Quo Vadis, Math Information Retrieval

Petr Sojka, Vít Novotný, Eniafe Festus Ayetiran,
Dávid Lupták, and Michal Štefánik

Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
sojka@fi.muni.cz, {witiko,ayetiran,dluptak,stefanik.m}@mail.muni.cz
https://mir.fi.muni.cz/

Abstract. With the exponential growth of information in the digital form, information retrieval and querying digital libraries is of paramount importance, and mathematical and technical STEM documents are not an exception. The key for precise searching is the adequate and unambiguous representation of documents, paragraphs, sentences and words, which we are going to evaluate. We are presenting a roadmap to tackle the problem of searching and question answering in the digital mathematical libraries, and discuss the pros and cons of promising approaches primarily for the key part, namely the document representation: several types of embeddings, topic mixtures and LSTM. The listed representation learning options will be evaluated at the next ARQMath evaluation lab of CLEF 2020 conference.

Keywords: math information retrieval, question answering, STEM, digital mathematical libraries, embeddings, MIaS, MIaSNG, DML

> "If not now, when?"
> Chapters of the Fathers (Pirkei Avot, 1:14)

1 Introduction

Content is king. Content expressed in math formalism and formulae are often crucial and non-negligible part of the content of science, technology, engineering, and mathematics (STEM) papers. Mathematical formulae and diagrams, mostly due to their inherent structure and complexity, used to be not taken into account when processing the language of documents. Mathematical discourse, as a niche market, was not supported by tools for indexing and searching, digitization, or question answering.

There is already more than a decade of attempts to tackle the problem of Math Information Retrieval (MIR) in the Digital Mathematical Libraries (DML):

Infty, 2003 first math optical character recognition (OCR) system. [45]
MathDex, 2006 first search engine ever indexing MathML. [26]
DML-CZ, 2005–2009 one of the first attempts to classify and categorize mathematical knowledge [43] by automated means and tools: Gensim library has been designed for the DML usage [34,35].

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2019, pp. 117–128, 2019. © Tribun EU 2019

DML workshop series, 2008–2011 first workshop series specifically targeted for MIR and related problems.

EuDML, 2010–2013 first DML that deployed the MIaS search engine [41,42,52] designed for searching math formulae, and developed tools like WebMIaS [20] specifically for the STEM domain.

NTCIR 10, 2013 first evaluation competition with a MIR task. [1]

Tangent, 2012–2014 first visual-based indexing (symbol layout tree) and search engine, that is able to find any two dimensional mathematical and diagrammatical structures. [28,38]

MCAT, Aizawa MIR group, 2016 first system to learn how to effectively combine text and math for reranking. [14]

Equation embeddings, 2018 first joint embedding model that represents formulae by taking into account surrounding texts. [15]

Tangent CFT, 2019 first formula embedding model that uses two hierarchical representations: Symbol Layout Trees (SLTs) for appearance, and Operator Trees (OPTs) for mathematical content. [21]

TopicEq, 2019 first topic model that jointly generates mathematical equations and their surrounding text (TopicEq). [53]

ARQMath lab, c/o CLEF 2020 first answer retrieval task for questions on math data[1].

The accelerated pace with which the research in MIR continues, and new machine-learning approaches that appeared recently show promise of a new generation of search engines. It will take into account not only similar words or phrases, but the disambiguated meaning of structured objects like equations, sentences, paragraphs, or trains of thoughts.

We are going to develop the Math Indexer and Searcher of New Generation (MIaSNG) that will take into account the latest state of the art in the area of document meaning representation. To this end, we start by studying and evaluating several approaches in document representations.

The structure of the paper is as follows. In Section 2, we present recent approaches based on embeddings, specifically those that are capable of representing structured objects such as equations. In Section 3, a method based on joint text and math topic modeling is discussed. Another possibility of representing sequences of tree structures is described in Section 4. Section 5 evaluates versions of transfer learning for our goals. This list of possibilities is by no means exhaustive, but presents approaches for evaluation during the preparation of MIaSNG.

