Mining Ancient Medicine Texts Towards an Ontology of Remedies – A Semi-automatic Approach

João Nunes[1], Orlando Belo[1(✉)], and Anabela Barros[2]

[1] ALGORITMI Research Centre/LASI, University of Minho, Campus of Gualtar, 4710-057 Braga, Portugal
a82300@alunos.uminho.pt, obelo@di.uminho.pt

[2] Centre for Humanistic Studies, CEHUM, University of Minho, Campus of Gualtar, 4710-057 Braga, Portugal
aldb@elash.uminho.pt

Abstract. Over the last years, ontology learning processes have gained a vast space for discussion and work, providing essential tools for discovering knowledge, especially from textual information sources. One of the most currently used techniques for extracting ontological elements from textual data is through the application of lexical-syntactic patterns, which aim to explore formalities of the language in which texts are written, for removing hyperonym/hyponym pairs that can be used to identify and characterize ontology concepts and create valuable semantic networks of terms. We applied a lexical-syntactic patterns approach in a set of medicine texts, written in classical Portuguese, during the 16th and 17th centuries, with the goal of extracting hyperonym/hyponym pairs to establish a medicine ontology of the time. In this paper, we discuss the most relevant aspects of an ontology learning system we implemented for extracting the referred ontology, which has the ability for characterizing the knowledge expressed in ancient medicament texts.

Keywords: Ontology Learning · Linguistic Ontologies · Extracting Knowledge from Textual Data · Natural Language Processing · Linguistic Patterns · Graph Databases

1 Introduction

In computer science, ontologies are data structures having a concise and explicit semantics in a given knowledge domain, which integrate different categories of data and metadata elements, such as classes, concepts, and relationships, among others. Ontologies [1, 2] are very useful instruments for supporting assumption or inference processes about their application domain, which greatly facilitates the understanding of its elements and the domain they represent and sustain. Additionally, allow for creating simple visualization methods on the data represented in an ontology's structure, providing a clear representation of all ontological elements and relationships. Inserting, updating or removing

S. Nandan Mohanty et al. (Eds.): ICISML 2022, LNICST 471, pp. 131–146, 2023.
https://doi.org/10.1007/978-3-031-35081-8_11

ontological elements are easy tasks to do, due to the flexibility of the structure of an ontology, being very scalable in this type of operations [3]. Furthermore, an ontology can easily receive and represent a model of a given domain of knowledge, easily to be share with other systems [4].

Nowadays, the popularity of ontologies, as well as the recognition of their utility, is due to the successful application of ontological systems in knowledge-based systems that provide information retrieval, natural language processing or machine learning services [5]. Thus, over the last years, ontology learning processes have gained a vast space for discussion and work, being essential tools for discovering knowledge, especially from textual information sources. Ontology learning [6, 7] can be defined as the set of methods or techniques used for constructing, maintaining or extending ontologies in an automatic way. To do this, a very diversified set of machine learning techniques is used on ontology learning processes, especially natural language processing, to identify, for example, terms, concepts, properties, and relationships that are present in data [8, 9]. Afterwards, all these elements can be used to compose an ontology, housing a given knowledge domain. However, the process of extracting an ontology can be very complex and demanding, both in terms of human and computational resources. Today, carrying out a process like this in a fully automatic way is already possible, but stills is very difficult to design and implement, requiring in most cases the attention of one or more specialists to validate, evaluate or adjust, in particular, the tasks of identification and extraction of the various elements of the knowledge involved. The difficulty of these processes is considered as complicated as the final knowledge granularity that we pretend include in the ontology structure and semantics. As referred, ontology learning processes are commonly used on unstructured texts. These texts are difficult to deal with, because most of the knowledge they contain can rise very different interpretations, from person to person, even when same words are used [10].

One of the techniques we can use for extracting ontological elements from textual data is through the application of lexical-syntactic patterns [11], which aim to explore formalities of the language in which texts are written, for removing hyperonym/hyponym pairs that can be used to identify and characterize ontology concepts and create valuable semantic networks of terms. We applied a lexical-syntactic patterns approach over a set of remedy texts [12, 13], written in classical Portuguese, during the 16th and 17th centuries, with the goal of extracting hyperonym-hyponym pairs to establish a remedy ontology of the time. After an exhaustive analysis of the remedy texts, edited in semidiplomatic version (respecting the classical language and orthography, with all its normal variation), and using ontological tools, it was possible to identify a significant set of linguistic patterns. These patterns allowed, with a very high degree of confidence, the automatic extraction of a large diversity of ingredients of the remedies and other relate data from those texts. All of this sustained the generation of a specialised ontology about remedies, prepared and used at the time the codex was written.

In this paper, we discuss the most relevant aspects of an ontology learning system we implemented for extracting a remedy ontology, characterizing the knowledge expressed in the remedy texts referred above. The ontology we got incorporates a large diversity of medicine elements that allow for knowing botanical specimens, animals and minerals, in addition to the medicinal, daily, agricultural and veterinary practices of the referred

period. It offers a very rich field of research for students, professors, researchers and individual users, offering a very interesting view of the remedy knowledge of 16th and 17th centuries to pharmacological and medicinal research in the 21st century [14]. The remaining part of this paper is organized as follows. Section 2 exposes the domain of ontologies, their methodologies and applications, Sect. 3 discusses ontology learning processes, approaching task models organization and functioning, Sect. 4 documents our application case and reveals how we extract the ontology of remedies on ancient remedy texts, and, finally, Sect. 5, presents conclusions and points out some research lines for future work.

