CIKM 2021 1-5 NOVEMBER

BEST APPLIED RESREACH AWARD

James Howison

Caifan Du

Johanna Cohoon

Mining Software Entities in Scientific Literature: Document-level NER for an Extremely Imbalance and Large-scale Task

Patrice Lopez
science-miner
France
patrice.lopez@science-miner.com

Caifan Du
Johanna Cohoon
University of Texas at Austin
USA

Karthik Ram
Berkeley Institute for Data Science
USA

James Howison
University of Texas at Austin
USA

ABSTRACT

We present a comprehensive information extraction system dedicated to software entities in scientific literature. This task combines the complexity of automatic reading of scientific documents (PDF processing, document structuring, styled/rich text, scaling) with challenges specific to mining software entities: high heterogeneity and extreme sparsity of mentions, document-level cross-references, disambiguation of noisy software mentions and poor portability of Machine Learning approaches between highly specialized domains. While NER is a key component to recognize new and unseen software, considering this task as a simple NER application fails to address most of these issues.

In this paper, we propose a multi-model Machine Learning approach where raw documents are ingested by a cascade of document structuring processes applied not to text, but to layout token elements. The cascading process further enriches the relevant structures of the document with a Deep Learning software mention recognizer adapted to the high sparsity of mentions. The Machine Learning cascade culminates with entity disambiguation to alleviate false positives and to provide software entity linking. A bibliographical reference resolution is integrated to the process for attaching references cited alongside the software mentions.

Based on the first gold-standard annotated dataset developed for software mentions, this work establishes a new reference end-to-end performance for this task. Experiments with the CORD-19 publications have further demonstrated that our system provides practically usable performance and is scalable to the whole scientific corpus, enabling novel applications for crediting research software and for better understanding the impact of software in science.

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**; • **Applied computing** → *Document analysis.*

CIKM '21, November 1–5, 2021, Virtual Event, Australia.

ACM ISBN 978-1-4503-8446-9/21/11.
https://doi.org/10.1145/3459637.3481936

KEYWORDS

Software; Scientific Literature; Entity Recognition; Entity Disambiguation; Document Analysis

ACM Reference Format:
Patrice Lopez, Caifan Du, Johanna Cohoon, Karthik Ram, and James Howison. 2021. Mining Software Entities in Scientific Literature: Document-level NER for an Extremely Imbalance and Large-scale Task. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21), November 1–5, 2021, Virtual Event, QLD, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3459637.3481936

1 INTRODUCTION

Software is increasingly used to support research activities. Many researchers include software with their results, as shown by services like rOpenSci [5] and Papers with Code[1]. Quality software facilitates scientific progress and research reproducibility. Software, however, is still relatively invisible in research, citation databases, and academic search engines. Because it is impossible to search for research software as easily as for articles, users miss relevant software and software is not easily found by potential users. In addition, because identifying and crediting contributions of software developers is difficult, researchers have less incentives to develop better and more reusable software. While some recent works have focused on improving software cataloging [11] and standards for software citation [15], mining software mentions in the scientific literature as currently published offers a factual approach, directly usable to increase research software visibility.

Named entity recognition (NER) in scholarly literature has been successfully applied across a range of entity types, including species names, chemical, and biomedical entities [6, 12, 13, 30]. Mining software entities has attracted a lot of interest in the recent years. A recent review article [16] offers a comprehensive analysis of prior works that extract software and data mentioned in research articles. From 48 reviewed studies, of which 15 cover software and packages, Krüger and Schindle categorize four extraction approaches: (1) term search, (2) manual extraction, (3) rule-based extraction, and (4) supervised learning.

Term search was employed in 12 studies, although only one study related to software extraction. Term search refers to searching bibliographical databases for known string identifiers, such as URL,

[1] https://paperswithcode.com