[1] `https://www.cs.rit.edu/~dprl/ARQMath/`

> "The day is short, the labor vast, the toilers idle, the reward great,
> and the Master of the house is insistent."
> Chapters of the Fathers (Pirkei Avot, 2:20)

2 Joint Embeddings for Text and Math: Equation Embeddings

Since the seminal work of Mikolov et al. [25], unsupervised word embeddings have become the preferred word representations for many natural language processing tasks. Document similarity measures extracted from unsupervised word embeddings, such as the Soft Cosine Measure (SCM) [39], are fast [27] and achieve strong performance on semantic text similarity [5], textual information retrieval [9], and entrance exam question answering [39] tasks.

In mathematical discourse, formulae are often more important than words for understanding. [16] Unlike words, formulae are deeply structured and most are unique. Unlike unsupervised embeddings of words, which are many and well-understood, unsupervised embeddings of mathematical formulae are only now beginning to be explored. The SCM with joint embeddings of words and formulae can form a basis of a fast and accurate mathematical search engine.

2.1 EqEmb and EqEmb-U Models

Krstovski and Blei [15] propose unsupervised joint embeddings of math and formulae: EqEmb and EqEmb-U. Their approach is based on a) symbol layout tree (SLT) visual encoding of mathematical formulae from Zanibbi et al. [54], and b) unsupervised embeddings of Koopman-Darmois family probability distributions [37] that generalize Skipgram with negative sampling of Mikolov et al. [24].

In the EqEmb model, every formula is represented as a single vocabulary entry and its internal structure is disregarded. For every input word i, the EqEmb model predicts the context words and formulae, and for every input formula m, the EqEmb model predicts the context words.

In the EqEmb-U model, formulae are tokenized using the SLT visual encoding. For every input word i, the EqEmb-U model predicts the context words and formula tokens, and for every input formula token m, the EqEmb-U model predicts the context formula tokens. After training, formula embeddings are obtained by averaging the embeddings of the formula tokens.

Unlike the FastText model of Bojanowski et al. [4], neither EqEmb nor EqEmb-U take subword information into account. Unlike EqEmb, EqEmb-U formula embeddings do not take context words into account. Unlike Mansouri et al. [21], both EqEmb and EqEmb-U use only a visual encoding of formulae (SLT) and not the Operator Tree (OPT), which encodes the meaning of formulae.

2.2 Evaluation

Krstovski and Blei compare EqEmb and EqEmb-U to existing word embeddings [18,25,29,37] using log-likelihood over arXiv articles from the NLP, IR, AI, and ML domains. They obtain best results using EqEmb-U closely followed by EqEmb across all domains. Qualitative evaluation shows that kNN on EqEmb formula embeddings can be used to recommend highly related words and formulae.

> "He who acquires a good name, has acquired himself something indeed."
> Chapters of the Fathers (Pirkei Avot, 2:8)

3 Joint Text and Math Topic Modelling

With the increasing number of documents and their archives available in our digital era, it becomes more difficult to find content that interests us. Search engines play a fundamental role in the area of information retrieval and help us find a set of documents based on given keywords. However, sometimes we might be out of keywords and want to explore similar materials related to the theme of other ones. Probabilistic topic modeling is designed for this purpose – a set of statistical methods that analyzes words in the texts and discovers topics based on their content.

3.1 TopicEq Model

In the context of STEM fields, the text itself is not the only part of papers that delivers the message. Mathematics is ubiquitous, and it comprehensively communicates the ideas. However, most works in natural language processing or machine learning study text and math separately. Yasunuga et al. [53] reason that these two components should be studied jointly together, as the text surrounding the mathematical equations can provide context for its better understanding and vice versa. They propose a new model called TopicEq that applies a topic model to the text context and jointly the same latent topic proportion vector to generate a sequence of math symbols in a recurrent neural network (RNN).

