M. Schubotz et al. "Introducing MathQA - A Math-Aware Question Answering System". In: *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL), Workshop on Knowledge Discovery.* Fort Worth, USA, June 2018

Introducing MathQA - A Math-Aware Question Answering System

Moritz Schubotz, Philipp Scharpf,
Kaushal Dudhat, Yash Nagar, Felix Hamborg, Bela Gipp
Information Science Group, University of Konstanz, Germany (first.last@uni-konstanz.de)

ABSTRACT

We present an open source math-aware Question Answering System based on Ask Platypus. Our system returns as a single mathematical formula for a natural language question in English or Hindi. This formulae originate from the knowledge-base Wikidata. We translate these formulae to computable data by integrating the calculation engine sympy into our system. This way, users can enter numeric values for the variables occurring in the formula. Moreover, the system loads numeric values for constants occurring in the formula from Wikidata. In a user study, our system outperformed a commercial computational mathematical knowledge engine by 13 %. However, the performance of our system heavily depends on the size and quality of the formula data available in Wikidata. Since only a few items in Wikidata contained formulae when we started the project, we facilitated the import process by suggesting formula edits to Wikidata editors. With the simple heuristic that the first formula is significant for the article, 80 % of the suggestions were correct.

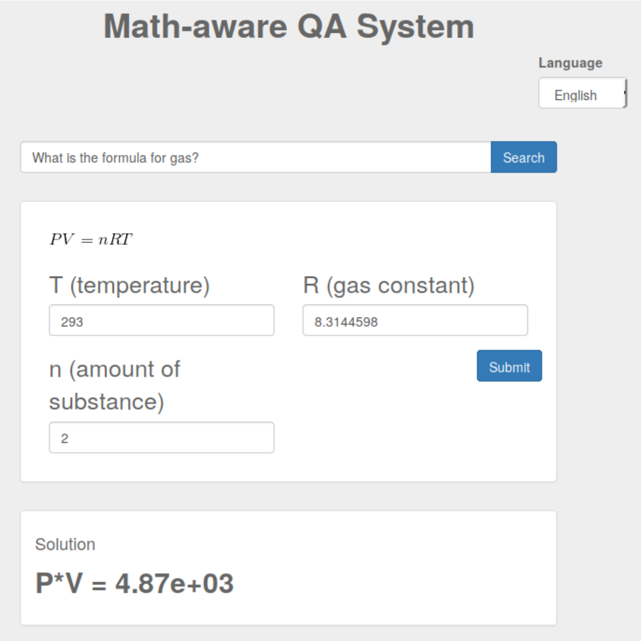
KEYWORDS

Mathematical Information Retrieval (MIR), Question Answering Systems (QA), Wikidata

ACM Reference Format:
Moritz Schubotz, Philipp Scharpf, Kaushal Dudhat, Yash Nagar, Felix Hamborg, Bela Gipp. 2018. Introducing MathQA - A Math-Aware Question Answering System. In *JCDL '18: The 18th ACM/IEEE Joint Conference on Digital Libraries, June 3–7, 2018, Fort Worth, TX, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/XXX

1 INTRODUCTION

Question Answering (QA) systems are Information Retrieval (IR) systems, allowing the user to pose questions in natural language to provide quick and succinct answers - in contrast to search engines which deliver ranked lists of documents. In this project, we developed an open source QA system, which is available at https://github.com/ag-gipp/MathQa. Our system can answer mathematical questions in the form of natural language, yielding a formula, which is retrieved from Wikidata. Wikidata is a free and open knowledge-base that can be read and edited by humans and machines. It stores common sources of other Wikimedia projects, especially for Wikipedia infoboxes. In addition, our system enables the user to perform arithmetic operations using the retrieved formula. We developed three modules: The *Question Parsing Module* (1) transforms questions into a triple representation and produces a simplified dependency tree. The *Formula Retrieval Module* (2) then queries the Wikidata knowledge-base for the requested formula and presents the result to the user. The user can subsequently choose values for the occurring variables and order a calculation that is done by a *Calculation Module* (3). If available, the system retrieves the identifier names and values from Wikidata, so that the user can understand their meaning (see Figure **??** above). Moreover, we developed a module which can answer questions in the Hindi language. In contrast to the English, which is exploiting the dependency graph of Stanford NLP, the Hindi module uses regular expressions to parse the questions and before passing the triple representation to the Formula Retrieval Module. Our QA system builds upon Ask Platypus [7], an existing QA engine that can answer English questions using Wikidata.



