Formula Concept Discovery and Recognition

Philipp Scharpf
Dept. of Computer and Information Science
Konstanz, Germany
philipp.scharpf@uni-konstanz.de

Abstract

In my dissertation, I will develop a method to discover (define) and recognize (identify) formula concepts in Wikipedia articles and STEM documents using Wikidata as a semantic knowledge-base. Both structural (syntax tree) and semantic (identifier names) formula information will be considered. The approach is expected to improve search engines, recommender systems, plagiarism and novelty detection and ontology learning.

Research Motivation

My research is motivated by 1) the need for Information Retrieval systems to match mathematical formulae when assessing semantic content and similarity of STEM documents, and 2) the challenge that a given mathematical formula concept usually appears in several variations or equivalent representations.

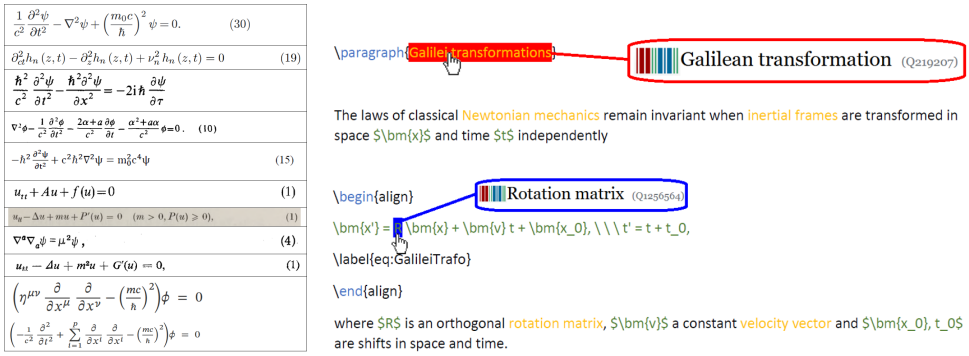

Figure 1: Various representations of the Klein-Gordon equation from arbitrarily selected sources (left) and LaTeX document annotation system "AnnotaTeX" (right).

Figure 1 (left) illustrates the observation that a selected formula concept - here the Klein-Gordon-equation from quantum physics - appears in the literature in a variety of different but equivalent representations. In numerous publications, it is at least slightly different from the rest. To make such a formula machine-interpretable, i.e., teach the computer how to understand its semantics, there needs to be a definition of the mathematical concept that comprises as many variations as possible.

Research Plans

My research goals are 1) development and optimization of Formula Concept Discovery to elaborate a definition of a formula concept by examining Wikipedia articles and STEM documents from the ArXiV, 2) seeding a scalable large number to Wikidata with suitable annotation of the formula parts (identifiers and operators) and as a preparation for 3) the development and optimization of Formula Concept Recognition techniques that match a given formula to a Wikidata concept item, independently from its specific representation.

Databases

arXiv, Wikipedia, and Wikidata In my research, I will focus on examples from mathematics and physics

In: O. Hasan, D. Gallois-Wong (eds.): Proceedings of CICM 2018, Hagenberg, Austria, 13-08-2018, published at http://ceur-ws.org

retrieved from the arXiv repository of electronic preprints (`http://arxiv.org/`) and Wikipedia. I am striving to develop a method that will be able to map, e.g., all of the formulae collected in figure 1 - in particular, linkable to the Wikidata entry `https://www.wikidata.org/wiki/Q868967`. I chose the semantic knowledge-base Wikidata because it is free, open and can be read and edited by humans and machines.

Research Method

Formula Feature Analysis

The first step in the formula feature analysis is tokenization, i.e., the decomposition into their components (identifiers, operators, numbers, etc.) and Part-Of-Math-tagging: a formula consists of different terms, which must be distinguished from each other. The Klein-Gordon equation $\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\psi - \nabla^2\psi + \frac{m^2c^2}{\hbar^2}\psi = 0$ from quantum physics, again used as an instructive example, contains a term $\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\psi$ with a double time derivative, one with a double space derivative $\nabla^2\psi$ as well one with a constant prefactor $\frac{m^2c^2}{\hbar^2}\psi$ -the first term can then be further decomposed into its characters (tokens), that is, the denominator c for the speed of light, the operator $\frac{\partial}{t}$ with an exponent (number) 2 and the identifier ψ for the physical (quantum) wave function.

When analyzing the semantics of a formula, we are faced with the problem of identifier ambiguity, which requires disambiguation with the help of the partial clarifications available in the text. A single identifier has a theoretically unlimited number of possible meanings, e.g., E in physics often refers to both an energy and an electric field, generally mathematically an expected value, etc. Thus, it is essential to improve the retrieval of the semantics from the surrounding text.

Research Questions

The aim of Formula Concept Discovery (FCD) is to 1) retrieve a large number of formula examples from Wikipedia articles and arXiv documents together with a mapping to formula concepts (Wikidata items), and 2) recover a general definition of a formula concept using feature analysis and abstract mathematical formalization.

The aim of Formula Concept Recognition (FCR) is to identify formulae in arXiv documents or Wikipedia articles as Wikidata formula concept items. Therefore, a measure of similarity that allows assigning a formula to a mathematical concept (equation) if it exceeds a defined threshold needs to be defined. A first rough approach could be a matching score = # recognized elements / # total elements. To successfully identify a single element, for example, the Laplace operator $\nabla^2 = \Delta$, it must be assigned to the corresponding concept in Wikidata, at `https://www.wikidata.org/wiki/Q203484`, i.e. to QID *Q203484*. The aim is to motivate active users of Wikidata to gradually build a hierarchical structure of the formula elements, assign elements to all available formulae (property *has part*) and create new items for formulae concepts directly including the parts.

Evaluation Plans

I will compare and discuss 1) several possible Formula Concept Discovery methods (e.g., taking the first formula from a Wikipedia article as defining formula of the concept, formula clustering, etc.), and 2) several possible Formula Concept Recognition methods (e.g., simple TeX string search vs. parts identification, recognition by identifier name, symbol and value, etc.).

Completed Research

In my first publication [SGPS+18], I significantly contributed to the creation of a Gold standard MathMLben for the evaluation of the conversion between different mathematical formats (LaTeX vs. Computer Algebra Systems). In my second publication [SSD+18], I presented the first math-aware QA system that can answer a natural language question yielding a mathematical formula using Wikidata. My third recent publication [SSG18] initiates my reasoning on a definition of a formula concept and its possible content representations in LaTeX, MathML, and Wikidata.

Remaining Research

In my next publication, I will provide a thorough literature review on formula feature analysis. Together with M. Schubotz and A. Greiner-Petter, I am planning to develop an annotation tool **AnnotaTeX** for LaTeX documents that will facilitate the annotation process by recommending identifier names to the user. Figure 1 (right) shows a proposed User Interface.

References

[SGPS+18] Moritz Schubotz, Andre Greiner-Petter, Philipp Scharpf, Norman Meuschke, Howard Cohl, and Bela Gipp. Improving the representation and conversion of mathematical formulae by considering their

textual context. In *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, Fort Worth, USA, Jun. 2018.

[SSD+18] Moritz Schubotz, Philipp Scharpf, Kaushal Dudhat, Yash Nagar, Felix Hamborg, and Bela Gipp. 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, Jun. 2018.

[SSG18] Philipp Scharpf, Moritz Schubotz, and Bela Gipp. Representing mathematical formulae in content mathml using wikidata. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2018.