NEARSIDE: Structured kNowledge Extraction frAmework from Species DEscriptions

Maya Sahraoui[‡], Marc Pignal[§], Régine Vignes Lebbe[|], Vincent Guigue[‡]

‡ ISIR, Paris, France
§ MNHN, Paris, France
| Institut de Systématique, Evolution, Biodiversité (ISYEB), Muséum national d'Histoire naturelle, CNRS, Sorbonne Université, EPHE, Université des Antilles, Paris, France

Corresponding author: Maya Sahraoui (sahraoui@isir.upmc.fr)

Abstract

Species descriptions are stored in textual form in corpora such as in floras and faunas, but this large amount of information cannot be used directly by algorithms, nor can it be linked to other data sources. The production of knowledge bases expressing structured data can benefit from collaborative and easy-to-use platforms like Xper3 (Vignes-Lebbe et al. 2017, Kerner and Vignes 2019, Saucède et al. 2021) but is very time-consuming at the human level. It is therefore mandatory for this task to make the information contained in species descriptions measurable and compatible with computer techniques.

One of the most used data structures on the web and by the deep learning community is the triplet structure. Each piece of information is represented by a set of 3 elements (subject, predicate, object). One of the first steps towards species information accessibility is developing a text-to-triplet model, also known as text-to-graph, for monograph descriptions.

In this work, we developed NEARSIDE, a text-to-graph model adapted to biology corpora to create normalized morphological characteristic knowledge bases for species descriptions.

In Natural Language Processing, deep learning models have proven to be effective in extracting knowledge from open domain corpora (Lample et al. 2016, Sutskever et al. 2014), especially since the emergence of attention-based models (Devlin et al. 2019b, Devlin et al. 2019a). Several works have been made also on biomedical corpora (Fries et al. 2017,Cho and Lee 2019). In our case, we propose a model adapted to floras.

Fully supervised deep learning models require a large amount of annotated data for training, nevertheless, the annotation process for the text-to-triplet task implies an expensive human intervention. Distant supervision is a technique that can be used to reduce this cost. This paradigm uses a small annotated glossary to project classes at the word level on a new complex and longer text (see Fig. 1).

Named Entity Recognition (NER) is an Natural Language Processing (NLP) task that consists of extracting and classifying words of interest from a text (Sutskever et al. 2014, Devlin et al. 2019b, Lample et al. 2016), while triplet extraction can be compared to the Relation Extraction task (RE) which consists of extracting the words and the semantic relations between pairs of words. Distantly supervised NER is an often studied subject in the literature in comparison to distantly supervised RE (Liang et al. 2020, Meng et al. 2021) simply because NER is a subtask to RE and distant annotations generation is less expensive for the NER task (see Fig. 2).

Our first contribution is creating a distantly annotated species description dataset for Named Entity Recognition with a well-balanced test set that allows us to bypass several biases that can be induced by the distant annotation and that are often observed in NER datasets (Taillé et al. 2021). In this dataset, each word of interest will be classified into one of 15 classes, each class being a specific kind of organ or descriptor.

Our second contribution is proposing a distantly supervised model trained on our dataset, since fauna and flora corpora are particularly long and use a very specific technical vocabulary. We develop a context-oriented model adapted to this data by pretraining the language model. Thus the encoder of our model provides contextualized vectors for each extracted word that can be used to measure description similarities between different species. Our model reaches 96% accuracy in named entity classification on the test set.

Our third contribution is the triplet construction module that can directly be applied to our model's outputs. This module is based on class dependency rules that are inspired by Xper3's data representation format (see Fig. 3).

Finally, NEARSIDE is an end-to-end structured knowledge extraction framework from unstructured species description corpora, that can be applied to several data sources. Thus making species descriptions from different corpora easily linked, compared and measured.

Keywords

Natural Language Processing, Artificial Intelligence, species identification, biodiversity

