information

Article

Event Extraction and Representation: A Case Study for the Portuguese Language

Paulo Quaresma *, **Vítor Beires Nogueira**, **Kashyap Raiyani** and **Roy Bayot**

Informatics Department, University of Évora, 7000-671 Évora, Portugal; vbn@uevora.pt (V.B.N.); kshyp@uevora.pt (K.R.); roybayot@gmail.com (R.B.)

* Correspondence: pq@uevora.pt

Received: 11 May 2019; Accepted: 5 June 2019; Published: 8 June 2019

Abstract: Text information extraction is an important natural language processing (NLP) task, which aims to automatically identify, extract, and represent information from text. In this context, event extraction plays a relevant role, allowing actions, agents, objects, places, and time periods to be identified and represented. The extracted information can be represented by specialized ontologies, supporting knowledge-based reasoning and inference processes. In this work, we will describe, in detail, our proposal for event extraction from Portuguese documents. The proposed approach is based on a pipeline of specialized natural language processing tools; namely, a part-of-speech tagger, a named entities recognizer, a dependency parser, semantic role labeling, and a knowledge extraction module. The architecture is language-independent, but its modules are language-dependent and can be built using adequate AI (i.e., rule-based or machine learning) methodologies. The developed system was evaluated with a corpus of Portuguese texts and the obtained results are presented and analysed. The current limitations and future work are discussed in detail.

Keywords: natural language processing; information extraction; text mining; events; ontologies population

1. Introduction

Text information extraction is an important natural language processing (NLP) task, aimed at automatically identifying, extracting, and representing information from text. Event extraction is an important and relevant sub-task in the NLP domain [1]. The conventional view of events is that, given a sentence, events denote an activity or a state of action. In the context of this work, the general assumption made is that the event structure is associated with the sentence predicates and their arguments. The argument structure is given by a set of arguments of the verb; namely, actors, objects, places, and time. The extracted information can be represented by specialized ontologies [2], supporting knowledge-based reasoning and inference processes. This topic has gained relevance with the exponential growth of social networks and the need to automatically identify and extract referred events [3,4]. In this paper, we will focus on the research about events, focusing on two questions: What are the primitive elements of events and how can they be automatically extracted?

On the other hand, for the Portuguese language, which is the sixth largest language in terms of number of native speakers and the fifth largest language in terms of number of internet users (Ethnologue: Languages of the World, 2019.), a set of computational processing tools have already been developed (see, for instance, the proceedings of the Computational Processing of the Portuguese Language (PROPOR) [5] however, the Portuguese language lacks an integrated architecture which allows the complete processing of Portuguese documents from text to the knowledge base population [6].

In this work, we will present and describe a proposal for event extraction from Portuguese texts, based on a pipeline of specialized natural language processing tools; namely, a part-of-speech tagger, a named entities recognizer, a dependency parser, semantic role labeling, and a knowledge extraction module. This architecture was designed to be language-independent but its modules are language-dependent, in the sense that they depend on specialized rules or that their models need to be created using machine learning approaches, requiring previously annotated Portuguese corpora.

The proposed system was evaluated with two Portuguese corpora, one being the publicly available corpus of PropBank [7], and the obtained results are presented and discussed. Due to the complexity of the task, there still exist many limitations and problems that need to be solved, but we believe this architecture can play an important tool in this domain and, in particular, in the context of the computational processing of the Portuguese Language. Moreover, this work is strongly related to the participation of the authors in the Portugal2020 Agatha project [8]. Basically, the aim of this project is to intelligently analyze open-source information for surveillance/crime control, following in the footsteps of similar open source information analysis, where author profiling [9], aggression identification [10] and hate-speech detection [11] over social media, as well as statute law retrieval and entailment for Japanese statutes [12] have already been done.

The remainder of this paper is organized as follows: In Section 2, we present an overview of the related work. Section 3 describes our proposed architecture and Section 4 presents the Portuguese modules for its computational processing. Finally, Section 5 evaluates the proposal, Section 6 discusses different design options, and, in Section 7, we provide our conclusions, together with some pointers for future work.

2. Related Work

Event detection from unstructured data, such as those obtained from the news wire, discussion forums, or social networks, is a challenging task. This statement can be supported, for instance, by inspecting the results of international contests like the Event Detection (Co-reference and Sequencing) track of the Text Analysis Conference [13].

From a broad point of view, we consider three main approaches for event extraction:

- Data-driven techniques which convert data into knowledge by means of statistics, machine learning, and so on;
- expert knowledge-driven methods which derive knowledge by resorting to experts, using, for instance, pattern-based approaches; and
- hybrid approaches, which combine the aforementioned approaches.

For an in-depth comparison between these approaches, see, for instance, [1] and, for an overview more focused on the Portuguese language, please refer to [6].

Collovini et al. [14] described a proposal for relation extraction in the Portuguese language using Conditional Random Fields (CRF), where they were able to obtain an f-measure of 0.45 for complete relation matching. In another work, Bonamigo et. al [15] proposed the use of pattern rules to identify relations between entities, but their approach is not easy to generalize and it was not able to deal well with the complexity and diversity of the language.

On the other hand, previous work done in the area of event extraction is mainly application-specific. For instance, Automatic Content Extraction (ACE) is a tool that extracts entities, relations, and events, but it is noteworthy that ACE takes input in the sgml format, which restricts user input [16].

Yuan et al. [17] proposed an event-based approach to visualize documents as a graph on different conceptual granularities. In [18], the authors treated events as undeniably temporal entities. In comparison to ACE, the event extraction task was done in modules, each of which was handled by a machine-learned classifier. The results of this approach [18] were better than those obtained by ACE, but the methodology was still domain-specific.

