Develop a mathematical e-dictionary and test it

Theresa Kruse[1]

[1]University of Hildesheim, Universitätsplatz 1, 31141 Hildesheim, Germany

Abstract

In the presented dissertation project, an electronic dictionary on graph theory is developed and evaluated. Student teachers are the target group of the dictionary. The information needed for the dictionary is automatically extracted from a domain-specific corpus. After implementation a dictionary prototype is evaluated with student teachers. The results are compared to similar use situations with Wikipedia.

Keywords

terminology, dictionary, term extraction, definition extraction

1. Research questions

The project examines how the usage of an electronic dictionary affects learning of terminology. We work with mathematic student teachers in the field of graph theory. We develop and test a domain-specific, bilingual e-dictionary. The project consists of two parts: create the dictionary in a highly automatic way and test its efficiency.

The first part will answer the following questions: Which tools help to create an e-dictionary in a highly automatic way? Do mathematical texts require special needs for term extraction? How can we use neural networks to extract definitions from mathematical texts?

The second part answers the following research questions: Which terminology do students use to describe graphs? Does their usage of terminology change when they use a domain-specific e-dictionary? Are there differences in the usage of the dictionary between the domain-specific ressource and Wikipedia? How do the students navigate through the access structure? Does the usage of terms derived from the general language differ from that of loan words? How can a domain-specific dictionary help when translating specialized texts?

2. Course of the research

The project began in 2018/19 and we plan to finish by the end of 2022. For the first part, we created two comparable corpora with texts on graph theory in German and English. We have about 700.000 German tokens and one million English tokens, with 30.000 types each.

We extracted terms with two different methods, both pattern-based. One took only frequencies into account, the other simply relied on frequencies and domain-specific patterns. Our working hypothesis was that due to the highly structured language in mathematics, it is not necessary

Woodstock'22: Symposium on the irreproducible science, June 07–11, 2022, Woodstock, NY

theresa.kruse@uni-hildesheim.de (T. Kruse)

0000-0003-4613-2060 (T. Kruse)

to include frequencies. Three experts on graph theory determine by rating which items are suitable for a dictionary on graph theory. The inclusion of frequency gave slightly better results. The final lemma list includes 1453 terms. The process of term extraction and lemma selection is described at length in Kruse and Heid [1].

In the next step, we used a neural network to extract definitions from the corpus. We used `bert-based-cased` and `bert-based-german-cased` from the Hugging Face library. For the English data, we used training data by Vanetik et al. [2]. The German training data had to be created manually from the corpus. For the evaluation, we calculated precision p, recall r, and F-score F for a random sample of 200 sentences from each category because we did not have the capacity to manually annotate the whole data set. For English results we got $p = 0.7054$, $r = 0.9100$, $F = 7948$ and for German $p = 0.7522$, $r = 0.8500$, $F = 0.7981$ [3]. As another reference, we intend to use term lists on graph theory provided by Wikipedia. We will compare which terms are included in both, our extracted definitions and the Wikipedia lists. Of course, we have to keep in mind that our tool is not able to find definitions for terms that are not mentioned in the corpus.

Parts of the extracted terms and definitions were implemented into a dictionary prototype using LexO, provided by Bellandi et al. [4]. To that end, the terms from the lemma list were combined with the extracted definitions. Further information for the dictionary entries was extracted from the corpus, e.g. terms with a semantic relation to each other.

For the second part of the project, we conducted a study with mathematics undergraduate students as they are the target group of the dictionary. The study was carried out in two parts: We did the first round with students using Wikipedia as an aid in summer 2020 ($N = 182$) and the second round with a comparable group in summer 2021 ($N = 113$) using our prototype. At this point of their studies, the participants already have a basic understanding of scholarly mathematics but do not have a particular knowledge of graph theory, yet.

We gave them five different tasks. In the first task, they were asked to describe given graphs. The second task asked them to recognize an isomorphism between two graphs. In the third task, we wanted to explore if there are peculiarities in the understanding of terminology derived from the general language. The fourth task asked them to list algorithms for a certain graph-theoretic problem. In the last task, they translated sentences from English to German.

Both studies are evaluated by a mixed-methods approach and we compare the results. Right now, the evaluation is an ongoing process for which the following methods are used: a qualitative content analysis of comments on the tasks, case studies for up to five single participants, Spearman correlation between the number of searches, the process time, and the number of points earned in the tasks. To compare both groups we use a t-test for the variables mentioned above and a U-test for the ordinal scaled data like study semester or language knowledge.

3. Publication plans

Unfortunately, we cannot make our corpora publicly available due to copyright restrictions. But the dictionary prototype itself will be available online after login. Preliminary results of the project were and are presented at conferences on lexicography and mathematical didactics.

References

[1] T. Kruse, U. Heid, From term extraction to lemma selection for an electronic lsp-dictionary in the field of mathematics, in: Electronic lexicography in the 21st century: post-editing lexicography, 2021, pp. 572–587.

[2] N. Vanetik, M. Litvak, S. Shevchuk, L. Reznik, Automated discovery of mathematical definitions in text, in: Proceedings of The 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 2086–2094. URL: https://www.aclweb.org/anthology/2020.lrec-1.256.

[3] T. Kruse, F. Kliche, Definition extraction from mathematical texts on graph theory in German and English, in: Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021), KONVENS 2021 Organizers, Düsseldorf, Germany, 2021, pp. 104–113. URL: https://aclanthology.org/2021.konvens-1.9.

[4] A. Bellandi, E. Giovannetti, S. Piccini, A. Weingart, Developing LexO: a collaborative editor of multilingual lexica and termino-ontological resources in the humanities, in: Proceedings of Language, Ontology, Terminology and Knowledge Structures Workshop (LOTKS 2017), Association for Computational Linguistics, Montpellier, France, 2017. URL: https://aclanthology.org/W17-7010.

A. Online Resources

The dictionary prototype is available online after login (gast / gast).