IADIS International Journal on Computer Science and Information Systems
Vol. 18, No. 1, pp. 30-45
ISSN: 1646-3692

ANNOTATING UNSTRUCTURED TEXTS FOR ENHANCING SEMANTIC ANALYSIS PROCESSES

Tiago Fraga[1], Orlando Belo[1] and Anabela Barros[2]
[1]*ALGORITMI Research Centre / LASI, University of Minho, Portugal*
[2]*CEHUM, Centre for Humanistic Studies, University of Minho, Portugal*

ABSTRACT

Annotation is a powerful instrument for enhancing knowledge containing in texts. When developing a text analysis process, we often make notes for identifying and characterizing concepts and relationships, or highlighting aspects in the text that could go unnoticed by some of its readers. In addition, text annotation can enrich the semantics of texts, giving them more value through the introduction of comments, explanations, references, among many other things. Today, most text annotation processes are carried out helped by computational tools, whose functionalities make it possible to simplify the most elementary annotation tasks and substantially reduce the annotation time. The annotation of old, unstructured texts is very relevant for all those who want to study and acquire knowledge about their contents. Annotating these texts makes them more accessible to people who are not experts in the domain or in the era in which they were produced. In this work we develop a specific annotation system, supported by natural language processing and machine learning tools, to reveal the knowledge contained in the Book of Properties – "Tombo da Mitra" –, a codex containing the inventory of the Archbishop's Table of Braga's properties (Portugal) in the 17th century. This codex contains a huge amount and a wide variety of elements, containing names, nicknames, settlements, professions, types of land and buildings, among many others. All these elements are very important for studying and learning of geography, culture, economy, architecture, religion and Portuguese language until the 17th century. Annotating the Book of Properties makes possible to maintain a tag database for indexing the most relevant information contained in the book and make its knowledge accessible to a wider range of people.

KEYWORDS

Annotation Systems, Natural Language Processing, Machine Learning, Automatic Tagging, Unstructured Texts, The Book of Properties

1. INTRODUCTION

Text semantic analysis (Goddard and Schalley, 2010) (Allen et al., 2008) involves the study of the meaning of words and the way they are incorporated into sentences, the analysis of expressions contained in texts or in their application contexts, among others things. When we promote the development of a semantic analysis process, we intend to know the meaning of the text and of its sentences for obtaining useful knowledge about all the elements integrated in the texts. The type of structure and organization of the texts – structured, semi-structured or unstructured – strongly influences how to apply and develop a semantic analysis process.

The area of semantic analysis of unstructured texts (Sinoara et al., 2017) has evolved a lot in recent years. An unstructured text does not have a uniform organization or format. Its content may adopt different forms of expression throughout the text. Unstructured texts are quite common and can be found in large quantities in most application domains. The characteristics of these texts, often written in natural language, raise many problems and challenges in the development of any semantic analysis initiative, whether manual or automatic.

The increasing interest of researchers from various scientific domains, in knowledge contained in documents available in archives, libraries, databases, or in the Web, promoted the development of much diversified solutions for the semantic analysis of these documents. The great emergence of tools and applications, particularly in the fields of text interpretation or content retrieval and extraction (Cornolti et al., 2013) (Chu et al., 2012) clearly proves this evolution. Some of these solutions involve researching and application of text annotation techniques, structured or not, as a way of revealing and tagging, manually or automatically, textual elements with a particular semantic meaning.

The need to make this annotation process more flexible, as well as to increase its performance and effectiveness, led to the development of systems to help the annotation of texts, both in initiatives that aimed at manual annotation and in other, more interesting and challenging ones, the use of automatic annotation systems. These systems (Moraes and Lima, 2008) include several mechanisms that are capable of identifying concepts and relationships using a combination of text mining, natural language processing and machine learning techniques. In practice, this makes possible to develop text analysis systems — written in natural language — that are capable of (semi)automatically annotating sets of words, as well as their respective application contexts, identifying concepts, characterizing them and establishing possible relationships. The extraction of concepts from texts is fundamental for any information retrieval process. However, it is difficult to perform, since obtaining annotated elements from a set of texts is a very complicated task. Automatic text annotation systems perform tagging tasks with a greater degree of speed and efficiency compared to conventional manual annotation processes.

It is very important that annotations (and correspondent tags) are meaningful, especially in the case of unstructured texts, for enhancing text semantics analysis as well as providing additional elements for reinforcing text's knowledge. Annotations must reflect the kind of semantics we are dealing with. Besides express explanations or comments, annotations must also categorize and classify what is annotated, and establish when possible relationships among the elements annotated and categorized. To perform all these tasks we need to have specific experts and sophisticated tools. For cases involving a large volume of texts, manual annotation, even carried out by specialists with great experience, has become a task too expensive in time and resources (Cai and Hofmann, 2003). This is not acceptable, because it is very slow and inefficient, taking into account the operational and analysis requirements of current specialists and scholars.

In research and analysis of texts content, the use of natural language processing and machine learning techniques allow for preparing texts, identifying sets of relevant words, establishing research and relationship patterns, or discovering concepts and their relationships. All these elements are very useful for establishing definition tags for text contexts. Furthermore, these techniques eliminate, or at least mitigate, most of the disadvantages associated with the preparation and annotation of texts manually, namely, subjective, time-consuming, inconsistent, and others. However, the use of these techniques does not remove all manual tasks in the general annotation process, such as, for example, checking and validating tags, for adjusting parameters and improving the quality of the annotation process.

In this work, we present an automatic annotation system for unstructured texts, which we developed for tagging texts of a very singular document: the Book of Properties (Barros, 2019) (Barros, 2021). This manuscript of the 17th century contains the inventory of the properties of the Archbishop's Table of Braga (Portugal) in the beginning of that period. It describes in detail all the rustic and urban properties of several districts located mostly in the north of Portugal, as well as presents the rents and payments due to their lease.

