Project: Named-Entity Recognition and Classification in Navy Documents

Gonçalo Azevedo Rodrigo

July 22, 2020

Abstract

An organization's information consists mostly of unstructured information. To transform it into useful information, techniques and tools for Information Extraction (IE) were developed. One of IE's tasks is Named-Entity Recognition and Classification (NERC). The named-entity concept was initially proposed by the MUC-6 Conference in 1996. Since then, multiple techniques have been developed to extract entities from different types of texts and for several languages. Even so, in the community of researchers, the interest to develop new approaches to identify and classify Named-Entities remains, since this operation allows to extract knowledge from the text. In this project, we carry out the treatment of Portuguese Navy documents, to produce a *Corpus*. Using *Corpus*, we also tested the NERC task of our Natural Language Processing chain.

Keywords: Natural Language Processing (NLP) Named-Entity Recognition and Classification (NERC)) .

1 Introduction

An organization's data can be divided into two categories: structured and unstructured. The unstructured data include: files, documentation, emails, project plans, product manuals, *WEB* pages, etc., and are created in different supports and formats. In 1998, Merrill Lynch introduced a rule that said 80% -90% of all an organization's data is unstructured [19]. The same percentage prevails until today[11]. Also, in 2017, IDC predicted that there will be 10 times more data in 2025[14].

The information *explosion* demanded the search for more efficient methods for processing unstructured documents. The branch of computer science that is concerned with solving this problem, as well as the interpretation and automatic generation of human language is Natural Language Processing (NLP).

Information extraction is a task of the NLP, which, in turn, has a sub-task called Named-Entity Recognition and Classification (NERC). In 1996, at the MUC-6 conference[7], the concept of *Named-Entity (NE)* was defined. This concept emerged after the recognition of the *information unit* as a fundamental element for the task of extracting information. NE is a linguistic expression used to designate real-world objects (people, places, organizations, etc.), usually corresponding to proper names. In addition to proper names, other types of expressions are used in the NERC, namely temporal expressions (dates, periods, ephemeris, etc.) and numeric expressions (number of taxpayers, car registrations, etc.). Also, the type of Named-Entity depends on the domain of interest. For example, in the Military domain, some relevant entities are the patent, the units, the mission, among others; on the other hand, in the general domain, the relevant entities are the person, the location, the organization, the numerical values, the temporal expressions, among others.

The task of correctly identifying and classifying these types of expressions is essential for a semantic analysis of the text and to facilitate subsequent syntactic processing. NERC also assists in some NLP tasks, such as Automatic Text Summarization[12], Machine Translation[1], Information Retrieval[9], Question-Answer System[13] and Speech Recognition.

The main contributions to NERC come from the techniques and tools developed for various scientific events, such as *Information Retrieval and Extraction*[18] (IREX), *Conference on Natural Language Learning 2002* [20]

(CONLL 2002) and *2003*[21] (CONLL 2003), *Automatic Content Extraction*[5] (ACE) Program, and, for Portuguese, the *Avaliação de Reconhecimento de Entidades Mencionadas*[17] (HAREM).

In this project, the NLP chain used is STRING [8]. STRING performs basic tasks of text processing in Portuguese, such as text segmentation and atomization, Part-of-Speech tagging , morphosyntactic disambiguation, syntactic analysis of the text in chunks and extraction of dependencies.

1.1 Problem

STRING was initially developed only to process text of a general nature (*e.g.* journalistic text). The *Corpus* used in this project is the Nacy correspondence, a particular textual domain, therefore, there are compound terms, Named-Entities and events that STRING did not identify and incorrectly classified.

Although NERC is generally considered a task with the objective achieved due to its high-performance rates at scientific conferences. In fact, these assessments use a limited set of types of NE, which rarely change over the years. In addition, they use *Corpus* of reduced dimensions, essentially when compared to other areas of Information Extraction. These factors lead to an overfitting of the tools and, consequently, to a limitation of the evolution in the area[10]. In other words, the NERC may be a challenge solved for the general textual domain, however, there is a deficit in the NERC for the domain of specific interest.

1.2 Objectives

The main objective of this project is to adapt a NLP system, in particular, its NERC module initially developed to process texts of a general nature (e.g. journalistic text), for a particular textual domain, the official correspondence of the Portuguese Navy.

With the processing of *Corpus* in the STRING, we intend to increase the number of named-entities and events identified and classified, for this, we will perform the following tasks:

- Constitution of an annotated Corpus of a specific domain;
- Identification of new compound terms;
- Identification and Classification of new named-entities;
- Identification of new events related to the Portuguese Armed Forces.

1.3 Contributions

Taking as a starting point unstructured information, we will deal with it, so that the NLP chain can process these documents. Then, rules will be added to STRING to recognize the new named-entities and classify them according to their type in the document. This project also aims to contribute to the progress in the NERC task. For this, we will use a particular textual domain, in contrast to the general textual domain used in NERC conferences and forums.

This project is the first step of two to carry out the automatic distribution of Navy documents. The second step is the classification of documents using machine learning techniques.

2 Architecture

This Section presents the general architecture of the NLP system, in which the task of recognizing Named-Entities is included. STRING is a NLP chain for Portuguese based on rules and statistical methods. This tool was developed by INESC-ID's Spoken Language Laboratory (L^2F) in Lisbon[8]. The processing of the chain is divided into three stages:

- Pre-processing;
- Disambiguation (rule-based and statistical);
- Syntactic analysis.

