

Improving Natural Product Automatic Extraction With Named Entity Recognition

Stefan Schmidt-Dichte^{1,*}, István J. Mócsy¹

¹AKSW, Leipzig University of Applied Sciences (HTWK), Gustav-Freytag-Straße 42a, Leipzig, 04277, Germany

Abstract

Knowledge graphs (KGs) play a vital role in providing structured data for various applications, but their creation is time-consuming and prone to errors. To address these challenges, automatic knowledge extraction methods using machine learning (ML) have gained attention. ML algorithms have shown promise in capturing subtle nuances in language data, offering comprehensive and robust solutions. In the field of biochemistry, knowledge extraction is crucial for advancing scientific research, product development, and policy-making. The First International Biochemical Knowledge Extraction Challenge focuses on extracting biochemical knowledge from scientific articles. This paper presents an updated approach that incorporates named entity recognition (NER) using scispaCy models to improve the accuracy and relevance of extracted entities. The evaluation of the approach utilizes the NatUKE benchmark and demonstrates improved performance in extracting bioactivity and isolation type. However, challenges remain in identifying compound names and species. Future research may explore hybrid approaches combining different techniques to address these specific challenges.

Keywords

Knowledge Extraction, NLP Pipelines, Natural Products, Knowledge Graphs, CEUR-WS

1. Introduction

Knowledge graphs (KGs) are important sources of structured data for various applications [1]. However, creating KGs can be time-consuming and prone to errors and incompleteness [2]. It involves complex natural language processing techniques and keeping the dataset updated is challenging [3]. Developing automatic knowledge extraction methods is crucial for easier KG curation and maintenance.

In recent times, there has been considerable progress in the field of machine learning (ML), particularly in its application to natural language tasks. ML methods have emerged as a promising approach, showcasing impressive results in various language-related endeavors. What sets ML apart from traditional, rule-based approaches, which are often created by humans based on limited data observations, is its ability to effectively capture and handle subtle intricacies within the data. ML algorithms can uncover nuances that might otherwise go unnoticed, offering a more comprehensive and robust solution to language challenges [4].


BiKE'23: First International Biochemical Knowledge Extraction Challenge, May 28 - Jun 1, 2023, co-located with Extended Semantic Web Conference (ESWC), Hersonissos, Greece

*Corresponding author.

✉ stefan.schmidt-dichte@stud.htwk-leipzig.de (S. Schmidt-Dichte); istvan.mocsy@stud.htwk-leipzig.de (I. J. Mócsy)



© 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)

When we talk about natural products, we are referring to chemical compounds that are produced by living organisms [5]. Natural products have had a substantial impact on the field of pharmacotherapy throughout history, particularly in the treatment of cancer and infectious diseases [6].

To make biodiversity knowledge more accessible, organizing and sharing it through knowledge graphs can aid scientific advancement, bio-friendly product development, and informed public policies. The Semantic Web architecture can facilitate this process, enabling researchers and institutions to publish, share, and utilize valuable information effectively. Automating the extraction of biochemical knowledge from scientific articles can accelerate research and increase awareness of biodiversity's significance, leading to the development of environmentally conscious products [7].

The First International Biochemical Knowledge Extraction Challenge aims to tackle the problem of extracting biochemical knowledge. To assess the progress in this area, the challenge utilizes a benchmark introduced in the NatUKE paper [8]. This work goes beyond NatUKE by incorporating Named Entity Recognition (NER) to enhance the results. Integrating NER into the existing framework has successfully achieved improved performance and accuracy in extracting biochemical knowledge. This extension demonstrates the commitment to advancing the field and finding effective solutions to the challenges of knowledge extraction in biochemistry.