The annotation system was developed to be integrated into a document management system, which was conceived to accommodate the contents of the Book of Properties, with the aim of improving the research and analysis processes of the book's texts, as well as revealing the various relationships that exist between its different textual elements. The annotation of the document database will allow for the creation of a set of relevant tags – indexed, discovered and established –, based on anthroponyms, toponyms, and degrees of kinship, properties and their location, among others. In this way, it is possible to maintain a very diverse set of tags especially oriented for indexing the most relevant information contained in the book. Additionally, based on the specification of the tags created, the annotation mechanisms will allow for the analysis of documents stored in the system and, similarly, it will be able to suggest a global annotation strategy for these tags, as well as generate a relationship map for discover similar contents or renewing their definitions over time.

In this paper, we will present and discuss the work carried out in the development of the automatic annotation system we referred, giving particular attention to all the aspects related to the automatic annotation process conceived and implemented. We will approach system's main operational tasks, namely selecting texts, tagging texts, validating tags and learning new terms, illustrating some of the results we achieved. The remaining part of this paper is structured as follows. Section 2 describes some related work in the area of text annotation, while section 3 exposes and describes the system we developed, and how it works in each execution stage. Finally, section 4 presents some conclusions and future work.

2. ANNOTATING TEXTS

A text can be seen as a set of sentences that convey some kind of information to its readers. However, if there are no additional information elements, markings or notes throughout the text, its interpretation will differ more than expected between different readers. This means that a particular word or sentence can convey different ideas to different readers, especially in manuscript texts from past centuries. Text annotation (Gosal, 2015) is a task that adds value and specification to texts. It is possible to provide additional elements about the text, using a well-defined set of tags for helping readers in its interpretation. For example, when analyzing a conventional text, numerous punctuation marks are easily to found. They correspond to a set of marks, whose function is to help readers of that text in their reading and interpretation; however, these marks are quite different in old unpublished manuscripts. Tags also have this function, being defined according to the content of the texts. Usually, we call them as "content-oriented annotations" (Ferreira, 2011). The use of tags intends to indicate the presence of different types of elements in the text, such as people names, nicknames, place names, professions, or products, among many other information elements. The greater the presence of labels in a text, the greater the percentage of annotated text. Consequently, the easier it is the interpretation and understanding of the text. Text annotation is often performed in an ad hoc manner, which obviously does not reveal to the reader the true usefulness of this technique. However, when we done the annotation process methodically, and guided by the context and content of the text, highlighting its main ideas, the understanding of the texts increases significantly. Thus, text annotation can help readers to analyze their own ideas and thoughts, allowing them to have a direct and faster access to the most pertinent ideas and elements of a text (Lynch, 2021).

Automatic text annotation systems carry out annotation tasks with a greater degree of speed and efficiency when compared to manual annotation processes. Text automatic annotation can be performed using different techniques and mechanisms, depending on the objectives of the work we intend to do. Some of the best-known variants are entity annotation, intent annotation, sentiment annotation, and semantic annotation (Khan, 2022). The process of annotating entities (Dozier et al., 2010) aims at labeling words related to proper names or other types of categorization of textual elements, in order to detect elements that correspond, for example, to names, dates or locations. This type of annotation is often used in social media environments applications. We have also the annotation of intentions (Maestre, 2022), whose objective is, as its name indicates, to discover the intentions of a customer, for example, that in some way may be expressed in a text and categorize it according to a certain set of previously established requirements. References to intentions in a text are very valid and useful elements, which help to determine the type of content of a text. Usually, this type of annotation is used in human-machine conversation systems for recognizing the purpose of a given conversation with a human agent. Other type of annotation is the annotation of feelings (Mohammad, 2016). It is also an area of great interest, in particular for companies that want to know the reaction of people who have purchased goods and products, or requested any service. It involves labeling opinions and feelings of people. However, this type of annotation is difficult to implement, since the identification and categorization of feelings are operations of great complexity. This type of annotation involves very sophisticated machine learning models, which act on texts extracted from review sites, discussion lists, social networks or emails. Finally, we have the annotation of texts with that is used for text semantic analysis (Allen et al., 2008). In this type of annotation, words, phrases or texts are associated with terms we want to annotate. Sometimes, we also associate references to entities or other objects, in order to enrich the contents expressed in the texts, giving them greater relevance and meaning.

In recent years, a lot of work was made in the field of semantic analysis of unstructured texts, which has given rise to many applications, in areas so diverse as text interpretation or content retrieval and extraction. Perhaps the main reason for this development was the high number of documents available in companies, libraries, databases and websites. For this volume of data, manual annotation of texts is not feasible, as it becomes too exigent in terms of time and money, requires the involvement of specialized human resources, and often is not carried out in a methodical and efficient way, which may cause errors and failures in annotation processes (Cai and Hofmann, 2003). For these reasons, many researchers have promoted the development processes of automatic annotation systems, as a viable way for a large part of the annotation problems in unstructured texts (Chu et al., 2012). These initiatives quickly revealed a significant increase in the speed of execution of the annotation process and in the efficiency rate of recognition of textual elements of interest, in the identification and extraction of entities in texts, as well as dispensing with a large part of non-specialized human work. The development that has taken place in recent years in the areas of natural language processing, machine learning and text mining also contributed to this. The combination and use of tools from these areas makes possible to develop systems for analyzing structured or unstructured texts, written in natural language, taking note of words, contexts of use, concepts and their relationships. In some cases, they evolve as they do more and more annotation processes, that is, they have the ability to learn. We are, therefore, in another dimension of text annotation. However, they remain complex systems and difficult to implement (Finlayson and Erjavec, 2017). Despite these difficulties, the interest in automatic annotation systems by the research community increased significantly, according to the results observed in the applications that were being developed (Cornolti et al., 2013). Today, we can say that the growth in the use of these tools and applications is remarkable.

The application spectrum of automatic annotation systems is huge and much diversified. For example, they can be used for:

- establishing user behavior patterns or discovering trends in the purchase of goods and services application of e-commerce sites or social networks;
- exploring discursive structure and identify discursive categories in texts in automatic summarization operations (Blais et al., 2007);
- signing annotation languages videos for reducing processing time and the subjectivity of manual annotations (Chaaban et al., 2021);
- annotating of adenomatous polyp pathology in the gastrointestinal tract (Selnes et al., 2022);
- recognizing and classifying entities in German-language texts (Benikova et al., 2010).

