Automated Event Extraction Model for Linked Portuguese Documents

Kashyap Raiyani, Teresa Gonçalves, Paulo Quaresma and Vitor Beires Nogueira
{kshyp,tcg,pq,vbn}@uevora.pt

Computer Science Department, University of Évora, Portugal

Abstract

In recent times, Machine Learning is booming and researchers are applying it to the most conceivable cases such as the area of linked documents. This article presents a process of automatic event extraction from Portuguese linked document whose accuracy (95.00%) was calculated by manual verification. With the help of an ontological structure, extracted events are mapped as a knowledge graph that represents the named entities and the events associated with each document. Such graphs are accessible through SPARQL queries. This way, the information existing in the linked documents can be easily accessed by resorting to a question-answering approach.

1 Introduction

Event extraction is an important and fundamental task in Natural Language Processing (NLP) and computational linguistics [JM09,MS99]. Event extraction is crucial in understanding user-generated text such as news, messages and blogs and is represented and stored in a structured way, e.g., in databases and graphs. A specific type of knowledge that can be extracted from the text by means of text mining, which can be represented as a complex combination of relations linked to a set of empirical observations from texts [BYZC15].

Natural language understanding involves identifying, classifying, and integrating information about events and other propositions mentioned in the text. While much effort has been invested in generic methods for analyzing single sentences and detecting the propositions they contain, little thought and effort have been put into the integration step: how to systematically consolidate and represent information contributed by propositions originating from multiple texts. Consolidating such information, which is typically both complementary and partly overlapping, is needed to construct multi-document summaries, to combine evidence when answering questions that cannot be answered based on a single sentence or document, and to populate a knowledge base while relying on multiple pieces of evidence. Yet, the burden of integrating information across multiple texts is currently delegated to downstream applications, leading to various partial solutions in different application domains. This paper suggests that a common consolidation step and a corresponding knowledge representation should be part of the "standard" semantic processing pipeline, to be shared by downstream applications. Specifically, we pursue an Open Knowledge Representation that captures the information expressed jointly in multiple texts while relying solely on the terminology appearing in those texts, without requiring predefined external knowledge resources or schemata. We do that by first extracting textual predicate-argument tuples,

In: A. Jorge, R. Campos, A. Jatowt, S. Bhatia (eds.): Proceedings of the Text2StoryIR'19 Workshop, Cologne, Germany, 14-April-2019, published at http://ceur-ws.org

each corresponding to an individual proposition mention. We then merge these mentions by accounting for proposition referencing, an extended notion of event coreference. This process yields consolidated propositions, each corresponding to a single fact, or assertion, in the described scenario. Similarly, entity coreference links are used to establish the reference to real-world entities. Taken together, our proposed representation encodes information about events and entities in the real world, similarly to what is expected from structured knowledge representations. Yet, being an open text-based representation, we record the various lexical terms (actor, place, time and object) used to describe the scenario. Further, we model information redundancy and containment among these terms through lexical entailment. Overall, our main contribution is in proposing to create a consolidated representation for the information contained in multiple texts, and in specifying how such representation can be created based on entity and event reference and lexical entailment. An accompanying contribution is our annotated dataset, which can be used to analyze the involved phenomena and their interactions, and as a development and evaluation set for the automated generation of Open Knowledge Representation structures. We further note that while this paper focuses on creating an open representation, by consolidating all the extracted information and storing them into the ontological structure. This way, the proposed method overcomes the limitation of referencing within multiple documents.

This paper continues with an elaboration of event extraction approaches in Section 2 and subsequently, Section 3 presents SRL parser, SVO extraction and the ontological structure in detail. Section 4 presents the evaluation phase of the proposed system and Section 5 concludes the paper.