2. Related Work

Automatic extraction of knowledge from text has been a topic of significant research interest in recent years. Several approaches have been proposed to tackle this task, aiming to automate the process of extracting valuable information from unstructured text data. The proposed PLUMBER [9] framework integrates disjoint Knowledge Graph (KG) information extraction efforts, offering reusable components and dynamically generated pipelines for effective KG triple extraction, outperforming baselines and providing analysis of failure cases, component similarities, and limitations. The t2kg framework [10] is designed to extract knowledge graphs from text, enabling the representation of structured information from unstructured sources. Furthermore, kgbert [11] utilizes a pre-trained BERT model to perform automatic extraction of knowledge from text. Additionally, seq2RDF [12] is a method that combines sequence labeling with semantic parsing techniques to extract structured information in the form of RDF triples. NatUKE [8] introduces a benchmark for natural product knowledge extraction from academic literature and evaluates unsupervised embedding methods.

Named entity recognition plays a crucial role in information extraction by identifying and classifying named entities within text. A comprehensive survey on deep learning approaches for NER is presented in [13], which explores various neural network architectures and techniques employed for effective entity recognition. Additionally, scispaCy [14] is a popular library specifically designed for biomedical NER, providing pre-trained models and tools to extract entities from scientific texts.

Knowledge graph embeddings have gained significant attention for representing

structured knowledge in a vector space. Several embedding models have been proposed to capture the semantic relationships between entities and relations in knowledge graphs. These models include TransE [15], TransH [16], TransD [17], TransR [18], DistMult [19], ComplEx [20], RotatE [21], TuckER [22], and Ephen [23]. Each of these models leverages different techniques, such as translation-based, projection-based, or tensor-based approaches, to embed entities and relations into a low-dimensional vector space.

3. Approach

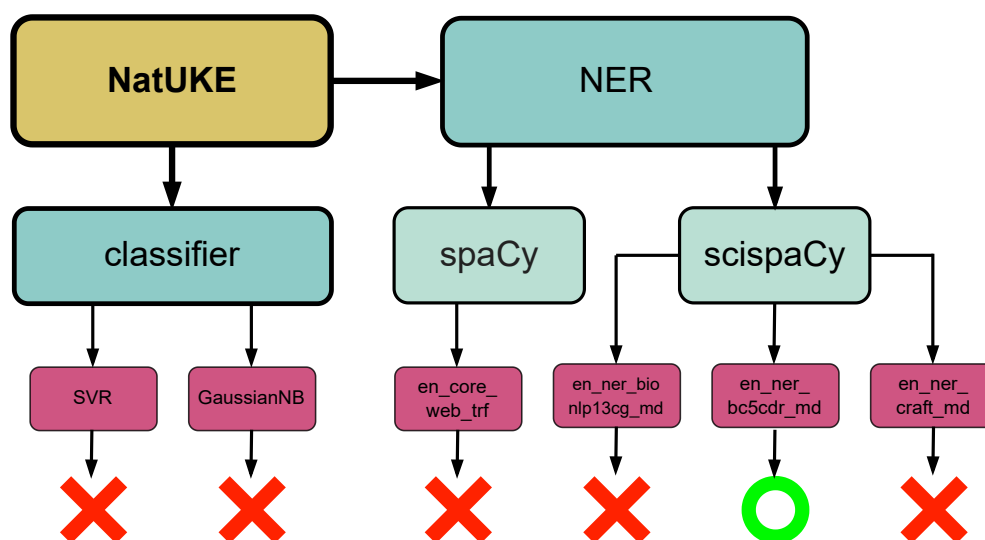


Figure 1: This graphic visualizes our journey of experimentation and highlights the successful ideas we eventually implemented. While exploring the idea of incorporating a classifier, we experimented with SVR and GaussianNB, but unfortunately, they did not yield the desired outcomes. In our pursuit of extracting locations, we tested the spaCy transformer model `en_core_web_trf`, but the results were unsatisfactory. Similarly, we explored scispaCy models such as `en_ner_bio_nlp13cg_md` and `en_ner_craft_md`, but they also produced unfavorable results. However, we finally achieved positive outcomes when utilizing the scispaCy model `en_ner_craft_md`, marking a breakthrough in our efforts.