Regardless of the development context and application area, an automatic annotation system must work according to a robust annotation scheme, which clearly defines the annotation strategy and rules, as well as the different tags to use. This work must be carried out with the support of specialists in annotation and in the field of application of the texts. In addition, we need to select suitable computational tools for the annotation process, which allow for the recognition and division of words, identifying and classifying later concepts and their relationships.

Several academic and business initiatives have made significant efforts in the development of programs that help automatic annotation systems and their application in areas of great interest. For example, the:

- EXACT system (Chen et al., 2019), developed at the University of Zhejiang, China, allows extracting entities from textual documents, through exploratory annotation operations that create attributes and associate them with tags; furthermore, the system may provide to a machine learning system all the entities identified, as well as present text annotation recommendations in real time, and support user interaction to validate the suggested annotations; after validating annotations, the system can improve its own annotation algorithm using new data to learn.
- Elketron system (Refinitiv, 2023), developed in an industrial context by Refinitiv, which has a concrete application in the areas of economics and financial markets, being used widely in banking systems and applications; the system has also a specific component for intelligent annotation of texts, which uses natural language processing techniques to get the meaning of large contents of unstructured texts – this component makes the process of identifying people, places, facts or events faster.

- Tagtop system (TagTop, 2022), which provides means for training a machine learning model for personalized annotation; this system also allows for the annotation of entities in texts, through an adaptable interface that can be used to visualize only the tools user needs for the annotation process; it also provides means for the implementation of automatic annotation processes.
- UBIAI system (UBIAI, 2022), which is capable of performing Named Entity Recognition (NER) (Marrero et al., 2013), relation extraction and document classification tasks; additionally, the system platform provides as well other services related to auto-labelling, document classification and multi-lingual annotation.

In addition to these examples, there are companies such as Uber, Apple or Microsoft using several types of annotation procedures to analyze their reputation in the field of social networks. Over the years the interest of companies for annotation system increased significantly, provoking the emergence of a large variety of systems and applications especially oriented for annotating text. Annotation has been used for establishing user behavior patterns, discovering trends, improving quality of service, especially in social networks and e-commerce sites (Win and Aung, 2018). The actors of these application areas are very competitive, being permanently interested in gaining advantages that allow them to impose themselves on the market in which they are operating. Annotating the texts produced by its users and customers in companies' web sites, for example, allows for enriching the knowledge about the activities they carry out and improve their business processes and their activities.

Regardless of the development context and application area, good annotation practice recommends the definition of a robust annotation scheme, which allows for a clear definition of how to annotate text or to select tags to be used. Furthermore, it is important to note that any text annotation process requires the involvement of several people and, in particular, of specialized tools capable of providing functionalities for recognizing entities, concepts and relationships, and for word division, just to name a few. In short, we must bear in mind that annotating a text is not a simple task.

3. THE ANNOTATION SYSTEM

Usually, we develop an annotation system for very specific application domains, in order to facilitate the processes of interpretation and research the contents of texts. Although, we can do manually a text annotation process, its automation is highly desired, since, when possible, it simplifies the annotation process and drastically reduces its completion time, as already referred. Furthermore, when properly configured, it automatically establishes relationships that exist between discovered (and annotated) tags and the various textual elements that may be associated with them. The construction of an automatic annotation system requires the use of natural language processing mechanisms, commonly combined with machine learning mechanisms. This combination of technologies enables the development of systems capable of analyzing texts in natural language, with the ability to identify sets of words and create contexts of use, according to an established set of prerequisites, as well as discover and manage tags through techniques of word processing, more expeditiously and with a very high level of correction.

3.1 The Application Domain

Over the last few years, we have been developing a document management system (Barros et al., 2020) to accommodate the content of the Book of Properties, an impressive manuscript from the 17th century, which contains a detailed inventory of the rural and urban properties of the archbishops of Braga (Portugal). According to Barros (2021), the manuscript, having 1288 large pages, presents in detail and precision the inventoried properties of the Portuguese counties of Valença, Vila Real, Chaves and Braga, and also some properties in Porto and Santarém, in Portugal, and Galicia, in Spain. Due to the detail of this information, it is possible that researchers or readers of the Book of Properties, if they have roots in these locations, can find references or properties of their ancestors.

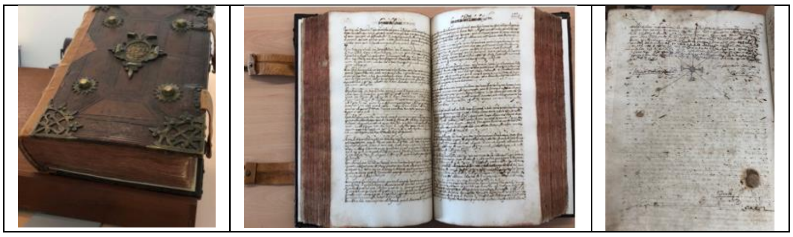


Figure 1. Some images of the Book of Properties

The codex (Figure 1) has a very impressive size, weight and binding, being handwrite in a very cultured and regular calligraphy, and presenting a rigorous, detailed and extensive description of all the properties, rents and pensions of the Archbishop's Table of Braga, as well as very interesting information about agricultural data, fauna and flora. Additionally, the Book of Properties provides information related to the properties inventoried – confrontations with the surrounding lands (and houses), owners, characteristics of the lands, types of cultivation, etc. –, which constitutes detailed knowledge about the country, the people and the Portuguese language in the 17th century.

Only philologists, or researchers with experience in reading this type of documents, can read and interpret the information contained in the book. Other scholars, without this experience, may have difficulties in reading and handling the manuscript. In addition, the fact that the codex is not reproduced in digital format, and can only be consulted in person at the Braga District Archive, means that only one person can study it at a time. On the other hand, people far from the city have to travel to consult the book, which can sometimes lead to the abandonment of their study. With the increase in the number of researchers and due to the importance of the codex, the opportunity arose to store the contents of the manuscript, edited since 2015, in a digital format, overcoming all the difficulties for the study of the book mentioned.