2 Related Work

Previous work done in the area of event extraction is mainly application specific. Automatic Content Extraction (ACE) is a tool that extracts Entities, Relations, and Events. ACE takes input in sgm/sgml format which makes user input restricted in ACE [DMP+04]. Yuan et al. [XYL+06] proposed an event-based approach to visualize documents as a graph on different conceptual granularities. Where, in [Ahn06], authors have treated events as undeniably temporal entities. In comparison to ACE, event extraction task is done in modules, each of which is handled by a machine-learned classifier. Result and methodology of this approach [Ahn06] is better than ACE but it is still domain specific. Halpin et al. [HM06] proposed that events are extracted for story rewriting work, as actions present in the text. Work done in [XUL06] has used domain ontology 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, the tweets from the Twitter network. Sakaki et al. [SOM10] describe a method to detect earthquake-related tweets. The method uses features specific to earthquake application. Benson et al. [BHB11] trained a relation extractor to identify artists and venues from tweets. The method is to develop a graphical model by learning records and records-message alignment. Ritter et al. [RMEC12] describe a method based on latent variable modeling to extract event types described in tweets. Here, features such as tweet popularity and the times of events referred to in the tweets are used. Zhao et al. [ZJH+11] describe a method to extract only the most "topical" keywords from tweets. Similar approaches have also been applied to mine relevant information from non-text sources. For instance, Zong et al. [ZWS+14] describes approximation algorithms to identify critical alerts from a large set of alert sequences. Our approach differs from above-mentioned methods as the system provides a representation of the events of interest only in the form of a set of terms describing representative events. Thus, arbitrary types of events can be discovered without utilizing event type-specific features such as application scenario.

3 SRL Parser, SVO Extraction and Knowledge Base Ontology

This section talks about the overview of the system flow. Starting from processing a Portuguese document to dependency parsing followed by Semantic Role Labeling (SRL) and finally extracting subject-verb-object (SVO). Apart from this, by using SRL tagged document, the proposed model is extracting SVO and storing them in form of triples over an ontology for further usage (information extraction). This way, system overcomes the limitation of referencing within multiple documents.

3.1 SRL Parser for Portuguese

Freeling [CCPP04] is an analysis tool based on the architecture of (Carreras and Padr [Car02]). Currently, it supports SRL tagging over five languages namely, Catalan, Croatian, English, German and Spanish. In this

work, a SRL parser for the Portuguese language is developed using Freeling as the baseline tool. This model is able to do SRL tagging based on dependency parsing. To train and develop the SRL Portuguese model (using machine learning algorithms), Universal Dependencies (UD) Portuguese treebank dataset [RCR$^+$17] was used and processed (tagged) manually. This process included converting of Universal POS tags to EAGLES tagset for the entire dataset as Freeling is only compatible with the EAGLES tagset. For this conversion, Universal Dependencies (UD) [1] was referred. Among 14 categories, the processed development dataset has 580 tags. In this paper, the empathize is on event extraction and knowledge base creation for linked documents.

3.2 SVO Extraction

The developed system is able to identify Actor, Event, Place, Time and Object in reference to passed sentence/file (also known as SVO). The SVO extraction process is described in Algorithm 1. After the extracting SVOs from single/multiple files, they are inserted into the ontology for the creation of the knowledge base, which is discussed in the subsection 3.3.

Algorithm 1 SVO Extraction Algorithm

```
1:  procedure SVO(sentence)
2:      SVO ← []
3:      while Predicates ≠ NULL do                ▷ For-each predicate in sentence
4:          Event ← Predicate
5:          while Arguments ≠ NULL do             ▷ For-each arguments associated with predicate
6:              switch Arguments.role() do
7:                  case A0
8:                      Actor ← Arguments
9:                  case A1
10:                     Object ← Arguments
11:                 case AM-LOC
12:                     Location ← Arguments
13:                 case AM-TMP
14:                     Time ← Arguments
15:         end while
16:         SVO.append(Event, Actor, Object, Location, Time)
17:     end while
18:     return SVO                                ▷ role() returns SRL tagging
19: end procedure
```

3.3 Knowledge Base Ontology

According to [GG95], 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 aimed at answering the question "What kind of objects exist and how are they interrelated?". Thus, describing the logical structure of a domain and the relations between them. The Figure 1 shows the ontology structure of the proposed system.To design this ontology, Simple Event Model [VHMS$^+$11] ontology was referred as a baseline model. The Table 1 shows the data property of the ontology. The entities of the ontology model are listed below.

