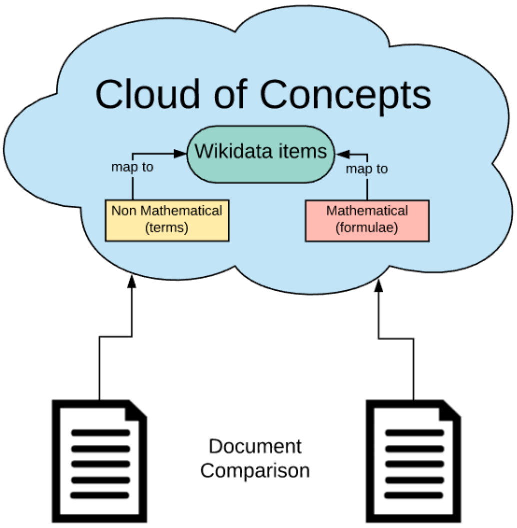
philipp.scharpf@uni-konstanz.de

STUDENT PROJECT PROPOSAL:

FORMULA CONCEPT DISCOVERY AND RECOGNITION USING WIKIDATA



Identification of Wikidata items as form- and language-independent markers in documents is gainful for document analysis to improve the assessment of semantic similarity and relatedness with applications for example in plagiarism detection, recommendation systems, literature exploration, novelty detection and much more.
While for non-mathematical terms, named entity recognition has already been developed [1-3], for mathematical formulae there is still an urgent need to elaborate effective methods to discover and recognize mathematical concepts using Wikidata. As the semantics of identifiers contained within a formula can often be retrieved from the surrounding text [4,5], we propose to use them for seeding and identifying a formula by its parts.

We offer two projects to ambitious students who are highly motivated to make a contribution to this innovative topic:

PROJECT 1: FORMULA CONCEPT DISCOVERY USING WIKIDATA

Research Goal

Effective seeding of formula concepts from Wikipedia articles or STEM documents to Wikidata items (Mathematical Language Processing task).

Research Tasks

- Propose three methods to discover semantic formula concepts in Wikipedia articles or STEM documents and seed them as items to Wikidata or add defining formulae (including the semantic of the identifiers) to existing items.
- Evaluate your methods in comparison.
  How many formula concepts were correctly seeded?

PROJECT 2: FORMULA CONCEPT RECOGNITION USING WIKIDATA

Research Goal

Effective mapping of formula concepts in STEM documents or Wikipedia articles to Wikidata items (Mathematical Language Processing task).

Research Tasks

- Propose three methods to identify semantic formula concepts in STEM documents or Wikipedia articles and match them (using the semantic of the identifiers) to Wikidata items.
- Evaluate your methods in comparison.
  How many formula concepts were correctly identified?

