From Textual Information Sources to Linked Data in the Agatha Project

Paulo Quaresma[0000−0002−5086−059X], Vitor Beires Nogueira[0000−0002−0793−0003], Kashyap Raiyani[0000−0002−6166−2038], Roy Bayot[0000−0002−1290−0239], and Teresa Gonçalves[0000−0002−1323−0249]

Universidade de Évora, Portugal
LISP - Laboratory of Informatics, Systems and Parallelism
{pq,vbn,kshyp,rkbayot,tcg}@uevora.pt

Abstract. Automatic reasoning about textual information is a challenging task in modern Natural Language Processing (NLP) systems. In this work we describe our proposal for representing and reasoning about Portuguese documents by means of Linked Data like ontologies and thesauri. Our approach resorts to a specialized pipeline of natural language processing (part-of-speech tagger, named entity recognition, semantic role labeling) to populate an ontology for the domain of criminal investigations. The provided architecture and ontology are language independent. Although some of the NLP modules are language dependent, they can be built using adequate AI methodologies.

Keywords: Linked Data, Ontology, Natural Language Processing, Events.

1 Introduction

The automatic identification, extraction and representation of the information conveyed in texts is a key task nowadays. In fact, this research topic is increasing its relevance with the exponential growth of social networks and the need to have tools that are able to automatically process them [11].

Some of the domains where it is more important to be able to perform this kind of action are the juridical and legal ones. Effectively, it is crucial to have the capability to analyse open access text sources, like social nets (Twitter and Facebook, for instance), blogs, online newspapers, and to be able to extract the relevant information and represent it in a knowledge base, allowing posterior inferences and reasoning.

In the context of this work, we will present results of the R&D project Agatha[1], where we developed a pipeline of processes that analyses texts (in Portuguese, Spanish, or English) and is able to populate a specialized ontology [17] (related to criminal law) for the representation of events, depicted in such texts. Events are represented by objects having associated actions, agents, elements, places and time. After having populated the event ontology, we have an automatic process linking the identified entities to external referents, creating, this way, a linked data knowledge base.

[1] http://www.agatha-osi.com/en/

It is important to point out that, having the text information represented in an ontology allows us to perform complex queries and inferences, which can detect patterns of typical criminal actions.

Another axe of innovation in this research is the development, for the Portuguese language, of a pipeline of Natural Language Processing (NLP) processes, that allows us to fully process sentences and represent their content in an ontology. Although there are several tools for the processing of the Portuguese language, the combination of all these steps in a integrated tool is a new contribution.

Moreover, we have already explored other related research path, namely author profiling [19], aggression identification [18] and hate-speech detection [21] over social media, plus statute law retrieval and entailment for Japanese [22].

The remainder of this paper is organized as follows: Section 2 describes our proposed architecture together with the Portuguese modules for its computational processing. Section 3 discusses different design options and Section 4 provides our conclusions together with some pointers for future work.

2 Framework for Processing Portuguese Text

The framework for processing Portuguese texts is depicted in Fig. 1, which illustrates how relevant pieces of information are extracted from the text. Namely, input files (Portuguese texts) go through a series of modules: part-of-speech tagging, named entity recognition, dependency parsing, semantic role labeling, subject-verb-object triple extraction, and lexicon matching.

The main goal of all the modules except lexicon matching is to identify events given in the text. These events are then used to populate an ontology.

The lexicon matching, on the other hand, was created to link words that are found in the text source with the data available not only on Eurovoc [2] thesaurus but also on the EU's terminology database IATE [6] (see Section 2.7 for details).

Most of these modules are deeply related and are detailed in the subsequent subsections.