In our research endeavor, we undertook the task of refining the existing NatUKE approach by introducing modifications to several of its components. NatUKE, in its original form, employed K-nearest neighbors (KNN) to sort by embedding similarity in an efficient way. However, we sought to enhance its performance by transforming it into a pair-to-pair binary classification approach, utilizing Support Vector Regression and Gaussian Naive Bayes models. Our experiments with these modifications yielded poor results, suggesting that the original KNN might not be the bottleneck of NatUKE’s performance.

The next idea we are exploring is to incorporate named entity recognition (NER) as an essential

component. Previously, the original approach relied on slicing phrases at the beginning of the pipeline according to the token size limit of 512 from DistilBERT [24]. Since slicing by token limit might separate and add irrelevant information we decided to leverage NER for the slicing process. By utilizing NER, we aim to identify and extract meaningful entities within the text. When an entity is found, the corresponding sentence containing that entity is extracted for further analysis. This approach not only ensures that important information is captured but also helps in maintaining the context of the identified entities. To accomplish NER, we opted to employ scispaCy. While scispaCy offers various NER models, we encountered suboptimal results when using SpaCy for location identification. Therefore, we embarked on experimenting with different built-in models. We discovered that the `en_ner_bc5cdr_md` model yielded the most promising results. This particular model, which is specifically designed for biomedical entity recognition, proved to be effective in identifying entities related to the biomedical domain [25]. By leveraging the strengths of `en_ner_bc5cdr_md`, we were able to enhance the accuracy and relevance of the extracted entities, thereby improving the overall performance of our pipeline.

In summary, our updated approach replaces the initial slicing of phrases with NER, allowing us to extract sentences based on identified entities. The utilization of scispaCy, particularly the `en_ner_bc5cdr_md` model, has proven to be valuable in achieving accurate and meaningful entity recognition within the context of our pipeline.

4. Evaluation

The evaluation was performed using the official BiKE challenge benchmark NatUKE [8]. It focuses on three aspects: (A) using the NuBBE_{DB} dataset to extract characteristics from papers, (B) obtaining knowledge extraction results through KG completion using graph embedding models, and (C) comparing the behavior of four different graph embedding models. The dataset used for evaluation and training was manually built from peer-reviewed scientific articles, and it includes information on various properties of natural products. The problem of knowledge extraction from unstructured data sources is addressed, and machine learning graph embeddings is proposed. The BERTopic model is employed to extract topics from papers, and the paper’s DOI is used as a central node in a knowledge graph [26]. The evaluation includes different stages with varying amounts of training data, and the accuracy of each approach is measured using hits@k metric. There are four graph embedding methods (DeepWalk [27], Node2Vec [28], Metapath2Vec [29], and EPHEN[23]) used for knowledge extraction. The goal is to create embeddings for each node in the graph without requiring a complete knowledge graph, predetermined weights, or ontology, and to ensure that the models can generate embeddings within a reasonable amount of time and computational resources.

Hits@k enables us to assess each feature extraction method based on our specific expectations by adjusting the value of k. In accordance with the methodology employed in NatUKE, the final k values ranging from 1 to 50, where only multiples of 5 are considered. Two thresholds are taken into account. First when a score of 0.50 or higher is attained, and second when a score of 0.20 or higher is attained.

The results of the original NatUKE pipeline are presented in table 1. For more comprehensive information, please consult the NatUKE benchmark paper.

Table 1

NatUKE results table for extracting: compound name (C), bioactivity (B), specie (S), collection site (L), and isolation type (T). The results consider different final k values corresponding to two different rules. The best results for each extraction are bold.