2 Ontologies, Methodologies and Applications

Technological advances taken place over the last few years have given rise to a large number of applications, somewhat dispersed across a large variety of knowledge fields, which raised an impressive volume of data, especially in the global network, the Internet. Almost proportionally to this volume of information, the types and models of data (and metadata) soared extraordinary, emerging large information sources of a very diverse nature. The vastness of this information, structured, semi-structured or unstructured, aroused the curiosity of many researchers and companies that saw that this information would bring them value for their innovation, development and business activities.

Currently, much of this information is available for exploration. However, given the large volume, diversity and variation over time, this information cannot be explored in a conventional way. It needs sophisticated and expeditious mechanisms of extraction, preparation, storage, and analysis, being capable of dealing with the nature of this data. This raised numerous questions and challenges. The representation of the knowledge contained in these data is just one of those problems, one of those challenges. Although not easy, it is a very important issue. The knowledge that data can provide, in order to be identified, needs special mechanisms capable of working with different types of data, recognizing concepts, properties and relationships, among other things, and then storing them in adequate structures that allow for a useful and expeditious exploration. The lack of an adequate storage structure usually makes impossible to explore the knowledge we acquired.

Nowadays, one of the most sophisticated structures we have access to accommodate this knowledge are ontologies [1, 2]. These can be seen as representative elements of a given knowledge domain [15]. When well defined, an ontology is a primordial structure in any knowledge-based system, allowing for exploring various strands of knowledge, sustained by the concepts and relationships acquired, taking into account the characteristics and semantics of knowledge elements. Through an ontology, we can share or reuse an explicit representation of knowledge for other purposes, partially or fully.

There are many other perspectives for defining ontologies, addressing many different aspects, but always having the same common denominator, the representation of knowledge. Uschold and Gruninger [16] defined an ontology for sharing knowledge about a given domain of interest, allowing for the encapsulation of a global perspective on that domain, involving the conceptualization of a set of concepts, in the form of

entities, attributes and relationships. In turn, Borst [17] defined an ontology as a formal specification of a shared conceptualization, emphasizing the fact that there has to be an agreement regarding the conceptualization of the ontology. Finally, the definition given by Gruber [18], in general terms, is the most consensual definition in the field of Computer Science.

Ontologies prove to be advantageous in terms of knowledge structures for making assumptions and inferencing methods over data, which may allow the creation of simple visualization methods for the data represented in the ontology's structure. Regarding the insertion of new elements in an ontology, due to the flexibility of its structure, an ontology can be easily extended, being scalable for any conventional data manipulation operations [3]. Additionally, different users or systems can easily share its structure [4].

Ontologies can be very useful tools for solving problems in the most diverse areas. The use of applications supported by ontologies can be divided into several categories, taking into account a very diverse range of aspects, such as, for example, the creation of data structures and associated semantics or the exchange and sharing of knowledge contained in an ontology. Today, we can easily found worldwide applications of ontologies, ranging from Music to Medicine and Biology. However, the process of developing an ontology [19] is not an easy process to execute. It requires the use of concrete strategies and methods that should be applied from its idealization, to its implementation and subsequent exploitation stages. Over the years, researchers in the field have been proposing and developing different types of approaches, not only for conceptualization but also for the implementation of ontologies. There are some ideas and aspects shared by the different ontology learning approaches, concerning the creation and development of ontologies. Considering alternatives in the idealization of the knowledge domain, knowing that the development of an ontology is an iterative process, or knowing that classes and relationships expressed in an ontology must always belong to a domain of knowledge, are only two examples of such ideas. In the next section, we will look at the ontology learning process in more detail.

3 Ontology Learning Processes

The design and implementation of an ontology learning process should not carried out in an *ad hoc* manner. By nature, they are quite complicated processes using different types of strategies and methods from different domains of working, for extracting knowledge from textual data, which often require natural language processing techniques [20], text mining [21], and machine learning mechanisms [22]. Due to their nature, all these domains are quite complicated and require a high level of knowledge for using and applying their tools and working methods. For overcoming ontology learning process implementation difficulties, we may use different types of approaches and methodologies. Despite their differences, they all share a set of common tasks, which regulates how an ontology learning process is built [4]. Basically, we need to take into account one must consider several alternatives in the idealization of the domain of knowledge, since there are also several possibilities, more or less complex, of conceptualizing an ontology. Furthermore, we must remember that an ontology development process is iterative, which according to the implementation requirements will allow to increase or decrease

the structure of the ontology, influencing its final complexity. The ontology's classes and relationships must always belong to the domain of knowledge we are dealing with, otherwise the ontology will not be useful since it did not respect the knowledge domain. Some common ontology development methodologies are Tove [23, 24], Uschold [25] and Methontology [4].

However, the ontology learning process is not limited to the choice of a methodology and its consequent application. In turn, the ontology has to be evaluated based on an established framework, knowledge representation, data types and extraction requirements. Reyes-Peña and Tovar-Vidal [25] argued that a methodology must be verified and evaluated in terms of the model it incorporates. For carrying on this verification, it is necessary to analyze, for example, all the lexical aspects related to the dictionaries of terms created and used, the semantic aspects referring to the consistency of the terms, and the structural aspects related to the ontology structure itself, to verify, for example, if there are loops in the ontology structure. Data driven and application driven [27] are two of the techniques that can be used to validate an ontology. In the first technique, we use textual data from the same domain of knowledge of the ontology for comparing the data with the elements defined and integrated in the ontology. In the second technique, we combine the ontology created with a given application for evaluating the performance of the ontological application, measuring the quality of the responses given by the application.

