Named-Entity Recognition for Portuguese Police Reports

Gonçalo Carnaz[1,2], Vitor Beires Nogueira[1,2], Mário Antunes[3,4], and N.M Fonseca Ferreira[5,6,7]

[1] Informatics Departament, University of Évora, Portugal
[2] LISP - Laboratory of Informatics, Systems and Parallelism, Portugal
[3] School of Technology and Management, Polytechnic Institute of Leiria, Portugal
[4] INESC-TEC, CRACS, University of Porto, Portugal
[5] Institute of Engineering of Coimbra, Polytechnic Institute of Coimbra, Portugal
[6] Knowledge Research Group on Intelligent Engineering and Computing for Advanced Innovation and Development (GECAD) of the Institute of Engineering, Polytechnic Institute of Porto, Portugal
[7] INESC TEC, Portugal

Abstract. During a criminal investigation several text documents are produced by police officers, creating a deluge of unstructured data obtained from heterogeneous sources. Therefore, identification and recognition of entities, i.e. places, organizations or persons, by a natural language pipeline, with named-entities recognition task, could help police officers to understand and find relevant information in data extracted. We aim to defined a natural language processing pipeline to identify and recognize entities from these police reports, supported by two trained corpus, namely Amazonia and a Portuguese News Corpus. Additionally, we evaluate named-entities recognition systems, focus in Portuguese language, with a dataset produced by the Portuguese police. We then evaluate the performance obtained on the information retrieval process applied to the dataset.

Keywords: natural language processing, named entity recognition, criminal investigation, police reports

1 Introduction

Criminal police must deal with a huge quantity of data acquired daily, or produced during investigations, from heterogeneous sources, like paper documents, digital reports, handwritten transcripts of interrogations, social media messages, transcripts or forensic logs. In every investigation, a final report is produced and includes analysis from different sources, made by texts and images. Therefore, we have unstructured data to be processed by natural language processing systems, through a procedure designated named-entity recognition (described in section 2).

The necessity of understanding and manipulation of texts and speech, produced by humans, determines Natural Language Processing (NLP) as computer science field applicable to our research. It is a huge challenge for this computer science field, sought by several computer scientists. Therefore, the NLP arises as the solution for that challenge, supported by other fields, e.g. linguistics, mathematics, artificial intelligence, robotics, psychology and others. By definition, Natural Language Processing (NLP) is defined as,

> *"...a field of computer science,artificial intelligence,and linguistics concerned with the interactions between computers and human (natural) languages. Many challenges in NLP involve natural language understanding, that is, enabling computers to derive meaning from human or natural language input, and others involve natural language generation."* [1] .

NLP is build under different tasks, such as sentence detection and tokenization [11], stemming [13], Part-of-Speech Tagging [1], named-entities recognition (NER) [20] [17], relation extraction and others.

The rest of the article is organized as follows: In section 2 we present an introduction to natural language processing and named entities recognition; in section 3 described related work about natural language processing (NLP) and named entities recognition related to crime domain. Additionally, we presented a review of related works for Portuguese language from different domains. In section 4 we defined our setup environment related to the selected frameworks and the performance measures obtained; in section 5 present our NLP with results obtained with two trained Corpus, which are Amazonia and Portuguese News. The paper ends with conclusion and future work in section 6.

2 Named-entity recognition

The Sixth Message Understanding Conference [8] (MUC-6), introduced the *Named-Entity Recognition* task, as an activity to extract terms related to different entities, i.e. persons, cities, date and time or other entities extracted from structured and unstructured documents. It was defined as a sub-task of information extraction [17]. In [17] authors defined NER as

> *"...is a sub problem of information extraction and involves processing structured and unstructured documents and identifying expressions that refer to peoples, places, organizations and companies. For us, humans,*

[8] http://www.itl.nist.gov/iaui/894.02/related_projects/muc/

> *NER (Named-Entity Recognition) is intuitively simple, because many named entities are proper names and most of them have initial capital letters and can easily be recognized by that way, but for machine, it is so hard. One might think the named entities can be classified easily using dictionaries, because most of named entities are proper nouns, but this is a wrong opinion. As time passes, new proper nouns are created continuously".*

Along years several approaches were made from the language factor, textual genre or domain factor to Entity type factor [17].