In their work, they apply neural variational inference technique [22,23,44] to train topic models. They employ an RNN to model equations as a sequence of LaTeX tokens [13]. They relate to and extend Latent Dirichlet allocation (LDA) [3] as they model two different modalities – word text and math equations. They also demonstrate that RNN-based models [7] are more effective than the bag of token-based models for equation processing. And finally, this work relates to equation embeddings [15] with additional modeling of each equation as a sequence of symbols.

3.2 Implementation and Evaluation

The correlated topic model [2] that uses a log-normal distribution is a baseline for the TopicEq model. The new model introduces the surrounding text of an equation as a context, and the generative process takes the same latent topic proportion vector for both this context and the math expression. An RNN generates equations as a sequence of mathematical symbols from the vocabulary of LaTeX tokens and the extension of the LSTM [11], named Topic-Embedded LSTM (TE-LSTM), embeds the proportional vector inside the LSTM cell to keep the topic knowledge.

They perform experiments on the dataset of context-equation pairs, constructed from sampled 100,000 arXiv articles. They define a context as five consecutive sentences both before and after the equation and kept equations only of a specific length, which yields the final 400,000 context-equation pairs. In the topic model evaluation, the results show that using the joint RNN equation model significantly improves the coherence of topics of scientific texts. In the equation model evaluation, TE-LSTM outperforms generic LSTM in reducing the perplexity and syntax error rate while also requiring less training time.

Qualitative analysis of the newly designed model shows its high capability to interconnecting the mathematics with the topics in several applications. In topic-aware equation generation, the generated equations reflect the characteristics of given topics, even if a mixture of topics is in question. In the equation topic inference, the TopicEq model performs better in precision and consistency than the bag-of-token baseline, and the topic-dependent alignment between mathematical tokens and words can predict results adequately based on the given topic.

TopicEq model can increase the interpretability of equations regarding their context and improve the exploring similar scientific documents. More experiments should be undergone for the very short as well as complex formulae to see the performance of this topic model. The selection of the proper word context length could also be crucial for the results.

> "What is the right path a man should choose?
> Whatever is honorable to himself, and honorable in the eyes of others."
> Chapters of the Fathers (Pirkei Avot, 2:1)

4 Mathematical Expressions Embedding Using Tree-Structured Bidirectional LSTM

The past few years have witnessed an upsurge in the use of deep learning architectures for semantic representation of sequential data due to their ability to capture long-range dependencies. Long short-term memory networks [11] addresses the problem of exploding or vanishing gradients, which had hitherto made traditional recurrent neural networks (RNNs) unable to capture long-range correlations in a sequence such as text. The long-range dependencies are preserved with a memory cell.

The bi-directional LSTM [10] is a variant of the traditional LSTM, which consists of two LSTMs that are run simultaneously, one for the input sequence and the other for the reverse of the input. This is to enable the network model to learn in both directions, taking into account the left and the right contexts i.e. the past and the future information using the hidden state of the LSTM at each time step.

The bi-directional LSTM architecture is suitable for strictly sequential information propagation and cannot handle information with hierarchical structure in mathematical formulae and expressions. The tree-structured LSTMs [17,46,55,56] have been introduced to model the syntactic structure of natural language text but they have not yet been applied to mathematics.