Ontologies can be extracted from texts manually or automatically. The manual construction of an ontology has costs of different nature, caused by the modeling and capture of new knowledge or by the need for an expert for mapping and sketching a certain domain of knowledge in an ontological structure. In small dimension and complexity domains, these costs do not prevent the manual construction of an ontology. However, in applications for wider domains, the manual preparation of an ontology may be unfeasible, and serious challenges and obstacles may be faced not allowing the extraction of knowledge and its materialization in an ontology [28]. As a response, automatic mechanisms were gradually introduced throughout the different stages of development of the ontology learning process, addressing tasks in which the application of natural language processing, and artificial intelligence programs brought great added value. Gradually a new expertise field emerged. Today, we recognize it as Ontology Learning [6, 7].

Afterwards, all these elements will be used to compose an ontology to house the knowledge we need. However, given the current state of technology in the domain, rarely a process of extracting an ontology can be considered fully automatic, as it often requires the attention of one or more users to validate, evaluate and adjust the parameters of the programs used, according to the results we want to get. In practice, human intervention is desirable for guaranteeing the quality of the ontology, being also dependent on the degree of difficulty of the extraction process. This is as complicated as the nature and complexity of the knowledge domain and consequently dependable of the granularity of the ontological structures required for its representation [8]. In short, supervised ontology extraction processes are usually referred to be semiautomatic processes.

The structure of a semi-automatic process was already defined for a long time. Somewhat inspired by the process of manual ontology construction, Brewster [29] used the image of a "layer cake" for representing the order in which extraction tasks would be

carried out, organizing them in an increasing order of execution difficulty – terms >> synonyms >> concepts >> hierarchies >> relationships >> axioms. At the base of the "cake" are the simplest tasks, such as the identification of terms, and at the top the more complicated ones, such as the definition of axioms, representing logical rules for the knowledge domain. Several authors discussed this organization, evaluating its structure and application. Tiwari and Jain [19] simplified the proposal of Brewster [29], considering that some tasks can be performed in the same process step. For example, we can execute the task of identifying terms and synonyms simultaneously, since both tasks are similar in terms of linguistics and extraction methods application. As a result, we have a layer cake with a smaller number of layers – terms and synonyms >> concepts >> relationships >> axioms.

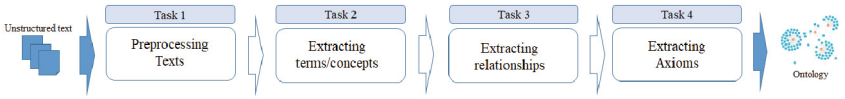

Fig. 1. Main tasks of an ontology learning process – adapted from [7].

More recently, Asim et al. [7] adapted the organization of Tiwari and Jain [19], adding two new tasks, namely data pre-processing and evaluation of the ontology extraction process (Fig. 1). In the first task, text pre-processing cleans and prepares text according to the selected extraction algorithm requirements, removing possible cases of ambiguity, or structural or semantic inconsistency that could compromise the final ontology [30]. In the next task, linguistic or statistical mechanisms are applied for extracting terms and concepts to be implemented in the ontology, grouping them into clusters, without any relationship established among them. Subsequently, these elements will be analyzed and combined in order to found possible relationships with each other, based on the dependencies between terms and concepts contained in the text [8]. The task of extracting axioms is the next. This is the most complex and demanding task of the entire process. Usually, it requires very specialized tools, which can use programming mechanisms in inductive logic [31] or Hidden Markov Models [32], among others. Finally, using of data driven or application driven techniques, all the identified ontological elements and their relationships are validated [26]. In short, this analysis and evaluates the adequacy level of the ontology. The approach of Asim et al. [7] served as a guide in the development process we will present and describe in the next section.

4 An Ontology for Ancient Remedies

4.1 The Application Case

The travels of the Portuguese around the world during the age of discovery allowed them allowed to access to a large variety of knowledge of different peoples and cultures, related to a wide variety of domains, such as medicine, biology, or gastronomy, among others. The knowledge they acquired was recorded in manuscripts and prints that are part of the history of the Portuguese. In many cases, some of these documents remain

unknown and without the recognition they deserve within the scope of their domains [14]. One of these manuscripts is the number 142, which is stored in the District Archive of Braga, in Portugal. It is an anonymous manuscript, which includes in one of its parts (notebook II) hundreds of texts, written in classical Portuguese, including receipts of remedies, advice and medicinal secrets, from Portugal, Europe but also from Brazil and Asia, collected during the 16th and 17th centuries. These remedies, prepared in convent apothecaries but also in the kitchens of each family house, still have a lot to offer to today's pharmacological and medicinal fields. They make known to us numerous botanical specimens, with geographically and linguistically varied names, animals and minerals, in addition to the medicinal, daily, agricultural and veterinary practices common in past centuries.

Pª faser Nascer Cabellos.

Tome' Ran's, e lagartigxas, ponhanas a torrar no forno dentro de hua panella pisenas como estiuere' be' torradas, e frigiaõ em aseite huas poucas de moscas, e depois ama anasando hua gemma de ouo deitemlhe deste aseite, e poós, e por 3es dias ponhaõ emprasto disto as noites donde quisere' q nasaõ os cabellos.