2.1 Information retrieval metrics

Information retrieval field defined metrics to measure entity recognition extraction systems performance. The metrics [14] established are: P : Precision, R : Recall and $F - Measure$.

Precision is defined by the ratio of correct answers (True Positives) among the total answers produced (Positives),

$$P(\ \textit{Precision}) = \frac{TP}{TP + FP}$$

where *TP* - *True Positive*, a predicted value was positive and the actual value was positive and *FP* - *False Positive*, predicted value was positive and the actual value was negative [12].

R - Recall is defined as a ratio of correct answers (True Positives) among the total possible correct answers (True Positives and False Negatives),

$$R(\ \textit{Recall}) = \frac{TP}{TP + FN}$$

where *FN* - *False Negative*, a predicted value was negative and the actual value was positive [12].

$F - Measure$ - is a harmonic mean of precision and recall,

$$\textit{F-Measure} = \frac{2 * \textit{precision} * \textit{recall}}{\textit{precision} + \textit{recall}}$$

3 Natural language processing system applied to crime domain - related work

In this section we describe named-entities recognition tasks and how these systems detect entities, i.e. places, persons, organizations or entities related to crime, from different languages. Additionally, we also describe systems developed for Portuguese language.

In 2010, [18] authors proposed an information extraction architecture to provide the input to a web-based system called WikiCrimes [10]. To analyze the extracted texts, they use a module called MorphoSyntactic Parser that performs a morphological and syntactic analysis creating a syntactic tree. In 2012, authors [24] proposed a system to extract Arabic named entities from crimes documents. The system used a standard preprocessing phrase, using a sentences splitting, tokenizer, Part-of-speech (POS) tagging (using a supervised statistical algorithm, trained with a corpus of crime related documents with 19800 words, in Arabic language) and a noun phrase chuncker. Follow by, a Named-Entity Identification and classification phrase with a Named-Entity Extraction, using a gazetteer (with a lexicon constituted by terms, i.e. Person, Personal properties, Location, Organizations and Indicative Words - crime terms), and a Pattern Rules module to train the NER to tag crime entities in documents.

In 2014, [8] proposed a NER system to extract entities from legal documents, e.g. judges, companies, courts or others. In [3] is proposed a system to extract crime information from online newspapers is proposed, fucused in the "hidden" information related to the theft crime.

In 2015 authors proposed crime information extraction from the Web, with crime NER task, using classification algorithms, e.g. Naive Bayes, Support Vector Machine and K-Nearest Neighbor. These classification algorithms are used to features extraction, through a voting combination module for features identification. Alongside, a indexing module that aims crime type identification, using the same classification algorithms [23] . In [25] authors proposed an approach based on raw text and extracts semi-structured information, in automatic way, using text mining techniques.

In 2016 authors proposed a system to extract verbs and their use, from crime clusters, using two data-sets, namely real data-sets from crime and industrial datasets with benchmarks. This system was defined with different tasks, removing not relating information, a stop words task with 571 words to delete from processed documents. Additional, the Porter stemmer for word stemming. For verbs identification, authors used a Word-Net identification method using the two datasets enumerated above [5].

Authors proposed in 2017 a named-entity recognition system for police documents for Dutch Police. The NER system is named *Frog*, using a traditional classification and evolution paradigm. With an annotated corpus, created from 250 criminal complaints reports, where domain experts identified entities, like location, person, organization, event, product and others [22]. In [2] methods are applied to discover criminal communities, analyzing their relations, and extract useful information from criminal text data.

There are several works related to natural language processing in native Portuguese language from different domains. Therefore, CaGE system [9] proposed a recognition and disambiguation system of geographic named entities mapping with geographic information, e.g. latitude and longitude coordinates. The main features are to identify and to disambiguate geographic entities, on a dictionary and a geographic dictionary. PorTexTO [7] proposed a system for named-entities processing, related to time. This system was created for HAREM [21] evaluation campaign. R3M [16] developed a NER system to identify and classify entities, e.g. people, organizations and locations. It is based on a semi-supervised learning approach rather them linguistic resources. Rembrant system [15] proposed a NER and relation detection between named entities in Portuguese texts, with source of knowledge the Wikipedia. SEI-Geo [6] is a NER system for identification and classification of named entities, e.g. Locations, based on geo-ontologies and patterns.