The document management system we implemented is a Web based application having the capacity to store, index and search the edited texts of the Book of Properties. It is a cross-platform system designed to run on the most relevant operating systems on the market. It is a client-server system, in which the backend received the native document management services, and the client sustain the user interaction and document loading and search services, running in conventional Web browsers. In practice, it is a web based application specially designed for the reception, processing and analysis of the texts contained in the Book of Properties, which has a set of specially created mechanisms for importing, cataloguing, modifying, removing, and analyzing texts stored in a document store. In addition to these functionalities, the document management system incorporates georeferencing mechanisms that use a set of specific tags, previously defined by a particular annotation module, which allows for indexing places, lands and properties referred in the Book of Properties, using a set of geographical coordinates of a given map (Gomes et al., 2021).

3.2 The Annotation Mechanisms

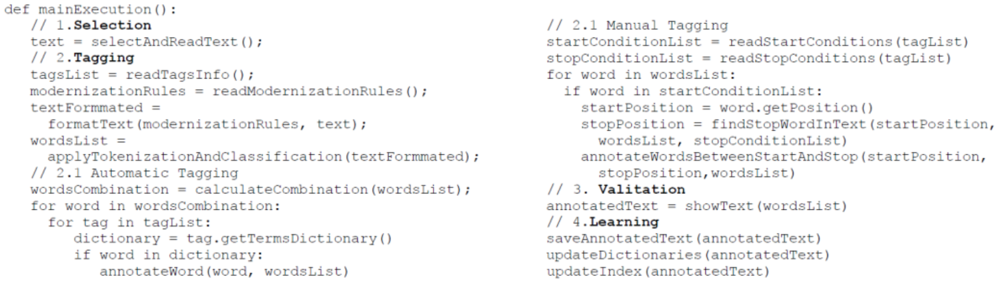
In order to improve the research and analysis of texts of the document management system developed, as well as to reveal the various relationships between its various textual elements, a set of mechanisms, specifically oriented for annotating the document database, was incorporated into its structure. With the introduction of such mechanisms, the system received new means for incorporating tags into texts, which we consider relevant for the study of the Book of Properties, discovering and cataloguing anthroponyms, toponyms and microtoponyms, degrees of kinship, products, plants, properties and their location, etc. In this way, it was possible to maintain a base of tags indexing the most relevant information contained in the book. Furthermore, based on the tag specifications, the annotation mechanisms allow for analyzing the texts that are contained in the system and, similarly, suggesting a global annotation strategy for other tags, as well as generating a map of tag relationships that can be used to discover similar content.

The quantity and diversity of the elements mentioned in the Book of Properties are surprising: all the names and surnames, nicknames, villages, professions, types of land and buildings, natural resources, cultures and ways of cultivation, among many others, are very important for the study and learning of geography, culture, agriculture, economy, architecture, religion and Portuguese language in the 17th century. The annotation of these elements expressively reveals their location in time and space, as well as their potential relationships, facilitating the study of the book and providing researchers, linguists, teachers and students with a valuable tool to reach and reinforce the knowledge about the codex.

When we made a detailed study of several existing automatic annotation systems – e.g. Benikova et al., 2010), Ahmadi and Moradi (2015), Chen et al. (2019) or Dias et al. (2020) – we analyzed their most relevant models and features, as well as identified their functional architectures. In general, all these systems had very similar elements. Combining some of these models and their respective features, we sketched the working model of our automatic annotation system. Despite of its design having in mind its application to the Book of Properties, we consider that the system obtained is of widespread application, incorporating the most relevant functionalities that should be present in an automatic annotation system. The automatic annotation process we developed considers four working stages, namely:

1) Selection, where we made the selection of the texts to be annotated for future analysis.
2) Tagging, which is responsible for processing and annotating the texts selected in the previous stage.
3) Validation, for verifying the annotations made in the previous stage with user supervision.
4) Learning, where the system updates the internal structure of the dictionaries, using machine-learning algorithms to readjust and enrich the annotation database.

Figure 2 presents an algorithm, in pseudo-code, which shows the most relevant operations performed by the system in each execution stage. This algorithm briefly describes the sequence of operations (modernization, tokenization, classification, dictionary analysis, etc.) involved in an annotation process of a text and, in some cases, the different parameters required by each of these operations. In the next section, we will describe in detail each one of the mentioned stages.



```
def mainExecution():
    // 1.Selection
    text = selectAndReadText();
    // 2.Tagging
    tagsList = readTagsInfo();
    modernizationRules = readModernizationRules();
    textFormmated =
      formatText(modernizationRules, text);
    wordsList =
      applyTokenizationAndClassification(textFormmated);
    // 2.1 Automatic Tagging
    wordsCombination = calculateCombination(wordsList);
    for word in wordsCombination:
      for tag in tagList:
         dictionary = tag.getTermsDictionary()
         if word in dictionary:
            annotateWord(word, wordsList)
    // 2.1 Manual Tagging
    startConditionList = readStartConditions(tagList)
    stopConditionList = readStopConditions(tagList)
    for word in wordsList:
      if word in startConditionList:
        startPosition = word.getPosition()
        stopPosition = findStopWordInText(startPosition,
          wordsList, stopConditionList)
        annotateWordsBetweenStartAndStop(startPosition,
          stopPosition,wordsList)
    // 3. Valitation
    annotatedText = showText(wordsList)
    // 4.Learning
    saveAnnotatedText(annotatedText)
    updateDictionaries(annotatedText)
    updateIndex(annotatedText)
```

Figure 2. The annotation process in pseudo-code

3.2.1 Selecting Texts

This first stage (selection) of the annotation process is the simplest part of the system. Here, the system simply displays the texts that are available for annotation. Texts were previously inserted in the document store of the document management system, being stored in a specific collection of documents in JSON format. In addition to the text, JSON documents also contain some other elements, namely the identification, a summary description, the type of the page, the version of the text (interpretive or semi-diplomatic) and a set of word indexes.

When selected, texts are loaded into memory from a specific data collection in the system's document store. This step is not very demanding in terms of computational resources (memory and processor). Despite the large number of texts in the Book of Properties, each one of them, individually, does not have a large dimension, which makes the annotation process simple and fast.