Fig. 2. A remedy's recipe to grow hair – "**Pª faser Nascer Cabellos**"[1]

In Fig. 2, we can see a fragment of the original text of a remedy's recipe, in classical Portuguese (semi diplomatic version). This is just a small example of the ancient medicine texts that we have available for analysis. With a brief analysis of the text it is possible to identify the various ingredients used in the recipe as well as to get to know their preparation and application processes. Obviously, the analysis of a text like this one it is easy to perform, as is the manual extraction process of the knowledge structures presented in the text. According to the content of the text presented in Fig. 1 and abstracting concepts and relationships, we can say that:

(1..1) RECIPE needs > (0..n) INGREDIENT
(1..1) RECIPE involves > (0..n) OPERATION
(1..1) RECIPE uses > (0..n) UTENSIL
(1..1) OPERATION uses > (0..n) INGREDIENT
(1..1) OPERATION requires > (0..n) UTENSIL

In fact, beyond knowing how to read and interpret a text written in classical Portuguese, it is not necessary a great expertize or knowledge to create an ontology for receiving such knowledge. However, when we have to deal with a few hundred texts, the process becomes a little more arduous and complex. In this case, opting for a manual approach for extracting knowledge is not feasible or achievable in useful time. Knowing this, we easily decided to design and develop a system for extracting the knowledge contained in remedies texts, which, in a supervised way, would provide ontology for representing such knowledge correctly and completely. The ontology will allow for a

[1] An English translation of the recipe: *"Take frogs and geckos; let roast them in the oven in, a pan; crush them when toasted; fry a few flies in olive oil; after kneading an egg yolk, put the oil and some powders; for 3 days, apply this at night, in the place where you want hair grow."*.

sustained analysis of the different concepts (and their relationships) related to different remedies and, in particular, to their ingredients, and preparation and application processes. Thus, it will be possible to demonstrate, in a simple and visual way, the process of making medicines used at the time, and study the herb, tree or plant, seed, diseases and symptoms, health practices, etc., that were used. For linguistics, the ontology will be a very useful instrument, allowing them to know the writing style followed in each of the recipes, the vocabulary and the way it was used at the time, the structure of the sentences, among many other things, allowing for analyzing and comparing with what is used today.

We define the scope of the ontology using a set of informal competency questions, involving the main aspects of a remedy (recipe, ingredients, utensils, operations, etc.) for validating the knowledge acquired during the ontology extraction process. Questions like "How to prepare the prescription of a remedy?", "How to apply a remedy?", "What ingredients are used in the preparation of a remedy?", or "What operations do you have to perform to prepare a remedy?, were the basis for evaluated the ontology across the several versions of the ontology learning process. In the next section, we will describe this process.

4.2 Extracting the Ontology

It is not necessary to make a detailed reading to verify that the texts of remedies mentioned previously contain immense information about the processes of making medicines that took place in the 16th and 17th centuries. From this, we are able to acquire very relevant knowledge about the medicines that were at that time produced, as well as their form of using and application. Providing an ontology with all this knowledge, will be useful for researchers, teachers, students and other users, to know, for example, the different names of an herb, tree or plant, or a seed used in convent apothecaries for treating diseases, or other health practices, from Portugal and Europe, but also from Brazil, and Asia, at that time. After assessing the effort that would be necessary to spend for treating and analyzing manually the referred texts, we found that this work would not be feasible, as it would not produce any results in the short term. Thus, we opted for developing a semi-automatic ontology learning process for extracting knowledge elements for the desired ontology of remedies.

To project and implement the ontology learning process, we adopted the model proposed by Asim et al. [7], as a working base and a general configuration for preparation and extraction tasks. However, given the specific characteristics of our application case, we adjusted or modified some of the tasks for using other text analysis approaches. For example, in the extraction of terms and concepts task we used Hearst patterns [11]) for making the extraction of hyperonym and hyponym relations in texts. Let see, then, how we proceed to get the ontology of old medicines that we will be present later.

Preprocessing Texts

According to the extraction plan we defined, we began preprocessing the texts containing the various preparation processes of remedies. However, before that we needed to analyze carefully the texts for understanding which models and techniques could be applied without distort the meaning of the text. This is fundamental for ensuring the reliability

of the ontology and guaranteeing that it is as close as possible of the domain's knowledge. Since texts are written in classical Portuguese, we decided to restrict preprocessing tasks as much as possible, in order to maintain intact the meaning of the text in.

Preprocessing began with the elimination of spaces, line changes and special characters presented in the texts, for ensuring that they did not interfere with the next tasks, text tokenization and analysis of part of speech. For accomplishing these two tasks we used spaCy [33], a natural language processing library, based on the performance it has shown in other similar applications, when compared to other similar alternatives, such as NLTK [34]. In addition, spaCy also offers pre trained and variable size pipelines ready to apply to any textual data, including tokenization, part of speech tagging, or named entity recognition (NER) tasks, among others. Additionally, we also removed other characters, such as apostrophes, which are identified by spaCy as individual tokens, as well as other cases, such as the inclusion of specific textual patters, such as the pattern "[3] " that does not add any meaning to the phrase in which it is inserted. We can see a concrete case of remedy recipe ("other"), which was worked according to the process we just explained. The text of the original recipe (in classical Portuguese) is:

"Outro.
[10]
Pª resoluer nascidas, farinha de semeas, oleo Rosado, vinho. Outro. Cosimento de barbasco, oleo rosado com hu' pouco de asafraõ, farinha de trigo galego feito huas papas. Outro pera faser arebentar em 24 horas. hua gemma de ouo anasada com asucar mto be' ate q' fique grosa; estendase em hu' paninho, e ponhase, renovandose como estiuer seca, e ella faz buraco, e chama a mala, e continuena depois de aberto, o tempo q'quisere', -pq' atte a rais chupa."