2.1 Part-Of-Speech Tagging

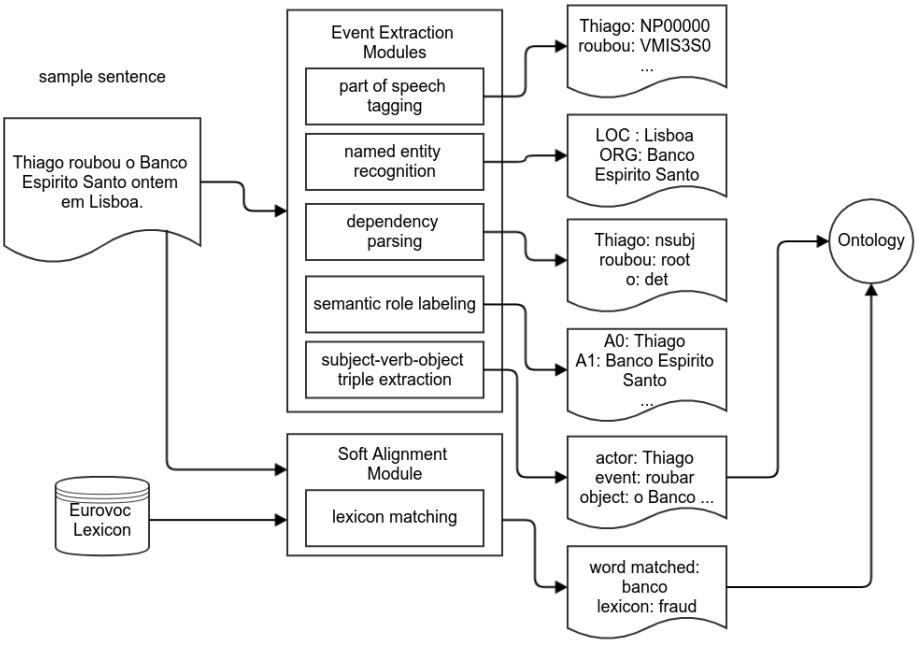
Part-of-speech tagging happens after language detection. It labels each word with a tag that indicates its syntactic role in the sentence. For instance, a word could be a noun, verb, adjective or adverb (or other syntactic tag). We used Freeling [14] library to provide the tags. This library resorts to a Hidden Markov Model as described by Brants [12]. The end result is a tag for each word as described by the EAGLES tagset [2].

2.2 Named Entity Recognition

We use the named entity recognition module after part-of-speech tagging. This module labels each part of the sentence into different categories such as "PERSON",

[2] `https://talp-upc.gitbook.io/freeling-4-0-user-manual/tagsets/tagset-pt`

Fig. 1. System Overview.



"LOCATION", or "ORGANIZATION". We also used Freeling to label the named entities and the details of the algorithm are shown in the paper by Carreras *et al* [15]. Aside from the three aforementioned categories, we also extracted "DATE/TIME" and "CURRENCY" values by looking at the part-of-speech tags: date/time words have a tag of "W", while currencies have "Zm".

2.3 Dependency Parsing

Dependency parsing involves tagging a word based on different features to indicate if it is dependent on another word. The Freeling library also has dependency parsing models for Portuguese. Since we wanted to build a SRL (Semantic Role Labeling) module on top of the dependency parser and the current released version of Freeling does not have an SRL module for Portuguese, we trained a different Portuguese dependency parsing model that was compatible (in terms of used tags) with the available annotated.

We used the dataset from System-T [8], which has SRL tags, as well as, the other preceding tags. It was necessary to do some pre-processing and tag mapping in order to make it viable to train a Portuguese model.

We made 589 tag conversions over 14 different categories. The breakdown of tag conversions per category is given by table 1. These rules can be further seen in the corresponding Github repository [1]. The modified training and development datasets are also available on another Github repository [10] for further research and comparison purposes.

Table 1. Training and Development - Tag Set Details.