PROPOSED TOOLS

python

Cornell University Library
arXiv.org

STEM
Science · Technology · Engineering · Math

WIKIDATA

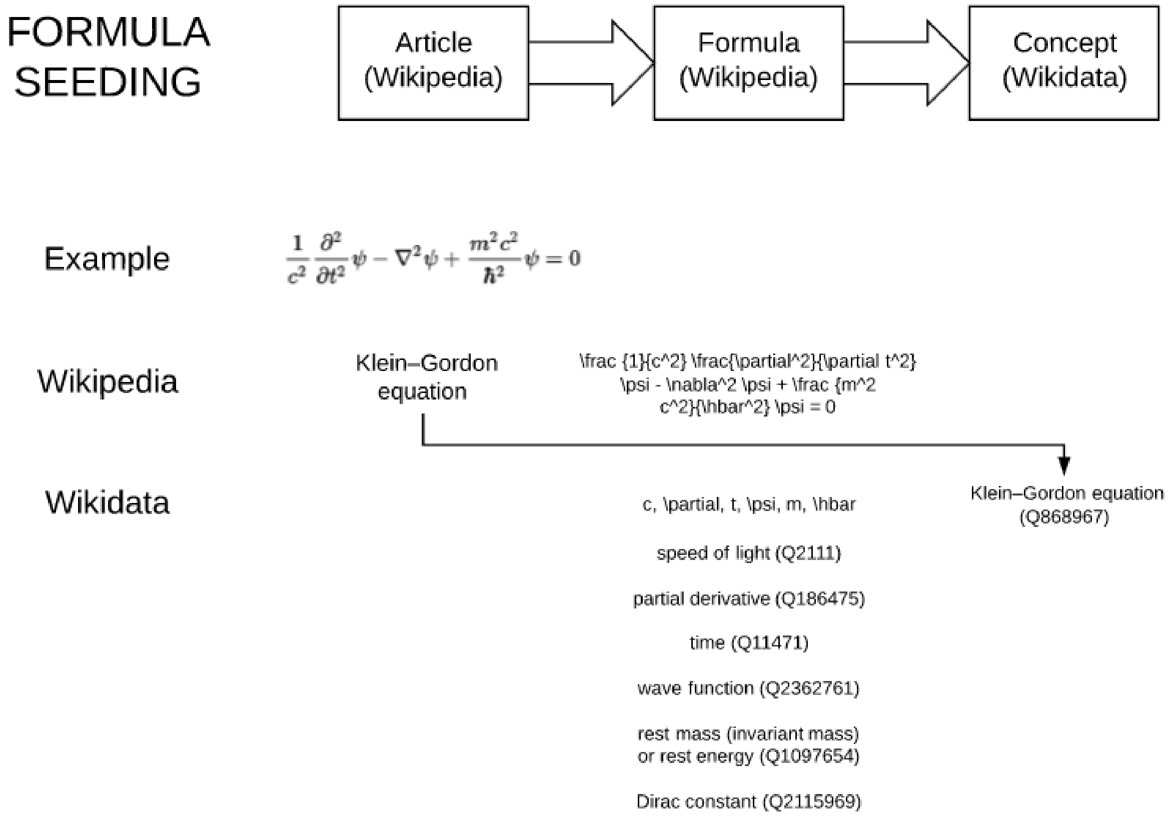
SPARQL

WIKIPEDIA
Die freie Enzyklopädie

Expected Background of the Applicant

✓ Programming skills ✓ Mathematical comprehension

PROPOSED SCHEMES
FOR FORMULA DISCOVERY AND RECOGNITION
(EXAMPLE FROM PHYSICS)



FORMULA SEEDING

Article (Wikipedia) → Formula (Wikipedia) → Concept (Wikidata)

Example

$$\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\psi - \nabla^2\psi + \frac{m^2c^2}{\hbar^2}\psi = 0$$

Wikipedia

Klein–Gordon equation

\frac {1}{c^2} \frac{\partial^2}{\partial t^2} \psi - \nabla^2 \psi + \frac {m^2 c^2}{\hbar^2} \psi = 0

Wikidata

c, \partial, t, \psi, m, \hbar

speed of light (Q2111)

partial derivative (Q186475)