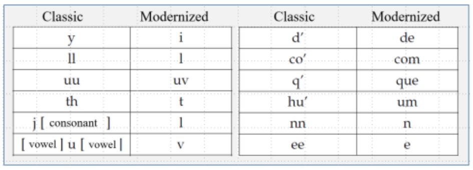
3.2.2 Tagging Texts

During the tagging stage, the system performs five very distinct specific tasks on the text being processed, namely its classification, modernization, annotation using dictionaries (automatic tagging), annotation using NER, and, finally, aggregation. In the first task, classification, the system divides the words and aggregates them in a specific data structure, for using them later in other annotation tasks. This data structure receives the classification of words and stores the position of each of them in the text, as well as their homogenized form (in lowercase and without accents). The homogenized form is used throughout the dictionary search process, along with a logical value, which informs whether the word has already been annotated or not. The division and aggregation of the words in the text were implemented with the natural language processing tool LinguaKit (Gamallo and Garcia, 2017). This tool provided us mechanisms for diving text into words and making its morphological classification, so that it was possible to verify if a word would be a noun, an adjective, a verb, an adverb, etc. – the identification of these elements is essential in any annotation process. Then, the system performs some modernization of the orthography.

As previously mentioned, the Book of Properties was written in classical Portuguese. The dictionaries of place names we used, for example, are in contemporary Portuguese, which imposes updating the spelling of the text. As most words do not follow current orthographic standards, and present multiple variants, or graphic forms, the detection of entities is difficult. To overcome this difficulty, we developed an automatic process for modernizing words, which works based on dictionaries and on a set of previously established lexical updating rules (Table 1). These rules match the classical patterns with the corresponding contemporary equivalents. The results of the word modernization process, the updated text and the data structure of the conversion dictionary, are stored in the system's document store, so that in all the time, whenever necessary, it will be possible to reverse the conversion (modernization) carried out and recover the original text. In Table 2 we can see an example of an original sentence extracted from a text (line 1) and the sentence that resulted from the modernization process (line 2).

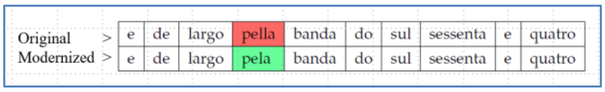
Tagging using dictionaries is the next task of the annotation stage. Here, the system annotates entities from the word dictionaries referring to each of the tags that we want to use. For example, for tagging locations, it was necessary to create a specific dictionary – a gazetteer – with all locations in the Portuguese territory, with particular emphasis on the initial definition of annotation of locations, so that, in the future, we could apply the same or similar strategy to other types of tags. In this particular case, the dictionary was built from a specific dataset obtained from the Portuguese Public Administration Open Data portal (Dados.Gov, 2022), which disposes information about all districts, municipalities and parishes in Portugal. For the other types of tags established as essential, such as people names, types of land and houses, products, and others, we implemented similar processes.

Table 2. Some examples of spelling update rules



Classic	Modernized	Classic	Modernized
y	i	d′	de
ll	l	co′	com
uu	uv	q′	que
th	t	hu′	um
j [consonant]	l	nn	n
[vowel] u [vowel]	v	ee	e

Table 2. A modernized sentence from the Book of Properties



Original >	e	de	largo	🟥 pella	banda	do	sul	sessenta	e	quatro
Modernized >	e	de	largo	🟩 pela	banda	do	sul	sessenta	e	quatro

Finally, the system compares the words contained in the text with the terms registered in the tag dictionaries that are stored in the system and maintained over time. The comparison process used only words in lower case, without accents. With this strategy, it was possible to cover a greater number of cases, which was essential for increasing the level of accuracy and effectiveness of the tagging using dictionaries. Next, we have the tagging using NER task. In this task, the system identifies elements that are not present in the system dictionaries and to which no tag has been associated. NER allows for identifying and recognizing entities in the text.

As mentioned, the Book of Properties is a manuscript from the early of 17th century. As such, its texts contains references to names of people and places, lands or professions, which have changed over the years. The book also contains words that we do not use anymore, and thus automatically escape from the previous annotation process. For example, there are references to locations in the manuscript whose names and writing have simply changed. Cases like these are not present in the dictionaries. To handle them, we had to develop a process that was able to detect and record them properly, in a manual way. Thus, in this new process, we implemented a mechanism to verify the beginning and the end of class indicators. These words use to precede or belong to the term that we need to identify. For example, when processing words from the text, as soon as the system detects a start-of-class indicator it is quite likely that following words can be categorized by a specific tag. Once detected the starting point for a possible annotation, the system determines the different stopping points. When the system finds, for example, a determiner, a verb or an adjective, it defines a breakpoint. If a preposition was detected, the system checks the next word, which, if it belongs to the grammatical class of the stopping cases, determines the stop of the annotation of that case. Then, the system reuses the classification mechanism from the previous task, organizing the words in the text according to their morphological class. This allows for finding stopping cases contained in the texts and detect potential entities to annotate. At the end, we got an annotated text, having notes such as: ". . . e do nascente e sul com <name>Dona Maria </name> de semeadura" (in Portuguese).

In this small example, the annotated name was identified based on the class start identifiers, which are the words that mark the beginning of the entity search, which in this case is "Dona". In turn, the stopping case is the preposition "de", which marks the end of the search. After identifying the referred limits, the system is able to identify the name "Dona Maria" and write it down correctly with the respective tag (<name>). After done the tagging tasks (automatic and manual), the system proceeds to the aggregation of annotated (and unannotated) words task. All the words were stored in a temporary JSON document, having a structure owning the attributes "word", "tag", "position", "annotated", "type" and "color". The "word" and "tag" attributes represent the form of the word, which may or may not contain a tag. For example, in an unannotated word, both attributes will have the same value. However, the same is no longer true for an annotated word. In this case, the "word" attribute will contain the word "surrounded" by the tag. As for the other attributes, they indicate, respectively, the position of the word in the text ("position"), whether or not the word is annotated ("annotated"), the type of tag used ("type") and the color that identify the tag in the text ("color").

