


# Knowledge Representation of Crime-Related Events: a Preliminary Approach

Gonçalo Carnaz 

Department of Informatics, University of Évora, Portugal  
d34707@alunos.uevora.pt

Vitor Beires Nogueira 

Department of Informatics, University of Évora, Portugal  
vbn@uevora.pt

Mário Antunes 

School of Technology and Management, Polytechnic Institute of Leiria, Portugal  
INESC-TEC, CRACS, University of Porto, Porto, Portugal  
mario.antunes@ipleiria.pt

---

## Abstract

The crime is spread in every daily newspaper, and particularly on criminal investigation reports produced by several Police departments, creating an amount of data to be processed by Humans. Other research studies related to relation extraction (a branch of information retrieval) in Portuguese arisen along the years, but with few extracted relations and several computer methods approaches, that could be improved by recent features, to achieve better performance results.

This paper aims to present the ongoing work related to SEM (Simple Event Model) ontology population with instances retrieved from crime-related documents, supported by an SVO (Subject, Verb, Object) algorithm using hand-crafted rules to extract events, achieving a performance measure of 0.86 (F-Measure).

**2012 ACM Subject Classification** Information systems → Information retrieval; Information systems → Ontologies

**Keywords and phrases** SEM Ontology, Relation Extraction, Crime-Related Events, SVO Algorithm, Ontology Population

**Digital Object Identifier** 10.4230/OASICS.SLATE.2019.13

## 1 Introduction

We are living in an era of data overloading, produced by machines and humans, and spread over the World Wide Web (WWW). This data has different formats, such as text documents, that is retrieved from heterogeneous sources. Therefore, different approaches have been released during the last decades that extract relevant information. These computer methods extract information in the form of named-entities (Named-Entity Recognition systems), relation extraction (Open Information Extraction or Traditional Information Extraction methods), or semantic roles (Semantic Role Labeling methods).

The extracted named-entities and relations/events could be represented by a knowledge base, such as ontologies that are conceptual models that aim to represent a particular domain, building concepts, notions or properties that represent the knowledge that exists in such domain.

In this paper, we present the ongoing work related to an approach to the SVO algorithm using hand-crafted rules to extract events and a posterior analysis regarding the SEM ontology population with the extracted events and named-entities.



© Gonçalo Carnaz, Vitor Beires Nogueira, and Mário Antunes;  
licensed under Creative Commons License CC-BY

8th Symposium on Languages, Applications and Technologies (SLATE 2019).

Editors: Ricardo Rodrigues, Jan Janoušek, Luís Ferreira, Luísa Coheur, Fernando Batista, and Hugo Gonçalves Oliveira; Article No. 13; pp. 13:1–13:8



OpenAccess Series in Informatics

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

The remainder of this paper is organized as follow: section 2 describes the related work regarding information retrieval and ontologies; in section 3 describes the background knowledge regarding the SEM ontology; in section 4 presents the Natural Language Processing (NLP) pipeline setup to support the relations/events extraction; in section 5 describes the relations/events extraction using the SVO algorithm; in section 6 discusses the results obtained and finally the section 7 explains the future work and conclusions.

## 2 Related Work

In this section we analyze the previous works related to Information Retrieval (IR) using the Portuguese language. In 2008, Mota et al. [14] proposed the SEI-Geo System to extract *part-of* relationships between geographic entities, using hand-crafted patterns to detect geographic entities. SeRELeP [2] proposed to recognize three different types of relationships (*occurred*, *part-of*, and *identity*) supported by heuristic rules applied to linguistic and syntactic features. The REMBRANDT [3] system aims to identify 24 different relations types using hand-crafted rules-based and supported by two knowledge bases: DBpedia and Wikipedia.

Garcia et al. [9] proposed in 2011, extracting *occupation* relationship instances over Portuguese texts. Training sentences using a Support Vector Machines classifier where each word evaluated (the lemma and the POS-tag) by computing the syntactic dependencies between words, using a syntactic parser.