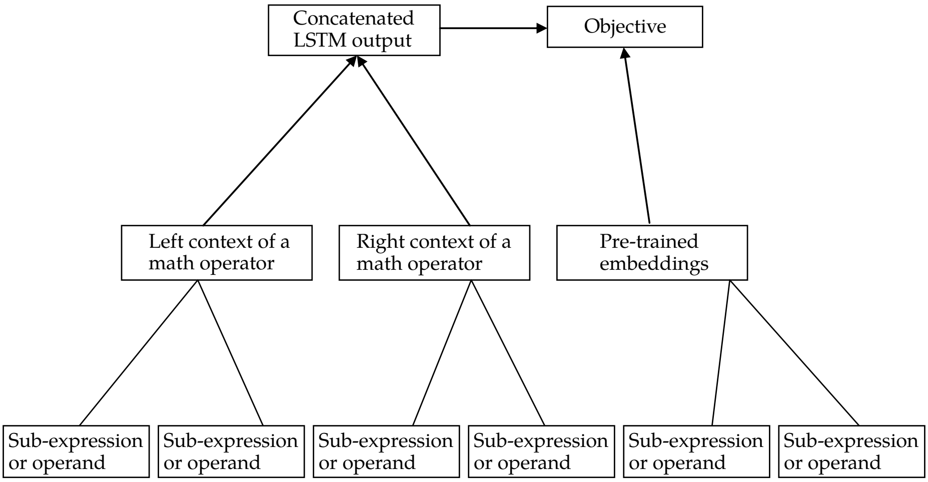
Concatenated LSTM output
Objective
Left context of a math operator
Right context of a math operator
Pre-trained embeddings
Sub-expression or operand
Sub-expression or operand
Sub-expression or operand
Sub-expression or operand
Sub-expression or operand
Sub-expression or operand

Fig. 1. Proposed Tree-Structured Bi-directional LSTM for Math Expressions Embedding

4.1 Proposed Bi-directional Tree-Structured LSTM

The general architecture of our proposed Bi-directional Tree-Structured LSTM is presented in Figure 1.

Recently, Thanda et al. [48], in their math information retrieval system for NTCIR-12 MathIR task used a bag-of-words version of Paragraph Vector (PV-DBOW) [18]. In their work, the math formulae in Wikipedia and arXiv papers are utilized. The math formulae are represented in the form of a tree, where the non-leaf nodes correspond to operators and the leaf nodes correspond to operands.

Our proposed method aims to use a similar approach but using the tree-structured bi-directional LSTM that can capture long-range dependencies. To represent any mathematical expression in a document, each mathematical operator will serve as the target token, and the two LSTMs will learn contexts in both directions. The operands will serve as the leaves, taking into account the structure of the formulae. First, the formulae or expressions will be parsed using ANTLR[2] or BISON[3] to recover their structure. The raw expressions will be converted to XML and MathML using the LaTeXML, the result of which will be canonicalized [40] and used to train a model using an adaptation of the Bi-directional Tree-structured LSTM [47] with a pre-trained embedding, using Word2Vec [24] as the objective.

As applicable to natural language texts, we hypothesize that the LSTM, apart from its capability to capture the long-range coherence in Math expressions will also be able to capture the order of combination of operators and operands alike, taking into account notational variations. However, we envisage an outcome

[2] https://www.antlr.org/

[3] https://www.gnu.org/software/bison/manual/bison.html

which may not be exact as the results reported for natural language texts due to the peculiarity of Math texts.

> “In a place where there are no worthy men, strive to be worthy.”
> Chapters of the Fathers (Pirkei Avot, 2:5)

5 Representation and Transfer Learning for Math

Lately, we have been the observers of the dramatic movement of the reached quality in solving some of the principal high-level NLP Tasks [6,33,50]. Some of the new approaches have, in fact, overreached not only the current state-of-the-art by even tens of percent [8,51], but also the measured human performance [33].

The new technologies are based on a distinct set of ideas which has driven the development of their architecture. ELMo [30] builds upon a character-level convolution and a joint optimization of sequential language models (bi-LM), while attention-based transformer architectures, such as GPT [32] or BERT [8], utilize forward, or bi-directional incremental pooling of so-called attention [49] in forward, or bi-directional manner, respectively.

Yet, there is an attribute that intersects this new stream of methods. It is a fact they are pre-trained without a supervisor on a vast amount of data in the form of general language corpus [8,30,32]. Subsequently, they can be fine-tuned on downstream tasks [30], or their internal representations can be even used in zero-shot manner [32]. Furthermore, the generalization properties of transformers in language modeling [32] has even been underscored by their performance on language-agnostic downstream tasks [31], where some multilingual models are documented to perform well on a zero-shot classification of previously-unseen languages, even on ones that do not share any vocabulary with the fine-tuned language [31].