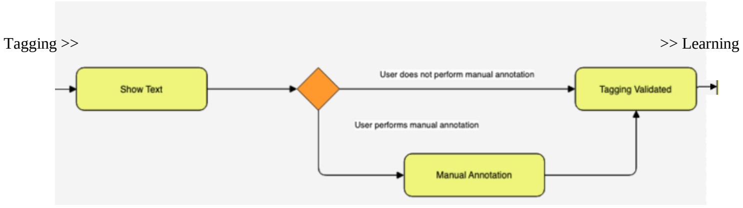

Tagging >>
>> Learning
Show Text
User does not perform manual annotation
Tagging Validated
User performs manual annotation
Manual Annotation

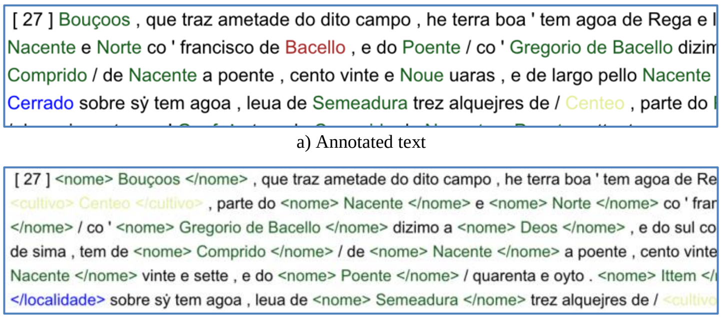
Figure 3. Validation tasks

3.2.3 Validating Tags

During the tag validation stage (Figure 3), the system verifies the annotation made in the previous stage, having user supervision. Based on the results of the previous stage, the system displays the tags defined and applied to the text (Show Text). The user confirm the tags that have been defined, or cancel the validation process, which will cause the system to discard all the annotation work performed. The reviewing tasks do not require large computational resources. However, it obviously requires some work from system user. Additionally, the user can make new manual annotations or change one or more annotations of entities made previously (Manual Annotation). This process happens mainly when user wants to correct a tagging error or add new tags, in case of annotation failure. After validating the annotations (Tagging Validated), with or without manual changes, the system proceeds to the last stage: learning.

3.2.4 Learning and Incorporating Tags

Learning is the last stage of the system. The first task of this step stores the text previously annotated in the system document store. Thus, any user will be able to view the annotated text, without having to perform any annotation task, in any access they make to the system. In a second task, the system updates the various tag indexes it has, identifying the occurrences of each of the tags in the texts. This task indicates the texts in which we can find references to relevant data elements, such as localities, professions, terrains, etc. Finally, the system updates tag dictionaries that it usually uses in the automatic annotation task. Updating dictionaries over time, after each annotation process carried out, we can say that the annotation system has some learning capabilities. This is because, during the annotation process, new entities can be discovered, from the manual annotation task or from annotations made by users in the validation state. As result, the system acquires and reflects appropriately in its storage structures that tagging actions, enriching the dictionaries and improving its own annotation process in future iterations. In Figure 4 we can see two tagging views for a (fragment of a) text: a) the annotated text, having the words annotated colored, and b) the text annotated, with the annotated words and their respective tags in sight. In this state of the annotation process, if necessary, it is possible to recover the original text. The system's document store maintains the various versions of the text created during annotation process.



[27] Bouçoos , que traz ametade do dito campo , he terra boa ' tem agoa de Rega e l
Nacente e Norte co ' francisco de Bacello , e do Poente / co ' Gregorio de Bacello dizim
Comprido / de Nacente a poente , cento vinte e Noue uaras , e de largo pello Nacente
Cerrado sobre sẏ tem agoa , leua de Semeadura trez alquejres de / Centeo , parte do

a) Annotated text

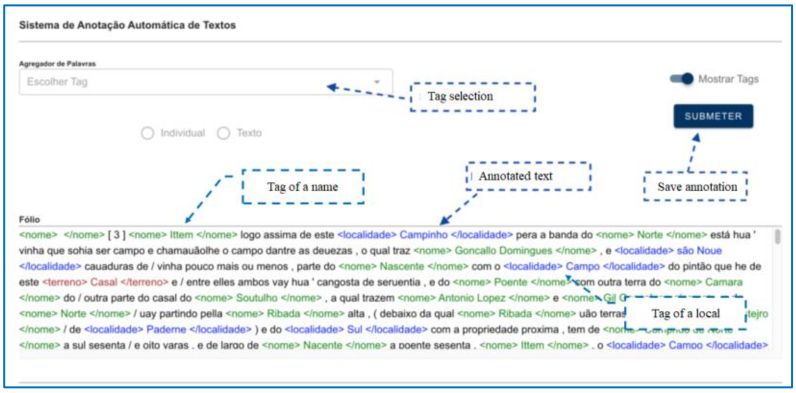
[27] <nome> Bouçoos </nome> , que traz ametade do dito campo , he terra boa ' tem agoa de Re
<cultivo> Centeo </cultivo> , parte do <nome> Nacente </nome> e <nome> Norte </nome> co ' frar
</nome> / co ' <nome> Gregorio de Bacello </nome> dizimo a <nome> Deos </nome> , e do sul co
de sima , tem de <nome> Comprido </nome> / de <nome> Nacente </nome> a poente , cento vinte
Nacente </nome> vinte e sette , e do <nome> Poente </nome> / quarenta e oyto . <nome> Ittem </
</localidade> sobre sẏ tem agoa , leua de <nome> Semeadura </nome> trez alquejres de / <cultivo

b) Annotated text showing tags

Figure 4. Two distinct views of a text fragment

3.3 Integrating the Annotation System

The integration of the annotation system into the document management system involved the creation of a new set of software components, which were specifically developed to host the new modules responsible for the annotation processes and to support interaction with system’s users. In particular, these last functionalities involved the creation of a new set of services to manage updating rules and tags implemented in the annotation system and the modernization of the texts. After the implementation of these services, it was possible to start and develop the annotation of the texts included in the Book of Properties.