4 Experimental setup

In the following paragraphs, we will describe the setup procedure for frameworks evaluation, the dataset and the frameworks used. Finally, a discussion about obtained results, following the information retrieval metrics (explain in section 2.1).

4.1 Setup procedure

We have designed a setup procedure, see picture 1, to explain the steps taken to obtain the metrics (precision, recall and F-measure) from the frameworks analyzed. In step one, we use as input a dataset (see section 4.2) created from a police report, that is a final report with documents elaborated during investigations, e.g., forensic reports or smartphone logs. The original police report, in MS Word format, was parsed to plain text (raw text), used in the NLP pipeline.

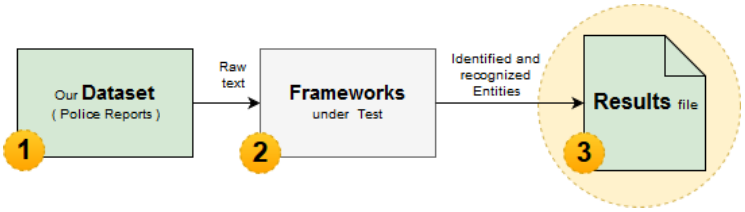

Fig. 1. *Test procedure workflow*

In step two, the text extracted will be processed by NER modules on each selected framework. Finally, the frameworks outputs are assessed. We will determine the *FN* (False Negatives), *FP* (False Positive) and *TP* (True Positives).

4.2 Dataset

We used a police report as a dataset, provided by a Portuguese police department and created during a drug crime investigation. the original file is in Microsoft Word format, with the following properties:

- Word count: 4657;
- Character count: 25152;
- Line count: 209;
- Paragraph count: 59;
- Other: tables and images.

The original file was processed, by a piece of code (supported by TIKA [9]), that parses the file into plain text (raw text), used as NLP pipeline input.

The focus of our experiment is to detect named entities, e.g. Person (PER), Organization (ORG), Locations (LOC) and Date (DAT) and how the selected frameworks process the dataset. Our dataset was annotated by domain specialist, using the following rules:

- Person (PER): identify persons names with a minimum of two words, i.e. politicians, scientists, artists or athletes;
- Organization (ORG): identified by full name or abbreviation, i.e. newspapers, banks, universities, schools, non-profits, companies or public services;
- Locations (LOC): identify by full address's or locals, i.e. countries, streets, cities or village's
- Date (DAT): identify by different formats, i.e. April 13 or 12/03/2013.

Entities extracted from our annotation procedure produce the following values: Person (PER) - 202; Organization (ORG) - 11; Locations (LOC) - 46 and Date (DAT) - 26.

4.3 Frameworks selected

In section 3, we described different approaches to NLP with NER tasks. The evaluation of these approaches could determined how our dataset will be processed by them, and what entities are identified and classified. We select approaches that are open source or trial versions for Portuguese or English language. Our focus is to identify named-entities in Portuguese, but in certain cases and because the framework was developed and trained for another language, we still evaluate them for discarding option. The selected frameworks are:

[9] https://tika.apache.org/

- KNIME (R) Analytics Platform [10]: is an open solution for data-driven innovation, that supports data mining and predicting. Among the predefined workflows in KNIME (R) Analytics Platform there is one for Natural Language Processing that is based on OpenNLP [11] and includes the process of Named-Entity Recognition;
- Linguakit [12]: created by the ProLNat@GE Group [13] (CITIUS, University of Santiago de Compostela) as a multilingual toolkit for NLP;
- RAPPort - A Portuguese Question-Answering System [19]: authors proposed a question answering system, supported by indices that store triples, related sentences and documents, using a NLP pipeline. This system is using CHAVE Corpus [14];

Each approach has somehow NLP toolkits that allow the development of the underlying tasks in an NLP pipeline, e.g. NLTK [15], Stanford CoreNLP [16], Pattern [17] and Polyglot [18].

