Project Proposal: Creating a Database of Definitions From Large Mathematical Corpora

Luis Berlioz

University of Pittsburgh
lab232@pitt.edu

We propose a method to gather large amounts of mathematical definitions from mathematical documents available online. Recent work indicates that well known text classification algorithms [2, 3] can have excellent accuracy at determining when a certain paragraph is in fact a definition [6]. These algorithms are trained on large math corpora available online like the arXiv website. The LaTeX source code of these documents is first converted into a more structured format like XML or HTML with the software package LaTeXML [10]. The content of the resulting files is then tokenized and fed into a word embedding algorithm like GloVe [12]. This has been implemented already and is available in [5].

As training data for the classifier, we use the passages of certain articles that are labeled as definitions by the author by placing them in certain LaTeX macro environments. These macros are normally defined in the preamble of the document using the `\newtheorem` macro. LaTeXML deals with the user defined macros and tags the corresponding text in the output. We have performed small experiments which show great promise. And these were confirmed with the results shown on the website https://corpora.mathweb.org/classify_paragraph.

The classifier takes the text of each paragraph of an article and outputs an estimate of the probability of it being a definition. Alternatively, a sliding window method can be used to obtain passages that produce a high probability. This method has the advantage of finding the definitions that are not expressed in precisely one paragraph, nevertheless it implies evaluating the classifier on a larger number of passages. In this situation, we consider the *fasttext* method in [8] which has a slightly lower accuracy but evaluates a passage much faster than any method previously considered.

Next, we plan to organize the definitions in an ordered tree structure where the nodes of the tree are definitions and the order represents the dependence between the nodes. In each definition we will identify the *definiendum* (i.e., the term being defined) by adapting a named entity recognition algorithm described in [13]. Moreover, by applying well established methods like [11, 4] to detect common phrases we can identify concepts with name spanning multiple words. We can also deal with the polysemy and synonymy [14, 7] which is very common in mathematical jargon by performing disambiguation on the cases polysemy and marking or merging the nodes that show synonymy.

We plan to produce a data set that would be useful in the formalization of mathematical theories, by giving a rough survey of the mathematical landscape. As another example, a database of virtually all the definitions in mathematics can be used to create user interfaces that allows authors to produce semiformal [9] versions of their work. This user interface would let authors browse all the alternative definitions of a given term, allowing them to reuse and improve on previous entries. We also plan to make all data freely available as part of the Formal Abstracts Project [1], in the hope of getting feedback from the interested community to improve and shape future iterations of this work.

References

[1] Formal abstracts. https://formalabstracts.github.io/, 2019.

[2] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

[3] Tao Chen, Ruifeng Xu, Yulan He, and Xuan Wang. Improving sentiment analysis via sentence type classification using bilstm-crf and cnn. *Expert Systems with Applications*, 72:221–230, 2017.

[4] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[5] Deyan Ginev. arxmliv:08.2018 dataset, an html5 conversion of arxiv.org, 2018. SIGMathLing – Special Interest Group on Math Linguistics.

[6] Deyan Ginev. A web demo for scientific paragraph classification. https://github.com/dginev/web-scipara-demo, 2018.

[7] Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882. Association for Computational Linguistics, 2012.

[8] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 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*, pages 427–431. Association for Computational Linguistics, 2017.

[9] Christoph Lange. *Enabling collaboration on semiformal mathematical knowledge by semantic web integration*, volume 11. IOS Press, 2011.

[10] Bruce Miller. Latexml: A latex to xml converter. url: http://dlmf. nist. gov. *LaTeXML/(visited on 03/12/2013)*, 2013.

[11] Richard C Murphy. Phrase detection and the associative memory neural network. In *Proceedings of the International Joint Conference on Neural Networks, 2003.*, volume 4, pages 2599–2603. IEEE, 2003.

[12] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[13] Horacio Saggion. Identifying definitions in text collections for question answering. In *LREC*, 2004.

[14] Yifan Sun, Nikhil Rao, and Weicong Ding. A simple approach to learn polysemous word embeddings. *arXiv preprint arXiv:1707.01793*, 2017.