Review and summary of articles in the special volumes

Logan Lomonaco

Editorial: Mining Scientific Papers: NLP-enhanced Bibliometrics. Iana Atanassova 1[*], Marc Bertin 2 and Philipp Mayr 3

Idea that bibliometrics can benefit from large-scale text analytics and sense mining of papers.

Recently, the ever growing availability of datasets and papers in full text and in machine-readable formats has made it possible to perform bibliometric studies considering both the metadata of papers and their full text content.

This has lead to growing interest in the field and the articles published in this Research Topic contribute to the literature through the use of AI tools to better analyze texts and further advance bibliometrics as a field.

Is the Abstract a Mere Teaser? Evaluating Generosity of Article Abstracts in the Environmental Sciences, Ermakova et al.

This paper examines the abstracts of scientific papers with a GEM score that measures the representativeness of an abstract or its “generosity” due to the abstract often being used as a proxy for the content of an article.

GEM scores were based on how different sections of the papers were scored based on importance to the reader, and sentences in the abstracts were matched to said sections based on similarity.

Overall it was found that GEM scores have increased over time as newer papers had higher scores. Another notable finding is that while there is not a perfect correlation between the GEM score and the mean citation rate, the lowest citations rates were for the articles with the lowest scores (≤0.4).

The Termolator: Terminology Recognition Based on Chunking, Statistical and Search- Based Scores, Meyers et al.

The Termolator is an open-source terminology extraction system, available on Github. This system utilizes chunking favoring chunks containing out-of-vocabulary words, nominalizations, technical adjectives, and other specialized word classes and term chunk ranking.

The Termolator system was about as accurate when analyzing Einstein's theory of relativity and more accurate on patent analysis when compared to a terminology extraction system called Termostat.

A Chinese language version of Termolator IS also in development and is currently being upgraded with the features in its english counterpart.

Overall development is still ongoing.

Deep Reference Mining From Scholarly Literature in the Arts and Humanities, Rodrigues Alves et al.

“Reference mining: the detection, extraction and classification of references within the full text of scholarly publications”.

This paper uses a deep learning architecture for reference mining from full scholarly texts. Paper discusses three main architectural components: word and character-level word embeddings, different prediction layers (Softmax and Conditional Random Fields) and multi-task over single-task learning.

Overall the paper finds that the adoption of deep learning methods can be a significant improvement for the general task of reference mining.

Temporal Representations of Citations for Understanding the Changing Roles of Scientific Publications, He and Chen

This paper analyzes the temporal characteristics of citations in order to represent the dynamic role of scientific publications by comparing different citation contexts to identify articles important to the development of science.

Despite the limitations of the study (information of how a citation was mentioned in a sentence was not used), It found that the temporal representation can be used to quantify how much the role of a publication changed as well as interpret how the role changed over time, that the metric for quantifying the changes of articles' roles is stable over time at the population level, and there is significant individual variability to distinguish individuals.

Resolving Citation Links With Neural Networks, Nomoto

Paper shows how neural network models (NNs) can be utilized to resolve citation links in the scientific literature by locating the passages in the source paper that author intended to cite in the paper.

Two kinds of models: The triplet network model ranks potential candidates using the triplet loss, and the Binary model tackles the issue by labeling a candidate as true or false based on how likely a target it is.

Over the work finds that NNs perform better on inputs expressed in binary format than on those encoded using the TFIDF metric or neural embeddings. The authors introduce the idea of approximately correct targets (ACTs) where the goal is to find a region which likely contains a true target rather than its exact location, and show that it allows NNs to outperform Ranking SVM and TFIDF.

The NLP4NLP Corpus (I and II): 50 Years of Publication, Collaboration and Citation in Speech and Language Processing by Mariani et al. and Mariani et al.

A two paper study of a dataset involving Natural Language Processing (NLP) and Spoken Language Processing (SLP) for the period 1956–2015 with each paper providing a different perspective of the analysis.

The first paper presents an overall analysis of the number of papers, authors, gender distributions, co-authorship, collaboration patterns and citation patterns. In contrast, the second paper looks at the research topics, their evolution over time, the key innovative topics and the authors that introduced them, and the reuse of papers and plagiarism.

The most significant obstacles in this study were errors due to the automatic processing of the full text such as scanned content, and s the lack of a consistent identification of authors, affiliations, conference titles, etc. Both of these issues required manual corrections.

Overall, the two papers provide a survey of the literature in NLP and SLP and the data to understand the trends and the evolution of research in this research community as well as provide a methodological framework for producing similar surveys.

Editorial: Mining Scientific Papers, Volume II: Knowledge Discovery and Data Exploitation by Atanassova et. al.

The topic of mining scientific papers, and more broadly text mining methods used in the fields of NLP-enhanced Bibliometrics and knowledge discovery, generate much interest from the community.

