input example. Each layer is described in detail below.

3.1 Token embedding layer

The input to the embedding model is a mathematical statement. This embedding layer is a



(a) Embedding model.



(b) Siamese Network

Figure 2: Architecture for STAR. Figure (A) presents the model used to generate a representation for each statement, where we combine two self-attention layers, one for each modality of token. Figure (B) presents the Siamese Network used for classifying the conjecture-premise pair, based on the representations obtained.

$W_E \in \mathbb{R}^{k \times v}$ where k is the dimension of the word embeddings, and v is given by $|\mathcal{W}| + |\mathcal{E}|$.

3.2 Word/Expression-specific Self-Attention Layer

Research on the human brain has shown that there is no overlap between the parts of the brain that are activated in math-related tasks (both simple and complex) and sentence comprehension and general semantic knowledge tasks (Amalric and Dehaene, 2016). This behaviour hints at how we should map distinct representations to these different modalities of symbols and linguistic structures (maths and natural language).

Inspired by the behaviour of the human brain, we introduce two layers of self-attention (Vaswani et al., 2017), one for each modality, attempting to approximate human reasoning. One layer captures specific natural language linguistic features, while the other represents particular mathematical formalism features. Given a matrix of queries Q and matrices of keys and values K and V. The attention head is defined as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

where d_k is the dimension of the keys.

These attention heads compose a multi-head attention mechanism, defined as:

$$MultiHead(Q, K, V) = Concat(head_1, \ldots, head_h)W^O \quad (2)$$

where:

$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$

and W_i^Q W_i^K and W_i^V are parameter matrices. In order to apply self-attention, we consider Q, K and V as the same values, obtained using a linear layer on top of the output of the embedding layer. Words and expressions tokens have a very distinct nature, and we hypothesise that these two layers allow learning and representing these differences.

3.3 Long Short-Term Memory Layer

LSTM networks (Hochreiter and Schmidhuber, 1997) are a complex activation unit, based on a

chain structure explicitly designed to capture long-term sequence dependencies. LSTM is an ideal candidate for treating sequential data such as mathematical statements. For the sake of brevity, we omit the description of this layer, as it is extensively described in the literature.

3.4 Training objective

Finally, in order to obtain the score between conjectures and premises, a siamese neural network setting is used (Figure 2b), where a pair of statements are simultaneously fed into two networks, with shared weights. This allows the model to learn the representation of each statement individually, while still being aware that the statements belong to the same semantic space.

The representation for each statement is obtained and combined, where the expected score is 1 if B is a premise to A, or 0 otherwise.

The used training objective function is the Cross Entropy Loss, defined as:

$$\mathcal{L} = -\frac{1}{N}\sum_{n=1}^{N}\left[Y_n \log \hat{Y}_n + (1-Y_n)\log(1-\hat{Y}_n)\right] \quad (3)$$

where Y is the predicted classification and $\hat{Y}_i$ is the expected classification.

4 Experiments

This section presents the experiments performed to test our hypotheses. We use the dataset PS-ProofWiki (Ferreira and Freitas, 2020a) for these experiments. This dataset is composed of pairs of conjectures and premises, framing the problem as a pair classification task. Each statement is written using a combination of words and LaTeXnotation. For each positive pair, where the statement is a premise to the conjecture, there can be n number of negative pairs. For testing the robustness to noise in the proposed model, we use $n \in \{1, 2, 5, 10\}$. The number of entries for Train, Validation and Test for each value of n is shown in Table 1.

The negative pairs are obtained using two different methods. The first collects random examples of statements that are not premises to form a new pair (**negative examples**). In the second technique, we use BM-25 to retrieve statements that are lexically similar to the premises, but that are not part of positive pairs (**similar examples**). For these experiments, we used 512 as a the size of the hidden units layer in the LSTM, embedding size and output statement vector in the embedding architecture. We used 50 epochs for each training round. As shown in Figure 3, with this number of epochs we achieve convergence for all values of n. For each epoch, the validation set was evaluated, and the best model was chosen for testing.

n	Train	Val	Test
1	32,758	10,798	10,112
2	49,137	16,197	15,168
5	98,274	32,394	30,336
10	180,169	59,389	55,616

Table 1: Number of entries for Training, Validation and Test for different values of n.

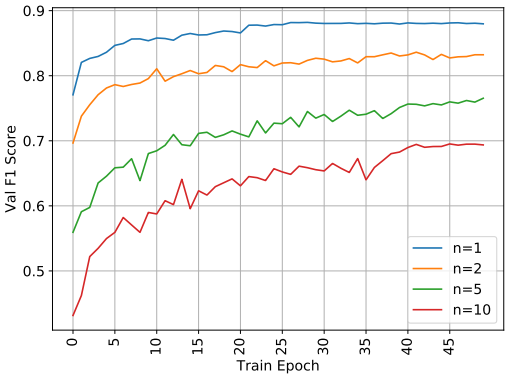


Figure 3: Number of training epochs and the obtained validation F1 score.