Category	Number of Tags
NOUN	20
VERB	101
PROPN	39
PRON	121
ADJ	70
DET	62
AUX	149
ADP	3
NUM	1
PUNCT	18
CCONJ	1
SCONJ	1
INTJ	1
ADV	2

2.4 Semantic Role Labeling

We execute the SRL (Semantic Role Labeling) module after obtaining the word dependencies. This module aims at giving a semantic role to a syntactic constituent of a sentence. The semantic role is always in relation to a verb and these roles could either be an actor, object, time, or location, which are then tagged as A0, A1, AM-TMP, AM-LOC, respectively. We trained a model for this module on top of the dependency parser described in the previous subsection using the modified dataset from System-T. The module also needs co-reference resolution to work and, to achieve this, we adapted the Spanish co-reference modules for Portuguese, changing the words that are equivalent (in total, we changed 253 words).

2.5 SVO Extraction

From the yield of the SRL (Semantic Role Labeling) module, our framework can distinguish actors, actions, places, time and objects from the sentences. Utilizing this extracted data, we can distinguish subject-verb-object (SVO) triples using the SVO extraction algorithm [20]. The algorithm finds, for each sentence, the verb and the tuples related to that verb using Semantic Role Labeling (subsection 2.4). After the extraction of SVOs from texts, they are inserted into a specific event ontology (see section 2.7 for the creation of a knowledge base).

2.6 Lexicon Matching

The sole purpose of this module is to find important terms and/or concepts from the extracted text. To do this, we use Euvovoc [2], a multilingual thesaurus that was developed for and by the European Union. The Euvovoc has 21 fields and each field is further divided into a variable number of micro-thesauri. Here, due to the application of this work in the Agatha project (mentioned in Section 1), we use the terms of the criminal law [3] micro-thesaurus. Further, we classified each term of the criminal law micro-thesaurus into four categories namely, actor, event, place and object. The term classification can be seen in Table 2.

Table 2. Eurovoc Criminal Law - Term Classification.

Classification	**# Terms**
Actor	9
Event	133
Place	22
Object	3

After the classification of these terms, we implemented two different matching algorithms between the extracted words and the criminal law micro-thesaurus terms. The first is an exact string match wherein lowercase equivalents of the words of the input sentences are matched exactly with lower case equivalents of the predefined terms. The second matching algorithm uses Levenshtein distance [7], allowing some near-matches that are close enough to the target term.

2.7 Linked Data: Ontology, Thesaurus and Terminology

In the computer science field, an ontology can be defined has:

- a formal specification of a conceptualization;
- shared vocabulary and taxonomy which models a domain with the definition of objects and/or concepts and their properties and relations;
- the representation of entities, ideas, and events, along with their properties and relations, according to a system of categories.

A knowledge base is one kind of repository typically used to store answers to questions or solutions to problems enabling rapid search, retrieval, and reuse, either by an ontology or directly by those requesting support. For a more detailed description of ontologies and knowledge bases, see for instance [16].

For designing the ontology adequate for our goals, we referred to the Simple Event Model (SEM) [23] as a baseline model. A pictorial representation of this ontology is given in Figure 2

Considering the criminal law domain case study, we made a few changes to the original SEM ontology. The entities of the model are:

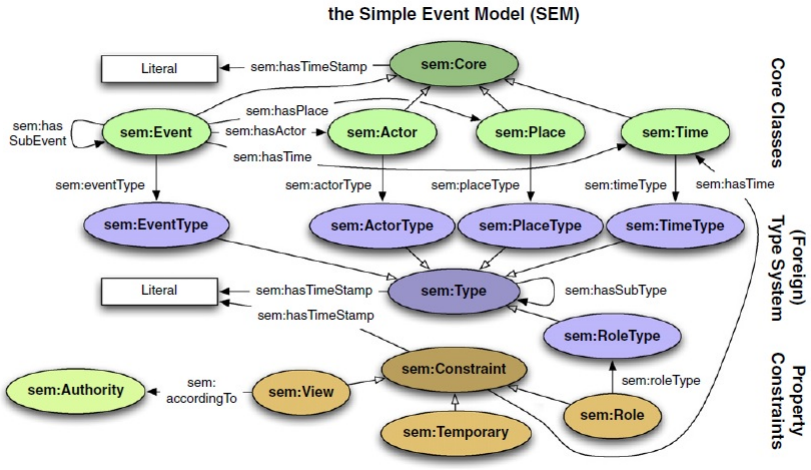


Fig. 2. The Simple Event Model [23]

