

Populating and Refining an Ontology of Cellulose Materials with Terms from Scientific Publications: Extended Abstract

Umayer Reza¹, Torsten Hahmann¹

¹*School of Computing and Information Science, University of Maine*

Abstract

Cellulose is a highly versatile biopolymer with numerous applications, such as paper and paperboard production, textiles, packaging, biofuels, and biomedical applications. Though, the scattered nature of cellulose knowledge with ambiguous terms and datasets presents significant obstacles to its optimal utilization. This project seeks to address these challenges by systematically accumulating scattered knowledge about cellulose, enabling it to be modifiable, extensible, and reusable. The objective of the project is to develop an automated system to extract relevant cellulosic terms from scientific publications which will show an improved performance in named entity classification by taking additional context and disambiguous information from an existing cellulose ontology. An incremental training process will be utilized to train a ScispaCy language model, which is specifically designed for analyzing scientific, clinical, and biomedical texts, in order to accomplish this task. The system will also generate new terms for the ontology by taking the existing ontology into account. Therefore, the proposed system will facilitate the extension of the ontology, while simultaneously benefiting from the ontology to enhance performance in named entity classification. By meeting these objectives, the project aims to contribute to the development of a sustainable bioproduct-based society by providing a resource of state-of-the-art knowledge in cellulose materials that can facilitate material science research.

Keywords

Named Entity Recognition, Cellulose Ontology, Knowledge Graph, Scientific Publication

1. Motivation

Cellulose, the most abundant and versatile biopolymer on earth found in plant cells and some bacteria, is the building block of cellulosic materials, which have many applications in various domains because of their sustainable, renewable, and biodegradable nature. One of the most significant applications of cellulosic materials is the production of paper and paperboard. The unique dimensions and characteristics of cellulose nanofibrils (CNFs) make them crucial in papermaking for enhancing the strength properties of paper [1]. In addition to their use in paper production, several nanocelluloses (NCs) are alternatives for the textile industry because of their higher mechanical resistance [2], and as a substitute for petroleum-based packaging [3],


FOIS 2023 Early Career Symposium (ECS), held at FOIS 2023, co-located with 9th Joint Ontology Workshops (JOWO 2023), 19-20 July, 2023, Sherbrooke, Québec, Canada

✉ a.reza@maine.edu (U. Reza); torsten.hahmann@maine.edu (T. Hahmann)

🌐 <https://umaine.edu/scis/people/rezaumayer> (U. Reza); <https://umaine.edu/scis/people/torsten-hahmann> (T. Hahmann)

🆔 0000-0003-4013-3513 (U. Reza); 0000-0002-5331-5052 (T. Hahmann)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

and as a natural polymer with low toxicity, high crystallinity, biocompatibility, and biosafety for biomedical applications [4].

However, the knowledge on cellulose, much of which is stored only in scientific publication and, even when available in digital formats like PDF or HTML, is not easily processed at a large scale due to the scattered and ambiguous nature of text. Information extraction (IE) approaches are needed to extract this information from scientific publications and make it accessible in structured formats. By putting the information into an ontology, the knowledge can be subsequently queried and reasoned with more efficiently. One of the key steps in IE is NER, or Named Entity Recognition, which is the task of extracting nouns and noun chunks, called entities or named entities, from text. Identification and classification of named entities is the key part of the information extraction process. Though, the current NER approaches are limited in their ability to recognize cellulosic named entities and to handle variations in the naming conventions of them. ChemSpot [5] is a hybrid chemical named entity recognizer that utilized a CRF (Conditional Random Field) model to identify chemical named entities in natural language texts. In the biomedical domain, tmChem [6] employed a model combination approach using two different CRF models to recognize chemical mentions, properties, and their relationships. Akkasi et al. (2016) introduced ChemTok [7], a rule-based tokenizer specifically designed for chemical named entity recognition. Swain et al. (2016) presented ChemDataExtractor [8], a toolkit capable of extracting chemical entities along with their properties, measurements, and relationships. Corbett and Boyle (2018) developed Chemlistem [9], a chemical named entity recognizer based on recurrent neural networks. Although there are other methods available for extracting material entities from text, they also struggle in recognizing the diverse range of entities encountered in the cellulosic domain. Zhao et al. (2021) introduced a fine-tuned BERT model [10] specifically designed for materials named entity recognition. Similarly, Miah and Sulaiman (2023) proposed a deep neural network-based model [11] tailored for materials named entity recognition. Shetty et al. (2023) presented an alternative approach [12] for extracting material property data. Furthermore, Weston et al. (2019) presented a comprehensive approach [13] that not only extracts material properties but also captures their applications and mentions of inorganic materials.

