Exposé

Rule Based vs. Machine Learning Based Named Entity Recognition Approaches in an Information Extraction Pipeline in the Context of Bibliographic Data on Sphecidae Wasp Family

Arsenii Gulevich

April 2023

1 Introduction

The following bachelor thesis is ought to cover multiple approaches to information extraction (IE) from PDF files containing semi-structured information. Each PDF file corresponds to a genus in a selection of families of Wasps. Inside a file, all species contained in the genus are listed. Each entry of the list contains valid, invalid and unavailable names of the species with corresponding bibliographical references sorted historically, as well as location of its holotype, etc. Other information about actions affecting nomenclature is provided, sometimes in the form of comments. Quick retrieval of relevant data from this collection is difficult in its current form. A form suggested in this bachelor thesis is a digital library that can be managed with bibliographical reference management software. The resulting data entries have to conform to the international code of zoological nomenclature[11].

We intend to explore approaches of information extraction on these documents. Details might be of great importance for an entry on a particular species of Wasps. Since the dataset contains loosely structured comments and typing errors, it would be expedient to try to extract the species names and bibliographical and historical information for each entry, where we are sure that the data were extracted correctly and facilitate manually going over the ambiguous cases, for example where an entry is particularly ill structured due to many past actions affecting its nomenclature, due to uncertainties about old publications or authorship, etc.

Rule-based methods and machine learning (ML), as standard approaches in the field, are of interest in this IE task.

2 Dataset Catalog of Sphecidae

The documents to be processed with the information extraction pipeline constitute a set of PDF files with taxonomic and bibliographical information about Sphecidae wasps family. These files have been carefully created and are curated by Dr. Wojcieh Pulawski at California Academy of Sciences[13]. Each file corresponds to a genus and contains a listing of species. Each species might have tens of different publications affecting its nomenclature. Each publication is referred to with an author name and a year inside the catalog, with a separate collection of PDF files with bibliographical references further specifying the publications. To understand these entries, domain knowledge is of great importance, since formatting, comments, and special notation play a big role (for example an author's name is only given in parentheses, if the current genus differs from the one originally used).

Therefore, for retrieving the relevant information on a given species out of the catalog, it would be beneficial to turn the semi-structured records in the documents into a structured digital library to enable search and further processing. In order to do that, we have to create a set of entities such as holotype location, species name, authors name, bibliographical reference, comment, etc. Then we have to find these entities in the text and store them.

3 Objectives, Epistemic Interest

The main aim of this bachelor thesis is to compare various techniques of information extraction[10], as applied to real-life data. Specifically, our focus will be on investigating the impact of replacing rule-based named entity recognition (NER) with machine learning-based NER, within a specific pipeline for processing a corpus of records on Sphecidae wasps and comparing the performance of resulting variants of the pipeline.

We expect that recent findings in fields like word embeddings would allow for better performance in both precision and recall compared to purely rule based approaches[6].

The following questions are to be answered in this Thesis:

- How do readily available rule-based and ML-based NER approaches compare, when applied on the dataset with zoological taxonomic and bibliographical information at hand?
- Which information is relevant for classifying entities (in the pipeline processing step after NER) in the dataset? It can be particularly challenging to differentiate between ambiguous instances, when parts of the entries have similar structure or spelling, but convey different meanings.
- What is the difference in performance metrics when substituting rule-based NER (using the ruleset provided with GATE) in ANNIE GATE plugin[4] with machine learning-based NER and comparing the two versions?
- Which IE ruleset (used for the final annotation in the pipeline) is suitable for semi-structured records in the catalog of Sphecidae?

4 State of Research and Background

Approaches on Information Extraction can be grouped into rule based, knowledge based or ML based. Technologies as finite state automata are widely used in rule based approaches. While being extensively utilized, rule based IE doesn't find a lot of coverage in recent academic literature. Although rule based IE might in specific circumstances outperform ML based solutions.[5].

Machine learning offers many research opportunities, although the explainability and maintainability of ML systems can be a downside[2] in comparison to rule based systems. The aspects of most importance in for this bachelor thesis are word embeddings and rule languages. We intend to use GATE[3] and FlairNLP[1] annotation frameworks.

4.1 Named Entity Recognition

In various Information Extraction applications it is of great use to recognize such information units like names, including person, organization and location names, numeric expressions including time, date, etc. Identifying mentions of these entities in text is called Named Entity Recognition. The approaches for NER vary between handcrafted

rules and machine learning. The need for annotated corpora for training supervised machine learning models can be very limiting. New ideas in the field of ML such as self supervision and weak supervision help to circumvent this problem[8].

4.2 GATE Java Annotation Patterns Engine (JAPE)

JAPE is a version of CPSL – common pattern specification language[4]. It allows annotation based on regular expressions with the ability of manipulating the annotations with the help of arbitrary java code. One specifies patterns via rules, that can include information about context on the left-hand side, and arbitrary java code on the right-hand side. The recognition power of JAPE is no more than regular. Subsequent runs of JAPE with different annotation rules enable complex annotations.