In 2014, Souza et al. [17] proposed a supervised OIE (Open Information Extraction) approach for extracting relational triples from Portuguese texts (using Corpus CETENFolha).

Collovini et al. [5] proposed in 2016, an evaluation of the Conditional Random Fields (CRF) classifier to extract relations between named-entities, such as Organizations, Locations, or Persons from Portuguese texts.

In 2017, Ricardo Rodrigues [16] proposed the RAPPORT system, a Portuguese Question-Answering System that uses a NLP pipeline with a fact extraction task (based on syntax and semantic patterns).

Along the years, several works were proposed related to ontologies applied to the criminal domain, and how to represent concepts/terms retrieved from crime-related documents. Despres et al. [18] proposed in 2004, the alignment of terms from a legal domain and a core ontology, generating a legal ontology from a European community legislation text (reuse of LRI-Core [1] and DOLCE<sup>1</sup>).

Casanovas et al. [4] developed in 2007 an Ontology of Professional Judicial Knowledge, called *OPJK*. Based on a manual selection of relevant terms from legal questions and modeled according to the the *DILIGENT* methodology. Using the *PROTON* [6] ontology as an upper-level ontology. Additionally, a methodology was presented to build a multilingual semantic lexicon for the law. Additionally, the Eurovoc thesaurus is integrated for project lexicon enrichment purposes [19]. Francesconi et al. [8] aimed to ensure that legal drafters and decision makers lead to control over a legal language, specified by *DALOS* Knowledge System. The *DALOS* project is divided into ontological and lexical layers. A domain ontology supports the Ontological Layer; and the *LOIS* database supports lexical Layer.

In 2009, Hoekstra et al. [20] proposed a legal core ontology that was part of the Legal Knowledge Interchange Format, known as *LKIF* Core Ontology, as a core in a legal knowledge system. Saskia et al. [20] described a system called *OWL Judge* using OWL 2 reasoning

---

<sup>1</sup> See <http://www.loa.istc.cnr.it/old/DOLCE.html> [Accessed: April 2019].

for legal assessments, where norms are represented in LKIF Core Ontology, associated with design patterns for norms and user cases definition. The use case used was the University Library Regulations.

Mehmet et al. [12] proposed in 2010, a class diagram to define an ontology, applied to money laundering schemes. Using class diagram objects to represent terms and relations used in money laundering schemes, e.g., people, organization, portfolios, messages, communication medium, invoice and identification documents.

Rajpu et al. [15] tried in 2014 to find suspicious financial transactions through an expert system, based on an ontology and a set of rules. The authors created a set of classes, objects, and properties that represent the transactions to be processed by the expert system — additionally, a set of rules, using SWRL (Semantic Web Rules Language), in order to infer new knowledge through existing knowledge. Also in 2014, and using Akoma Ntoso XML schema, LKIF-Core, Legal Case Ontology, JudO and Carneades Argument Format, [10] proposed a basic and semantic annotation approach for complaints, using a Serbian Judiciary use case for validation.

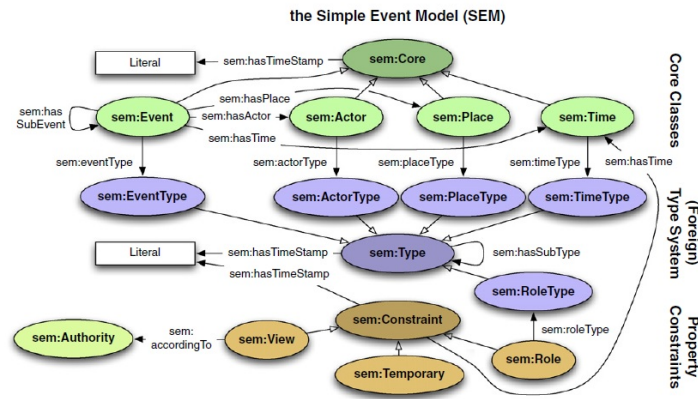
In 2017, Oliveira Rodrigues et al. [7] proposed a reuse of *UFO-B* and LKIF ontology, as a concept model for property crimes representation applied to the Brazilian Criminal Code, called *OntoPropertyCrime*. Additionally, a theory of crime is formalized, called *OntoCrimeAlpha*. Mezghanni et al. [13] proposed *CrimAr* ontology is defined by a handcrafted approach, for the Arabic legal domain, supported by LRI-Core as top-level ontology. McDaniel et al. [11] proposed a framework, based on an ontology for physical evidence from a crime scene. They are adding an identity judgment (in an id-situation) aligned to legal cases. The ontology includes a situation ontology, focus on physical evidence.