After applying the prediction models and techniques we selected, the text above was transformed in the following:

"Outro.P resoluer nascidas, farinha de semeas, oleo Rosado, vinho. Outro. Cosimento de barbasco, oleo rosado com hu pouco de asafraõ, farinha de trigo galego feito huas papas. Outro pera faser arebentar em 24 horas. Hua gemma de ouo anasada com asucar mto be ate q fique grosa; estendase em hu paninho, e ponhase, renovandose como estiuer seca, e ella faz buraco, e chama a mala, e continuena depois de aberto, o tempo q quisere, pq atte a rais chupa."

This new version of the text is easier to interpret and maintains its original meaning. The modifications made did not have a major impact on the meaning of the text, providing a clear text, easier to interpret and process by the tools we chose. After this cleansing step, we fragmented the texts accordingly their sentences. Although spaCy offers components to detect and segment the sentences of a text, the result we got with it was not always correct, revealing some incorrectly segments of sentences. We believe this situation occurred because texts are written in classical Portuguese. Thus, we had to segment sentences using the '.' character as segmentation reference. At the end, each processed text was stored in a document store supported by Mongodb. The use of MongoDB was essentially due to its ability for scaling and rapid processing of queries, in inserting and exploring data operations [35]. Figure 3 shows all the tasks performed during text preprocessing.

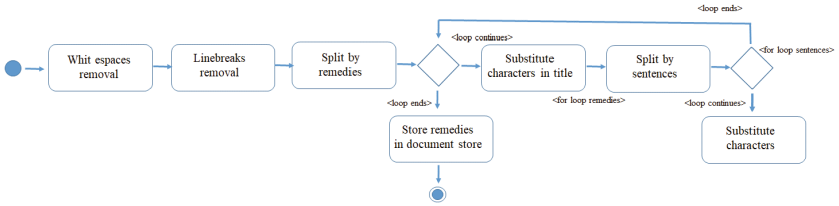

Fig. 3. Text preprocessing tasks

Extracting Concepts and Relationships

After preprocessing texts, we proceeded the extraction of concepts and relationships tasks, createing a specific implementation of a pipeline for NER. We trained the NER model using a data set extracted from remedy texts after being annotatted appropriately. The pipeline was personalized (Fig. 4) and is feed with the sentences of the texts. The pipeline structure closely follows the approach by Honnibal and Montani [33].

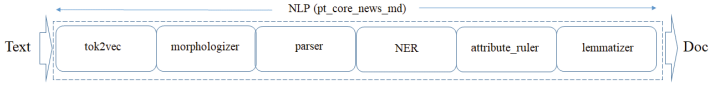

Fig. 4. Pipeline for concepts and relations extraction [33]

The first component of the model, tok2vec (token-to-vector), is responsible for separating the text into individual tokens. Then, each of the tokens is represented in a vector of a representation space for word embedding. Using a CNN neural network (Convulutional Neural Network) we got the vector representation. The neural network and the embeddings will be shared by the successive components, in order to reduce the size of the models, since the embeddings are only calculated and copied once. Furthermore, this solution allowed reducing the system's processing time. Next, the morphologizer module identifies and classifies the part of speech tags for each token founded. This component uses tok2vec for adding a new layer to the network, containing a softmax activation function to produce part of speech tags predictions. After identifying the tags, the parser component learn dependencies between the different vectors. Once again, we used tok2vec again to create a new model called the transition-based parser. This model can be used either to perform named entity recognition or dependency parsing. Finally, the lemmatization module simplifies the text by reducing words to their basic form. At the end of the process, we got a doc object, containing the sequence of the token objects related to each word. For each token object, it will be possible to consult information regarding the output of each component of the pipeline, being able to access to the part of speech tags or the lemmatized form of the word, among other things.

Lexical-Syntactic Patterns

For extracting concepts and relationships referred in the texts, we used Hearst patterns [11], which are lexical-syntactic patterns specially oriented to extract hypernym and hyponym relationships from the text. Hearst applied these patterns to encyclopedic and

journalistic corpora, obtaining 63% of good quality relationships. The work carried out by Hearst was a pioneer in the use of lexical patterns in the extraction of semantic relationships. There are other relevant works concerning Hearst patterns application to texts written in Portuguese. Authors such as Freitas and Quental [36], Baségio [37], Taba and Medeiros [38] or Machado and Strube de Lima [39], developed very interesting works in this domain, in particular Baségio, which directed its research to the Portuguese language used in Brazil.

The concept (and their relationships) process began with a manual analysis of the texts, for interpreting and identifying possible lexical-syntactic patterns and then extract the hypernym and hyponym pairs. Taking into account the requirements and the goal of the ontology, we applied extraction using a survey of the ingredients that made up a recipe for a remedy, and, later, for identifying ingredients' quantities and recipes' preparation methods, if any. We defined the lexical-syntactic patterns using part of speech tags. The spaCy tool provides a matcher object that can be used to find words in texts that instantiate some of the defined patterns. Then, we used these patterns in conjunction with the matcher object for improving concept and relationship identification. Table 1 shows some of the patterns we used in the ontology learning process.

Table 1. Examples of Hearst patterns used in the ontology learning process.