1. Actor - person involved with event
2. Place - location of the event
3. Time - time of the event
4. Object - that actor act upon
5. Organization - organization involved with event
6. Currency - money involved with event

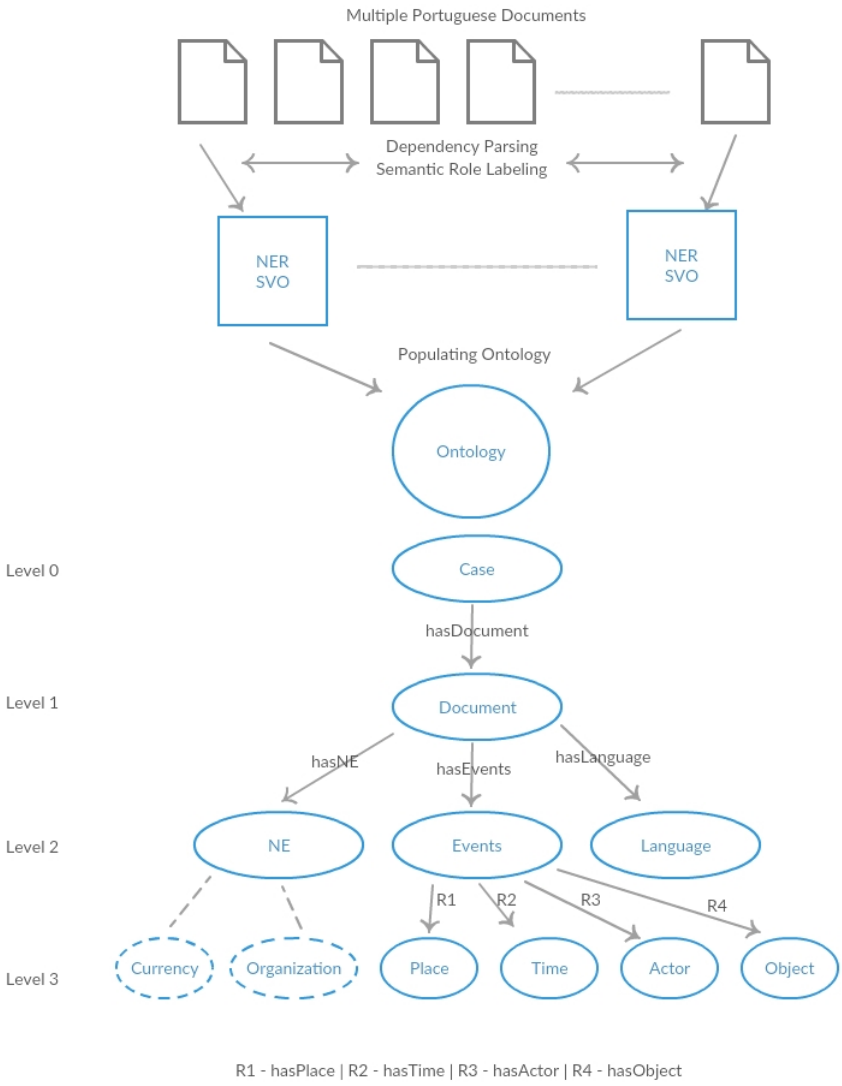

Figure 1: Connection Between Documents and Ontology Structure

Table 1: Ontology Data Property

Node	**Property**
Case	CaseID
Document	Document ID foreach Case
Language	Language foreach Document
NER	NER foreach Document
NER Sub-Classes	Time, Place, Person, Organization, and Currency
Event	Events foreach Document
Event Sub-Classes	Actor, Place, Time, and Object

The Figure 1 shows the ontology, which is populated with 6178 events entries from 51 documents of evaluation dataset (discussed in section 4). For simplicity, populated ontology only consists of the document-event combination. From the figure, the connection between all the document and events is quite visible. Here, other information such as actor, place, time, object, organization and currency associated with the event could also be added. In addition to that, ontology is designed in such a manner that it can incorporate documents of other cases. For example, assuming that the source of linked documents are legal police documents, where each document are under the hood of a particular case. In addition to that, 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 connection among the documents of a single case but rather among all the cases with all the combine 200 documents. This way, the proposed method is able to produce detailed and well-connected knowledge base. For creation of this ontology Protege[2] tool was used and for populating & querying the data, GraphDB[3] was used. The Section 4 discusses the evaluation of the proposed system.

4 Evaluation

The evaluation dataset was created with documents from several online sources, aiming to illustrate and evaluate the performance of the developed systems. It is composed of 51 documents with 1221 sentences and 48914 words. Figure 2 shows the number of sentences and words per document and Figure 3 shows the tagged event by system and manually per document. The Figure 4 shows the missing event tags by the system per document vs. manual and the Table 2 provides the performance overview over evaluation dataset. Some parts of the evaluations are not comprehensive due to the lack of Portuguese gold-standard data. Specifically, the identification of the time and place from the documents. The Training, Development, Evaluation, and System Extracted Events datasets are accessible[4] for further research and comparison purpose.