Property	k	DeepWalk				Node2Vec				Metapath2Vec				EPHEN			
		1st	2nd	3rd	4th	1st	2nd	3rd	4th	1st	2nd	3rd	4th	1st	2nd	3rd	4th
C	50	0.08	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.10	0.08	0.09	0.20	0.09	0.02	0.03	0.04
B	5	0.41	0.12	0.10	0.07	0.41	0.07	0.03	0.03	0.27	0.17	0.13	0.12	0.55	0.57	0.60	0.64
S	50	0.37	0.24	0.27	0.25	0.36	0.22	0.25	0.24	0.40	0.41	0.42	0.44	0.36	0.24	0.29	0.30
L	20	0.56	0.41	0.38	0.29	0.57	0.36	0.28	0.23	0.40	0.42	0.42	0.40	0.53	0.52	0.55	0.55
T	1	0.25	0.14	0.14	0.09	0.10	0.07	0.05	0.01	0.28	0.22	0.19	0.19	0.71	0.66	0.75	0.75

5. Results

The outcomes of extracting five distinct natural product characteristics from biochemical academic papers are presented in Table 2. We utilize the NuBBE[KG] ontology and dataset to make predictions about properties¹. The selection of values for k , which determines the difficulty level in property-value prediction, varies proportionally. For example, it is more challenging to accurately predict the name of a natural product compared to predicting the type of isolation. There are significantly fewer distinct possible characteristics for isolation type compared to compound name. Consequently, predicting the correct compound name poses a significantly greater challenge.

Overall, EPHEN demonstrates the highest performance in extracting bioactivity and isolation type, progressively improving accuracy across the evaluation stages. For instance, in the first evaluation stage, EPHEN achieves a hits@5 score of 0.60, which steadily increases to 0.69 in the fourth evaluation stage. In contrast, DeepWalk obtains the best results in the first evaluation stage with a score of 0.39 but experiences a decline, reaching the second-worst results of 0.07 in the fourth evaluation stage.

Comparing NatUKE with NatUKE + NER in Table 1 and 2 reveals notable differences. The inclusion of NER in NatUKE has led to significant improvements in the EPHEN model concerning bioactivity, collection site, and isolation type. However, it is worth mentioning that there was no improvement observed in the identification of compound names and species for EPHEN when using NatUKE + NER. Furthermore, when considering the utilization of NatUKE + NER in conjunction with DeepWalk, Node2Vec, and Metapath2Vec, no significant enhancements were observed, indicating that these embedding methods did not benefit considerably from the inclusion of NER in the model.

6. Conclusion & Future Works

The utilization of NER within a pipeline has shown potential for improving results. However, it is important to note that NER alone may not be sufficient for enhancing the recognition of compound names and species, as these areas have presented significant challenges. Therefore, it is necessary to explore alternative methods to address these specific issues.

¹<https://nubbekg.aksw.org>

Table 2

NatUKE + NER results table for extracting: compound name (C), bioactivity (B), specie (S), collection site (L), and isolation type (T). The results consider different final k values corresponding to two different rules. The best results for each extraction are bold.

Property	k	DeepWalk + NER				Node2Vec + NER				Metapath2Vec + NER				EPHEN + NER			
		1st	2nd	3rd	4th	1st	2nd	3rd	4th	1st	2nd	3rd	4th	1st	2nd	3rd	4th
C	50	0.09	0.00	0.00	0.00	0.09	0.00	0.00	0.00	0.10	0.08	0.09	0.17	0.09	0.00	0.00	0.00
B	5	0.39	0.10	0.05	0.07	0.38	0.07	0.05	0.03	0.26	0.15	0.13	0.15	0.60	0.65	0.65	0.69
S	50	0.37	0.24	0.25	0.26	0.40	0.23	0.25	0.20	0.40	0.42	0.43	0.43	0.34	0.23	0.27	0.29
L	20	0.54	0.40	0.32	0.27	0.56	0.36	0.26	0.22	0.40	0.36	0.37	0.32	0.55	0.58	0.60	0.57
T	1	0.15	0.12	0.12	0.13	0.10	0.06	0.08	0.07	0.26	0.20	0.18	0.20	0.73	0.74	0.77	0.76

One suggested approach for future research is to adopt a hybrid approach, combining different techniques or models to leverage their respective strengths. This could involve integrating NER with other strategies or technologies to create a more comprehensive and robust pipeline. Such an approach has the potential to yield better outcomes by capitalizing on the strengths of different methods.