Halpin et al. [19] proposed the extraction of events for story-rewriting. In [20], domain ontology was used as a method for extracting events. However, updating the domain ontology with new terms is crucial when dealing with contemporary dynamic data.

Recently, several works have been published on information extraction from social media; in particular, using tweets from the Twitter network. Sakaki et al. [21] described a method to detect earthquake-related tweets. This method used features specific to earthquakes. Benson et al. [22] trained a relation extractor to identify artists and venues from tweets. This method was designed to develop a graphical model by learning records and record-message alignments. Ritter et al. [23] described a method, based on latent variable modeling, to extract the event types described in tweets, where features, such as tweet popularity and the times of events referred to in the tweets, were used. Zhao et al. [24] described a method to extract only the most "topical" keywords from tweets. In [25], the authors resorted to un-supervised methods to extract real-world events from Twitter data streams. Amato et al. [3,4] proposed the use of a hypergraph-based approach to exploit influence analysis methodologies and to identify the most important entities in social media networks.

Similar approaches have also been applied to mining relevant information from non-text sources. For instance, Zong et al. [26] described approximation algorithms to identify critical alerts from a large set of alert sequences.

Our approach differs from the above-mentioned methods, as it is a complete pipeline of specialized modules for the Portuguese language which receives general-purpose sentences, where the output populates an event ontology.

3. Proposed Architecture

The proposed system is described in Figure 1, which illustrates how relevant pieces of information are extracted from the text. Input files (Portuguese texts) pass through a series of modules: Language detection, 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 for lexicon matching) is to identify the events in the text. These events are, then, used to populate an ontology. The lexicon matching module, on the other hand, was created to generate a soft alignment between the words that are found in the text and words that are found in Eurovoc [27].

Where available, we used the Freeling framework [28] for the pipeline of event extraction. It was necessary, however, to create specific modules or to train existing ones for dependency parsing, semantic role labeling, and subject–verb–object extraction. Table 1 shows an example of the output of each module for the following sample sentence 'Thiago roubou o Banco Espirito Santo ontem em Lisboa.' (i.e., 'Thiago robbed the Bank of the Holy Spirit yesterday in Lisbon.') We will be referring to this table later, when we discuss each module.

Table 1. Sample output for the different modules.

ID	Form	Lemma	Postag	Cpostag	Ner	Head	Dprel	Srl
1	Thiago	thiago	NP00000	NP	B-PER	2	nsubj	A0
2	roubou	roubar	VMIS3S0	VMI	-	0	root	-
3	o	o	DA0MS0	DA	-	4	det	-
4	Banco _Espirito_Santo	banco_espirito_santo	NP00000	NP	B-ORG	2	dobj	A1
5	ontem	ontem	RG	RG	-	2	advmod	AM-TMP
6	em	em	SP	SP	-	7	case	-
7	Lisboa	lisboa	NP00000	NP00000	B-LOC	2	nmod	AM-LOC
8	.	.	Fp	Fp	-	2	punct	-

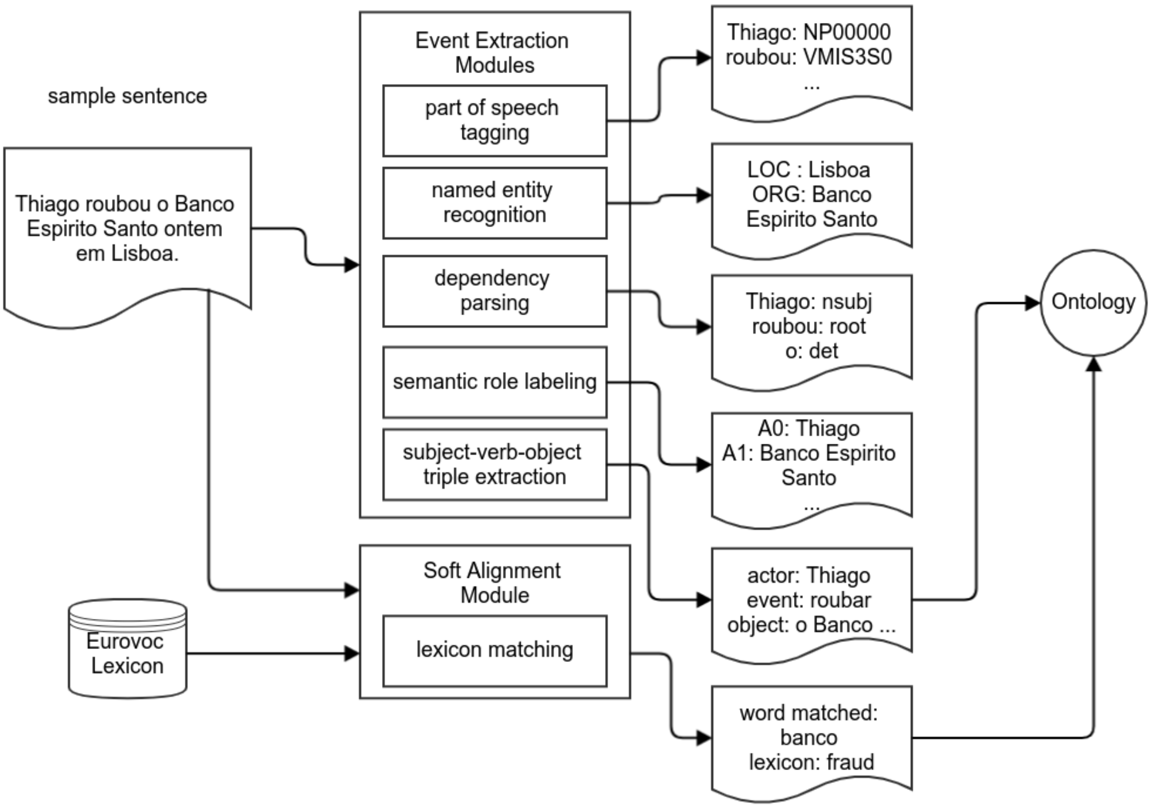

Figure 1. System Overview.