- Actor: person involved with event
- Place: location of the event
- Time: time of the event
- Object: that actor act upon
- Organization: organization involved with event
- Currency: money involved with event

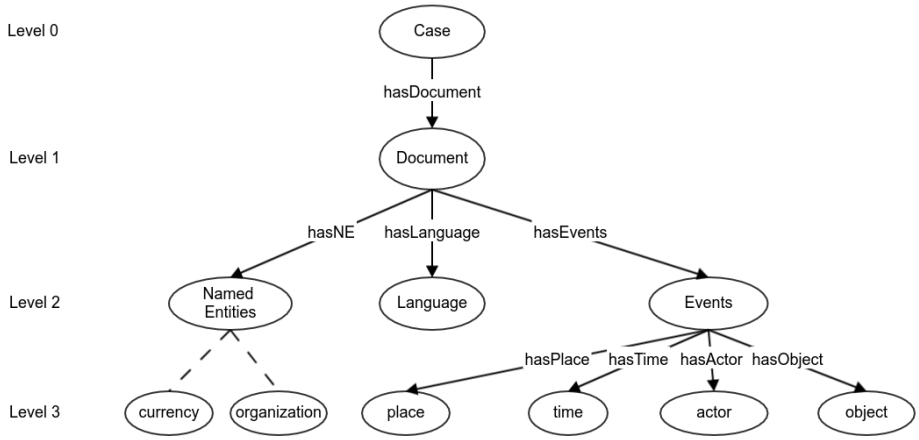
The proposed ontology was designed in such a manner that it can incorporate information extracted from multiple documents. In this context, suppose that the source of documents is a police department, where each document is under the hood of a particular case/crime; furthermore, a single case can have documents from multiple languages. Now, considering case 1 has 100 documents and case 2 has 100 documents then there is not only a connection among the documents of a single case but rather among all the cases with all the combined 200 documents. In this way, the proposed method is able to produce a detailed and well-connected knowledge base.

Figure 3 shows the proposed ontology, which, in our evaluation procedure, was populated with 3121 events entries from 51 documents.

Protege [9] tool was used for creating the ontology and GraphDB [5] for populating & querying the data. GraphDB is an enterprise-ready Semantic Graph Database, compliant with W3C Standards. Semantic Graph Databases (also called RDF triple-stores) provide the core infrastructure for solutions where modeling agility, data integration, relationship exploration, and cross-enterprise data publishing and consumption are important. GraphDB has a SPARQL (SQL-like query language) interface for RDF graph databases with the following types:

- SELECT: returns tabular results
- CONSTRUCT: creates a new RDF graph based on query results
- ASK: returns "YES", if the query has a solution, otherwise "NO"

Fig. 3. Ontology Diagram.



- DESCRIBE: returns RDF data about a resource. This is useful when the RDF data structure in the data source is not known
- INSERT: inserts triples into a graph
- DELETE: deletes triples from a graph

Furthermore, we have extended the ontology [4] to connect the extracted terms with Eurovoc criminal law (discussed in subsection 2.6) and IATE [6] terms. IATE (**I**nter**a**ctive **T**erminology for **E**urope) is the EU's general terminology database and its aim is to provide a web-based infrastructure for all EU terminology resources, enhancing the availability and standardization of the information. The extended ontology has a number of sub-classes for Actor, Event, Object and Place classes detailed in Table 3.

3 Discussion

We have defined a major design principle for our architecture: it should be modular and not rely on human made rules allowing, as much as possible, its independence from a specific language. In this way, its potential application to another language would be easier, simply by changing the modules or the models of specific modules. In fact, we have explored the use of already existing modules and adopted and integrated several of these tools into our pipeline.