All experiments and data can be found in our Github repository[1].

4.1 Quantitative Analysis

In order to verify our hypothesis, we compare the proposed approach, i.e., using different self-attention layers for each modality (mathematical elements and natural language) with a modified model, using only one self-attention layer for all parts of the text. This modified model is obtained by replacing the layers inside the dotted rectangle from Figure 2a with a single self-attention layer. This modified model is referred here as **Self-attention + BiLSTM**.

[1] http://github.com/ai-systems/crossmodal_embedding

4.1.1 Results

Table 2 presents the results for the premise selection task using the random examples.

The aggregate scores obtained using STAR is consistently higher than the baseline. Even though there is an expected degradation in the score with the addition of more negative examples, STAR still outperforms the baseline in all cases, demonstrating robustness to noise. These results support our hypothesis that different modalities inside the mathematical text should be represented in different linguistic spaces.

Similarly, we re-run both models, but this time using the similar examples. The results can be found in Table 3.

We can notice that STAR precision decreases when compared with the results obtained using the random examples. However, once more, STAR outperforms the baseline for all values of n. The results of the baseline model do not change significantly from the previous result improving it in some cases. We hypothesise that this is due to the fact that the use of lexical similarity for the generation of similar examples does not provide reliable discriminators (due to the limited intrinsic semantics of variables). Variables can have the same lexical form across mathematical statements, without sharing the same meaning.

4.1.2 Transferring Knowledge across mathematical domains

Another targeted hypothesis is that STAR performs better than the baseline in the task of transferring knowledge between different mathematical domains. In order to verify this hypothesis, we train the baseline and our model using one topic and test it in a different one, the topics used are Abstract Algebra (AA), Topology (TP) and Set Theory (ST). Table 4 presents the number of statements for Train/Val/Test for each topic.

Table 5 shows the experimental results for the different mathematical topics. Initially, we expected that training using the largest dataset would allow both models to obtain the best performance. However, training using the Topology dataset topic did not achieve the highest results. This is likely because of the distinctive nature of its symbolic space, more focused on the properties of geometric objects. On the other hand, the best performing training and test dataset, Abstract Algebra, is heavily based on the algebraic notation that our model is capable of capture using cross-modal attention.

In terms of transferable knowledge, Set Theory is the tested dataset with the highest score, confirming the expectation that Set Theory is an important component of both Abstract Algebra and Topology, being an intrinsic part of the mathematical argumentation on these topics. Therefore, such knowledge is more natural to transport. Our proposed approach outperforms the baseline in all cases. However, both models see substantial performance degradation when trying to transfer the knowledge from one topic to another, indicating both the need for better abstractive mathematical models and an intrinsic domain-specificity mathematical inference.

4.1.3 Other baselines

In order to verify the model performance, we test our model against two state-of-the-art models. The first baseline is a Transformer-based model, BERT. We fine-tune BERT (Vaswani et al., 2017) using the same configuration as the one used for Natural Language Inference (Jiang and de Marneffe, 2019) since this task carries similarities with the premise selection task. The other baseline is MathSum (Yuan et al., 2019): an encoder-decoder model used to represent mathematical content found in online forums. We use only the encoder part of this model, together with the same siamese network as STAR and the same parameter configuration. The results can be found in Table 6.

Considering the F1-Score obtained, BERT was placed second in the test set evaluation. Even though BERT is not explicitly trained for the Mathematical domain, it presents an excellent performance for the premise selection task. BERT is a large-scale model that was also trained on sources containing mathematical notation, including latex notation, therefore it partially encodes mathematical notation. Our model outperforms BERT for the test set, even though it employs a significantly smaller set of parameters (5x less parameters) and is not pre-trained on a large corpus as BERT is.

4.2 Qualitative analysis

We present examples of predicted pairs in Table 7. When analysing the obtained classified pairs, we found that STAR not only can deal with heavily equational statements, such as the second pair from the table, but it can also handle statements that contain a high level of entanglement between mathematical and natural language terms, such as the first pair.

		Val			Test		
	n	**F1**	**P**	**R**	**F1**	**P**	**R**
STAR	1	**.885**	.854	.917	**.882**	.865	.899
	2	**.836**	.803	.871	**.829**	.793	.870
	5	**.765**	.693	.853	**.765**	.706	.835
	10	**.695**	.614	.799	**.684**	.603	.791
Self-att + LSTM	1	.651	.550	.796	.631	.573	.703
	2	.514	.406	.702	.514	.420	.663
	5	.493	.372	.728	.461	.344	.700
	10	.408	.283	.731	.406	.276	.766

Table 2: Comparison of STAR with a model containing a single self-attention layer. In this experiment, we test for a different number of random negative examples (n). The metrics used are F1-score (F1), precision (P) and recall (R).