The Research Topic discusses approaches that focus on the processing and exploitation of data extracted from scientific literature and the possibility to enrich metadata by the full-text processing of papers.

NLP4NLP+5: The Deep (R)evolution in Speech and Language Processing by Mariani et. al.

Continues the series of on the NLP4NLP corpus in the previous Research Topic with similar methods, but adds to the dataset 5 more years of publications between 2016 and 2020.

This addition accounts for new research in the field of speech and language processing during these years. The analysis shows a shift in research to novel topics such as Neural Networks and Word Embeddings. This combined with the acceleration of the publication process and the growth in language resources usage, account for some important transformations in this field of research as well as in the emergence of a new generation of authors and the appearance of new publications around artificial intelligence, neural networks, machine learning, and word embedding.

Language Bias in Health Research: External Factors That Influence Latent Language Patterns by Valdez and Goodson

Peer-reviewed research is facing unprecedented retraction rates for published work due to an ongoing replicability crisis. While the inability to replicate findings is generally assumed to be the fault of study data itself or the methods used to analyze it, the language used in scientific reporting can misrepresent research findings in a manner analogous to falsifying or incorrectly analyzing numeric data.

This study aimed to explore language framing and its effect on the presentation of scientific findings and frame language as an equally important contributor to problematic science.

This paper argues for the need to critically assess findings in the literature both linguistically and numerically and argues that readers should be better equipped to identify biasing factors through the use of novel tools and methodologies such as the topic modeling approach.

Large Scale Subject Category Classification of Scholarly Papers With Deep Attentive Neural Networks by Kandimalla et al.

The paper proposes a method for classifying scientific articles based on their abstracts. For this purpose, the authors propose to use a deep attentive neural network (DANN) trained on abstracts obtained from the Web of Science (WoS) and its categories.

The results obtained suggest that the new methods scale better than than existing clustering-based methods relying on citation networks with the combination of word vectors with TFIDF outperforms character and sentence level embedding models.

SYMBALS: A Systematic Review Methodology Blending Active Learning and Snowballing by van Haastrecht et al.

SYMBALS blends the traditional method of backward snowballing with the machine learning method of active learning with this method validated using a replication study with ASReview, where SYMBALS could accelerate the title and abstract screening. This lead to the expedition of the systematic review process, while at the same time making systematic reviews accessible with the approach allowing for researchers to accelerate title and abstract screening by a factor of 6.

Future research includes the completion of the full cybersecurity metric review case study, investigating which choices in the selection of active learning tools, classification models, and stopping criteria are optimal in which scenarios, and Optimising SYMBALS in these areas can certainly benefit researchers performing systematic reviews, although they should take care to not reduce the reproducibility of their results.

Enhancing Knowledge Graph Extraction and Validation From Scholarly Publications Using Bibliographic Metadata by Turki et al.

Opinion article exploring how bibliographic metadata can provide useful insights to enhance the automatic enrichment and fact-checking of knowledge graphs from scholarly publications.

Looks at Title, Abstract, Controlled Keywords, Citation Analysis, Section Titles, and other Metadata

Authors suggest for future work to explore how bibliographic metadata can enhance information retrieval algorithms, how bibliometric-enhanced information retrieval can enhance knowledge graph construction and validation, other interesting computational tasks such as predicting future scientific breakthroughs and major prize winners, natural language generation and translation of scholarly texts, the automation of the creation and update of various kinds of research outputs, the adaptation of BIR algorithms to support the augmentation of university-level courses, building a framework for explainable artificial intelligence that returns explanations of the use of machine learning models for a given task based on what is currently available about the matter in research papers.

Visual Summary Identification From Scientific Publications via Self-Supervised Learning by Yamamoto et al.

Constructs a novel benchmark data set for visual summary identification from scientific publications using a self-supervised learning approach to learn a heuristic matching of in-text references to figures with figure captions.

The results show that the self-supervised learning approach is effective for central figure identification without manually annotating data and outperforms the existing supervised approach indicated that model performances and attention patterns stay roughly consistent across the subdomains. However, qualitative analysis revealed that different types of figures are ranked higher in different subdomains.

Discussion questions

1. Do you consider the abstract of a paper as a summary/proxy for the content of an article or as a teaser of its contents? Do you think a GEM score that measures the representativeness of an abstract or its "generosity" like that in Ermakova et al. would be useful to you when researching?
2. What is the value of observing the bibliometric trends over time such in The NLP4NLP Corpus (I and II) an NLP4NLP+5 by Mariani et al?
3. Do you find the argument raised in Valdez and Goodson that language used in scientific reporting can misrepresent research findings in a manner analogous to falsifying or incorrectly analyzing numeric data to be compelling?