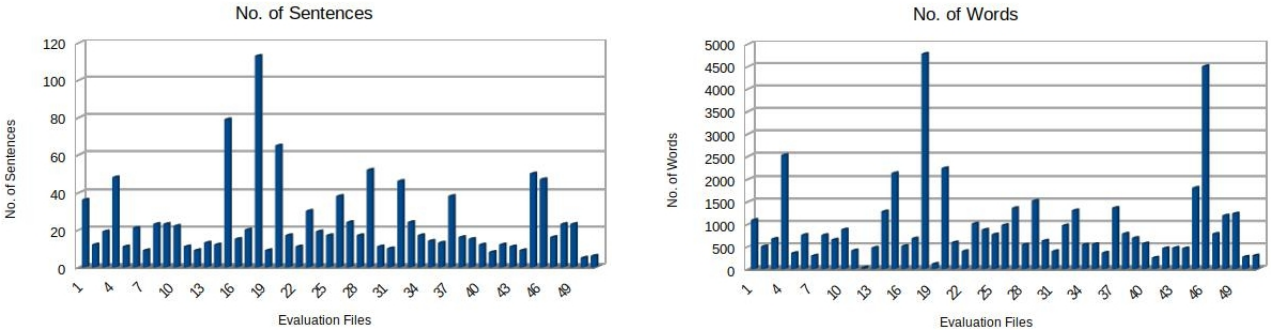

Figure 2: No. of Sentences and Words

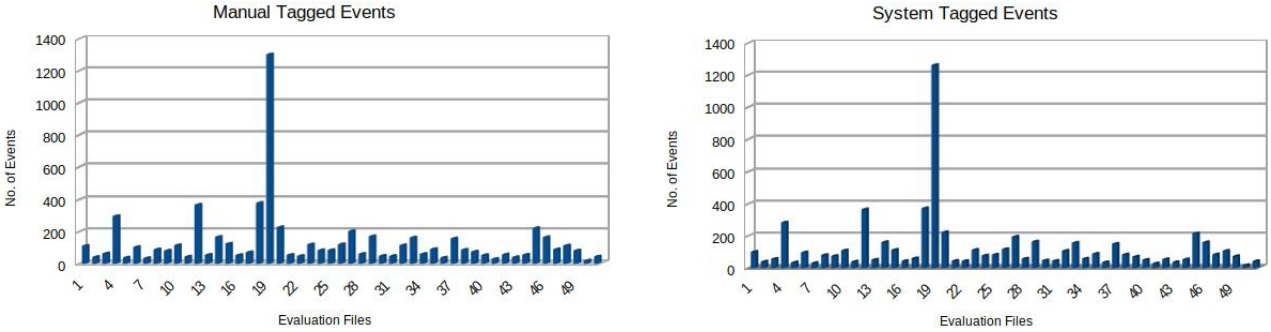

Figure 3: No. of Tagged Events: Manual and System

[1] https://universaldependencies.org/tagset-conversion/pt-freeling-uposf.html
[2] https://protege.stanford.edu/
[3] http://graphdb.ontotext.com/
[4] https://github.com/kraiyani/Automated-Event-Extraction-Model-for-Multiple-Linked-Portuguese-Documents

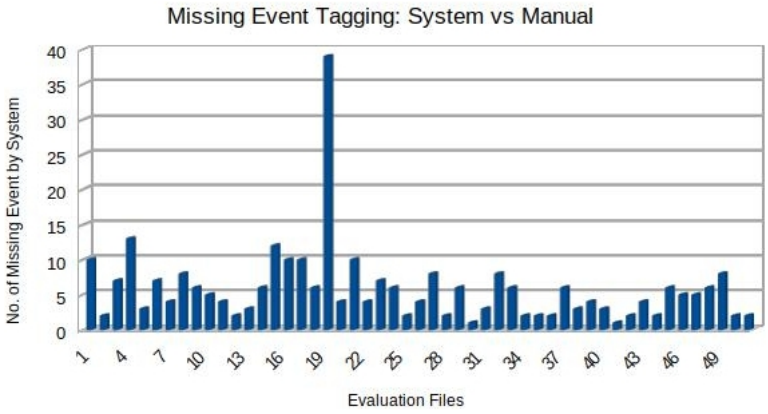

Figure 4: No. of Missing Event Tags by System

Table 2: Computational Results Over Evaluation Dataset

Event Tagging	**Number**
System	6178
Manual	6471
Difference	296
Min. Missing	1
Max. Missing	39
Avg. Missing	5.75
Accuracy	**95.00%**

5 Conclusion and Future Work

In this paper, we propose a pipelined system for event extraction from Linked Portuguese Documents. Evaluation results show that our model is able to extract the majority of events (95.00%). Additionally, due to resorting to ontologies, the system is able to create a knowledge base connection among documents using extracted events. Resulting in an easier platform for information retrieval over linked documents.