4. Computational Processing of Portuguese Modules

As in the previous section, suppose we have the sample sentence 'Thiago roubou o Banco Espirito Santo ontem em Lisboa.' (i.e., 'Thiago robbed the Bank of the Holy Spirit yesterday in Lisbon.'). This sentence will pass through the seven main modules of the proposed system: language detection, part-of-speech tagging, named entity recognition, dependency parsing, semantic role labeling, subject–verb–object triple extraction, and lexicon matching. Some of these modules involve others, these are detailed in the subsequent sub-sections.

4.1. Language Detection

Language detection is the first text-processing module. We used the Polyglot [29] library to detect the text language, which depends on the Compact Language Detector 2 (CLD2 [30]) which uses a Naïve Bayesian classifier. This step serves, mainly, as an initial check, so that the text that is further processed will only be Portuguese and not belong to another language.

4.2. Part-of-Speech Tagging

Part-of-speech (POS) tagging happens after the language detection. Basically, it marks the words in the sentence to their corresponding parts of speech, be it noun, verb, adjective, or otherwise. We used the Freeling [28] library to do the POS tagging. We (currently) use the default setting, which uses a Hidden Markov Model, as described by Brants in [31].

The sample sentence 'Thiago roubou o Banco Espirito Santo ontem em Lisboa.' (i.e., 'Thiago robbed the Bank of the Holy Spirit yesterday in Lisbon.') has the parts-of-speech shown in Table 1, indicated in the fourth and fifth column. The fifth column is more coarse-grained, while the fourth column also encodes language-specific properties. For instance, the verb *roubou* is associated with VMIS3S0, which indicates that it is a main verb in the past tense for the third person singular and the indicative mood. The coarse-grained tag (i.e., VMI) simply indicates that it is the main verb for the indicative mood.

4.3. Named Entity Recognition

After part-of-speech tagging, the named entity recognition module is executed. This module is used to identify which words in the sentence are named entities. Examples of these are names of people, organizations, and locations. We also used Freeling to extract these three named entity types. We extracted date, time, and currency, as well. Details of the algorithm are shown by Carreras et al. in [32]. Looking at the example shown in Table 1, the column with "ner" indicates if the word is part of a named entity. In this case, Thiago is indicated as "B-PER" (for person), Lisboa is indicated as "B-LOC" (for location), and Banco Espirito Santo is indicated as "B-ORG" (for organization).

4.4. Dependency Parsing

Dependency parsing involves tagging a word, based on different features, to indicate if it is dependent on another word. We had to train a different Portuguese dependency parsing model for Freeling, even though the software already had an available model. The new model was needed because we wanted to build a semantic role-labeling module on top of the dependency parser and as the available annotated corpus had different tags. Moreover, the present version of Freeling does not have an SRL (Semantic Role Labeling) module for Portuguese.

We used the data set from System-T [33], which has SRL (Semantic Role Labeling) tags, as well as the other preceding tags. It was necessary to do some pre-processing and tag mapping, so that it was viable to use for training a Portuguese model. Table 2 shows the conversion of the tags of one word. The coarse-grained POS tags (cpos) were the same as the POS tags obtained by the previous module. We converted each POS tags to its EAGLES [34] tag set form. We did this by inferring from the morphological features column. We also converted the contracted form to an explicit form. In this case, "Ind" was turned to "indicative", "Sing" to "singular", and so on. These conversions are presented in Table 2.

Table 2. Sample of tag mapping.

	Form	Cpos	Pos	Morphological Features
before	ficou	VERB	VERB	Mood=Ind | Number=Sing | Person=3 | Tense=Past | VerbForm=Fin
after	ficou	VMI	VMIS3S0	pos=verb | type=main | mood=indicative | tense=past | person=3 | num=singular

In total, we had to make 589 tag conversions over 14 different categories. The breakdown of tag conversions per category is given by Table 3. These rules can be further seen on this specific Github page [35]. The modified training and development data sets are also available on the Github repository [36], for further research and comparison purposes.

Going back to our example in Table 1, the outputs of the dependency parser are indicated with columns "head" and "dprel". Head indicates the dependency of each word. For example, the 2nd word 'roubou' has a dependency of 0, since it is regarded as the main verb and, hence, the root. Other words like 'Thiago', 'Banco Espirito Santo', 'ontem', and 'Lisboa' all had a dependency of 2, which indicate that they are dependent on 'roubou'. The column dprel indicates the relations. For instance 'Thiago' is the subject, as indicated by "nsubj"; 'o' is the determiner "the" indicated by "det"; and 'Banco Espirito Santo' is the object "Bank of the Holy Spirit" to the determiner "the".

Table 3. 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

4.5. Semantic Role Labeling

After inferring the dependencies, we execute the SRL (Semantic Role Labeling) module. This module, essentially, tags words with A0, A1, AM-TMP, or AM-LOC which indicates if the word is an actor, an object, a time, or a location, respectively. They also indicate which verbs are associated with which event and the corresponding actors, objects, and so on. We trained a model for this module on top of the dependency parser described earlier using the modified data set from System-T. The module also needed co-reference resolution to work and, in order to achieve this, we adapted the Spanish co-reference modules for Portuguese, changing words that were equivalent, such as fazer, ser, estar, and haver. In total, there were 253 different words that were changed. The performance of the semantic role labeling module over the development set is shown in the results, in Section 5.