In order to comprehensively capture cellulosic knowledge, it is crucial to extract a wide range of relevant entities beyond just chemicals and materials. This includes extracting properties associated with materials and chemicals, manufacturing processes, as well as names of products and equipment. Therefore, existing methods often fail to accurately identify a significant portion of cellulosic entities due to their limited familiarity with cellulosic data. The ultimate goal of the project is to contribute in growing an ontology-guided knowledge body about cellulose by extracting relevant terms from the scientific literatures which are the preferred source of knowledge. Initially, a manually made cellulose ontology will play a significant role in NER by providing a structured representation of the knowledge and relationships among entities. It can also assist to improve NER performance by providing additional context and disambiguous information, as well as enabling more sophisticated reasoning and inference. Later on the cellulose ontology itself will grow by adding newly identified cellulosic entities to speed up ontology development. The manual amendment of ontologies can be a time-consuming and costly process which limits their usefulness in practice. In contrast, an automated process can help to overcome these limitations and enable more efficient and effective use of ontologies in

NER and other NLP tasks. Additionally, the automatic amendment of ontologies can help to ensure that they are up-to-date and reflect the latest developments in the domain, but there is little work in leveraging the synergies between NER and ontologies: in (1) utilizing ontologies for NER and (2) using NER to amend and populate ontologies.

In scientific domains where accurate organization of terms is important to avoid misrepresentation of knowledge, relying solely on an automatic ontology construction method that builds the ontology from scratch may not be effective. Instead, a semi-automated method, where domain experts contribute initial concepts and relationships to establish a core domain ontology, can be utilized to amend the ontology with additional terms. The purpose of incorporating an ontology in the named entity recognition process is to improve its performance in the cellulosic domain. This integration will enable the system to identify the named entities that align with the concepts and relationships of the ontology and classify them accordingly.

2. Research Questions

The proposed dissertation aims to leverage the synergies between ontologies and NER by specifically addressing the following three research questions:

1. Under what conditions can pre-trained language models be incrementally re-trained for improved named entity recognition of terms in the cellulosic domain?
2. How can cellulose-related terms that are identified by such NER approaches be categorized more effectively and precisely by leveraging a small hand-curated ontology of cellulose materials?
3. What methods are suitable for determining whether a particular identified term refers to a concept that already exists in the ontology, or a new concept that requires amending the ontology?

3. Objective

The objective of the project is to develop an automated system that will identify cellulosic terms from given text with a higher accuracy. Additionally, the system aims to classify these identified terms, as much as possible, with the most relevant concepts available in a cellulose ontology which is currently being developed. To accomplish this, the proposed system will establish internal communication with the ontology, enabling the identification and classification of new cellulosic terms that are not yet incorporated in the ontology. The resulting set of new terms will be shared with domain experts for review, allowing them to assess the relevance of those terms and determine their appropriate placement within the taxonomy. If a new term is found to carry a more refined semantic meaning than an existing term in the ontology, the ontology will be amended accordingly. Furthermore, the identification and exclusion of irrelevant terms in every NER process will accelerate the continual assessment process for recognized cellulosic terms and their association with the ontology over time.

4. Research Methodology

A ScispaCy [14] language model will be selected considering its current performance on the cellulosic data. This performance will be measured based on the number of correctly recognized cellulosic terms from a curated corpus of evaluation data. The best performing model will go through an incremental training process using spaCy [15] model training pipeline. In the first phase of training, the selected model will be introduced with CHEMDNER corpus [16] which is currently the largest corpus for chemical terms. Then the model will undergo training using various corpus of training data, including materials, properties, processes, and more, across different training phases. This iterative process will enable the model to progressively enhance its understanding and knowledge on diverse terms of the cellulosic domain. After each phase of training the performance of the model will be evaluated using standard metrics such as precision, recall, and F1-score. To assess the improvement, a comparison will also be conducted between the performance of the model in the current training phase and the performance of the models in the previous training phases using an identical evaluation dataset.

Finally, the improved language model will be employed to extract cellulosic named entities from a vast collection of text documents. The extracted cellulosic terms will be comprehensively compared to the existing terms in the ontology, generating a set of candidate terms to be forwarded for further verification by domain experts. The domain experts will assess the suitability of these terms and make informed decisions regarding their rejection or integration into the ontology. This collaborative process will ensure that the ontology is enriched with relevant and accurate terms, enhancing its overall effectiveness and comprehensiveness. The effectiveness of the approach will be measured by calculating the percentage of terms accepted by domain experts. This metric will provide valuable insights into the success and acceptance of the proposed method in enriching the ontology with relevant and authoritative terms.

Acknowledgments

This research was supported in part by the U.S. Department of Agriculture:

- Forest Service, Project 20-JV-11111124-055
- USDA Agricultural Research Service (ARS), Project 0204-41510-001-98S
- National Institute of Food and Agriculture (NIFA), Award 2021-67022-34366

References

- [1] T. B. Jele, P. Lekha, B. B. Sithole, Role of cellulose nanofibrils in improving the strength properties of paper: a review, *Cellulose* 29 (2021) 55–81. doi:10.1007/s10570-021-04294-8.
- [2] C. Felgueiras, N. G. Azoia, C. M. A. Gonçalves, M. Gama, F. Dourado, Trends on the cellulose-based textiles: Raw materials and technologies, *Frontiers in Bioengineering and Biotechnology* 9 (2021). doi:10.3389/fbioe.2021.608826.
- [3] Y. Su, B. Yang, J. Liu, B. Sun, C. Cao, X. Zou, R. Lutes, Z. He, Prospects for replacement of some plastics in packaging with lignocellulose materials: A brief review, *Bioresources* 13 (2018) 4550–4576. doi:10.15376/biores.13.2.Su.