Language: English

What is the formula for gas?

$PV = nRT$

T (temperature): 293

R (gas constant): 8.3144598

n (amount of substance): 2

Solution

P*V = 4.87e+03

Figure 1: Screenshot of MathQA

JCDL '18, June 3–7, 2018, Fort Worth, TX, USA
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5178-2/18/06...$15.00
https://doi.org/XXX

We chose Ask Platypus as the best among other Wikidata-based QA systems and extended its functionality to include mathematical questions. Finally, we evaluated the system's performance and the quality of results in comparison to a commercial computational mathematical knowledge-engine. Our system outperforms the reference engine on some definition and geometry type questions, and we conjecture that the validity can be expanded to the whole domain. Before building the QA system, we performed a seeding of all currently available mathematical formulae (labeled by math-tags) from Wikipedia into Wikidata. Each section of this paper is divided into two parts: the first part describes the process of seeding the Wikidata knowledge-base with mathematical formulae from Wikipedia as a separate project, laying the foundation for the second part, its application in the QA system.

Vision. The mathematical QA system is a first motivating application that exploits the mathematical knowledge seeded into Wikidata. It is a first step towards our long-term goal of building a *collaborative, semi-formal, language independent math(s) encyclopedia* hosted by Wikimedia at *math.wikipedia.org* [3]. Using the popular Wikipedia framework as frontend will help popularize the project and motivate many experts from the mathematical sciences to contribute. We envisage a future centralized, machine-readable repository for mathematical world-knowledge that can be utilized to enable cross-article queries, e.g., to automate proofs of mathematical theorems. A crucial foundation for a path towards this long-term goal is having a large amount of mathematical data in Wikidata. This paper is a starting point for the development of effective methods to automatically seed Wikidata with mathematical formulae from Wikipedia or STEM documents.

Problem Setting. Wikipedia consists of many pages related to mathematics. However, promptly grasping the essence of an article can be a difficult task as many pages contain a lot of information. Using Wikipedia means reading articles, and there currently is no way to automatically gather information scattered across multiple articles [5]. To overcome this problem, Wikidata can be used as a source. Wikidata provides machine-readable content that can automatically be interpreted by a computer and queried to access specific information. Thus, there is a huge potential in adding formulae related to all mathematical topics as items to Wikidata, enabling direct access to the defining formula of a requested mathematical concept. The first goal of this project was to enrich Wikidata with mathematical knowledge it currently lacks. Adding this information into Wikidata will not only increase the content of these items but also make them more meaningful and useful. Furthermore, these formulae will be machine-interpretable and can be used in many applications in the STEM disciplines. Most importantly, we are then able to develop the mathematical QA system which can directly answer mathematical questions provided by the user, using the mathematical formulae and relations available on Wikidata. As a result, instead of retrieving a whole Wikipedia page which is full of text, users will directly get the desired piece of information, the formula they are looking for.

Research Objectives. Motivated by the lack of mathematical knowledge in Wikidata, the following research objective was defined:

Identify and extract defining formulae from all the available mathematical articles on Wikipedia to seed them into the Wikidata knowledge-base.

To achieve this objective, the following tasks were performed: 1.) Identification of mathematical articles from the Wikipedia data dump. 2.) Manual analysis to determine the defining formula of an individual article. 3.) Seeding of the retrieved formulae into Wikidata using the *Primary sources tool* [1]. 5.) Evaluation of the overall correctness and accuracy of the data migration by precision, recall, and f-measure.

Subsequently, we capitalized on the formulae seeded into Wikidata to

Build a math-aware QA system, processing a mathematical natural language question to retrieve a formula from Wikidata and allow a calculation based on input values for the occurring variables provided by the user.