Additionally, the use of Large Language Models (LLMs) has demonstrated promising results in other related work. Considering this, it would be intriguing to explore how LLMs can be effectively integrated into the pipeline. This integration could enhance the overall performance by leveraging the contextual understanding and language capabilities of LLMs.

In summary, while NER has shown potential in improving pipeline results, addressing the challenges related to the compound name and species recognition requires further exploration. A hybrid approach that combines different methods and incorporates LLMs into the pipeline could be a worthwhile avenue for future investigation. By continuously refining and evolving the pipeline, we can strive for improved accuracy and efficiency in recognizing and extracting relevant information. In future work, we plan to explore extraction on multilingual [30] and also evaluate the models' information bias [31].

References

- [1] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. d Melo, Knowledge graphs [J]. Synthesis Lectures on Data Semantics and Knowledge (2021).
- [2] A. Zaveri, D. Kontokostas, M. A. Sherif, L. Böhmann, M. Morsey, S. Auer, J. Lehmann, User-Driven Quality Evaluation of DBpedia, in: Proceedings of the 9th International Conference on Semantic Systems, I-SEMANTICS '13, Association for Computing Machinery, New York, NY, USA, 2013, p. 97–104. URL: <https://doi.org/10.1145/2506182.2506195>. doi:10.1145/2506182.2506195.
- [3] S. Hellmann, C. Stadler, J. Lehmann, S. Auer, DBpedia live extraction, in: On the Move to Meaningful Internet Systems: OTM 2009: Confederated International Conferences, CoopIS, DOA, IS, and ODBASE 2009, Vilamoura, Portugal, November 1-6, 2009, Proceedings, Part II, Springer, 2009, pp. 1209–1223.
- [4] C. C. Aggarwal, Machine learning for text, volume 848, Springer, 2018.

- [5] D. J. Newman, G. M. Cragg, Natural products as sources of new drugs from 1981 to 2014, *Journal of natural products* 79 (2016) 629–661.
- [6] A. G. Atanasov, S. B. Zotchev, V. M. Dirsch, C. T. Supuran, Natural products in drug discovery: advances and opportunities, *Nature reviews Drug discovery* 20 (2021) 200–216.
- [7] BiKE: First International Biochemical Knowledge Extraction Challenge, 2023. URL: <https://aksw.github.io/bike/>.
- [8] P. V. Do Carmo, E. Marx, R. Marcacini, M. Valli, J. V. Silva e Silva, A. Pilon, NatUKE: A Benchmark for Natural Product Knowledge Extraction from Academic Literature, in: 2023 IEEE 17th International Conference on Semantic Computing (ICSC), 2023, pp. 199–203. doi:10.1109/ICSC56153.2023.00039.
- [9] M. Y. Jaradeh, K. Singh, M. Stocker, A. Both, S. Auer, Better Call the Plumber: Orchestrating Dynamic Information Extraction Pipelines, in: M. Brambilla, R. Chbeir, F. Frasincar, I. Manolescu (Eds.), *Web Engineering*, Springer International Publishing, Cham, 2021, pp. 240–254.
- [10] N. Kertkeidkachorn, R. Ichise, T2kg: An end-to-end system for creating knowledge graph from unstructured text, in: *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- [11] L. Yao, C. Mao, Y. Luo, KG-BERT: BERT for Knowledge Graph Completion, 2019. arXiv:1909.03193.
- [12] Y. Liu, T. Zhang, Z. Liang, H. Ji, D. L. McGuinness, Seq2RDF: An end-to-end application for deriving Triples from Natural Language Text, 2018. arXiv:1807.01763.
- [13] D. Nadeau, S. Sekine, A survey of named entity recognition and classification, *Linguisticae Investigationes* 30 (2007) 3–26.
- [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, 2019. URL: <https://doi.org/10.18653/v1/w19-5034>. doi:10.18653/v1/w19-5034.
- [15] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, *Advances in neural information processing systems* 26 (2013).
- [16] Z. Wang, J. Zhang, J. Feng, Z. Chen, Knowledge graph embedding by translating on hyperplanes, in: *Proceedings of the AAAI conference on artificial intelligence*, volume 28, 2014.
- [17] G. Ji, S. He, L. Xu, K. Liu, J. Zhao, Knowledge graph embedding via dynamic mapping matrix, in: *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)*, 2015, pp. 687–696.
- [18] Y. Lin, Z. Liu, M. Sun, Y. Liu, X. Zhu, Learning entity and relation embeddings for knowledge graph completion, in: *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.
- [19] B. Yang, W.-t. Yih, X. He, J. Gao, L. Deng, Embedding entities and relations for learning and inference in knowledge bases, arXiv preprint arXiv:1412.6575 (2014).
- [20] T. Trouillon, J. Welbl, S. Riedel, É. Gaussier, G. Bouchard, Complex embeddings for simple link prediction, in: *International conference on machine learning*, PMLR, 2016, pp.