- [4] S. Gopi, P. Balakrishnan, D. Chandradhara, D. Poovathankandy, S. Thomas, General scenarios of cellulose and its use in the biomedical field, *Materials Today Chemistry* 13 (2019) 59–78. doi:10.1016/j.mtchem.2019.04.012.
- [5] T. Rocktäschel, M. Weidlich, U. Leser, Chemspot: a hybrid system for chemical named entity recognition, *Bioinformatics* 28 (2012) 1633–1640. doi:10.1093/bioinformatics/bts183.
- [6] R. Leaman, C.-H. Wei, Z. Lu, tmchem: a high performance approach for chemical named entity recognition and normalization, *Journal of Cheminformatics* 7 (2015) S3. doi:10.1186/1758-2946-7-S1-S3.
- [7] A. Akkasi, E. Varoğlu, N. Dimililer, Chemtok: A new rule based tokenizer for chemical named entity recognition, *BioMed Research International* 2016 (2016) 1–9. doi:10.1155/2016/4248026.
- [8] M. C. Swain, J. M. Cole, Chemdataextractor: A toolkit for automated extraction of chemical information from the scientific literature, *Journal of Chemical Information and Modeling* 56 (2016) 1894–1904. doi:10.1021/acs.jcim.6b00207.
- [9] P. T. Corbett, J. Boyle, Chemlistem: chemical named entity recognition using recurrent neural networks, *Journal of Cheminformatics* 10 (2018) 59. doi:10.1186/s13321-018-0313-8.
- [10] X. Zhao, J. Greenberg, Y. An, X. T. Hu, Fine-tuning bert model for materials named entity recognition, in: *2021 IEEE International Conference on Big Data (Big Data)*, IEEE, 2021, pp. 3717–3720. doi:10.1109/BigData52589.2021.9671697.
- [11] M. S. U. Miah, J. Sulaiman, Material named entity recognition (mner) for knowledge-driven materials using deep learning approach, in: M. S. Kaiser, S. Waheed, A. Bandyopadhyay, M. Mahmud, K. Ray (Eds.), *Proceedings of the Fourth International Conference on Trends in Computational and Cognitive Engineering*, Springer Nature Singapore, Singapore, 2023, pp. 199–208. doi:10.1007/978-981-19-9483-8_17.
- [12] P. Shetty, A. C. Rajan, C. Kuenneth, S. Gupta, L. P. Panchumarti, L. Holm, C. Zhang, R. Ramprasad, A general-purpose material property data extraction pipeline from large polymer corpora using natural language processing, *npj Computational Materials* 9 (2023) 52. doi:10.1038/s41524-023-01003-w.
- [13] L. Weston, V. Tshitoyan, J. Dagdelen, O. V. Kononova, A. Trewartha, K. A. Persson, G. Ceder, A. Jain, Named entity recognition and normalization applied to large-scale information extraction from the materials science literature, *Journal of chemical information and modeling* 59 (2019) 3692–3702. doi:10.1021/acs.jcim.9b00470.
- [14] M. Neumann, D. King, I. Beltagy, W. Ammar, Scispacy: Fast and robust models for biomedical natural language processing, in: *Proceedings of the 18th BioNLP Workshop and Shared Task*, Association for Computational Linguistics, Florence, Italy, 2019, pp. 319–327. doi:10.18653/v1/W19-5034.
- [15] M. Honnibal, I. Montani, S. V. Landeghem, A. Boyd, *spacy: Industrial-strength natural language processing in python* (2020). doi:10.5281/zenodo.1212303.
- [16] M. Krallinger, O. Rabal, F. Leitner, M. Vázquez, D. Salgado, Z. Lu, R. Leaman, Y. Lu, D.-H. Ji, D. M. Lowe, R. A. Sayle, R. T. Batista-Navarro, R. Rak, T. Huber, T. Rocktäschel, S. Matos, D. Campos, B. Tang, H. Xu, T. Munkhdalai, K. H. Ryu, S. V. Ramanan, P. S. Nathan, S. Žitnik, M. Bajec, L. Weber, M. Irmer, S. A. Akhondi, J. A. Kors, S. Xu, X. An, U. K. Sikdar, A. Ekbal, M. Yoshioka, T. M. Dieb, M. Choi, K. M. Verspoor, M. Khabsa, C. L. Giles, H. Liu, K. E. Ravikumar, A. Lamurias, F. M. Couto, H.-J. Dai, R. T.-H. Tsai, C. Ata, T. Can, A. Usie,

R. Alves, I. Segura-Bedmar, P. Martínez, J. Oyarzábal, A. Valencia, The chemdner corpus of chemicals and drugs and its annotation principles, *Journal of Cheminformatics* 7 (2015) S2. doi:10.1186/1758-2946-7-S1-S2.