<nome> </nome> [3] <nome> Ittem </nome> logo assima de este <localidade> Campinho </localidade> pera a banda do <nome> Norte </nome> está hua ' vinha que sohia ser campo e chamauãolhe o campo dantre as deuezas , o qual traz <nome> Goncallo Domingues </nome> , e <localidade> são Noue </localidade> cauaduras de / vinha pouco mais ou menos , parte do <nome> Nascente </nome> com o <localidade> Campo </localidade> do pintão que he de este <terreno> Casal </terreno> e / entre elles ambos vay hua ' cangosta de seruentia , e do <nome> Poente </nome> com outra terra do <nome> Camara </nome> do / outra parte do casal do <nome> Soutulho </nome> , a qual trazem <nome> Antonio Lopez </nome> e <nome> Gil G <nome> Norte </nome> / uay partindo pella <nome> Ribada </nome> alta , (debaixo da qual <nome> Ribada </nome> uão terras tejro </nome> / de <localidade> Paderne </localidade>) e do <localidade> Sul </localidade> com a propriedade proxima , tem de <no </nome> a sul sesenta / e oito varas . e de larqo de <nome> Nacente </nome> a poente sesenta . <nome> Ittem </nome> . o <localidade> Campo </localidade>

Figure 5. A screenshot of the annotation environment

The annotation process proceeds as described in the previous section. However, to start, the user must select the text to annotate from a list of documents that the system presents on demand. The selected text is then processed by the automatic annotation system. Then, the result is revealed to the user in a specific environment (Figure 5). In this environment, the user can see and analyze all the elements that the system automatically annotated. These elements are revealed through their respective tags, which appear in the preview of the selected text. Tags are represented using their respective names and colored according to their categorization.

Tag categorization is carried out in advance, using a specific parameterization program that allows to define the state of the tag (active or inactive), the color to apply to the tag in the annotation environment, or the modification of the dictionary of words or the class indicators, among other things. These last two parameterization options involve the insertion of new words in the system that can be useful for the annotation processes, whether they are automatic or manual. In the annotation environment, the user can add new tags or change other tags previously defined by the System automatically. If new tags are defined by the user, they will be saved in the system and applied, if justified, in the next annotation processes. If the user wants to apply these new tags to other texts that have been previously annotated, he can ask the system to re-annotate these texts.

4. CONCLUSIONS AND FUTURE WORK

Automatic text annotation systems are sophisticated tools for enriching text contents. They can be used for helping users to extract and annotate relevant information in unstructured texts. Usually, these systems are developed for very specific application domains, being designed and configured accordingly the nature of the texts and the annotation process we want to apply. In this paper, we presented an automatic annotation system specially designed and implemented for tagging the texts of the Book of Properties. The system provides a set of services for the automatic identification and annotation of various data elements referred in the book, such as names, locations, professions, terrains, as well as other elements with high historical interest of a large Portuguese territory. Such elements are very important to researchers, professors and students, for studying sociocultural, economic, architectural, agricultural, linguistic and religious aspects of the 17th century in the north of Portugal. At this point, we have available an operational version of the system, working as planned. However, as expected, the implementation system was not simple. It raised a very diverse range of challenges, ranging from modelling to tagging. It was difficult to integrate a supervised machine-learning module into the system, due to the lack of data for training the model we designed.

The Book of the Properties was written in classical Portuguese. This circumstance means that its texts are not suitable for training the systems. Furthermore, the editing of this codex, from handwritten to digital form, is still in progress. It is a very time-consuming task, and at this moment the edition, already resulted in four books having more than two thousand pages. During the design and development phase of the system, the number of texts stored in the system was not large. Even so, we developed the bases for training and implementing the tagging system, obtaining a nice set of annotated texts. This allowed us to demonstrate the usability and utility of the system.

As future work, we intend to increase significantly the number of texts in the system's document store (a task wich is recomended only when the last edition criteria will be definitely fixed, after the completion of the codex edition, since it will affect all the text orthography), in order to improve the automatic annotation techniques, with new tagging services and learning mechanisms, increasing system robustness and performance. With this, we expect to open the use of the system to a selected community of researchers, professors, and students.

ACKNOWLEDGMENT

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

REFERENCES

Ahmadi, F. and Moradi, M. (2015). A Hybrid Method for Persian Named Entity Recognition. *7th Conference on Information and Knowledge Technology, IKT 2015*, May. DOI: https://doi.org/10.1109/IKT.2015.7288806.

Allen, J., Swift, M. and Beaumont, W. (2008). Deep semantic analysis of text. *Proceedings of the 2008 Conference on Semantics in Text Processing (STEP '08)*. Association for Computational Linguistics, USA, 343–354, 2008.

Barros, A. (2019). Apontamentos lexicais sobre o Livro das Propriedades ou Tombo da Mitra Arquiepiscopal de Braga: designações de terras e outros aspectos das propriedades. In *Estudos de linguística histórica: mudança e estandardização*, Coimbra: Imprensa da Universidade de Coimbra, pp. 393-428.

Barros, A. (2021). A edição do Livro das Propriedades ou Tombo da Mitra Arquiepiscopal de Braga. *Os sete castelos. Congresso de Homenagem a D.Rodrigo de Moura Teles*, Braga.

Barros, A., Belo, O., Gomes, J., Fraga, T., Martins, R. and Carvalho, J.P. (2020). A Computational Instrument for Students Accessing and Exploring The Book of Properties of The Braga Archbishop's Table (17th Century), *Proceedings of 13th Annual International Conference of Education, Research and Innovation*" (ICERI'2020), 9th-10th November.

Benikova, D., Yimam, S., Santhanam, P., and Biemann, C. (2010). *Germa-NER: Free Open German Named Entity Recognition Tool*. 1(1): 31–38.

Blais, A., Atanassova, I., Descles, J., Zighem, L., Zhang, M. (2007). Discourse Automatic Annotation of Texts: an Application to Summarization. *Proceedings of the 20th International FLAIRS Conference*, January.

Cai, L. and Hofmann, T. (2003). Text Categorization by Boosting Automatically Extracted Concepts. *SIGIR '03 Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, edited by ACM, 8. Toronto, Canada.

Chaaban, H., Gouiffès, M., and Braffort, A. (2021). Automatic Annotation and Segmentation of Sign Language Videos: Base-level Features and Lexical Signs Classification. *Proceedings of 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2021*), February, France.