The surprising generalization capabilities of the attention-based technologies motivate us to evaluate their performance in the context of relevant tasks of math understanding. For our objectives, we propose several adaptations of the Transformer architecture, which reaches state-of-the-art results on other tasks, e.g. a question answering task [33] with an objective of explaining math formula variables, based on the surrounding context. Here, the variable denotation can be interpreted as a question, and the figure context as the answering paragraph. Similarly, we propose the use of sequence classification architecture used for paraphrase detection for the MIR formula-based search [20]: the similarity of their contexts can determine the disambiguation of the parts of the formulae.

We believe that the mentioned math understanding tasks could be fine-tuned from the weights pre-trained on general language modeling tasks [19], just like their native paired tasks. Subsequently, in cases where we can directly interpret our tasks in a framework of the native ones (SQuAD and MRPC, in the mentioned cases), we might be able to utilize the rich data sets of these native tasks to additionally fine-tune the pre-trained models in favour of our objectives.

Perhaps the main drawback of the methods above is their computational complexity. In any area related to the information retrieval, it is crucially important to be able to either compute the representations of the documents on-the-fly, or to pre-index them and compare pairwise to the query in the real time.

It remains an open question whether the transferability of the models can also be successfully utilized in MIR. There are reports that the methods can be used to directly infer the context-dependent embeddings [8] of the input, or the internal representations can be fine-tuned to provide the similar embeddings of related paragraph pairs [36].

We believe that all the aforementioned methods are not necessarily bound to natural language applications: the successful inter-lingual applications [31] suggest that as long as the meaning of mathematical documents can be captured both in the natural language, just as in the math formulae, their general representation can be modeled.

We plan to design the experiments to evaluate whether the bidirectional sequence embeddings [12] or Transformers' internal representation [49] can be adapted to a general math representation. It is possible that the meaning shared among the math formulae and its interpretation in a natural text is still too latent to follow using a single model. Another eventual threat is that joint representations will, in comparison to standard, separate representations, miss some important low-level (e.g. morphological) features that are, however, crucial for relevant math information retrieval.

> "It is not incumbent upon you to complete the work,
> but neither are you at liberty to desist from it."
> Chapters of the Fathers (Pirkei Avot, 2:21)

6 Conclusion

This essay records the possibilities of representing the complex meaning of mathematical structures in STEM documents. Discussed representations reach state-of-the-art performance in text-only versions of NLP tasks, and their adaptation to cope with math seems feasible.

The paper thus serves as an outline of the prototypical implementation and evaluation of MIaSNG, the Math Indexing and Searching system of New Generation, which we are about to develop in the following years. We believe that having a joint representation of the meaning of both text and math will allow a new level of querying mathematical corpora such as arXiv or EuDML. Even though it seems that the Pareto principle holds – 80% of the conveyed message is in the form of text, the discussed approaches consistently show that coupling the text and math meaning increases the quality of language models, representations, and thus the future MIR via MIaSNG.

Acknowledgements This publication was written with the support of the Specific University Research provided by the Ministry of Education, Youth and Sports of the Czech Republic. Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum provided under the programme "Projects of Large Research, Development, and Innovations Infrastructures" (CESNET LM2015042) is greatly appreciated.

