Application of Artificial Intelligence to Knowledge Discovery for the Maintenance and Conservation of Botanical Gardens

Claudia Aguilar-Rajme[1,2]

[1]GPLSI research group, University of Alicante, Spain
[2]Department of Computer Science, Technical Sciences Faculty, Agricultural University of Havana, Cuba

Abstract

Botanical gardens play a key role in the conservation of biodiversity and generate large amounts of data related to the management and maintenance of plants on a daily basis. In this sense, the Network of Botanical Gardens in Cuba stands out, made up of 12 institutions that, among other things, have in common that they manage these data by dividing them into three main registers: Introduction Register, Living Plant Register and Herbarium. The diversity of formats and the heterogeneity of the information make it difficult to manage by the Center's specialists, as well as to make it accessible to researchers and students in search of references. In this context, there is an opportunity to use artificial intelligence and natural language processing techniques to facilitate access to the implicit and explicit information contained in these records. To this end, it is proposed to evaluate the results of their use in order to finally obtain a product capable of using the extracted knowledge to provide recommendations and guidelines for the management and conservation of plants.

Keywords

Natural Language Processing, Information Extraction, Botany, Gardens

1. Justification of the proposed research

According to Smith and Harvey-Brown [1], the most widely accepted definition of botanic gardens is that expressed by Jackson in 1999 [2], who stated that they are "institutions containing documented collections of living plants for the purposes of scientific research, conservation, exhibition and education". Among the main functions of these centers are :

- proper documentation of collections, including wild origin,
- monitoring of collected individuals,
- communication and information to and from other Gardens, other institutions, and the public; and
- promotion of conservation through extension and environmental education activities [3].

These tasks involve the management of large amounts of data, which then become sources of frequent consultation for researchers, students, and the general public. The work of botanic gardens is essential to the preservation of the planet's biodiversity, and this is one of the reasons why botanic garden institutions exist all over the world.

In Cuba, there is a network of botanical gardens made up of 12 institutions, including the Botanical Garden of Cienfuegos, the oldest in the country, and the National Botanical Garden of Havana, the largest in the Cuban territory. In these institutions, as in the rest of the network, large amounts of data related to the processes that take place in them are generated daily. Of these data, those related to the management and maintenance of the plants are mainly divided into 3 groups, the first of which is the introduction record, where the information about the plant is stored when it arrives at the garden, by whatever means, and is kept there for a period of time after which its destination within the institution

Doctoral Symposium on Natural Language Processing, 26 September 2024, Valladolid, Spain.
claudia.aguilarrajme@gmail.com (C. Aguilar-Rajme)
0000-0002-7447-4458 (C. Aguilar-Rajme)
© 2024 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 ISSN 1613-0073

CEUR-WS.org/Vol-3797/paper23.pdf

is decided. Those specimens that are selected to become part of the collections are transferred to the Register of Living Plants, where they are tracked for the rest of their lives in the Garden. The third main repository of the Botanical Garden is the Herbarium, defined by the RAE as "a collection of dried and classified plants used as material for the study of botany". [4]. In addition to all this information generated within the institutions, various bibliographic sources are constantly consulted in search of information on the identifying characteristics of the plants, both from the physical point of view for their correct identification in nature, as well as the best practices in the management and conservation of specimens and the different uses that can be given to each variety.

All this results in a large amount of information, all related to botany and the plant species present in the garden, but in different formats and very heterogeneous, which makes it difficult to handle by the specialists of the center and even more difficult for researchers and students who are looking for references for consultation and research. This is where artificial intelligence and the various techniques available for handling large amounts of information come into play, since the use of models and algorithms can facilitate access to the implicit and explicit information contained in these records.

2. Background and related work

Natural Language Processing (NLP) is a branch of artificial intelligence that focuses on the interaction between computers and human language. In the last decade, NLP techniques have experienced significant growth and are being applied in a variety of fields, including botany and plant data management. These applications include information extraction (IE) from scientific texts, databases, and other relevant sources to facilitate the integration and analysis of large amounts of data related to plant species. In this research, we will focus on the task of Named Entity Recognition (NER), as it is a fundamental step to identify species names, morphological characteristics, geographic locations, and other key concepts in botanical texts.

There are several techniques for extracting information from texts, one of which is the use of rule-based models. These models rely on specific dictionaries and linguistic rules to identify entities. For example, projects such as FloraQuest [5] and Planteome [6] have used dictionaries of botanical terms to perform NER. These approaches are highly accurate in specific domains, but lack scalability and flexibility. Another technique used for information extraction is deep learning based models. In recent years, models such as Recurrent Neural Networks (RNN), especially LSTM (Long Short-Term Memory) architectures, and transformer-based models such as BERT (Bidirectional Encoder Representations from Transformers) have demonstrated superior performance in NER tasks. An example of the use of large language models can be found at [7], where they are being tested for efficiency of different language models for the identification of single and multi-word names of flowers and plants. The analysis was done for both Spanish and English, with a significantly smaller dataset in Spanish, but still obtaining relevant results. Thirteen models were tested for English and four for Spanish, demonstrating in both cases the superiority of the discriminative models over the generative ones in the task to be evaluated. For English the model with the best results was BERT-LARGE-CASED and for Spanish BERT-BASE-MULTILINGUAL-CASED. A similar study [8] was also carried out for the identification of plant names, but in this case metaphorical names, not textual ones, and the superiority of the discriminative models over the generative ones was maintained.