In our example, the output for semantic role labeling are given in the column "srl" in Table 1. We can see that *Thiago* is indicated by A0 (for actor), *Banco Espirito Santo* is indicated by A1 (for object), *ontem* is indicated by AM-TMP (for time), and, finally, *Lisboa* is indicated by AM-LOC (for location).

4.6. Subject–Verb–Object Extraction

From the output of the SRL (Semantic Role Labeling) module, our system was able to identify actors, actions, places, times, and other objects in the processed sentences. Using this information, we identified subject–verb–object (SVO) triples, which constitute the basis of events. The SVO extraction process [37] is described in Algorithm 1. In the algorithm, `Predicate` stands for the verb of the sentence and the associated arguments are actor, object, time, and place, and `Argument.role()` is the function which returns the "srl" tags (discussed in Section 4.5) for the sentence. After the extraction of SVOs from the text, they are inserted into a specific event ontology (see Section 4.7) for the creation of a knowledge base.

Algorithm 1 SVO Extraction Algorithm

```
procedure SVO(sentence)
    SVO ← []
    for each Predicate in sentence do
        Event ← Predicate
        for each Argument associated with Event do
            switch Argument.role() do                    ▷ Argument.role() returns SRL tagging
                case A0
                    Actor ← Argument
                case A1
                    Object ← Argument
                case AM-LOC
                    Location ← Argument
                case AM-TMP
                    Time ← Argument
        end for
        SVO.append(Event, Actor, Object, Location, Time)
    end for
    return SVO
end procedure
```

4.7. Ontology: Knowledge Base

According to [38], an ontology can be understood as an intentional semantic structure which encodes the implicit rules constraining the structure of a piece of reality. Ontologies are, thus, aimed at answering the question "What kind of objects exist in one domain of the real world and how are they inter-related?". An ontology, thus, describes the logical structure of a domain, its concepts, and the relations between them.

To design an ontology adequate for our goals, the Simple Event Model (SEM) [39] ontology was referred to as a baseline model. Taking into account the goals of our work and in order to simplify it, we made some changes to the SEM ontology; the entities of the model are listed below.

1. Actor: Person involved with the event.
2. Place: Location of the event.
3. Time: Time of the event.
4. Object: That which the actor acts upon.
5. Organization: Organization involved with the event.
6. Currency: Money involved with the event.

The proposed ontology was designed in such a manner that it can incorporate information extracted from multiple documents.

In order to better evaluate our approach, we decided to implement a case study for the legal domain. In this context, suppose the the source of documents is a police department, where each document is part of a particular case/crime. Furthermore, a single case can have documents from multiple languages. Now, if 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 within the combined 200 documents. In this way, the proposed method is able to produce a detailed and well-connected knowledge base. Furthermore, we are also working on the extension of the ontology to connect the extracted terms with Eurovoc criminal law (discussed in Section 4.8) and IATE (Interactive Terminology for Europe) terms. Here, IATE [40] is the general terminology database of the EU. The aim of IATE is to provide a web-based infrastructure for all EU terminology resources, enhancing the availability and standardization of information.

Figure 2 shows the proposed ontology, which, in our evaluation procedure, was populated with 3121 event entries from 51 documents (from Data set 2, which is further explained in Section 5). The Protege [41] tool was used for the ontology creation and GraphDB [42] was used for populating and querying the data.

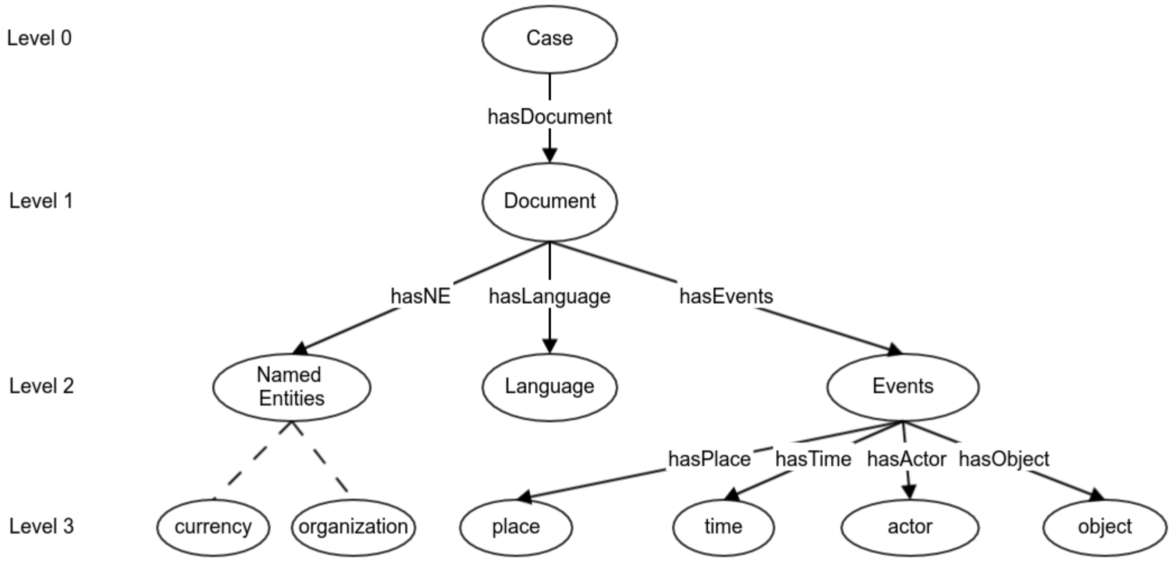

Figure 2. Ontology Diagram.