References

1. Aizawa, A., Kohlhase, M., Ounis, I.: NTCIR-10 Math Pilot Task Overview. In: Proc. of the 10th NTCIR Conference. pp. 654–661. NII, Tokyo, Japan (2013)
2. Blei, D.M., Lafferty, J.D.: A correlated topic model of science. The Annals of Applied Statistics **1**(1), 17–35 (06 2007). https://doi.org/10.1214/07-AOAS114
3. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet Allocation. The Journal of Machine Learning Research **3**, 993–1022 (2003)
4. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017)
5. Charlet, D., Damnati, G.: SimBow at SemEval-2017 task 3: Soft-cosine semantic similarity between questions for community question answering. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 315–319. Association for Computational Linguistics (2017). https://doi.org/10.18653/v1/S17-2051
6. Chen, S.F., Beeferman, D., Rosenfeld, R.: Evaluation Metrics for Language Models. In: DARPA Broadcast News Transcription and Understanding Workshop. pp. 275–280. CMU (1998)
7. Deng, Y., Kanervisto, A., Ling, J., Rush, A.M.: Image-to-markup Generation with Coarse-to-fine Attention. In: Proceedings of the 34th International Conference on Machine Learning – Volume 70. pp. 980–989. ICML'17, JMLR.org (2017), http://dl.acm.org/citation.cfm?id=3305381.3305483
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805 (2018), https://arxiv.org/abs/1810.04805
9. González Barbosa, J.J., Frausto-Solis, J., Villanueva, D., Valdés, G., Florencia, R., González, L., Mata, M.: Implementation of an Information Retrieval System Using the Soft Cosine Measure, vol. 667, pp. 757–766. Springer (12 2017). https://doi.org/10.1007/978-3-319-47054-2_50
10. Graves, A., Jaitly, N., Mohamed, A.: Hybrid speech recognition with deep bidirectional LSTM. In: Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). pp. 273–278. Olomouc, Czech Republic (2013)
11. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (Nov 1997). https://doi.org/10.1162/neco.1997.9.8.1735
12. Józefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., Wu, Y.: Exploring the Limits of Language Modeling. arXiv preprint arXiv:1602.02410 (2016), https://arxiv.org/abs/1602.02410
13. Karpathy, A.: The Unreasonable Effectiveness of Recurrent Neural Networks, http://karpathy.github.io/2015/05/21/rnn-effectiveness/, Andrej Karpathy blog
14. Kristianto, G.Y., Topić, G., Aizawa, A.: Combining effectively math expressions and textual keywords in math IR. In: Proceedings of the 3rd International Workshop on Digitization and E-Inclusion in Mathematics and Science 2016 (DEIMS2016). pp. 25–32. sAccessNet (Nonprofit organization), Japan (2016), http://workshop.sciaccess.net/DEIMS2016/articles/p02_Kristianto&Aizawa.pdf
15. Krstovski, K., Blei, D.M.: Equation embeddings. arXiv preprint (2018), https://arxiv.org/abs/1803.09123
16. Larson, R.R., Reynolds, C., Gey, F.C.: The abject failure of keyword IR for mathematics search: Berkeley at NTCIR-10 math. In: Kando, N., Kato, T. (eds.) Proceedings of the

10th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-10, National Center of Sciences, Tokyo, Japan, June 18-21, 2013. National Institute of Informatics (NII) (2013)