Despite significant progress, information extraction in botany presents unique challenges. The diversity and complexity of the language used in botanical texts, the ambiguity of scientific and common names, and the scarcity of annotated datasets limit the effectiveness of traditional approaches.

3. Description of the proposed research

The main objective of this research is the identification and application of NLP techniques that contribute to the extraction of knowledge from different scientific sources available in Spanish in the botanical field.

A number of specific objectives are planned to achieve the main objective:

1. Determine the state of the art of information extraction in the field of botany.
2. Identify reliable data sources in Spanish in the field of botany.
3. Select NLP techniques for information extraction.
4. Design a knowledge structure where the knowledge will be stored.
5. Apply the identified techniques to the data to extract and store the information.
6. Implement a recommendation system based on the extracted knowledge.

At the end of the thesis, it is expected to have identified and evaluated NLP techniques that allow obtaining efficient results in the discovery of knowledge in the field of botany in Spanish. This knowledge will later be used for a better management and conservation of the plant specimens within the collections of each garden, through a product that allows users to make queries about certain elements of the plants and provide recommendations for their conservation and maintenance.

4. Methodology

The methodology proposed to achieve the objectives set in the research is based on the fulfillment of tasks planned throughout the years of the doctoral program.

We will start with a search and reading of the state of the art to know the techniques currently used to extract information from unstructured texts in the field of botany or similar domains and that can be applied to it.

Then, since there are data from different sources (described in the next section), it is necessary to organize, integrate and standardize them to achieve a correct curation process and that the resulting dataset is as complete as possible and of great value in the field of botany. For this purpose, it will be necessary to validate the sources with experts in this field of knowledge and, once the sources to be used have been confirmed, to apply web scraping techniques to obtain the data available online.

The next step would be to process this data to extract and represent the knowledge it contains, for which we can use Named Entity Recognition (NER) techniques and semantic representation of the data. We will do this by following the next steps:

1. Text pre-processing: This stage removes unwanted words or characters, performs tokenisation and text normalisation, such as converting everything to lower case.
2. Grammatical tagging: This step involves tagging each word in the text with its appropriate grammatical category, using techniques such as POS (part-of-speech) tagging or dependency analysis. We should also explore ways of working with untagged text, as most of the data sources we have identified are untagged.
3. Entity identification: In this phase, we search for and identify named entities in the text. We will do this using machine learning models, rule systems and a combination of both to see which gives us better results.
4. Entity classification: Once entities have been identified, they are assigned a specific category or classification. For example, entities can be classified as plant name, life cycle, medicinal use, etc.
5. Feature extraction: Relevant features are extracted from the identified entities for storage, analysis and processing.
6. Validation and refinement: This is where the results of the analysis of the named entities are evaluated and adjustments or refinements are made where necessary. This involves verifying the accuracy and quality of the identified and classified entities.

This will allow us to build a knowledge structure that we can then use to support decision making. In this way, we will be able to build a recommendation system based on questions and answers that can interact with users and provide them with information and recommendations on various areas of botany and agriculture. This can be very useful for the staff of the botanical garden, as well as for

students and researchers, or even for the general public, since it will be a tool to accompany the design, creation and maintenance of a garden, with a scientific basis on the different plant species, soil types, seasons of the year and any other knowledge that we are able to integrate into the knowledge base from the data sources we are working with.

5. Data available

The data initially available came from the institutions of the Network of Botanical Gardens of Cuba, mostly records of the specimens present in each center, as an inventory or control. There were no records with textual descriptions of the plants, nor of the norms to be followed for the care and conservation of the different types of plants, nor of the uses that can be given to them, which is the information that a system intended to help the user in the process of creating and maintaining gardens would need. However, this information is available in several online sources that are consulted daily by professionals and that have scientific validity, which guarantees the quality of the knowledge extracted from them. The sources identified in Spanish are:

- Revista del Jardín Botánico Nacional (https://revistas.uh.cu/rjbn): aims to disseminate the results of Cuban and foreign scientific work in the field of botany and mycology, and publishes original articles and short communications in Spanish and English. It has 28 volumes available online and is published annually with high international visibility.
- Records of adventitious plants in the province of Alicante [9]: This is a reference handbook with the necessary information to design ecological gardens with adventitious plants. It contains a total of 45 plant files with scientific and common names of the species, life cycle, as well as textual descriptions of their physical characteristics (leaves, stems, flowers, fruits), agronomic needs and possible uses (Figure 1a).
- Flora Ibérica (http://www.floraiberica.es/): a website that collects definitions and synthesizes current knowledge about the vascular plants that grow spontaneously in the Iberian Peninsula and the Balearic Islands. Its aim is to facilitate the identification of the plants, and for this purpose it has a .pdf file for each species that contains the correct scientific name and its synonyms, a description that highlights the morphological peculiarities, the habitat in which it can be found, its geographical distribution in the world, its flowering period, its chromosome number, etc.
- Virtual Herbarium of the Western Mediterranean (http://www.floraiberica.es/): contains information and an extensive gallery of images of the vascular plants of the countries of the Western Mediterranean. It is structured in tabs or pages for each plant species treated, the main reason for each tab are the images of the plants, but also a brief information in text form about it (which is why it is a valuable source of data in this research) (Figure 1b).
- Spanish Invasive Alien Species Catalog - Plants (https://www.miteco.gob.es/es/biodiversidad/temas/conservacion-de-especies/especies-exoticas-invasoras/ce_eei_flora.html): official website of the Spanish government, where the list of species considered invasive in the territory can be found, and for each of them can be obtained a .pdf file with textual descriptions of the physical appearance of the plant, the main distinguishing characteristics compared to other similar plants, the ecological and health impact and the main routes of entry (Figure 1c).

All of these sources have been reviewed by experts in the field of botany and have been found to be a reliable source of reference in this area. They are therefore considered suitable for the purposes of the research and web scraping techniques will be used to extract the information they contain to create the dataset from which the information extraction will be carried out.

Atriplex halimus (salada blanca o *salat blanc*). (www.apatita.com d, 2018).

Familia: *Chenopodiaceae*.

Ciclo: perenne.

Descripción: es un arbusto que puede crecer hasta 2,5 metros, con ramas desde la base y que su corteza es grisáceo-blanquecina. Las hojas son muy variables, de deltoideoorbiculares a lanceoladas y de corto pecíolo. La inflorescencia es de flores poco vistosas.

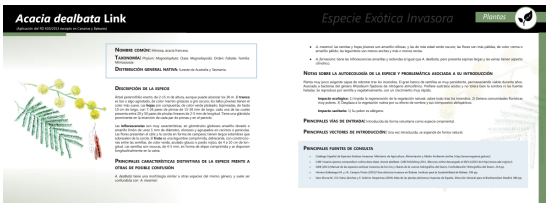
Necesidades agronómicas: crece en suelos arcillosos, limosos o arenosos en los que siempre hay un cierto grado de salinidad. Presenta resistencia a la sequía, a la intensa insolación y a la salinidad.

Época de floración: julio-noviembre.

Usos: se cultiva frecuentemente como planta ornamental para formar setos y como planta forrajera. www.asturnatura.com e, 2018).

(a) Records of adventitious plants in the province of Alicante

(b) Virtual Herbarium of the Western Mediterranean



(c) Spanish Invasive Alien Species Catalog - Plants

Figure 1: Data Source Examples

6. Specific Issues of Research to be Discussed

In this section, after explaining the proposed research, some questions are posed for discussion:

Q1. Ambiguity in the data: Since we have different sources of data and the same species may appear in more than one of them, it is possible that we will come across cases where it is necessary to disambiguate the information for one or more characteristics; in these cases, how should this issue be approached? would it be wise to choose one of the different options found or to reflect all of them with the appropriate reference?

Q2. Evaluation: What metrics would be most appropriate to evaluate the results of the information extraction task?

The conclusions of the debate generated by the questions presented, as well as other aspects that may emerge, would be of great value for the development of research.

References

[1] P. P. Smith, Y. Harvey-Brown, BGCI Technical Review: Defining the botanic garden, and how to measure performance and success, volume 2 of *kkhui*, jhhghhgj ed., Botanic Gardens Conservation International, hjhbjh, 2017. Uygfyuu.

[2] P. S. W. Jackson, Experimentation on a large scale-an analysis of the holdings and resources of botanic gardens, Botanic Gardens Conservation News (1999).

[3] W. IUCN-BGCS, The botanic gardens conservation strategy, IUCN-BGCS, WWF Gland, Swit-zerland (1989).

[4] R. A. ESPAÑOLA, Diccionario de la lengua española, 23.ª ed. URL: https://dle.rae.es.

[5] N. C. B. Garden, Floraquest (????).

[6] L. Cooper, J. Elser, M.-A. Laporte, E. Arnaud, P. Jaiswal, Planteome 2024 update: Reference ontologies and knowledgebase for plant biology, Nucleic Acids Research 52 (2024). doi:10.1093/nar/gkad1028.

[7] D. Premasiri, A. Haddad, T. Ranasinghe, R. Mitkov, Deep learning methods for identification of multiword flower and plant names, Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing (2023).

[8] T. Ranasinghe, R. Mitkov, A. Haddad, D. Premasiri, Métodos de aprendizaje profundo para la extracción de nombres metafóricos de flores y plantas, Sociedad Española para el Procesamiento del Lenguaje Natural, Section: Procesamiento del lenguaje natural (2023).

[9] J. A. Mateu Brotons, Fichas de plantas adventicias de la provincia de Alicante para su uso en el diseño de jardines ecológicos, Master's thesis, Universidad Miguel Hernandez de Elche Escuela Politécnica Superior de Orihuela, 2018.