4.8. Lexicon Matching

The goal of this module is to find words in the input files that may be linked to important terms and/or concepts. The terms we used are part of Eurovoc: [27], a multilingual thesaurus developed for and by the European Union. This thesaurus has 21 fields and each field is further divided into a variable number of micro-thesauri. Due to application of this work in the Agatha project (mentioned in Section 1), we narrowed the terms to the one of micro-thesauri for criminal law [43], part of the law field. Finally, one interesting feature of Eurovoc is that its terms are also tagged for inter-relation to other terms, if there is some connection.

The 204 terms of criminal law micro-thesauri were manually annotated with actor, event, place, and object tags, whenever that could be the case. Table 4 shows the distribution of terms between events, actors, places, and objects. With the help of this annotation, we can find possible relations between the extracted actors, events, places, and objects and Eurovoc criminal law terms.

Table 4. Eurovoc criminal law term classification.

Classification	Number of Terms
Actor	9
Event	133
Place	22
Object	3

We implemented two ways to find matches between the words in the input and the terms and concepts in Eurovoc. The first was an exact string match, wherein lower-case equivalents of the words of the input sentences were matched exactly with lower-case equivalents of the pre-defined terms. The second matches by using Levenshtein distance [44]. This allowed some near-matches that were close enough to the target term.

Going back to the example 'Thiago roubou o Banco Espirito Santo ontem em Lisboa.', each word will be compared to a list. An exact match would be *banco*, which can also be found in Eurovoc. It is also associated with terminologies related to fraud. Due to the manual annotations, the term can also be associated with a place.

5. Results

Regarding the evaluation of our system, we will only present the results obtained with the new modules and with the overall performance of the system (i.e., we do not evaluate NLP modules which are already existent and have been previously evaluated).

The dependency parser was trained over a development set [45] and evaluated over a validation set [46]. The resulting accuracy, for 50 iterations, is given in Figure 3. The peak accuracy is at the 9th iteration, at about 94.72%.

Semantic Role Labeling was also evaluated on the validation set. This is shown in Figure 4. The highest $F1$ measure was observed at the 34th iteration, with a value of 71.06%.

The accuracy of extracting the SVO by the proposed model was evaluated over two different data sets.

1. Data set 1: The Propbank-Br v.1.1 Corpus [47] is a Brazilian Portuguese corpus, annotated in conll format, with Verb (event), A0 (subject), A1 (object), Place (location), and Time. Table 5 details the properties of this data set.

Table 5. Properties of Data set 1.

Properties	Details
Sentence	3348
Word	69,234
Verb	5931
Subject (A0)	2829
Object (A1)	5269

2. Data set 2: This was created by the consortium of the project Agatha, using 51 documents from several online sources, aiming to illustrate and evaluate the performance of the developed modules. This data set was manually verified by a research team of the Portuguese Department of the University of Macau, lead by Professor Ana Leal and with the help of three MSc students. In their work, they focused on the accuracy and precision of the system and did not evaluate recall. Table 6 details the properties of the extracted system.

Table 6. Properties of Data set 2.

Properties	Details
Sentence	941
Word	30,028
Extracted Verb	3773

The results of Data set 1 were compared with the existing tool LinguaKit [48]. LinguaKit is a Natural Language Processing tool containing several NLP modules, such as a POS Tagger, Named Entity Recognition, a Dependency Parser, Semantic Role Labeling, and Co-reference Resolution of Named Entities. Currently, LinguaKit supports four different languages; namely, English, Galician, Portuguese, and Spanish. The Semantic Role Labeling module is not publicly available, at present, on the Linguakit toolkit GitHub page. For comparison purposes, the developers of Linguakit sent us a prototype of the Role Labeling module, based on written rules.

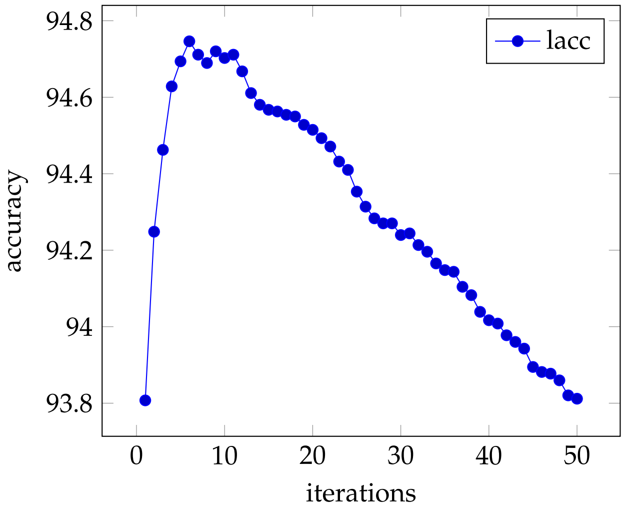

Figure 3. Accuracy of dependency parsing on the validation set.

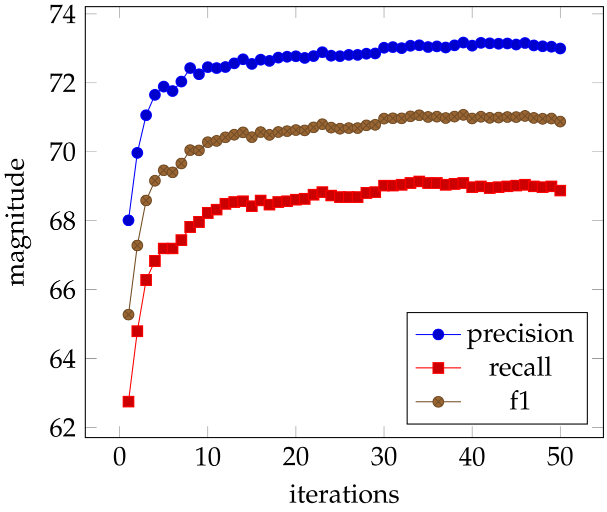

Figure 4. Precision, recall, and F1 measure of semantic role labeling on the validation set.