Presenting author

Maya Sahraoui

Conflicts of interest

References

- Cho H, Lee H (2019) Biomedical named entity recognition using deep neural networks with contextual information. BMC Bioinformatics 20 (1). https://doi.org/10.1186/s12859-019-3321-4
- Devlin J, Chang M, Lee K, Toutanova K (2019a) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv. Number: arXiv:1810.04805 arXiv: 1810.04805 [cs]. URL: http://arxiv.org/abs/1810.04805
- Devlin J, Chang M, Lee K, Toutanova K (2019b) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv. Number: arXiv:1810.04805 arXiv: 1810.04805 [cs]. URL: http://arxiv.org/abs/1810.04805
- Fries J, Wu S, Ratner A, Ré C (2017) SwellShark: A Generative Model for Biomedical Named Entity Recognition without Labeled Data. arXiv. Number: arXiv:1704.06360 arXiv: 1704.06360 [cs]. URL: http://arxiv.org/abs/1704.06360
- Kerner A, Vignes R (2019) Multi-context Knowledge Base using Calculated Descriptors from Xper3: the Archaeocyaths Knowledge Base example. Biodiversity Information Science and Standards 3 https://doi.org/10.3897/biss.3.37083
- Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C (2016) Neural Architectures for Named Entity Recognition. arXiv. Comment: Proceedings of NAACL 2016. URL: http://arxiv.org/abs/1603.01360
- Liang C, Yu Y, Jiang H, Er S, Wang R, Zhao T, Zhang C (2020) BOND: BERT-Assisted Open-Domain Named Entity Recognition with Distant Supervision. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. [ISBN 978-1-4503-7998-4]. https://doi.org/10.1145/3394486.3403149
- Meng Y, Zhang Y, Huang J, Wang X, Zhang Y, Ji H, Han J (2021) Distantly-Supervised Named Entity Recognition with Noise-Robust Learning and Language Model Augmented Self-Training. arXiv. Comment: EMNLP 2021. (Code: https://github.com/yumeng5/RoSTER). URL: http://arxiv.org/abs/2109.05003
- Saucède T, Eléaume M, Jossart Q, Moreau C, Downey R, Bax N, Sands C, Mercado B, Gallut C, Vignes-Lebbe R (2021) Taxonomy 2.0: computer-aided identification tools to assist Antarctic biologists in the field and in the laboratory. Antarctic Science 33 (1): 39-51. https://doi.org/10.1017/S0954102020000462
- Sutskever I, Vinyals O, Le Q (2014) Sequence to Sequence Learning with Neural Networks. arXiv. Comment: 9 pages. URL: http://arxiv.org/abs/1409.3215
- Taillé B, Guigue V, Scoutheeten G, Gallinari P (2021) Separating Retention from Extraction in the Evaluation of End-to-end Relation Extraction. arXiv. Comment: Accepted at EMNLP 2021. URL: http://arxiv.org/abs/2109.12008
- Vignes-Lebbe R, Bouquin S, Kerner A, Bourdon E (2017) Desktop or remote knowledge base management systems for taxonomic data and identification keys: Xper2 and Xper3. Biodiversity Information Science and Standards 1 https://doi.org/10.3897/tdwgproceedings.1.19911

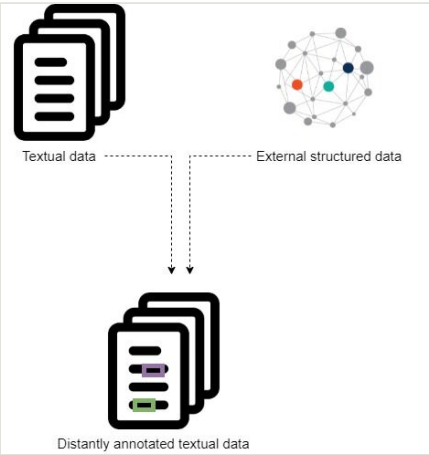

Figure 1.

Illustration of the distant annotation technique applied on textual data.

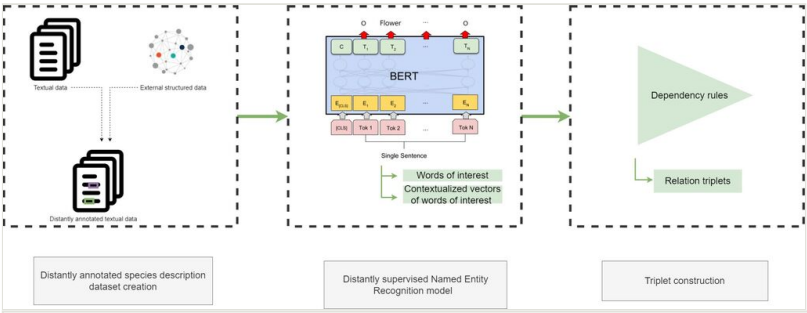

Figure 2.

Illustration of the structured knowledge extraction pipeline.

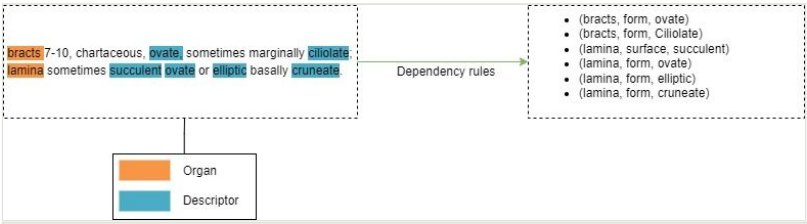

Figure 3.

Illustration of the triplet construction based on dependency rules applied on the extracted word of interest.