We performed the following subtasks: 1.) Development of a *Question Parsing Module* that determines a triple representation of the user's input. 2.) Development of a *Formula Retrieval Module* to query Wikidata using pywikibot [2]. 3.) Development of a *Calculation Module* that performs a calculation based on the retrieved formula for the question and input values for the variables provided by the user. 4.) Evaluation of the overall performance and comparison to a commercial computational mathematical knowledge-engine. 5.) Development of regular expressions to maximize the number of answerable questions provided by the user in the Hindi language.

Section Outline. This paper is organized as follows: Section *Background* contains details about the Wikimedia sister projects Wikipedia and Wikidata and the concept of QA systems. Subsection *Implementation* describes our approach of transferring formulae from Wikipedia to Wikidata and the structure of the QA system which uses the seed. In subsection *Evaluation* we describe the construction of a random sample to assess the quality of the data transfer by precision, recall and f-measure. Subsequently, we evaluate the performance of the QA system and discuss its limitations. Finally, we conclude with a summary and suggested improvements for future work.

2 BACKGROUND

2.1 Wikipedia and Wikidata

Started in 2001, mainly as a text-based resource, Wikipedia[1] is the world's largest online encyclopedia which allows its users to edit articles and add new information into it [4].

[1] http://www.wikipedia.org

Wikipedia has collected an rapidly increasing amount of information, including numbers, coordinates, dates and other types of relationships among different domains of knowledge. Denny Vrandecic, ontologist at Google, claims that *It has become a resource of enormous value, with potential applications across all areas of science, technology and culture* [9].

Wikipedia is open and welcomes everyone who wants to make a positive contribution. Ward Cunningham, the inventor of Wiki, describes Wikipedia as *The simplest online database that could possibly work* [6].

End of the free preprint. Contact moritz@schubotz.de for further information.

ACKNOWLEDGMENTS

We would like to thank Akiko Aizawa for her advice and for hosting us as visiting researchers in her lab at the National Institute of Informatics (NII) in Tokyo. Furthermore, we thank Wikimedia Foundation and Wikimedia Deutschland providing cloud computing facilities and us for a research visit. Besides many Wikimedians, Lydia Pintscher and Jonas Kress were a great help in getting started with Wikidata. This work was supported by the FITWeltweit program of the German Academic Exchange Service (DAAD) as well as the German Research Foundation (DFG grant GI-1259-1).

REFERENCES

[1] S. S. (Google). *Wikidata:Primary sources tool.* https://www.wikidata.org/wiki/Wikidata:Primary_sources_tool. Accessed: 2018-04-11.

[2] 1. G. contributors. *Pywikibot: Python library to automate work on MediaWiki sites.* https://www.mediawiki.org/wiki/Manual:Pywikibot. Accessed: 2018-04-11.

[3] J. Corneli and M. Schubotz. "math. wikipedia. org: A vision for a collaborative semi-formal, language independent math (s) encyclopedia". In: *Proc. CAITP* (2017).

[4] W. Foundation. *Wikipedia, The Free Encyclopedia.* Accessed: 2018-04-11.

[5] M. Krötzsch et al. "Semantic Wikipedia". In: *J. Web Sem.* 5.4 (2007), pp. 251–261. DOI: 10.1016/j.websem.2007.09.001.

[6] B. Leuf and W. Cunningham. "The Wiki way: quick collaboration on the Web". In: (2001).

[7] S. by Lexistems SAS and E. de Lyon. *Ask Platypus.* https://askplatyp.us/. Accessed: 2018-04-11.

[9] D. Vrandecic and M. Krötzsch. "Wikidata: a free collaborative knowledgebase". In: *Commun. ACM* 57.10 (2014), pp. 78–85. DOI: 10.1145/2629489.

Listing 1: Use the following BibTeX code to cite this article

```
@InProceedings{Schubotz2018a,
  author    = {Moritz Schubotz and Philipp Scharpf and Kaushal Dudhat and Yash Nagar and Felix
      Hamborg and Bela Gipp},
  title     = {Introducing MathQA - A Math-Aware Question Answering System},
  booktitle = {Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL),
      Workshop on Knowledge Discovery},
  year      = {2018},
  month     = {6},
  address   = {Fort Worth, USA},
}
```