Nr	Hearst Patterns
i	com(erlaõla) (B1,)* (B2(elou))* B3
ii	such NP as {NP,}* {(orland)} {NP}
iii	(B1,)* tomado
iv	beb(erleralaõlalidalido) (B1,)* (B2(elou))* B3
v	NUM (onsaslonças) de (B1,)* (B2(elou))* B3

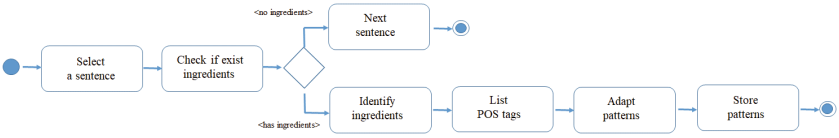

Fig. 5. Identification of patterns for ingredients

Additionally, we assumed the existence of a common hypernym, having the value of the ingredient, and then extract the hyponyms related to the existing hypernym from the text being processed. Figure 5 illustrates this process. To demonstrate this process, we will analyze a specific sentence from one of the remedy prescriptions used in the ontology extraction process. The sentence we selected is:

*"**Macela**, **Coroa de Rei**, **Maluaisco**, ou **Maluas**, cosido tudo com agoa e metido este cosimento em hua bexiga de boi posta sobre a pontada."*

Thus, after we have selected the phrase, we identify the ingredients making possible to define a lexical-syntactic pattern. In the sentence presented, we identify four distinct

ingredients (“Macela”, “Coroa de Rei”, “Maluaisco” and “Maluas”), which will be integrated later into the composition of the lexical-syntactic pattern. Next, we identify the syntactic elements necessary for defining a pattern that allows us to extract the ingredients identified previously. As we can see, in the selected sentence the ingredients identified were founded before the key words “cosido tudo”, which refers a sewn operation. This allows us to verify and generalize, with some certainty, that before these words we can find ingredients. After identifying the pattern, we analyze other recipes to adjust its definition, if occurred some variations caused by the contexts of other recipes. Having the keywords of the pattern, we dispose the doc object information, coming from the spaCy analysis of the sentence used, and we consulted the part of speech tags. In order to generalize the creation of patterns, they were created using a block containing the most common tags for supporting the extraction process. These blocks comprise four tags, related to nouns (NOUN), to proper nouns (PROPN), to adjectives (ADJ) and to verbs (VERB). This format was created based on the patterns presented and described in [39].

The Ontology

After finished the definition of lexical-syntactic pattern, we applied them to each of the sentences of the texts used in the extraction process. Whenever we detected an instantiation of a pattern in a sentence, we recorded the occurrence ensuring that we did not keep repeated patterns, whether they were partially or fully instantiated. Then, pattern instantiations were analyzed for extracting the ingredients they referenced. It was necessary to apply some filters for selecting and extracting the ingredients from the texts, in order to get a consistent list, without repetitions. The instantiation of each pattern returned a list of strings, containing the names of the ingredients and other words included in the sentence. According to the position of the ingredients in the string list, we checked for the existence of other specific strings (or just characters) that were not part of the ingredient name. For example, occurrences of commas, or conjunctions such as “and” and “or” were removed. From this process, we got the ingredients of the recipes processed, as well as the quantities used and the method of preparation of the remedies. All these elements are distinct. They were obtained by different patterns, with different unification models. At the end of the process, the system generated a list of tuples containing all the identified ontological elements. Each type of elements has a specific tuple having a particular structure. For example, based on the sentence presented above, we got tuples like:

– (‘macela’, ‘is-a’, ‘INGREDIENT’)
– (‘macela’, ‘has-operation’, ‘cosido’)
– (‘coroa de rei’, ‘is-a’. ‘INGREDIENT’)
– (‘coroa de rei’, ‘has-operation’, ‘cosido’)

Then, this process was repeated for all the sentences of the texts. At the end, the tuples were prepared and exported to a graph-oriented database created in Neo4J. To visualize the ontology, we chose to use the interrogation and visualization mechanisms Neo4J provides. The native characteristics of this system, similarly to other graph-oriented database management systems, allows for receiving and exploring easily structures like

ontologies [40]. In Fig. 6, we show a fragment of the ontology graph that was generated, revealing four remedy nodes and their ingredients.

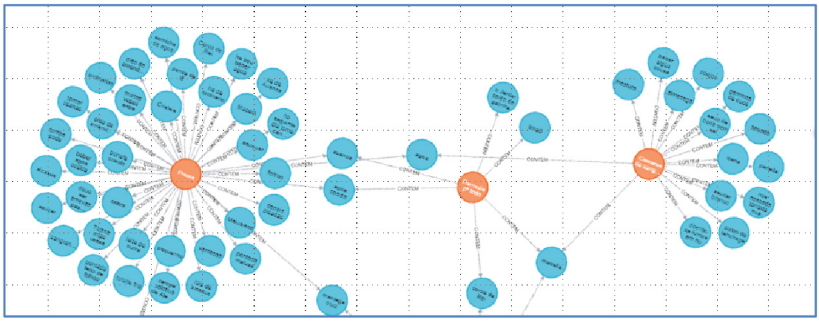

Fig. 6. A small view of the ontology create – ingredients segment

5 Conclusions and Future Work

The inception and implementation of ontology learning applications are not easy to perform. During the development of an ontology learning process a wide range of challenges raise, often creating barriers in textual data processing and problem solving tasks. However, today we have already available very specialized methods and tools that simplify a lot the process of establishing and characterizing an ontology based on knowledge extracted form texts. During the last few years, the evolution of these techniques and tool was extraordinary. Semiautomatic extraction of ontologies has great advantages when we have to deal with a large number and volume of textual data, whose treatment would not be viable with a manual approach. Automatic approaches allow for reducing the time of knowledge analysis and extraction, contributing to the reduction of the cost of the process of creating an ontology, especially in terms of human resources. However, they are not easy to apply, requiring often the use of very complex tools that require the intervention of experts of natural language processing, text mining and machine learning. Moreover, the nature of the texts, especially the content, text formats, type of writing and vocabulary used, requires a lot of work for configuring, training and tuning, in particular, the algorithms responsible for the extraction of concepts and their relationships, as well as of its subsequent characterization and semantic enrichment.