17. Le, P., Zuidema, W.: Compositional distributional semantics with long short term memory. In: Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics. pp. 10–19. Denver, Colorado, USA (2015)
18. Le, Q., Mikolov, T.: Distributed Representations of Sentences and Documents. In: Proceedings of International Conference on Machine Learning. pp. 1188–1196. Beijing, China (2014)
19. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L.S., Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692 (2019), https://arxiv.org/abs/1907.11692
20. Líška, M., Sojka, P., Růžička, M.: Math indexer and searcher web interface: Towards fulfillment of mathematicians' information needs. In: Watt, S.M., Davenport, J.H., Sexton, A.P., Sojka, P., Urban, J. (eds.) Intelligent Computer Mathematics CICM 2014. Proceedings of Calculemus, DML, MKM, and Systems and Projects. pp. 444–448. Springer International Publishing Switzerland, Zurich (2014). https://doi.org/10.1007/978-3-319-08434-3_36, https://arxiv.org/abs/1404.6476
21. Mansouri, B., Rohatgi, S., Oard, D.W., Wu, J., Giles, C.L., Zanibbi, R.: Tangent-CFT: An Embedding Model for Mathematical Formulas. In: Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval. pp. 11–18. ICTIR '19, ACM, New York, NY, USA (2019). https://doi.org/10.1145/3341981.3344235
22. Miao, Y., Grefenstette, E., Blunsom, P.: Discovering discrete latent topics with neural variational inference. In: Proceedings of the 34th International Conference on Machine Learning – Volume 70. pp. 2410–2419. ICML '17, JMLR.org (2017), http://dl.acm.org/citation.cfm?id=3305890.3305930
23. Miao, Y., Yu, L., Blunsom, P.: Neural variational inference for text processing. In: International conference on machine learning. pp. 1727–1736. ICML '16, JMLR.org (2016), http://dl.acm.org/citation.cfm?id=3045390.3045573
24. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed Representations of Words and Phrases and their Compositionality. In: Burges, C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K. (eds.) Advances in Neural Information Processing Systems 26, pp. 3111–3119. Curran Associates, Inc. (2013)
25. Mikolov, T., Chen, K., et al.: Efficient estimation of word representations in vector space. arXiv preprint (2013), https://arxiv.org/abs/1301.3781v3, accessed 22 October 2019
26. Munavalli, R., Miner, R.: MathFind: a math-aware search engine. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 735–735. SIGIR '06, ACM, New York, NY, USA (2006). https://doi.org/10.1145/1148170.1148348
27. Novotný, V.: Implementation Notes for the Soft Cosine Measure. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management. pp. 1639–1642. CIKM '18, ACM, New York, NY, USA (2018). https://doi.org/10.1145/3269206.3269317
28. Pattaniyil, N., Zanibbi, R.: Combining TF-IDF Text Retrieval with an Inverted Index over Symbol Pairs in Math Expressions: The Tangent Math Search Engine at NTCIR 2014. In: Kando, N., Joho, H., Kishida, K. (eds.) Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies. pp. 135–142. National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430 Japan, Tokyo (2014), http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings11/pdf/NTCIR/Math-2/08-NTCIR11-MATH-PattaniyilN.pdf

29. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
30. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the ACL: Human Language Technologies, Volume 1 (Long Papers). pp. 2227–2237. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/N18-1202
31. Pires, T., Schlinger, E., Garrette, D.: How multilingual is Multilingual BERT? (2019)
32. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI Blog **1**(8) (2019)
33. Rajpurkar, P., Jia, R., Liang, P.: Know what you don't know: Unanswerable questions for SQuAD. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 784–789. Association for Computational Linguistics, Melbourne, Australia (Jul 2018). https://doi.org/10.18653/v1/P18-2124, `https://www.aclweb.org/anthology/P18-2124`
34. Řehůřek, R., Sojka, P.: Automated Classification and Categorization of Mathematical Knowledge. In: Autexier, S., Campbell, J., Rubio, J., Sorge, V., Suzuki, M., Wiedijk, F. (eds.) Intelligent Computer Mathematics—Proceedings of 7th International Conference on Mathematical Knowledge Management MKM 2008. Lecture Notes in Computer Science LNCS/LNAI, vol. 5144, pp. 543–557. Springer-Verlag, Berlin, Heidelberg (Jul 2008)
35. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks. pp. 45–50. ELRA, Valletta, Malta (May 2010). https://doi.org/10.13140/2.1.2393.1847, `http://is.muni.cz/publication/884893/en`, software available at `http://nlp.fi.muni.cz/projekty/gensim`
36. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks (2019)
37. Rudolph, M., Ruiz, F., Mandt, S., Blei, D.: Exponential family embeddings. In: Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems 29, pp. 478–486. Curran Associates, Inc. (2016), `http://papers.nips.cc/paper/6571-exponential-family-embeddings.pdf`
38. Schellenberg, T., Yuan, B., Zanibbi, R.: Layout-based substitution tree indexing and retrieval for mathematical expressions. In: Viard-Gaudin, C., Zanibbi, R. (eds.) Document Recognition and Retrieval XIX. vol. 8297, pp. 126–133. International Society for Optics and Photonics, SPIE (2012). https://doi.org/10.1117/12.912502
39. Sidorov, G., et al.: Soft similarity and soft cosine measure: Similarity of features in vector space model. CyS **18**(3), 491–504 (2014). https://doi.org/10.13053/cys-18-3-2043
40. Sojka, P.: Exploiting semantic annotations in math information retrieval. In: Kamps, J., Karlgren, J., Mika, P., Murdock, V. (eds.) Proceedings of ESAIR 2012 c/o CIKM 2012. pp. 15–16. Association for Computing Machinery, Maui, Hawaii, USA (2012). https://doi.org/10.1145/2390148.2390157
41. Sojka, P., Lee, M., Řehůřek, R., Hatlapatka, R., Kucbel, M., Bouche, T., Goutorbe, C., Anghelache, R., Wojchiechowski, K.: Toolset for Entity and Semantic Associations – Final Release (Feb 2013), deliverable D8.4 of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library
42. Sojka, P., Líška, M.: The Art of Mathematics Retrieval. In: Proceedings of the ACM Conference on Document Engineering, DocEng 2011. pp. 57–60.