### 3 Background Knowledge

Developing ontologies from scratch can be costly, lengthy and with several points of view for the same concept. Therefore, the reuse of existing ontologies, with slight adaptations to the study domain, could reduce time and cost regarding ontology construction. At this stage of the work, the purpose is to represent existing events in crime-related documents, such as persons, locations, organizations or time/date. Ontology could support the knowledge representation that could answer the following questions: *Who did what?*, *Where?*, *When?*, *How?* *Why?*.

Figure 1 shows the SEM ontology, that was created to model events that are present in various application domains, without making assumptions about the domain-specific vocabularies and without no connection to any domain, for example historical, cultural heritage or geographical domains. If we look carefully at newspaper news, such as crime-related news, we can easily denote that events are also central elements, because news is based on events that occurred in a fixed or extended time, with entities (such as persons, locations or objects).

The SEM ontology is based on four main classes: Events, Actor, Place and Time. The events on the SEM ontology are represented by the class *sem:Event*, this being the central class where the ontology is based. Having as properties: *eventProperty*, *eventType*, *hasSubEvent* and *SubEventOf*. The *sem:Actor* class was proposed to describe “*who or what participated, who is doing something*”. This class (a powerful entity of the domain that can activate or perform events) holds instances (retrieved from corpus) that are part of a given event, actively or passively. We can not see the actors only as persons, but also as objects, which are animate or inanimate and physical or not physical. The *sem:Place* is the class meant

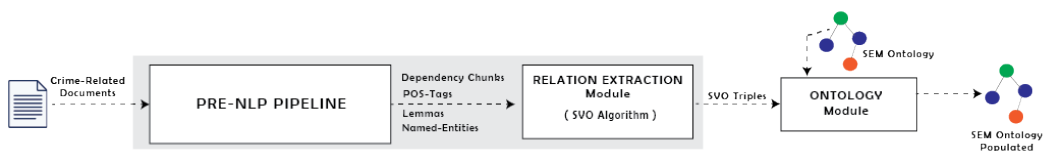


■ **Figure 1** SEM Ontology. [Retrieved from <https://semanticweb.cs.vu.nl/2009/11/sem/>]

to describe “where” is something happening. Places are locations where an Event happens. Neither do they need to have any significance apart from them being the location of an Event. Finally, the *sem:Time* is the class meant to describe “when” is something happening.

#### 4 Natural Language Processing Pipeline Setup

Figure 2 describes a typical NLP (Natural Language Processing) pipeline. First, as data input, we have the crime-related documents retrieved from a Portuguese newspaper; in second, the first phase of the proposed pipeline (Stop-Words Removal, Sentence Detection, Tokenization, Named-Entity Recognition, POS Tagging, Lemmatization and a Dependency Parser); followed by Relation Extraction Module supported by the SVO algorithm and finally, the Ontology Module that have as input (SEM Ontology) and as output, the SEM ontology populated with instances retrieved from crime-related documents.



■ **Figure 2** NLP pipeline (high-level design) proposal.

For pipeline prototyping, we used the RAPPORT [16] system, with some tweaks in the NLP pipeline, such as:

- The Stop-Words Removal task was introduced to remove the words that are not relevant for the NLP processing, using an external file with Stop-Words<sup>2</sup>;
- The Named-Entity Recognition task was training with the following corpus<sup>3</sup>: Amazonia + CETEMPúblico;
- Dependency parser model for Portuguese (MaltParser), using Bosque<sup>4</sup> (contains both European (CETEMPúblico) and Brazilian (CETENFolha) variants) ConLL treebank.