2071–2080.

- [21] Z. Sun, Z.-H. Deng, J.-Y. Nie, J. Tang, Rotate: Knowledge graph embedding by relational rotation in complex space, arXiv preprint arXiv:1902.10197 (2019).
- [22] I. Balažević, C. Allen, T. M. Hospedales, Tucker: Tensor factorization for knowledge graph completion, arXiv preprint arXiv:1901.09590 (2019).
- [23] P. do Carmo, R. Marcacini, Embedding propagation over heterogeneous event networks for link prediction, in: 2021 IEEE International Conference on Big Data (Big Data), 2021, pp. 4812–4821. doi:10.1109/BigData52589.2021.9671645.
- [24] V. Sanh, L. Debut, J. Chaumond, T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108 (2019).
- [25] J. Li, Y. Sun, R. J. Johnson, D. Sciaky, C.-H. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wieggers, Z. Lu, BioCreative V CDR task corpus: a resource for chemical disease relation extraction, Database 2016 (2016). URL: <https://doi.org/10.1093/database/baw068>. doi:10.1093/database/baw068. arXiv:<https://academic.oup.com/database/article-pdf/doi/10.1093/database/baw068/8224483/baw068.pdf>, baw068.
- [26] M. Grootendorst, BERTopic: Neural topic modeling with a class-based TF-IDF procedure, arXiv preprint arXiv:2203.05794 (2022).
- [27] B. Perozzi, R. Al-Rfou, S. Skiena, Deepwalk: Online learning of social representations, in: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014, pp. 701–710.
- [28] A. Grover, J. Leskovec, node2vec: Scalable feature learning for networks, in: Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, 2016, pp. 855–864.
- [29] Y. Dong, N. V. Chawla, A. Swami, metapath2vec: Scalable representation learning for heterogeneous networks, in: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, 2017, pp. 135–144.
- [30] L. Hinguruduwa, E. Marx, T. Soru, T. Riechert, Assessing Language Identification over DBpedia, in: 2021 IEEE 15th International Conference on Semantic Computing (ICSC), 2021, pp. 296–297. doi:10.1109/ICSC50631.2021.00084.
- [31] E. Marx, Assessing Bias on Entity Retrieval Models through Conjunctive Fallacies, in: 2023 IEEE 17th International Conference on Semantic Computing (ICSC), IEEE, 2023, pp. 260–261. doi:10.1109/ICSC56153.2023.00050.