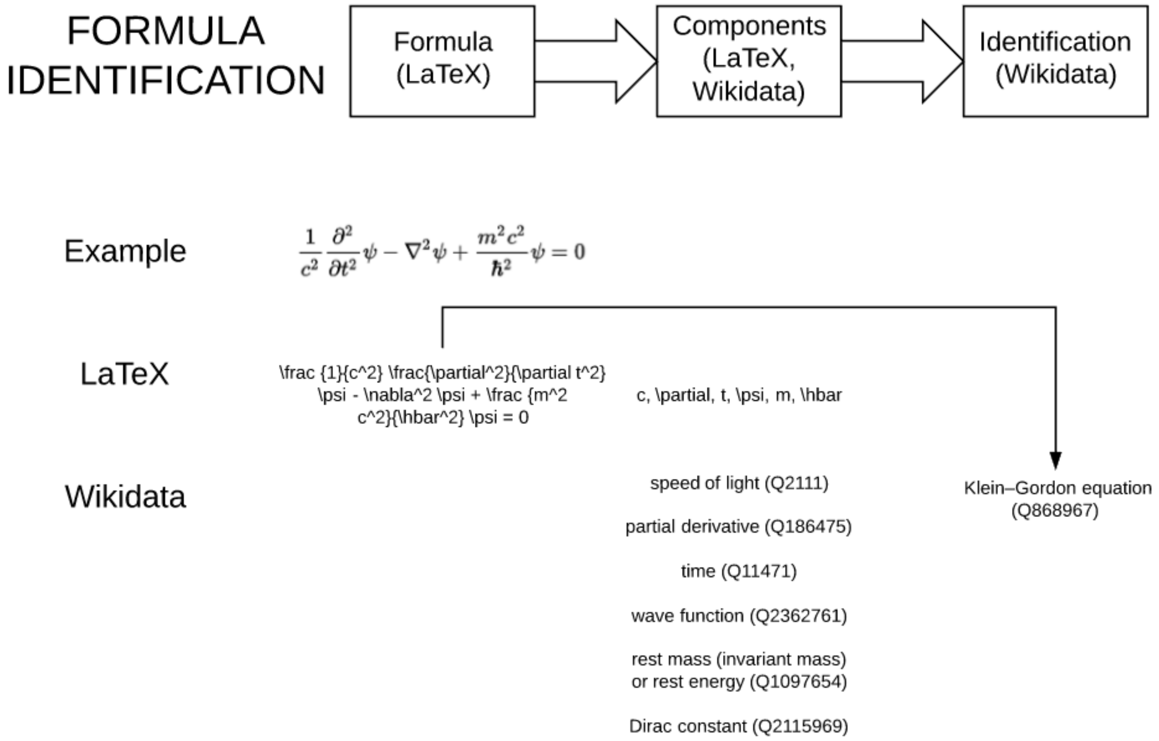
time (Q11471)

wave function (Q2362761)

rest mass (invariant mass) or rest energy (Q1097654)

Dirac constant (Q2115969)

Klein–Gordon equation (Q868967)



FORMULA IDENTIFICATION

Formula (LaTeX) → Components (LaTeX, Wikidata) → Identification (Wikidata)

Example

$$\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\psi - \nabla^2\psi + \frac{m^2c^2}{\hbar^2}\psi = 0$$

LaTeX

\frac {1}{c^2} \frac{\partial^2}{\partial t^2} \psi - \nabla^2 \psi + \frac {m^2 c^2}{\hbar^2} \psi = 0

c, \partial, t, \psi, m, \hbar

Wikidata

speed of light (Q2111)

partial derivative (Q186475)

time (Q11471)

wave function (Q2362761)

rest mass (invariant mass) or rest energy (Q1097654)

Dirac constant (Q2115969)

Klein–Gordon equation (Q868967)

Proposed Literature

[1] Geiß, Johanna, Andreas Spitz, and Michael Gertz. *NECKAr: A Named Entity Classifier for Wikidata.* International Conference of the German Society for Computational Linguistics and Language Technology. Springer, Cham, 2017.

[2] Spitz, Andreas, et al. *State of the Union: A Data Consumer's Perspective on Wikidata and Its Properties for the Classification and Resolution of Entities*. Wiki@ ICWSM, 2016.

[3] Taufer, Pavel. *Named Entity Recognition and Linking*. Faculty of Mathematics and Physics, Charles University, Prague Czechia. Master’s Thesis, 2017.

[4] Schubotz, Moritz, et al. *Semantification of Identifiers in Mathematics for Better Math Information Retrieval*. Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. ACM, 2016.

[5] Schubotz, Moritz. *Augmenting Mathematical Formulae for More Effective Querying & Efficient Presentation*. Epubli, 2017.

Contact

philipp.scharpf@uni-konstanz.de

https://www.isg.uni-konstanz.de/people/doctoral-researchers/philipp-scharpf/