<sup>2</sup> See <https://github.com/stopwords-iso/stopwords-pt> [Accessed: April 2019]

<sup>3</sup> See <https://www.linguateca.pt/Floresta/corpus.html> [Accessed: April 2019]

<sup>4</sup> See [https://github.com/UniversalDependencies/UD\\_Portuguese-Bosque](https://github.com/UniversalDependencies/UD_Portuguese-Bosque) [Accessed: April 2019]

## 5 SVO Algorithm Proposal for Event Extraction

Our approach is based on SVO (Subject, Verb, Object) sentence analysis to construct triples data by parsing crime-related sentences, using the Maltparser<sup>5</sup> (dependency parser) tool, and then extracting SVO triples from parser sentences. Because the Portuguese language is supported by a word order language, such as SVO, VSO, VOS, OVS or OSV, the algorithm must be adapted to identify all variations.

The algorithm 1 describes the instructions to retrieve the subjects, verbs, and objects. The entities types are identified and collected to populate the ontology, such as a person, locations, places or time/date. It is using hand-crafted rules based on syntactic and semantic features to extract relations (verbs) between entities.

---

### Algorithm 1: SVO Algorithm.

---

```

1 Subject, Object, Verb ← NULL;
  // Extracted named-entities with the Named-Entity Recognition module
  // by its tokens
2 Entities ← NamedEntityRecognitionModule(tokens);
  // Extract the tokens in CoNLL format with Dependency Parser
  // (Maltparser tool)
3 CoNLLToken ← DependencyParser ();
  // Identify the number of verbs by its verbal tense
4 NumVerbs ← RetrieveNumberOfVerbsBySentence ();
  // For every Named-Entity Detected
5 for entities ← 0 to n do
6   for CoNLLToken ← 0 to n do
7     if CoNLLToken ← matches the (VerbTense) and NumVerbs > 0 then
8       Verb ← EventDetected;
9       decrement NumVerbs;
10    end
11    if CoNLLToken ← contains a (NamedEntity) then
12      if CoNLLToken ← matches the Subject then
13        Subject ← SubjectDetected;
14        SubEntType ← NamedEntityType;
15      end
16      if CoNLLToken ← contains the (RelationDependency) then then
17        Object ← ObjectDetected;
18        ObjEntType ← NamedEntityType;
19      end
20    end
21    SEMOntologyInstances(Subject, Verb, Object, SubEntType, ObjEntType);
22  end
23 end

```

---

Aforementioned, the algorithm is based on the extraction of the subject, verb and object, each identified with the help of the dependency parser in ConLL format (token). Also, the NER task detected the named-entities (subject or object), used to delimit the verbs,

<sup>5</sup> See <http://maltparser.org/> [Accessed: April 2019]

that could be combined to create a relation extraction tagged sentence with the following format: *<Entity as Subject><Verb as event><Entity as Object>*.

We aim to detect the highest number of relations, by the verbs in their different forms (tense), detecting the total number of verbs in a sentence, and cyclically detecting the parallel entities. Finally, the subjects, verbs, objects, and named-entities types are used to populated the SEM ontology, using the method (SEMontologyInstances).

## 6 Preliminary Results

As preliminary results, we evaluated a set of sentences related to crime extracted from a Portuguese newspaper. As an example, the following sentence was evaluated with SVO algorithm: “*Arminda Marta foi detida a 21/08/1976.*” (in Portuguese) or “*Arminda Marta was arrested in 21/08/1976.*” (in English), obtaining the following results:

- In Portuguese: “<Subject>**Arminda Marta**</Subject> <Verb>**deter**</Verb> <Object>**21/08/1976**</Object>”;
- In English: “<Subject>**Arminda Marta**</Subject> <Verb>**arrested**</Verb> <Object>**21/08/1976**</Object>”;

Also we have obtained the candidate instances to populate the SEM ontology:

- Events: “**deter, ser**” (in Portuguese), “**arrest, to be**” (in English);
- Actor: “**Arminda Marta**”;
- Time: “**21/08/1976**”.