We consider that our approach can be improved, mainly due to the following facts: (1) The system lacks discovering time categories associated with events with many others(actor, object and place) left unidentified. (2) Due to this, a proper timeline creation for linked Portuguese documents is not feasible. Lastly, how to address these aspects and generate a more accurate, comprehensive and fine-grained event list with a timeline for linked documents constitutes our further work.

5.0.1 Acknowledgements

The authors would like to thank COMPETE 2020, PORTUGAL 2020 Programs, 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).

References

[Ahn06] David Ahn. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning About Time and Events*, ARTE '06, pages 1–8, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[BHB11] Edward Benson, Aria Haghighi, and Regina Barzilay. Event discovery in social media feeds. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 389–398, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[BYZC15] Jingwen Bian, Yang Yang, Hanwang Zhang, and Tat-Seng Chua. Multimedia summarization for social events in microblog stream. *IEEE Transactions on Multimedia*, 17:216–228, 2015.

[Car02] Xavier Carreras. A flexible distributed architecture for natural language analyzers. In *In Third International Conference on Language Resources and Evaluation, LREC-02*, pages 1813–1817, 2002.

[CCPP04] Xavier Carreras, Isaac Chao, Lluís Padró, and Muntsa Padro. Freeling: An open-source suite of language analyzers. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, 01 2004.

[DMP+04] George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 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, May 2004. European Language Resources Association (ELRA). ACL Anthology Identifier: L04-1011.

[GG95] Nicola Guarino and Pierdaniele Giaretta. Ontologies and knowledge bases: Towards a terminological clarification. In *Towards very Large Knowledge bases: Knowledge Building and Knowledge sharing*, pages 25–32. IOS Press, 1995.

[HM06] Harry Halpin and Johanna D. Moore. 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*, ACL-44, pages 857–864, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[JM09] Daniel Jurafsky and James H. Martin. *Speech and Language Processing (2Nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2009.

[MS99] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA, 1999.

[RCR+17] Alexandre Rademaker, Fabricio Chalub, Livy Real, Cláudia Freitas, Eckhard Bick, and Valeria de Paiva. Universal dependencies for portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling)*, pages 197–206, Pisa, Italy, September 2017.

[RMEC12] Alan Ritter, Mausam, Oren Etzioni, and Sam Clark. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 1104–1112, New York, NY, USA, 2012. ACM.

[SOM10] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 851–860, New York, NY, USA, 2010. ACM.

[VHMS+11] Willem Robert Van Hage, Véronique Malaisé, Roxane Segers, Laura Hollink, and Guus Schreiber. 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.

[XUL06] Feiyu Xu, Hans Uszkoreit, and Hong Li. Automatic event and relation detection with seeds of varying complexity. In *AAAI Workshop Event Extraction and Synthesis*, pages 12–17, Boston, 7 2006. ,.

[XYL+06] Wei Xu, Chunfa Yuan, Wenjie Li, Mingli Wu, and Kam-Fai Wong. Building document graphs for multiple news articles summarization: An event-based approach. In Yuji Matsumoto, Richard W. Sproat, Kam-Fai Wong, and Min Zhang, editors, *Computer Processing of Oriental Languages. Beyond the Orient: The Research Challenges Ahead*, pages 181–188, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[ZJH+11] Xin Zhao, Jing Jiang, Jing He, Yang Song, Palakorn Achananuparp, Wayne Xin, Zhao Jing, Jiang Jing, He Yang, Song Palakorn Achananuparp, Ee peng Lim, and Xiaoming Li. Topical keyphrase extraction from twitter. In *In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL-2011*, 2011.

[ZWS+14] Bo Zong, Yinghui Wu, Jie Song, Ambuj K. Singh, Hasan Cam, Jiawei Han, and Xifeng Yan. Towards scalable critical alert mining. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 1057–1066, New York, NY, USA, 2014. ACM.