Table 7 details the proposed system and the Linguakit-extracted properties, respectively, using Data set 1.

Table 7. Data set 1: Proposed system versus Linguakit-extracted properties.

Extracted Properties	Proposed System	Linguakit
Verb	7767	6159
Subject (A0)	3605	2259
Object (A1)	5645	6018

Tables 8 and 9, respectively, show the proposed system evaluation and Linguakit evaluation on Data set 1.

Table 8. Proposed system evaluation on Data set 1.

System	Verb	Subject (A0)	Object (A1)
Propbank	5931	2829	5269
Match/True Positive	5378	2151	3254
Missing/False Negative	551	678	2015
Extra/False Positive	2389	1454	2391

Table 9. Linguakit evaluation on Data set 1.

System	Verb	Subject (A0)	Object (A1)
Propbank	5931	2829	5269
Match/True Positive	3457	1449	3172
Missing/False Negative	2474	1380	2097
Extra/False Positive	2702	810	2846

Tables 10 and 11, respectively, detail the performance of the proposed system and Linguakit on Data set 1.

Table 10. Performance of proposed system on Data set 1.

System	Verb	Subject (A0)	Object (A1)
Precision	0.6924	0.5966	0.5764
Recall	0.9070	0.7603	0.6175
F1 Score	0.7853	0.6685	0.5962

Table 11. Performance of Linguakit on Data set 1.

System	Verb	Subject (A0)	Object (A1)
Precision	0.5612	0.6414	0.5270
Recall	0.5828	0.5121	0.6020
F1 Score	0.5717	0.5695	0.5620

A comparison of F1 score performance between the proposed system and Linguakit on Data set 1 is shown in Table 12.

Table 12. Comparison of F1 score between the proposed system and Linguakit on Data set 1.

System	Verb	Subject (A0)	Object (A1)
Proposed System	**0.7853**	**0.6685**	**0.5962**
Linguakit	0.5717	0.5695	0.5620

The proposed system, with an F1 score of 0.7853 for verb/event detection, out-performed the existing LinguaKit NLP processing tool by quite a good margin (0.2136). Moreover, our stated approach used deep learning methods and trained over the data set, whereas LinguaKit is a rule-based "srl labeling" tool. From the results shown in Table 12, we can clearly state that our proposed system out-performed the existing LinguaKit NLP processing tool.

Finally, Table 13 shows the system evaluation and performance over Data set 2. Note that, as referred to before, false negative verbs for this data set were not evaluated.

Table 13. System evaluation and performance over Data set 2.

System	Verb
Match/True Positive	3121
Extra/False Positive	553
Precision	0.8494

6. Discussion

During the design of our system, we explored other modules/systems, faced various options, and made several decisions. One of the key points in the design process was the fact that we wanted to have a more dynamic approach, one which does not rely on hand-made rules. The main motivation for this was that it would be easier to apply the same approach to other languages, using different data sets for training.

Although there are a wide range of NLP systems which include most, or even all, of the steps of our proposed pipeline, the majority of them lack support for the Portuguese language. Nevertheless, we carried out some experiments with systems such as Rembrandt [49] or LinguaKit; the former only had the initial steps of our proposal (i.e., up until NER) and the latter performed worse than our system, as described in Section 5.

The framework developed within the Agatha project (as described in Section 1) resorts to our system for all processing that includes textual data. In this context, we tested our system with a data set specifically developed by this project and proposed three levels of processing: The first performs all tasks until named entity recognition; the second computes until subject–verb–object triple extraction; and the last level executes the pipeline until lexicon matching. All levels insert the corresponding output into the ontology developed. This way, we provide a scale from speedier (although superficial) to thorough (and, consequently, slower) language processing.

7. Conclusions and Future Work

As concluding remarks, we would like to summarize the main contributions of this work:

1. An end-to-end pipeline process for event extraction and representation for the Portuguese language.
2. Tag set conversion of Universal Dependencies (UD) for a Portuguese tree-bank data set (discussed in Section 4.4).
3. Development of an integrated dependency parser (with an accuracy of 94.72%) and semantic role labeling (with an $F1$ measure of 71.06%) for the Portuguese Language.
4. Creation of an ontology (knowledge base) for the criminal law domain.
5. Association between Eurovoc and IATE terms with the extracted terms.
6. Creation and annotation of a new Portuguese corpus for event analysis (Data set 2, as discussed in Section 5).
7. Event extraction, with:
    - Precisions of 0.6924, 0.5966, and 0.5764; recalls of 0.9070, 0.7603, and 0.6175; and $F1$ scores of 0.7853, 0.6685, and 0.5962, for verbs, subjects, and objects, respectively, for PropBank-BR (Data set 1).
    - Precision of 0.8494 for verbs for Data set 2.

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 of 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 is achievable. For instance, using Levenshtein distance on a word level is quite a naïve approach. We plan to explore other methods of checking for similarity, such as those outlined by Gali et al. [50].

As future work, we intend to continue to try to improve the individual modules, with a specific focus on the SRL and event extraction modules. We also plan to extend this work to the automatic creation of event timelines and to apply the work to open-access texts, such as online news.

Author Contributions: Conceptualization, P.Q.; Investigation, K.R. and R.B.; Methodology, P.Q.; Supervision, P.Q. and V.N.; Validation, K.R. and R.B.; Writing—original draft, P.Q., V.N., K.R. and R.B.; Writing—review and editing, P.Q. and V.N.