The crime event is the result of criminal behavior and consists of the offense (an actor) to an interested protected by Law. Some of the sentences analyzed enumerated crime types, such as “*Pedro é suspeito de violar, sequestrar e agredir uma jovem em Braga.*”(in Portuguese), or “*Pedro are suspect of raping, kidnapping and assaulting a young woman in Braga.*”(in English). Therefore, the criminal domain has the main event - the crime (a sequence of events that lead to crime type), in its different crime types, such as violation, abduction, or aggression. Moreover, these different types are not reproduced by verbs (in some cases, because ”kill” or ”matar” in Portuguese, is a verb), such as homicide (in Portuguese, ”homicidio”) that is a male noun.

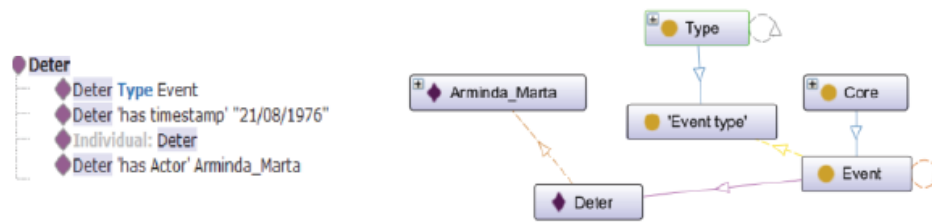
The table 1 shows the results obtained by evaluating the measures, such as the Correct Events (CE), Identified Events (EI) and Total Events (TE). The SVO algorithm obtained an F-measure result of 0.86, that is a trade-off between precision and recall measures.

■ **Table 1** Results of the SVO algorithm evaluation.

Correct Events	Identified Events	Total Events	P	R	F1
209	221	267	0,95	0,78	0,86

Figure 3 represents the extracted entities manually populated, using the Protege<sup>6</sup> tool. As we can see, the SEM Ontology could represent the event extracted and the named-entities: hasActor, an object property (Arminda Marta) and hasTimeStamp, a data property (21/08/1976).

<sup>6</sup> See <https://protege.stanford.edu/> [Accessed: April 2019]



■ **Figure 3** SEM Ontology populated example, using the sentence above enumerated.

## 7 Conclusion and Future Work

Concluding, the knowledge representation of already developed ontologies allows us to remove the lack of time and resources, like adapting the SEM ontology to our domain. Therefore, the representation of the events, by themselves, and the named-entities extracted from the crime-related documents, allow the representation of the event (crime and others) in SEM ontology.

The SVO approach allows us to extract, even to a limited extent, the verbs as events and the named-entities. There is a way to improve the extraction of events and named-entities and the relationship between them, where our work must continue to be improved.

For future work, we enumerated the following items to improve our work:

- improve the SVO algorithm to detect the variations, such as VSO, VOS, OVS or OSV;
- use a large dataset related to crime, created or reused, to test our approach and evaluate the performance measures (Precision, Recall, and F-Measure);
- extract crime related concepts that denotes events, such as homicide, abduction or others;
- extract relations that are important to detect geo-localization (like “District-of”, “County-of”, “Street”, “Country”) of named-entities, such as persons, objects or organizations;
- adapt the SEM ontology regarding the crime related concepts and properties;