It is important to point out that, as far as we know, there is no integrated architecture supporting the full processing pipeline for the Portuguese language. We evaluated several systems like Rembrandt [13] or LinguaKit: the former only has the initial steps of our proposal (until NER) and the later performed worse than our system.

This framework, developed within the context of the Agatha project (described in Section 1) has the full processing pipeline for Portuguese texts: it receives sentences

Table 3. Extended Ontology [4] - Sub-Classes.

Class	No. of Sub-Classes	Terms
Actor	6	Victim,Inmate,Prisoner,Hostage,Hijacker,Accomplice
Event	64	Slavery,Trade,Tax,Evasion,Spoofing,Slander,Shady, Violence,Sexual,Scam,Repentance,Rehabilitation, Refoulement,Rape,Punishment,Ponzi,Piracy,Aggression, Phishing,Trafficking,Pardon,Harassment,Mobbing, Misdemeanour,Libel,Trading,Imprisonment, Restraint,Theft,Arrest,Homicide,Hit-and-run,Hijacking, Forgery,Forfeiture,Fraud,Fight,Falsification,Extradition, Expulsion,Elimination,Offence,Drug,Detention, Deprivation,Deportation,Defamation,Penalty,Negligence, Execution,Counterfeit,Corruption,Confiscation, Conditional,Con,Order,Complicity,Campaign, Bully,Breach,Banish,Aggravate,Crime,Abduction
Object	3	Fine,Invoice,Bill
Place	11	Facility,Institution,Center,Confinement,Reformatory Penitentiary,Penal,Prison,Jail,Isolation,Banco

as input and outputs ontological information: a) first performs all NLP typical tasks until semantic role labelling; b) then, it extracts subject-verb-object triples; c) and, then, it performs ontology matching procedures. As a final result, the obtained output is inserted into a specialized ontology.

We are aware that each of the architecture modules can, and should, be improved but our main goal was the creation of a full working text processing pipeline for the Portuguese language.

4 Conclusions and Future Work

Besides the end–to–end NLP pipeline for the Portuguese language, the other main contributions of this work can be summarize as follows:

- Development of an ontology for the criminal law domain;
- Alignment of the Eurovoc thesaurus and IATE terminology with the ontology created;
- Representation of the extracted events from texts in the linked knowledge base defined.

The obtained results support our claim that the proposed system can be used as a base tool for information extraction for the Portuguese language. Being composed by several modules, each of them with a high level of complexity, it is certain that our approach can be improved and an overall better performance can be achieved.

As future work we intend, not only to continue improving the individual modules, but also plan to extend this work to the:

– automatic creation of event timelines;
– incorporation in the knowledge base of information obtained from videos or pictures describing scenes relevant to criminal investigations.

Acknowledgments

The authors would like to thank COMPETE 2020, PORTUGAL 2020 Program, the European Union, and ALENTEJO 2020 for supporting this research as part of Agatha Project SI & IDT number 18022 (Intelligent analysis system of open of sources information for surveillance/crime control). The authors would also like to thank LISP - Laboratory of Informatics, Systems and Parallelism.