		Val			Test		
	n	**F1**	**P**	**R**	**F1**	**P**	**R**
STAR	1	**.798**	.725	.886	**.793**	.723	.879
	2	**.716**	.624	.840	**.707**	.593	.875
	5	**.620**	.485	.857	**.626**	.493	.854
	10	**.546**	.412	.809	**.528**	.387	.834
Self-att + BiLSTM	1	.648	.561	.767	.538	.699	.437
	2	.537	.444	.679	.540	.448	.678
	5	.389	.261	.760	.379	.251	.773
	10	.289	.179	.759	.286	.174	.799

Table 3: Results for STAR and baseline for different number of negative examples (n) using similar examples.

Topic	**Train**	**Val**	**Test**
AA	2,246	633	580
ST	1,897	618	590
TP	2,539	810	788

Table 4: Distribution of dataset wih different topics.

However, we found that STAR can sometimes struggle with variable names. For example, in pair 3, the variable T appears several times. STAR infers that this implies that there is a relation between both statements. The relationship exists since both statements refer to the concept *spaces*; however, this does not define a dependency relationship. This result provides evidence for the need of an architecture which better captures variable semantics.

Figure 4 presents a comparison of our model with the single attention model. This graph shows the percentage of mathematical elements in the statement versus the percentage of the statements in the dataset that the model was able to predict correctly.

STAR has an consistent performance throughout different distributions of mathematical and natural language terms. Such results demonstrate a need of an attention layer for each term modality. On the other hand, we can observe that the baseline struggles to predict statements that are mostly mathematical (right-end of the graph), finding it easier to predict the statements which have the prevalence of natural language terms (left-end of the graph). The results show that our model is better suitable for dealing with this type of entangled text.

5 Related Work

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 mathe-

Topic		STAR			Self-att + BiLSTM		
Train	**Test**	**F1**	**P**	**R**	**F1**	**P**	**R**
AA	AA	**.862**	.823	.906	.629	.581	.684
TP	TP	**.752**	.692	.825	.722	.680	.769
ST	ST	**.787**	.763	.813	.613	.654	.578
AA	ST	**.662**	.595	.747	.627	.595	.664
AA	TP	**.595**	.520	.693	.570	.539	.605
TP	AA	**.654**	.536	.836	.649	.602	.704
TP	ST	**.673**	.588	.787	.628	.561	.714
ST	TP	**.535**	.539	.531	.578	.535	.627
ST	AA	**.644**	.598	.697	.625	.591	.663

Table 5: Testing how different mathematical areas are transportable to other areas. The areas considered here are Abstract Algebra(AA), Topology (TP) and Set Theory (ST). For these experiments, we use random examples with $n = 1$.

	Val			Test		
	F1	**P**	**R**	**F1**	**P**	**R**
BERT	**.886**	.871	.901	.877	.925	.834
MathSum	.644	.512	.869	.459	.562	.388
Self-attention + BiLSTM	.651	.550	.796	.631	.573	.703
STAR	.885	.854	.917	**.882**	.865	.899

Table 6: Comparison of our model with other baselines, using n=1 and random examples.

matical 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. (2019) 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 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

Conjecture	Premise	Predicted	Label
Let $T = (S, \tau)$ be a topological space. Let A, B be subsets of S. Then: $\partial(A \cap B) \subseteq \partial A \cup \partial B$ where ∂A denotes the boundary of A.	Let S, T_1, T_2 be sets such that T_1, T_2 are both subsets of S. Then, using the notation of the relative complement: $ST_1 \cap T_2 = ST_1 \cup ST_2$	1	1 ✓
$\int \frac{\mathrm{x}}{x(x^2-a^2)} = \frac{1}{2a^2}, \ln \frac{x^2-a^2}{x^2} + C$ for $x^2 > a^2$.	$\int \frac{dx}{x} = \ln x + C$ for $x \neq 0$.	1	1 ✓
Let $T = S, \tau$ be a compact space. Then T is countably compact.	Let $T = (S, \tau_{a,b})$ be a modified Fort space. Then T is not a T_3 space, T_4 space or T_5 space.	1	0 ✗

Table 7: Some of the premises existing in the dataset, together with the predictions from STAR.