4.4 Comparison results

To measure the performance of enumerated frameworks a set of metrics were used, i.e. Precision (P), Recall (R) and F-Measure (F1). The table 1 shows the performance measures:

Table 1. Frameworks performance metrics

	PER			**ORG**			**LOC**			**DAT**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Knime (NLP Workflow)	53%	18%	13%	30%	30%	10%	5%	5%	3%	-	-	-
RAPPort (DEI-UC)	51%	59%	55%	8%	50%	13%	48%	58%	53%	98%	87%	92%
Linguakit	67%	28%	39%	12%	82%	21%	50%	78%	60%	90%	90%	90%

Globally, the RAPPort approach reached the highest F-measure result for each entities (approximately 55%, 53% and 92%, respectively for Organization, Location and Date entities) for the detected entities, having the best trade-off

[10] https://www.knime.com

[11] opennlp.apache.org/

[12] https://github.com/citiususc/Linguakit

[13] http://gramatica.usc.es/pln/

[14] http://web.letras.up.pt/traducao/index_files/Page2219.htm

[15] www.nltk.org

[16] stanfordnlp.github.io/CoreNLP/

[17] www.clips.ua.ac.be/pattern

[18] polyglot.readthedocs.io/

regarding both measures (precision and recall). The best result obtained, related to F-measure for Organization entity was obtain by Linguakit, giving the best trade-off between precision and recall.

5 Our NLP Pipeline proposal

Our NLP pipeline proposal, is based on RAPPort [19], supported by three phases, that follows a standard NLP pipeline. In the first phase, we defined the data source and a parser module, the output of this phase is the processed original file (police report) into raw text, an important feature is data cleaning, e.g. removing formatting, images and tables. In the second phase, a pre-processing pipeline with a sentence boundary, tokenizer, stemming and POS Tagging tasks will prepare data for the next phase, the named-entities recognition module.

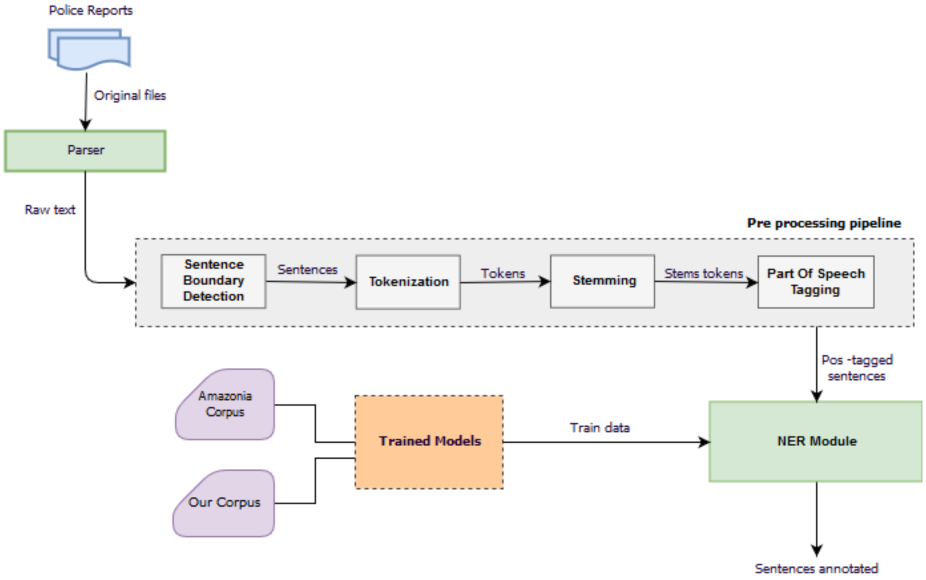

Fig. 2. *Proposed NLP Pipeline*

To complete these framework, we have a NER module supported by a trained corpus. We trained our model for named-entity recognition with two different corpus, e.g. Amazonia Corpus and Our News Crime Corpus. First, the Amazonia Corpus [19] has 4.6 millions of words (about thousand sentences) retrieved from Overmundo website, in Portuguese - Brazilian language, annotated by PALAVRAS [4]. The table 2 describes the trained corpus and a result of OpenNLP tool for training models:

[19] https://www.linguateca.pt/Floresta/corpus.html

Table 2. Amazonia Corpus data summary

Amazonia Corpus	
Sentences	81049
Tokens	1542622
Named-Entities	
Person	25237
Time	5490
Organization	20523
Place	15612

Regarding the second corpus, our motivation was to create a new corpus from portuguese online news about crime, e.g. Publico [20], Diário de Noticias [21] or Diário de Coimbra [22]. According to the domain experts the syntax and semantics of these news are similar to police language presented in the police reports. After that, we trained the corpus with OpenNLP [23] training tool, with the results described in table 3:

Table 3. Our News Corpus data summary

PT News Crime Corpus	
Sentences	310
Tokens	12078
Named-Entities	
Person	40
Time	57
Organization	82
Place	45

5.1 Results obtained

Following the setup procedure, we evaluated our dataset with our proposal for each corpus, obtaining the results described in table 4:

[20] https://www.publico.pt/

[21] https://www.dn.pt/

[22] http://www.diariocoimbra.pt/

[23] https://opennlp.apache.org/

Table 4. Evaluation results

	PER			ORG			LOC			DAT		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Amazonia Corpus	66%	79%	72%	6%	67%	11%	27%	42%	33%	33%	92%	48%
Our Corpus	-	-	-	68%	13%	22%	97%	29%	44%	73%	56%	64%

Globally, with Our Corpus we reached the highest F-measure result for each entities (approximately 22%, 44% and 64%, respectively for Organization, Location and Date entities) for the detected entities, having the best trade-off regarding both measures (precision and recall). There is an entity that was not detected using Our Corpus, the Person entity, the reason for this failure detection is because our corpus does not have sufficient data for a fine train.

6 Conclusion and future work

The work developed was focused on the evaluation of open source frameworks for named-entity recognition retrieved from unstructured data, and a framework proposal for named-entity recognition with two trained corpus. In both cases, performance measures were performed, but other conclusions and investigation paths will be considered. We have obtained promising results with the frameworks analyzed, in almost all entities.. But our focus is the crime domain, therefore the obtained results were weak, no entity related to crime domain was identified. There are approaches, described in section 3, that identify and recognize entities related to crime for English or other languages. Our NLP framework proposal for named-entity recognition retrieved from Portuguese police reports, tries to increase named-entities detection adding two trained corpus, supporting police reports parsing to a common format. The preliminary results encourage the approach taken, but with improvements to be realized, e.g. a better trained corpus or identify and recognize entities related to crime.

Future work will consist of the framework improvement, clarifying the possibility of extracting named entities and relations from police reports, recognizing the entities related to crime, e.g. crime or narcotics. Additionally, we plan to increase our corpus quality to improve performance measures of our framework proposal.

References

1. N. Adhvaryu and P. Balani. Survey : Part-Of-Speech Tagging in NLP. *International Journal of Research in Advent Technology*, 1(1):102–107, 2015.