Chen, K., Feng, L., Chen, Q., Chen, G. and Shou, L. (2019). EXACT: Attributed Entity Extraction By Annotating Texts. In ACM, editor, *SIGIR'19 Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 4, Paris, France. DOI: https://doi.org/10.1145/3331184.3331391.

Chu, B., Zahari, F. and Lukose, D. (2012). Benchmarking T-ANNE: Text Annotation System. In ACM, editor, *i-KNOW '12 Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies*, page 5, Graz, Austria. DOI: https://doi.org/10.1145/2362456.2362464.

Cornolti M., Ferragina, P. and Ciaramita, M. (2013). A Framework for Benchmarking Entity-Annotation Systems. *WWW '13 Proceedings of the 22nd international conference on World Wide Web*, page 11, Pisa, Italy. University of Pisa, Italy, ACM. DOI: https://doi.org/10.1145/2488388.2488411.

Dados.Gov, 2022, *Portal de Dados Abertos da Administração Publica*, Web Site [online] <https://dados.gov.pt/en/> [Accessed in 25 August 2022].

Dias, M., Boné, J., Ferreira, J., Ribeiro, R. and Maia, R. (2020). Named entity recognition for sensitive data discovery in portuguese. *Applied Sciences* (Switzerland), 10(7). DOI: https://doi.org/10.3390/app10072303.

Dozier, C., Kondadadi, R., Light, M., Vachher, A., Veeramachaneni, S., and Wudali, R. (2010). Named Entity Recognition and Resolution in Legal Text. In: Francesconi, E., Montemagni, S., Peters, W., Tiscornia, D. (eds) *Semantic Processing of Legal Texts. Lecture Notes in Computer Science*, vol. 6036. Springer, Berlin, Heidelberg. DOI: https://doi.org/10.1007/978-3-642-12837-0_2

Finlayson, M., Erjavec, T. (2017). Overview of Annotation Creation: Processes and Tools, pages 167–191. 06 2017. DOI: https://doi.org/10.1007/978-94-024-0881-2 5.

Gamallo, P., Garcia, M. (2017). LinguaKit: uma ferramenta multilingue para a análise linguística e a extração de informação. Linguamática, 9(1), pages 19–28, jul. DOI:10.21814/lm.9.1.243.

Goddard, C. and Schalley, A. (2010). Semantic Analysis. In *Handbook of Natural Language Processing*, Chapter 5, Eds Indurkhya, N., Damerau, F., 2nd Edition, CRC Press, Taylor & Francis, January.

Ferreira, L. (2011). *Medical Information Extraction in European Portuguese*. PhD Thesis, Universidade de Aveiro.

Gomes, J., Barros, A. and Belo, O. (2021). Georeferencing Toponyms in Didactic Contexts for Multidisciplinary Researching and Teaching, *Proceedings of 13th International Conference on Education and New Learning Technologies (EDULEARN21),* 5-6 July. DOI: https://doi.org/10.21125/edulearn.2021.

Gosal, G. (2015). A Survey on Semantic Annotation of Text. *International Journal of Advanced Research in Computer Science and Software Engineering*, Volume 5, Issue 9, September.

Khan, A. (2022). *How Text Annotation Is Helping Companies Succeed by Monitoring Social Noise*. August 2022. [online] <https://datasaur.ai/blog-posts/annotation-is-helping-companies-succeed> [Accessed in 16 January 2023]

Lynch, E. (2021). *Annotating Text Strategies That Will Enhance Close Reading*, 2021. [online] < https://www.sadlier.com/school/ela-blog/teaching-annotation-to-students-grades-2-8-annotating-text-strategies-that-will-enhance-close-reading> [Accessed in 25 August 2022]

Maestre, M. (2022). Communicative Intentions Annotation Scheme for Natural Language Generation. *Proceedings of the Doctoral Symposium on Natural Language Processing from the PLN.net network 2022 (RED2018-102418-T)(PLNnet-DS-2022) co-located with the XXXVIII edition of the International Conference of the Spanish Society for Natural Language Processing (SEPLN 2022),* A Coruña, Spain, September 21-23, 2022. CEUR Workshop Proceedings, Vol-3270, 31-39.

Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J. and Gómez-Berbís, J. (2013). Named Entity Recognition: Fallacies, challenges and opportunities, *Computer Standards & Interfaces*, Volume 35, Issue 5, Pages 482-489. DOI: https://doi.org/10.1016/j.csi.2012.09.004.

Mohammad, S. (2016). A Practical Guide to Sentiment Annotation: Challenges and Solutions. *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 174–179, San Diego, California. Association for Computational Linguistics.

Moraes, S., Lima, V. (2008). Abordagem nao Supervisionada para Extração de Conceitos a partir de Textos. In ACM, editor, *WebMedia'08 Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web*, page 5, Vila Velha, Espírito Santo, Brazil, DOI: https://doi.org/10.1145/1809980.1810066.

Refinitiv (2023). *Intelligent Tagging*. [online] <https://www.refinitiv.com/content/dam/marketing/en_us/documents/fact-sheets/intelligent-tagging-fact-sheet.pdf > [Accessed in 18 January 2023]

Selnes, O., Bjørsum-Meyer, T., Histace, A., Baatrup, G., Koulaouzidis, A. (2022). Annotation Tools in Gastrointestinal Polyp Annotation, Diagnostics 12, *Diagnostics*, no. 10: 2324. DOI: https://doi.org/10.3390/diagnostics12102324

Sinoara, R., Antunes, J., Rezende, S. (2017). Text mining and semantics: a systematic mapping study. *Journal of the Brazilian Computer Society*, 23, 9. DOI: https://doi.org/10.1186/s13173-017-0058-7

TagTop (2022). *The Text Annotation Tool to Train AI*. [online] <https://www.tagtog.com/> [Accessed in 25 August 2022]

UBIAI (2022). *Transform Your Unstructured Data Into Intelligence*. [online] <https://ubiai.tools/> [Accessed in 25 August 2022].

Win, S., Aung, T. (2018). Automated Text Annotation for Social Media Data during Natural Disasters, *Advances in Science, Technology and Engineering Systems Journal*, vol. 3, no. 2, pp. 119-127. DOI: https://doi.org/10.25046/aj030214.