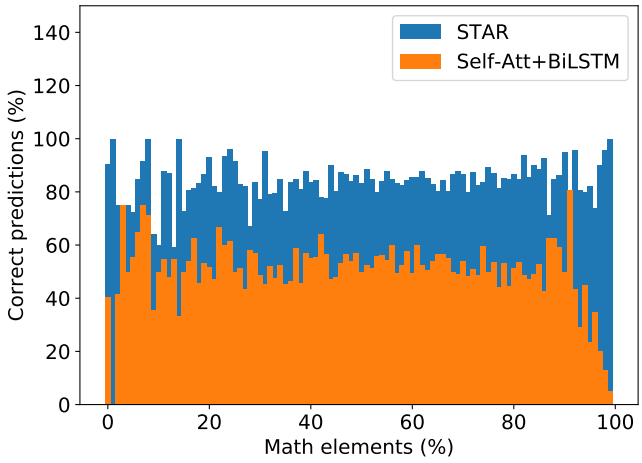


Figure 4: Comparison of our model and the baseline on the capability of predicting correctly statements with different levels of entanglement.

automatically constructed proof of c, where P and c are both written using a formal language. (Irving 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. Approaches for dealing with such type of challenge in the science domain include unification retrieval (Valentino et al., 2020b,a) and abductive reasoning (Thayaparan et al., 2020).

In this work, we propose a new representation that distinctively captures both language modalities present in the mathematical discourse in order to solve the premise selection task.

6 Conclusion

In this work, we introduced STAR, a model to represent mathematical statements for the task Natural Language Premise Selection. In this model, we used two layers of self-attention, one for each language modality present in the mathematical text.

In order to test STAR's ability to capture the different aspects of each modality, verifying if it can interpret that expressions and words belong to different lexical spaces, we compared our performance with other baselines. We found that having one layer for each modality significantly increases the performance for premise selection. We also compared our approach with state-of-the-art models and found that STAR achieves the highest results for the Test set. STAR was also tested for transfer learning, revealing that cross-modal attention improves the transportability between different mathematical areas.

However, we discovered that STAR is still limited regarding variable modelling. There is still a gap in how to handle variable typing in latent models, considering its meaning instead of its lexi-

cal symbol. As future work, this issue will be addressed using latent representations trained specifically for variable modelling.

References

Marie Amalric and Stanislas Dehaene. 2016. Origins of the brain networks for advanced mathematics in expert mathematicians. *Proceedings of the National Academy of Sciences*, 113(18):4909–4917.

Brian Butterworth. 2002. Mathematics and the brain. *Opening address to the Mathematical Association, Reading*.

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.

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)*.

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*, pages 1–10.

Mohan Ganesalingam. 2013. The language of mathematics. In *The Language of Mathematics*, pages 17–38. Springer.

André Greiner-Petter, Terry Ruas, Moritz Schubotz, Akiko Aizawa, William Grosky, and Bela Gipp. 2019. Why machines cannot learn mathematics, yet. *arXiv preprint arXiv:1905.08359*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Geoffrey Irving, Christian Szegedy, Alexander A Alemi, Niklas Eén, François Chollet, and Josef Urban. 2016. Deepmath-deep sequence models for premise selection. In *Advances in Neural Information Processing Systems*, pages 2235–2243.

Nanjiang Jiang and Marie-Catherine de Marneffe. 2019. Evaluating bert for natural language inference: A case study on the commitmentbank. 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)*, pages 6088–6093.

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)*, pages 271–281.

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*, pages 11–18.

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.

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)*, pages 2474–2484.

Moritz Schubotz, David Veenhuis, and Howard S Cohl. 2016. Getting the units right. In *FM4M/MathUI/ThEdu/DP/WIP@ CIKM*, pages 146–156.

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)*, pages 303–312.

Mokanarangan Thayaparan, Marco Valentino, and André Freitas. 2020. Explanationlp: Abductive reasoning for explainable science question answering. *arXiv preprint arXiv:2010.13128*.

Marco Valentino, Mokanarangan Thayaparan, and André Freitas. 2020a. Explainable natural language reasoning via conceptual unification. *arXiv preprint arXiv:2009.14539*.

Marco Valentino, Mokanarangan Thayaparan, and André Freitas. 2020b. Unification-based reconstruction of explanations for science questions. *arXiv preprint arXiv:2004.00061*.

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*, pages 5998–6008.

Michihiro Yasunaga and John D Lafferty. 2019. Topiceq: A joint topic and mathematical equation model for scientific texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7394–7401.

Ke Yuan, Dafang He, Zhuoren Jiang, Liangcai Gao, Zhi Tang, and C Lee Giles. 2019. Automatic generation of headlines for online math questions. *arXiv preprint arXiv:1912.00839*.

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*, pages 145–154.

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*, pages 3928–3937.

Claus Zinn. 2004. Understanding informal mathematical discourse. *PhD thesis, Institut fur Informatik, Universitat Erlangen-Nurnberg*.