OntoClue, a framework to compare vector-based approaches for document relatedness using the RELISH corpus

Rohitha Ravinder [1,2], Tim Fellerhoff [1,3], Vishnu Dadi [1,4], Lukas Geist [1,4], Guillermo Rocamora [1,5], Muhammad Talha [1,4], Dietrich Rebholz-Schuhmann [1,6] and Leyla Jael Castro [1]

[1] *ZB MED Information Centre for Life Sciences, Gleueler Str. 60, Cologne, 50931, Germany*
[2] *Bonn-Aachen International Centre for Information Technology (B-IT), University of Bonn, Friedrich-Hirzebruch-Allee 6, Bonn, 53115, Germany*
[3] *Heinrich-Heine University Düsseldorf, Universitätsstraße 1, Düsseldorf, 40225, Germany*
[4] *Hochschule Bonn-Rhein-Sieg, Grantham-Allee 20, Sankt Augustin, 53757, Germany*
[5] *Universidad de Murcia, Avda. Teniente Flomesta 5, Murcia, 30003, Spain*
[6] *University of Cologne, Albertus-Magnus-Platz, Cologne, 50923, Germany*

Abstract

The continuous increase of biomedical scholarly publications makes it challenging to construct document recommendation algorithms to navigate through literature, an important feature for researchers to keep up with relevant publications. Understanding semantic relatedness and similarity between two documents could improve document recommendations. The objective of this study is performing a comparative analysis of vector-based approaches to assess document similarity in the RELISH corpus. Here we present our approach to compare five different techniques to generate vectors representing the text in the documents. These techniques employ a combination of various Natural Language Processing frameworks such as Word2Vec, Doc2Vec, dictionary-based Named Entity Recognition as well as state-of-the-art models based on BERT.

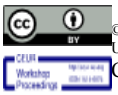
Keywords

Document similarity, Word embeddings, Named Entity Recognition

1. Introduction

Recommendation systems are a successful method to cope with information overload wrt scientific publications [1]. For biomedical publications, PubMed Related Articles (PMRA) [2] is still considered the de facto standard; however, Natural Language Processing (NLP) advances, including word-embeddings, offer alternative paths to improve the state of the art and explore further similarity, relatedness and relevance. The RELISH [3] dataset corresponds to a document-to-document relevance assessment (definitely relevant, partially relevant, non-relevant) that can be used for comparing, improving and translating newly developed literature search techniques, including recommendation systems. Here we present OntoClue, a framework to compare different approaches to generate vectors for articles in the RELISH corpus.

Proceedings Semantic Web Applications and Tools for Healthcare and Life Sciences, February 13–16, 2023, Basel, Switzerland
EMAIL: ljgarcia@zbmed.de (A. 8)
ORCID: 0000-0002-8725-1317 (A.2); 0000-0002-3082-7522 (A.3); 0000-0002-2910-7982 (A.4); 0000-0002-1018-0370 (A.7); 0000-0003-3986-0510 (A.8)



2. OntoClue framework

OntoClue can be summarized in the following steps: (i) retrieve title and abstract for the RELISH articles in XML recording those that cannot be retrieved, (ii) trim the RELISH corpus so it includes only retrieved articles, (iii) reduce the RELISH corpus so only relevance assessment for which there is a clear consensus are kept, (iv) connect approaches to be compared by OntoClue in a workflow fashion, (v) optimize the approaches using an Area Under the Curve (AUC) approach, (vi) evaluate precision and cumulative gain for each approach using the optimal parameters, (vii) provide comparison tables for the different approaches. The hyperparameter optimization follows a multi-classification approach using Cosine Similarity intervals from 0 to 1 with increments of 0.1 and counting the number of definitely relevant, partially relevant and non-relevant RELISH pairs for each interval. The optimization is based on the best AUC score obtained from different hyperparameter combinations for each participating approach. The optimization can also be simplified to two classes by combining definitely and partially relevant into one single class "relevant".

We are testing and tuning our OntoClue framework with five approaches: (i) Doc2Vec [4], existing approach for document vectors; (ii) word2doc2vec, in-house approach to document vectors; (iii) whatizit-dictionary, using Whatizit [5], a dictionary-based named entity recognition approach; (iv) hybrid-doc2vec, combination of Doc2Vec and Whatizit; and (v) a BERT-based approach using BERT pre-trained models (all of the others are trained with the RELISH articles only).

3. Future Work

We plan to use our OntoClue framework to compare the five mentioned approaches so we can select the best approach to propose a new recommendation system that should cover not only the biomedical domain but also the agricultural one as they correspond to our use case LIVIVO, the ZB MED literature portal. The recommendation system should also integrate multilingualism as LIVIVO contains publications in English, German, French, Portuguese and Spanish. In addition, we want to support coverage for non-traditional, e.g., data and software, and non-peer-reviewed journal publications, e.g., conference papers and preprints.

4. Acknowledgements

This work was partially supported by the STELLA project funded by DFG (project no. 407518790), the NFDI4DataScience project funded by GWK and DFG (no. NFDI 34/1), and the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, 031A538A)

5. References

1. Zhu J, Patra BG, Yaseen A. Recommender system of scholarly papers using public datasets. AMIA Jt Summits Transl Sci Proc. 2021 May 17;2021:672-679. PMID: 34457183; PMCID: PMC8378599.
2. Lin J, Wilbur WJ. PubMed related articles: a probabilistic topic-based model for content similarity. BMC Bioinformatics. 2007 Oct 30;8:423. doi: 10.1186/1471-2105-8-423. PMID: 17971238; PMCID: PMC2212667.
3. Brown P; RELISH Consortium, Zhou Y. Large expert-curated database for benchmarking document similarity detection in biomedical literature search. Database (Oxford). 2019 Jan 1;2019:baz085. doi: 10.1093/database/baz085. PMID: 33326193; PMCID: PMC7291946.
4. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. InAdvances in neural information processing systems. 2013:3111–3119.
5. Dietrich Rebholz-Schuhmann, Miguel Arregui, Sylvain Gaudan, Harald Kirsch, Antonio Jimeno. Text processing through Web services: calling Whatizit, Bioinformatics, Volume 24, Issue 2, 15 January 2008, Pages 296–298, https://doi.org/10.1093/bioinformatics/btm557.