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

Acknowledgments: The authors would like to thank Portuguese Department of the University of Macau, lead by Professor Ana Leal for helping us during the annotation of Data Set 2.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hogenboom, F.; Frasincar, F.; Kaymak, U.; de Jong, F.; Caron, E. A Survey of Event Extraction Methods from Text for Decision Support Systems. *Decis. Support Syst.* **2016**, *85*, 12–22, doi:10.1016/j.dss.2016.02.006. [CrossRef]
2. Guarino, N.; Oberle, D.; Staab, S. What Is an Ontology? In *Handbook on Ontologies*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 1–17. doi:10.1007/978-3-540-92673-3_0. [CrossRef]
3. Amato, F.; Moscato, V.; Picariello, A.; Sperlí, G. Extreme events management using multimedia social networks. *Future Gener. Comput. Syst.* **2019**, *94*, 444–452, doi:10.1016/j.future.2018.11.035. [CrossRef]
4. Amato, F.; Castiglione, A.; Moscato, V.; Picariello, A.; Sperlí, G. Multimedia Summarization Using Social Media Content. *Multimed. Tools Appl.* **2018**, *77*, 17803–17827, doi:10.1007/s11042-017-5556-2. [CrossRef]
5. International Conference on the Computational Processing of Portuguese Language. Available online: http://www.propor.org/ (accessed on 6 May 2019).
6. de Abreu, S.C.; Bonamigo, T.L.; Vieira, R. A review on Relation Extraction with an eye on Portuguese. *J. Braz. Comput. Soc.* **2013**, *19*, 553–571, doi:10.1007/s13173-013-0116-8. [CrossRef]
7. Duran, M.S.; Aluísio, S.M. Propbank-Br: A Brazilian Treebank annotated with semantic role labels. In Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, 23–25 May 2012; Calzolari, N., Choukri, K., Declerck, T., Dogan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Eds.; European Language Resources Association (ELRA): Paris, French, 2012; pp. 1862–1867.
8. Agatha an Intelligent Open Source Analysis System. Available online: http://www.agatha-osi.com/ (accessed on 6 May 2019).
9. 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 Proceedings of the Working Notes of CLEF 2018-Conference and Labs of the Evaluation Forum, Avignon, France, 10–14 September 2018.
10. 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). Association for Computational Linguistics, Santa Fe, NM, USA, 20–21 August 2018; pp. 28–41.
11. 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), Minneapolis, MN, USA, 6–7 June 2019; pp. 520–524.
12. Raiyani, K.; Quaresma, P. Keyword & Machine Learning Based Japanese Statute Law Retrieval and Entailment Task at COLIEE-2019. In Proceedings of the 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), Montréal, QC, Canada, 17–21 June 2019.
13. Mitamura, T.; Liu, Z.; Hovy, E.H. Events Detection, Coreference and Sequencing: What's next? Overview of the TAC KBP 2017 Event Track. In Proceedings of the 2017 Text Analysis Conference, TAC 2017, Gaithersburg, MD, USA, 13–14 November 2017.

14. Collovini, S.; Pugens, L.; Vanin, A.A.; Vieira, R. Extraction of Relation Descriptors for Portuguese Using Conditional Random Fields. In *Advances in Artificial Intelligence—IBERAMIA 2014*; Bazzan, A.L., Pichara, K., Eds.; Springer: Berlin/Heidelberg, Germany, 2014; pp. 108–119.
15. Bonamigo, T.L.; Vieira, R. A Model for Information Extraction in Portuguese Based on Text Patterns. In Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing-Volume 2, Samos, Greece, 24–30 March 2013; Springer: Berlin/Heidelberg, Germany, 2013; pp. 359–368, doi:10.1007/978-3-642-37256-8_30. [CrossRef]
16. Doddington, G.; Mitchell, A.; Przybocki, M.; Ramshaw, L.; Strassel, S.; Weischedel, R. The Automatic Content Extraction (ACE) Program Tasks, Data, and Evaluation. In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC-2004), Lisbon, Portugal, 26–28 May 2004; European Language Resources Association (ELRA): Lisbon, Portugal, 2004.
17. Xu, W.; Yuan, C.; Li, W.; Wu, M.; Wong, K.F. Building Document Graphs for Multiple News Articles Summarization: An Event-Based Approach. In *Computer Processing of Oriental Languages. Beyond the Orient: The Research Challenges Ahead*; Matsumoto, Y., Sproat, R.W., Wong, K.F., Zhang, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2006; pp. 181–188.
18. Ahn, D. The Stages of Event Extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2006; pp. 1–8.
19. Halpin, H.; Moore, J.D. Event Extraction in a Plot Advice Agent. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia, 17–18 July 2006; Association for Computational Linguistics: Stroudsburg, PA, USA, 2006; pp. 857–864, doi:10.3115/1220175.1220283. [CrossRef]
20. Xu, F.; Uszkoreit, H.; Li, H. *Automatic Event and Relation Detection with Seeds of Varying Complexity*; In Proceedings of the AAAI Workshop Event Extraction and Synthesis, Boston, MA, USA, 16–17 July 2006; pp. 12–17.
21. Sakaki, T.; Okazaki, M.; Matsuo, Y. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In Proceedings of the 19th International Conference on World Wide Web, Raleigh, NC, USA, 26–30 April 2010; ACM: New York, NY, USA, 2010; pp. 851–860, doi:10.1145/1772690.1772777. [CrossRef]
22. Benson, E.; Haghighi, A.; Barzilay, R. Event Discovery in Social Media Feeds. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, Portland, OR, USA, 19–24 June 2011; Association for Computational Linguistics: Stroudsburg, PA, USA; pp. 389–398.
23. Ritter, A.; Mausam; Etzioni, O.; Clark, S. Open Domain Event Extraction from Twitter. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 12–16 August 2012; ACM: New York, NY, USA, 2012; pp. 1104–1112, doi:10.1145/2339530.2339704.
24. Zhao, X.; Jiang, J.; He, J.; Song, Y.; Achananuparp, P.; Xin, W.; Jing, Z.; Jing, J.; Yang, H.; Achananuparp, S.P.; et al. Topical keyphrase extraction from twitter. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL-2011), Portland, OR, USA, 19–24 June 2011.
25. Zhou, Y.; De, S.; Moessner, K. Real World City Event Extraction from Twitter Data Streams. *Proced. Comput. Sci.* **2016**, *98*, 443–448, doi:10.1016/j.procs.2016.09.069. [CrossRef]
26. Zong, B.; Wu, Y.; Song, J.; Singh, A.K.; Cam, H.; Han, J.; Yan, X. Towards Scalable Critical Alert Mining. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 24–27 August 2014; ACM: New York, NY, USA; pp. 1057–1066, doi:10.1145/2623330.2623729. [CrossRef]
27. EU Vocabularies. Available online: https://publications.europa.eu/en/web/eu-vocabularies (accessed on 6 May 2019).
28. Carreras, X.; Chao, I.; Padró, L.; Padro, M. FreeLing: An open-source suite of language analyzers. In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04), Lisbon, Portugal, 26–28 May 2004.
29. Polyglot a natural language pipeline that supports massive multilingual applications. Available online: https://pypi.org/project/polyglot/ (accessed on 6 May 2019).
30. Compact Language Detector 2. Available online: https://github.com/CLD2Owners/cld2 (accessed on 6 May 2019).