2. R. Al-Zaidy, B. C. M. Fung, and A. M. Youssef. Towards discovering criminal communities from textual data. In *Proceedings of the 2011 ACM Symposium on Applied Computing*, SAC '11, pages 172–177, New York, NY, USA, 2011. ACM.
3. R. Arulanandam, B. T. R. Savarimuthu, and M. A. Purvis. Extracting crime information from online newspaper articles. In *Proceedings of the Second Australasian Web Conference - Volume 155*, AWC '14, pages 31–38, Darlinghurst, Australia, Australia, 2014. Australian Computer Society, Inc.
4. E. Bick. *THE PARSING SYSTEM "PALAVRAS" Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework "*. PhD thesis, University of Århus, 2000.
5. Q. Bsoul, J. Salim, and L. Q. Zakaria. Effect Verb Extraction on Crime Traditional Cluster. *World Appl. Sci. J.*, 34(9):1183–1189, 2016.
6. M. S. Chaves. Geo-ontologias e padrões para reconhecimento de locais e de suas relações em textos: o SEI-Geo no Segundo HAREM. In *Desafios na avaliação conjunta do reconhecimento entidades mencionadas O Segundo HAREM*, pages 231–245. 2008.
7. O. Craveiro. PorTexTO: sistema de anotação/extracção de expressões temporais. In *Desafios na avaliação conjunta do reconhecimento entidades mencionadas O Segundo HAREM*, pages 159–170. 2008.
8. C. Dozier, R. Kondadadi, M. Light, A. Vachher, S. Veeramachaneni, and R. Wudali. Semantic processing of legal texts. In E. Francesconi, S. Montemagni, W. Peters, and D. Tiscornia, editors, *Semantic Processing of Legal Texts*, chapter Named Entity Recognition and Resolution in Legal Text, pages 27–43. Springer-Verlag, Berlin, Heidelberg, 2010.
9. B. Emanuel. *Geographically Aware Web Text Mining.* PhD thesis, Universidade de Lisboa, 2008.
10. V. Furtado, L. Ayres, M. D. Oliveira, E. Vasconcelos, C. Caminha, J. D. Orleans, and M. Belchior. Collective intelligence in law enforcement – The WikiCrimes system. *Inf. Sci. (Ny).*, 180(1):4–17, 2010.
11. V. Gupta, L. C. Science, and G. S. Lehal. A Survey of Text Mining Techniques and Applications. *J. Emerg. Technol. Web Intell.*, 1(1):60–76, 2009.
12. I. M. Konkol. *Named Entity Recognition.* PhD thesis, University of West Bohemia, 2015.
13. B. Lovins. Development of a Stemming Algorithm. *Mech. Transl. Comput. Linguist.*, 11(June):22–31, 1968.
14. A. Mansouri, L. S. Affendey, and A. Mamat. Named Entity Recognition Approaches. *Journal of Computer Science*, 8(2):339–344, 2008.
15. E. Mencionadas. REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em Relações e ANálise Detalhada do Texto. In *Desafios na avaliação conjunta do reconhecimento entidades mencionadas O Segundo HAREM*, pages 195–211. 2008.
16. C. Mota. R3M, uma participação minimalista no Segundo HAREM. In *Desafios na avaliação conjunta do reconhecimento entidades mencionadas O Segundo HAREM*, pages 181–193. 2008.
17. D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

18. V. Pinheiro, V. Furtado, T. Pequeno, and D. Nogueira. Natural language processing based on semantic inferentialism for extracting crime information from text. In *2010 IEEE International Conference on Intelligence and Security Informatics*, pages 19–24, May 2010.
19. R. Rodrigues and P. Gomes. Rapport- a portuguese question-answering system. In F. Pereira, P. Machado, E. Costa, and A. Cardoso, editors, *Progress in Artificial Intelligence*, pages 771–782, Cham, 2015. Springer International Publishing.
20. C. J. Saju and A. S. Shaja. A survey on efficient extraction of named entities from new domains using big data analytics. In *2017 Second International Conference on Recent Trends and Challenges in Computational Models (ICRTCCM)*, pages 170–175, Feb 2017.
21. D. Santos, N. Seco, N. Cardoso, and R. Vilela. HAREM : An Advanced NER Evaluation Contest for Portuguese.
22. M. Schraagen. Evaluation of Named Entity Recognition in Dutch online criminal complaints. *Comput. Linguist. Netherlands J.*, 7:3–15, 2017.
23. H. A. Shabat and N. Omar. Named Entity Recognition in Crime News Documents Using Classifiers Combination. *Middle-East J. Sci. Res.*, 23(6):1215–1221, 2015.
24. A. I. Technology, M. Asharef, N. Omar, and M. Albared. Arabic Named Entity Recognition in Crime. *J. Theor. Appl. Inf. Technol.*, 44(1):1–6, 2012.
25. Y. Yang, M. Manoharan, and K. S. Barber. Modelling and analysis of identity threat behaviors through text mining of identity theft stories. In *2014 IEEE Joint Intelligence and Security Informatics Conference*, pages 184–191, Sept 2014.