4.3 Word Embeddings and Self Supervision

In the 1950s the idea was proposed that a meaning of a word can be represented as vector with coordinates corresponding to the word ratings on some chosen scales[12]. Furthermore, it was proposed to define a meaning of a word by its distribution in language use. These revolutionary ideas gave the start to the field of vector semantics. In vector semantics, words are represented as vectors. The computation of these vectors coordinates can be based on a co-occurrence matrix, a way of representing how often words co-occur in a corpus. Linear algebra operations can be applied to such coordinates, and similarity of words can be thought of mathematically. Sentences and entire documents can be assigned vectors representing their meaning by computing a centroid of all vector representations of words in the sentence or the document. These vectors are called embeddings[8], they can be computed using self supervised learning. The idea behind self supervised learning for computing word embeddings is as follows: The text utilized for training the model by itself can serve as a labeled dataset (that would be required fur supervised learning), since it already answers the question of which words can co-occur with a given word. A logistic regression classifier for example can be then trained on the data using neighbors of a word in the text as true instances and random sample words from the text for false instances. Word embeddings can be put to use for the task of Named entity Recognition[1].

5 Methodology

First, a sample annotated corpus has to be created with the help of a domain expert. We are going to manually annotate a randomly chosen set of entries from the catalog of Sphecidae. We will use GATE to develop an Information Extraction pipeline using a JAPE[4] Grammar in the ANNIE plugin. The plugin includes a tokenizer module, sentence splitter, a lookup module, a NER module and a so-called transducer, which as well as the NER module makes use of a JAPE Grammar to automate annotations. It will be applied for extracting the entities. The ruleset (JAPE Grammar) used in the NER module is going to be used as is, the ruleset for the final annotations, that uses

the annotations made in the previous steps to classify final entities has to be developed accordingly to the catalog of Sphecidae.

The NER part of the readily available pipeline is rule based. Our intention is to introduce a machine learning NER approach into the pipeline using FlairNLP[1] and compare resulting pipelines. In the end, ensuing differences in performance metrics are to be evaluated.

6 Performance Measure

As mentioned above, an annotated selection of entries from the catalog of Sphecidae representing the ground truth is going to be available for measuring the performance. It is going to serve as a reference, against which a hypothesis produced by an information extraction pipeline is going to be evaluated. Hypothesis here denoting the annotations produced by the Pipeline. In an Information extraction task at hand, 3 types of errors can occur:

S: substitutions (label differs from the one in the reference)

D: deletions (label is missing in the hypotheses)

I: insertions (label is not present in the reference)

By counting all the labels that are identical in the hypothesis and the reference, we get:

C: number of correct labels

To be considered correct, the spans with the correct labels don't always have to be entirely identical for the result to be sufficient in the context of retrieving information from the catalog of Sphecidae. Some characters in the retrieved fields may differ.

As a performance metric, we chose the slot error rate (SER)[9]. In comparison to the F-measure, it doesn't weigh different types of errors differently[9].

$$SER = \frac{S + D + I}{C + D + S} \tag{1}$$

The SER is equal to the total number of errors divided by the number of labels in the reference, and should be minimized[7].

7 Time Plan

Duration: 3 Months (04.23-07.23).

References

[1] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL: https://aclanthology.org/N19-4010, doi:10.18653/v1/N19-4010.

[2] Nadia Burkart and Marco F Huber. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70:245–317, 2021.

[3] Hamish Cunningham, Allan Hanbury, and Stefan Rüger. Scaling up high-value retrieval to medium-volume data. In Hamish Cunningham, Allan Hanbury, and Stefan Rüger, editors, *Advances in Multidisciplinary Retrieval (the 1st Information Retrieval Facility Conference)*, Lecture Notes in Computer Science, Volume 6107, Vienna, Austria, May 2010. Springer.

[4] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, 2002.

[5] Andreas Grivas, Beatrice Alex, Claire Grover, Richard Tobin, and William Whiteley. Not a cute stroke: Analysis of rule- and neural network-based information extraction systems for brain radiology reports. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 24–37, Online, November 2020. Association for Computational Linguistics. URL: https://aclanthology.org/2020.louhi-1.4, doi:10.18653/v1/2020.louhi-1.4.

[6] Michel Hahn, Udo; Oleynik. Medical information extraction in the age of deep learning. *Yearb Med Inform*, 29(01):208–220, Aug 2020. URL: http://www.thieme-connect.com/products/ejournals/abstract/10.1055/s-0040-1702001, doi:10.1055/s-0040-1702001.

[7] Julia Hirschberg. A corpus-based approach to the study of speaking style. In Merle Horne, editor, *Prosody: Theory and Experiment*, volume 14 of *Text, Speech and Language Technology*, pages 335–350. Springer Netherlands, Dordrecht, 2000. doi:10.1007/978-94-015-9413-4_12.

[8] Dan Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. In preparation. URL: http://web.stanford.edu/~jurafsky/slp3/.

[9] John Makhoul, Francis Kubala, Richard Schwartz, and Ralph Weischedel. Performance measures for information extraction. *Proceedings of DARPA Broadcast News Workshop*, 08 2000.

[10] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008. doi:10.1017/CBO9780511809071.

[11] International Commission on Zoological Nomenclature., W. D. L. Ride, International Trust for Zoological Nomenclature., International Union of Biological Sciences. General Assembly, and England) Natural History Museum (London. *International code of zoological nomenclature = Code international de nomenclature zoologique*, volume 1999. London, International Trust for Zoological Nomenclature, c/o Natural History Museum, 1999, 1999. https://www.biodiversitylibrary.org/bibliography/50608. URL: https://www.biodiversitylibrary.org/item/107142.

[12] Charles E. Osgood, George J. Suci, and Percy H. Tannenbaum. *The measurement of meaning*. The measurement of meaning. Univer. Illinois Press, Oxford, England, 1957. Pages: 342.

[13] Wojciech J. Pulawski. Catalog of sphecidae. https://www.calacademy.org/scientists/projects/catalog-of-sphecidae.