In fact, in our application case, the main challenge was placed by the nature of the texts, as they were written in classical Portuguese, using a type of writing and vocabulary that we do not use today. This makes the preparation of texts very demanding, both in terms of time, computational resources and expertise in the field of knowledge represented in the texts - XVI-XVII century's remedies recipes, collected by Portuguese during the time of the discoveries. After this phase, the ontology learning process was executed regularly. We adopted a regular work configuration, and iteratively we did all the necessary knowledge extraction tasks (concepts, relationships and axioms). As we

going discovered ontological structures, we were also correcting and improving extraction mechanisms, having particular emphasis on lexical-syntactic patterns, refining the process gradually. The current version of ontology, although not complete, reflects adequately the knowledge contained in the recipes of medicines, giving a very concrete image of what remedies contain and represent. Analyzing the texts processed and the results we got, it was possible to demonstrate the reliability and usefulness of the ontology the system provided. Now, we need to incorporate all the remedies texts we had access, extending current ontological elements as well as the representation spectrum of the domain's knowledge. Throughout the various remedies' recipes, we have verified the sporadic existence of new knowledge representation patterns. To treat and incorporate them in the process, we would need to define new lexicon-syntactic patterns for capturing and representing the knowledge gathered using them. This will be studied and implemented in a new version of our ontological system for old remedies recipes.

Acknowledgements. This work has been supported by FCT – Fundação para a Ciência e Tecnologia within the R&D Units Project Scope: UIDB/00319/2020.

References

1. Guarino, N., Oberle, D., Staab, S.: What Is an Ontology? In: Staab, S., Studer, R. (eds.) Handbook on Ontologies. IHIS, pp. 1–17. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-540-92673-3_0
2. Keet, M.: An Introduction to Ontology Engineering, University of Cape Town (2018). https://people.cs.uct.ac.za/~mkeet/OEbook/. Accessed 15 Oct 2022
3. El Kadiri, S., Terkaj, W., Urwin, E.N., Palmer, C., Kiritsis, D., Young, R.: Ontology in engineering applications. In: Cuel, R., Young, R. (eds.) FOMI 2015. LNBIP, vol. 225, pp. 126–137. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-21545-7_11
4. Noy, N., Mcguinness, D.: Ontology Development 101: A Guide to Creating Your First Ontology. Knowl. Syst. Lab. **32** (2001)
5. Drumond, L., Girardi, R.: A survey of ontology learning procedures. In: Proceedings of the 3rd Workshop on Ontologies and their Applications, Salvador, Bahia, Brazil, October 26 2008
6. Cimiano, P., Mädche, A., Staab, S., Völker, J.: Ontology Learning. In: Staab, S., Studer, R. (eds.) Handbook on Ontologies. IHIS, pp. 245–267. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-540-92673-3_11
7. Asim, M., Wasim, M., Khan, M., Mahmood, W., Abbasi, H.: A survey of ontology learning techniques and applications. Database **2018**, 1758–0463 (2018). https://doi.org/10.1093/database/bay101
8. Wong, W., Liu, W., Bennamoun, M.: Ontology learning from text. ACM Comput. Surv. **44**(4), 1–36 (2012). https://doi.org/10.1145/2333112.2333115
9. Biemann, C.: Ontology learning from text: a survey of methods. LDV Forum **10**, 75–93 (2005)
10. Belhoucine, K.: Mourchid, pp. 113–123. M, A Survey on Methods of Ontology Learning from Text (2020)
11. Hearst, M.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the Fourteenth International Conference on Computational Linguistics, Nantes France, July 1992. https://doi.org/10.3115/992133.992154
12. Barros, A.: As receitas de cozinha de um frade português do século XVI, com Prefácio de Raquel Seiça e Colaboração de Joana Veloso e Micaela Aguiar. Coimbra: Imprensa da Universidade de Coimbra, jun 2013