31. Brants, T. TnT: A statistical part-of-speech tagger. In Proceedings of the Sixth Conference on Applied Natural Language Processing, Seattle, WA, USA, 29 April–4 May 2000; pp. 224–231.
32. 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, Edmonton, Canada, 31 May–1 June 2003.
33. Portuguese Universal Propositions. Available online: https://github.com/System-T/UniversalPropositions/tree/master/UP_Portuguese-Bosque (accessed on 6 May 2019).
34. FreeLing 4.1 User Manual. Available online: https://talp-upc.gitbook.io/freeling-4-1-user-manual/v/master/tagsets/tagset-pt (accessed on 6 May 2019).
35. Automated Event Extraction Model for Multiple Linked Portuguese Documents. Available online: https://github.com/kraiyani/Automated-Event-Extraction-Model-for-Multiple-Linked-Portuguese-Documents/blob/master/Universal_to_eagle_tagset.xlsx (accessed on 6 May 2019).
36. Training and Development Dataset for Automated Event Extraction Model for Multiple Linked Portuguese Documents. Available online: https://github.com/kraiyani/Automated-Event-Extraction-Model-for-Multiple-Linked-Portuguese-Documents (accessed on 6 May 2019).
37. 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, 14 April 2019.
38. Guarino, N.; Giaretta, P. Ontologies and knowledge bases: Towards a terminological clarification. In *Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing*; IOS Press: Amsterdam, The Netherlands, 1995; pp. 25–32.
39. Van Hage, W.R.; Malaisé, V.; Segers, R.; Hollink, L.; Schreiber, G. Design and use of the Simple Event Model (SEM). *Web Semant. Sci. Serv. Agents World Wide Web* **2011**, *9*, 128–136. [CrossRef]
40. IATE (Interactive Terminology for Europe). Available online: https://iate.europa.eu/home (accessed on 6 May 2019).
41. Protege. Available online: https://protege.stanford.edu/ (accessed on 6 May 2019).
42. GraphDB. Available online: http://graphdb.ontotext.com/ (accessed on 6 May 2019).
43. EU Vocabularies, Thesauri, 1216 criminal law. Available online: https://publications.europa.eu/en/web/eu-vocabularies/th-concept-scheme/-/resource/eurovoc/100180?target=Browse (accessed on 6 May 2019).
44. Levenshtein Distance. Available online: https://en.wikipedia.org/wiki/Levenshtein_distance (accessed on 6 May 2019).
45. Development Dataset of Automated Event Extraction Model for Multiple Linked Portuguese Documents. Available online: https://github.com/kraiyani/Automated-Event-Extraction-Model-for-Multiple-Linked-Portuguese-Documents/blob/master/pt_devel.txt (accessed on 6 May 2019).
46. Validation Dataset of Automated Event Extraction Model for Multiple Linked Portuguese Documents. Available online: https://github.com/kraiyani/Automated-Event-Extraction-Model-for-Multiple-Linked-Portuguese-Documents/blob/master/pt_train.txt (accessed on 6 May 2019).
47. PortLEX Project, PropBank.Br Dataset. Available online: http://www.nilc.icmc.usp.br/portlex/index.php/en/projects/propbankbringl (accessed on 6 May 2019).
48. Gamallo, P.; Garcia, M.; Pineiro, C.; Martinez-Castaño, R.; Pichel, J.C. LinguaKit: A Big Data-based multilingual tool for linguistic analysis and information extraction. In Proceedings of the 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS), Valencia, Spain, 15–18 October 2018; pp. 239–244.
49. Cardoso, N. Rembrandt—A named-entity recognition framework. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, 21–27 May 2012; European Language Resources Association (ELRA): Istanbul, Turkey, 2012; pp. 1240–1243.
50. Gali, N.; Mariescu-Istodor, R.; Hostettler, D.; Fränti, P. Framework for syntactic string similarity measures. *Expert Syst. Appl.* **2019**, *129*, 169–185. [CrossRef]