References

1. Automated event extraction model for multiple linked portuguese documents, https://github.com/kraiyani/Automated-Event-Extraction-Model-for-Multiple-Linked-Portuguese-Documents/blob/master/Universal_to_eagle_tagset.xlsx, [Available online: accessed on 06 05 2019]
2. Eu vocabularies, https://publications.europa.eu/en/web/eu-vocabularies, [Available online: accessed on 06 05 2019]
3. Eu vocabularies, thesauri, 1216 criminal law, https://publications.europa.eu/en/web/eu-vocabularies/th-concept-scheme/-/resource/eurovoc/100180?target=Browse, [Available online: accessed on 06 05 2019]
4. Extended ontology, http://owlgred.lumii.lv/online_visualization/e9fh#, [Available online: accessed on 25 06 2019]
5. Graphdb, http://graphdb.ontotext.com/, [Available online: accessed on 06 05 2019]
6. Iate (interactive terminology for europe), https://iate.europa.eu/home, [Available online: accessed on 06 05 2019]
7. Levenshtein distance, https://en.wikipedia.org/wiki/Levenshtein_distance, [Available online: accessed on 06 05 2019]
8. Portuguese universal propositions, https://github.com/System-T/UniversalPropositions/tree/master/UP_Portuguese-Bosque, [Available online: accessed on 06 05 2019]
9. Protege, https://protege.stanford.edu/, [Available online: accessed on 06 05 2019]
10. Training and development dataset for automated event extraction model for multiple linked portuguese documents, https://github.com/kraiyani/Automated-Event-Extraction-Model-for-Multiple-Linked-Portuguese-Documents, [Available online: accessed on 06 05 2019]
11. Amato, F., Moscato, V., Picariello, A., Sperlì, G.: Extreme events management using multimedia social networks. Future Generation Comp. Syst. **94**, 444–452 (2019). https://doi.org/10.1016/j.future.2018.11.035, https://doi.org/10.1016/j.future.2018.11.035
12. Brants, T.: Tnt: a statistical part-of-speech tagger. In: Proceedings of the sixth conference on Applied natural language processing. pp. 224–231. Association for Computational Linguistics (2000)

13. Cardoso, N.: Rembrandt - a named-entity recognition framework. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012). pp. 1240–1243. European Language Resources Association (ELRA), Istanbul, Turkey (May 2012), http://www.lrec-conf.org/proceedings/lrec2012/pdf/409_Paper.pdf
14. Carreras, X., Chao, I., Padró, L., Padro, M.: Freeling: An open-source suite of language analyzers. Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04) (01 2004)
15. Carreras, X., Màrquez, L., Padró, L.: A simple named entity extractor using adaboost. In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 (2003)
16. Guarino, N., Giaretta, P.: Ontologies and knowledge bases: Towards a terminological clarification. In: Towards very Large Knowledge bases: Knowledge Building and Knowledge sharing. pp. 25–32. IOS Press (1995)
17. Guarino, N., Oberle, D., Staab, S.: What Is an Ontology?, pp. 1–17 (05 2009)
18. Raiyani, K., Gonçalves, T., Quaresma, P., Nogueira, V.B.: Fully connected neural network with advance preprocessor to identify aggression over facebook and twitter. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018). pp. 28–41. Association for Computational Linguistics (2018), http://aclweb.org/anthology/W18-4404
19. Raiyani, K., Gonçalves, T., Quaresma, P., Nogueira, V.B.: Multi-language neural network model with advance preprocessor for gender classification over social media: Notebook for pan at clef 2018. In: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018. (2018), http://ceur-ws.org/Vol-2125/paper_105.pdf
20. Raiyani, K., Gonçalves, T., Quaresma, P., Nogueira, V.B.: Automated event extraction model for linked portuguese documents. In: Proceedings of Text2Story — Second Workshop on Narrative Extraction From Texts co-located with 41th European Conference on Information Retrieval (ECIR 2019) Cologne, Germany, April 14th (2019), http://ceur-ws.org/Vol-2342/paper2.pdf
21. Raiyani, K., Gonçalves, T., Quaresma, P., Nogueira, V.B.: Vista.ue at semeval-2019 task 5: Single multilingual hate speech detection model. In: Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019). pp. 520–524. Association for Computational Linguistics (2019)
22. Raiyani, K., Quaresma, P.: Keyword & machine learning based japanese statute law retrieval and entailment task at coliee-2019. In: Proceedings of Competition on Legal Information Retrieval and Entailment Workshop (COLIEE 2019) in association with the 17th International Conference on Artificial Intelligence and Law 2019 (ICAIL 2019). Easychair (2019)
23. Van Hage, W.R., Malaisé, V., Segers, R., Hollink, L., Schreiber, G.: Design and use of the simple event model (sem). Web Semantics: Science, Services and Agents on the World Wide Web **9**(2), 128–136 (2011)