13. Barros, A.: Remédios vários e receitas aprovadas. Segredos vários - Edição semidiplomática e edição interpretativa do Caderno II do manuscrito 142 do Arquivo Distrital de Braga. Imprensa da Universidade de Coimbra / Fundação Calouste Gulbenkian, dec 2016. https://doi.org/10.14195/978-989-26-1282-9
14. Barros, A.: Remédios vários e receitas aprovadas (ms. 142 do Arquivo Distrital de Braga): entre a História da Medicina e a História da Língua, a Ecdótica. In: As Humanidades e as Ciências: Disjunções e Confluências. XV Colóquio de Outono, Publisher: Húmus & Centro de Estudos Humanísticos da Universidade do Minho, Eds Macedo, A., Sousa, C., Moura, V., 2014
15. Chandrasekaran, B., Josephson, J., Benjamins, V.: What are ontologies, and why do we need them? IEEE Intell. Syst. **14**(1), 20–26 (1999). https://doi.org/10.1109/5254.747902
16. Uschold, M., Gruninger, M.: Ontologies : Principles , methods and applications Ontologies : Principles, Methods and Applications Mike Uschold Michael Gruninger AIAI-TR-191 February 1996 To appear in Knowledge Engineering Review. vol. 11 no. 2, June 1996 Mike Uschold Tel : Mi. Knowledge Engineering Review (1996)
17. Borst, W.: Construction of engineering ontologies for knowledge sharing and reuse. Twente, sep 1997
18. Gruber, T.R.: Toward principles for the design of ontologies used for knowledge sharing? Int. J. Hum Comput Stud. **43**(5–6), 907–928 (1995). https://doi.org/10.1006/ijhc.1995.1081
19. Tiwari, S., Jain, S.: Automatic ontology acquisition and learning. Int. J. Res. Eng. Technol. **03**(26), 38–43 (2014). https://doi.org/10.15623/ijret.2014.0326008
20. Sharma, A.: Natural language processing and sentiment analysis. Int. Res. J. Comput. Sci. **8**(10), 237 (2021). https://doi.org/10.26562/irjcs.2021.v0810.001
21. Maheswari, M.: Text Mining: survey on techniques and applications. Int. J. Sci. Res. (IJSR) **6**(6), 1660–1664 (2017)
22. Zhang, L., Wang, S., Liu, B.: Deep learning for sentiment analysis: a survey. WIREs Data Min. Knowl. Discov. **8**(4), e1253 (2018). https://doi.org/10.1002/widm.1253
23. Jones, D., Bench-Capon, T., Visser, P.: Methodologies for Ontology Development (1998)
24. Cristani, M., Cuel, R.: A survey on ontology creation methodologies. Int. J. Semant. Web Inf. Syst. **1**(2), 49–69 (2005). https://doi.org/10.4018/jswis.2005040103
25. Uschold, M., Gruninger, M.: Ontologies: principles, methods and applications. Knowl. Eng. Rev. **11**(2), 93–136 (1996). https://doi.org/10.1017/S0269888900007797
26. Cecilia Reyes-Peña, C., Tovar-Vidal, M.: Ontology: components and evaluation, a review. Res. Comput. Sci. **148**(3), 257–265 (2019). https://doi.org/10.13053/rcs-148-3-21
27. Brank, J., Grobelnik, M., Mladenic, D.: A survey of ontology evaluation techniques. In: Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD 2005), Ljubljana, Jan 2005
28. Maedche, A., Staab, S.: Ontology learning for the semantic web. IEEE Intell. Syst. **16**(2), 72–79 (2001). https://doi.org/10.1109/5254.920602
29. Brewster, C.: Ontology Learning from Text: Methods, Evaluation and Applications. Buitelaar, P., Cimiano, P., Magnini, B. (eds.), (DFKI Saarbrücken, University of Karlsruhe, and ITC-irst), Amsterdam: IOS Press (Frontiers in artificial intelligence and appl. Computational Linguistics, **32**(4), 569–572 (2006). https://doi.org/10.1162/coli.2006.32.4.569
30. El Ghosh, M., Naja, H., Abdulrab, H., Khalil, M.: Ontology Learning Process as a Bottom-up Strategy for Building Domain-specific Ontology from Legal Texts. In: Proceedings of the 9th International Conference on Agents and Artificial Intelligence, pp. 473–480. SCITEPRESS - Science and Technology Publications, Jan 2017. https://doi.org/10.5220/0006188004730480
31. Hawthorne, J.: Inductive Logic. The Stanford Encyclopedia of Philosophy (Spring Edition), Edward N. Zalta (ed.) (2021)
32. Rabiner, L, Juang, B.: An introduction to hiden markov models. IEEE ASSP Mag. **3**(1), 4–16 (1986)

33. Honnibal, M., Montani, I.: spaCy Industrial-strength Natural Language Processing in Python (2017). https://spacy.io/api. Accessed 07 Oct 2022
34. Bird, S.: NLTK: the natural language toolkit. In: Proceedings of ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Australia, 17–21 July 2006. https://doi.org/10.3115/1225403.1225421
35. Gyorodi, C., Gyorodi, R., Pecherle, G., Olah, A.: A comparative study: MongoDB vs. MySQL. In: 2015 13th International Conference on Engineering of Modern Electric Systems (EMES), pages 1–6. IEEE, jun 2015. https://doi.org/10.1109/EMES.2015.7158433
36. Freitas, M.: Quental, pp. 1585–1594. V, Subsídios para a Elaboração Automática de Taxonomias (2007)
37. Baségio, T.: Uma Abordagem Semi-automática para Identificação de Estruturas Ontológicas a partir de Textos na Língua Portuguesa do Brasil Uma abordagem Semi-Automática para Identificação de Estruturas Ontológicas a partir de Textos na Língua Portuguesa do Brasil. pages 1–124 (2007)
38. Taba, L.: Medeiros, pp. 2739–2746. C, Automatic semantic relation extraction from Portuguese texts (2014)
39. Machado, P., Strube de Lima, V.: Extração de relações hiponímicas em um corpus de língua portuguesa, Revista de Estudos da Linguagem 23(3):599, December 2015. https://doi.org/10.17851/2237-2083.23.3.599-640
40. López, F., De La Cruz, E.: Literature review about Neo4j graph database as a feasible alternative for replacing RDBMS. Ind. Data **18**(2), 135–139 (2015). https://doi.org/10.15381/idata.v18i2.12106