Association of Computing Machinery, Mountain View, CA, USA (Sep 2011), https://doi.org/10.1145/2034691.2034703

43. Sojka, P., Řehůřek, R.: Classification of Multilingual Mathematical Papers in DML-CZ. In: Sojka, P., Horák, A. (eds.) Proceedings of Recent Advances in Slavonic Natural Language Processing—RASLAN 2007. pp. 89–96. Masaryk University, Karlova Studánka, Czech Republic (Dec 2007)
44. Srivastava, A., Sutton, C.A.: Autoencoding Variational Inference For Topic Models. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference Track Proceedings. OpenReview.net (2017), https://openreview.net/forum?id=BybtVK9lg
45. Suzuki, M., Tamari, F., Fukuda, R., Uchida, S., Kanahori, T.: INFTY — An Integrated OCR System for Mathematical Documents. In: Vanoirbeek, C., Roisin, C., Munson, E. (eds.) Proc. of ACM Symposium on Document Engineering 2003. pp. 95–104. ACM, Grenoble, France (2003)
46. Tai, K.S., Socher, R., Manning, C.D.: Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. pp. 1556–1566. Beijing, China (2015)
47. Teng, Z., Zhang, Y.: Head-lexicalized bidirectional tree LSTMs. Transactions of the Association for Computational Linguistics **5**, 163–177 (2017). https://doi.org/10.1162/tacl_a_00053
48. Thanda, A., Agarwal, A., Singla, K., Prakash, A., Gupta, A.: A Document Retrieval System for Math Queries. In: Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies. pp. 346–353. Tokyo, Japan (2016)
49. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is All you Need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 5998–6008. Curran Associates, Inc. (2017), https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf
50. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding (2018)
51. Warstadt, A., Singh, A., Bowman, S.R.: Neural Network Acceptability Judgments. arXiv preprint arXiv:1805.12471 (2018), https://arxiv.org/abs/1805.12471
52. Wojciechowski, K., Nowiński, A., Sojka, P., Líška, M.: The EuDML Search and Browsing Service – Final (Feb 2013), deliverable D5.3 of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library, revision 1.2 https://project.eudml.eu/sites/default/files/D5_3_v1.2.pdf
53. Yasunaga, M., Lafferty, J.D.: TopicEq: A Joint Topic and Mathematical Equation Model for Scientific Texts. Proceedings of the AAAI Conference on Artificial Intelligence **33**, 7394–7401 (07 2019). https://doi.org/10.1609/aaai.v33i01.33017394
54. Zanibbi, R., Davila, K., Kane, A., Tompa, F.W.: 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. pp. 145–154. ACM (2016)
55. Zhang, X., Lu, L., Lapata, M.: Top-down Tree Long Short-Term Memory Networks. In: Proceedings of NAACL-HLT. pp. 310–320. San Diego, California (2016)
56. Zhu, X., Sobhani, P., Guo, H.: Long short-term memory over recursive structures. In: Proceedings of the 32nd International Conference on Machine Learning. pp. 1604–1612. Lille, France (2015)