## References

- 1 JAPJ Breukers and RJ Hoekstra. Epistemology and ontology in core ontologies: FOLaw and LRI-Core, two. *Citeseer*, 2004.
- 2 José Guilherme Mírian Bruckschen, Re-nata Vieira Souza, and Sandro Rigo. Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, page 436, 2008.
- 3 Nuno Cardoso. Rembrandt-reconhecimento de entidades mencionadas baseado em relações e análise detalhada do texto. *quot; Encontro do Segundo HAREM (Universidade de Aveiro Portugal 7 de Setembro de 2008)*, 2008.
- 4 Pompeu Casanovas, Núria Casellas, Christoph Tempich, Denny Vrandečić, and Richard Benjamins. OPJK and DILIGENT: ontology modeling in a distributed environment. *Artificial Intelligence and Law*, 15(2):171–186, 2007.
- 5 Sandra Collovini, Gabriel Machado, and Renata Vieira. A Sequence Model Approach to Relation Extraction in Portuguese. In *LREC*, 2016.
- 6 John Davies. Lightweight ontologies. In *Theory and Applications of Ontology: Computer Applications*, pages 197–229. Springer, 2010.
- 7 Cleyton Mário de Oliveira Rodrigues, Frederico Luiz Goncalves De Freitas, and Ryan Ribeiro De Azevedo. An ontology for property crime based on events from ufo-b foundational ontology.



- In *2016 5th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 331–336. IEEE, 2016.
- 8 Enrico Francesconi, Pier-Luigi Spinosa, and Daniela Tiscornia. A Linguistic-ontological Support for Multilingual Legislative Drafting: the DALOS Project. In *LOAIT*, pages 103–111, 2007.
  - 9 Marcos Garcia and Pablo Gamallo. Evaluating various linguistic features on semantic relation extraction. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 721–726, 2011.
  - 10 Marko Marković, Stevan Gostojić, and Zora Konjović. Structural and semantic markup of complaints: Case study of Serbian Judiciary. In *2014 IEEE 12th International Symposium on Intelligent Systems and Informatics (SISY)*, pages 15–20. IEEE, 2014.
  - 11 Marguerite McDaniel, Emma Sloan, William Nick, James Mayes, and Albert Esterline. Ontologies for situation-based crime scene identities. In *SoutheastCon 2017*, pages 1–8. IEEE, 2017.
  - 12 Murad Mehmet and Duminda Wijesekera. Ontological Constructs to Create Money Laundering Schemes. In *STIDS*, pages 21–29. Citeseer, 2010.
  - 13 Imen Bouaziz Mezghanni and Faiez Gargouri. CrimAr: A Criminal Arabic Ontology for a Benchmark Based Evaluation. *Procedia Computer Science*, 112:653–662, 2017.
  - 14 Cristina Mota and Diana Santos. Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. In *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, chapter : Geo-ontologias e padrões para reconhecimento de locais e de suas relações em textos: o SEI-Geo no Segundo HAREM, page 436. Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM, 2008. doi:10.1103/PhysRevB.82.193405.
  - 15 Quratulain Rajput, Nida Sadaf Khan, Asma Larik, and Sajjad Haider. Ontology based expert-system for suspicious transactions detection. *Computer and Information Science*, 7(1):103, 2014.
  - 16 Ricardo Manuel da Conceição Rodrigues. *RAPPORT: A Fact-Based Question Answering System for Portuguese*. PhD thesis, Universidade de Coimbra, 2017.
  - 17 Erick Nilsen Pereira Souza and Daniela Barreiro Claro. Extração de relações utilizando features diferenciadas para português. *Linguamática*, 6(2):57–65, 2014.
  - 18 Sylvie Szulman Sylvie Despres. Construction of a legal ontology from a european community legislative text. In *Legal Knowledge and Information Systems: JURIX 2004, the Seventeenth Annual Conference*, volume 120, page 79. IOS Press, 2004.
  - 19 Daniela Tiscornia. The LOIS project: Lexical ontologies for legal information sharing. In *Proceedings of the V Legislative XML Workshop*, pages 189–204. Citeseer, 2006.
  - 20 Saskia Van De Ven, Rinke Hoekstra, Joost Breuker, Lars Wortel, Abdallah El Ali, et al. Judging Amy: Automated Legal Assessment using OWL 2. In *OWLED*, volume 432, 2008.