2.1 Pre-processing

Pre-processing corresponds to the Morphosyntactic Analyzer. This stage is mainly responsible for dividing the entry into segments, also known as *tokens*. Because the individual segments are called *tokens*, this step can also be called a *tokenizer*.

Also, this stage is also responsible for identifying punctuation marks and symbols, as well as alphanumeric expressions, some of which correspond to certain types of Entities mentioned.

2.2 Disambiguation

The next step in the processing chain is the disambiguation process, which comprises two steps:

- Rule-based disambiguation, performed by RuDriCo[3, 4];
- Statistical disambiguation carried out by MARv[16].

2.2.1 Rule-based Disambiguation

The main objective of *RuDriCo*, according to Diniz[3], is to provide an adjustment in the results produced by the morphosyntactic analyzer for the specific needs of each parser. To achieve this goal, *RuDriCo* modifies the segmentation previously performed by *LexMan*. For example, you can join two tokens into one, expressions like *ex-* and *namorada* in *ex-namorada*; or on the contrary, expand an expression, *à* contraction to two segments, *a*. Changing the segmentation is also useful for number and date recognition tasks.

2.2.2 Statistical Disambiguation

The main objective of *MARv*[15] is to analyze the Part-of-Speech (POS) tagging assigned to each token in the previous step of the processing chain, and then choose the most likely annotation for each one. To achieve this objective, we used a statistical model known as the Hidden Markov Model (HMM).

2.3 Syntax Analisys

The third and final step in the processing chain is the parsing performed by the XIP. In this stage, the identification and classification of NEs takes place, so the main work of this project occurs in this module. XIP is a rule compiler that adds linguistic information (syntactic and semantic) to the return of the Part-of-Speech (POS) tagging. This tool accesses the circulating context, as well as allowing to represent and manipulate several linguistic characteristics. The system is independent of the language, and new rules can be created incrementally over existing ones. XIP also uses parsing functionality to divide the text into chunks such as the nominal chunk (NC) and verbal chunk (VC). Then, the syntactic relations between the heads of the chunks are extracted. These relations represent the main functionality of parsing between syntactic dependencies (E.g. Subject-Direct Complement, etc.), but also include auxiliary dependencies between different chunks and words, for example the connection between verbal segments and auxiliary words[2].

3 Corpus

This section describes the structure of Navy documents. The document handling process is also described, from the format provided (*.pdf*) to a format that STRING can process.

3.1 Documents

The *Corpus* was extracted from documents received and sent from the four units of the Information Technology Superintendence (ITS). Documents are organized by year, from 2015 to 2019; within the year it is divided into three groups: Entries, Interns and Exits; finally, within the origin of the documents are the TIS units.

The *Corpus* has a total of 7,302 documents. As for their content, the documents are very varied, with 64.89% of the documents having the same structure, called standardized. The remaining documents are non-standard and may be invoices, receipts, faxes, memos, diplomas or e-mails from the Office of the Chief of the Armed Forces.

This *Corpus* will be used for machine learning in two projects with different classifiers. One project will use the process number, present only in the standardized documents, however, the other project will use the distribution table, present in 90.85% of the documents. In short, the *Corpus* is made up of different types of documents in which all have a distribution table, however, about 65% of the documents have a document number.

3.1.1 Structure of Documents

Standard documents have a predefined structure and their structure is organized into six segments: *header*, *document number*, *subject*, *reference*, *recipient*, *body* and *signature*. There is an optional element, the *attachment*.

3.2 Processing of Documents

The processing of documents is divided into four phases: conversion, filtering, removal and segmentation. This treatment chain aims to correct conversion errors, clear unnecessary information and segment the information by sections. In this way, the data present in the Corpus improved its quality, getting closer to the information present in the documents provided. Thus, the

final format of the documents is the text format, this way STRING can process them. This treatment was carried out through programs developed in Python 3.7.1 with the library of regular expressions (*Lib/re.py*). Figure 1 illustrates the various phases of the documents, as well as the corresponding tasks.

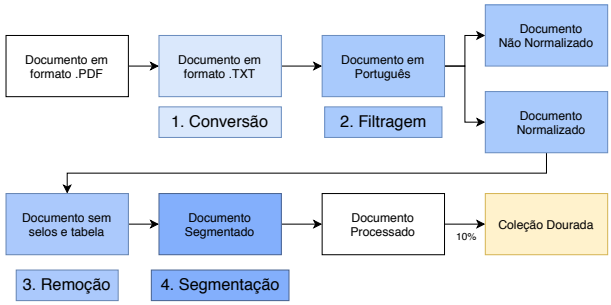

Figure 1: Evolution of documents throughout the processing (In Portuguese).

In the Conversion phase, the documents provided by the Portuguese Navy were initially in the Portable Document Format (*pdf*). As the NLP tool used only supports text in text format (*txt*), one of the steps in the *Corpus* treatment was to convert these documents from *pdf* to *txt*.

The types of documents are varied, therefore, it was also necessary to separate standardized and non-standardized documents to carry out a specific treatment for each case, in the Filtration phase.

The standardized documents have unnecessary information for the NERC, therefore, in the Removal phase, they were removed. The following elements were identified as unnecessary: Entry and Exit Stamps, Recipients' Stamp, Distribution Table and Handwriting.

Finally, segmentation consists of dividing data by segments using tags, increasing knowledge of the document's internal structure. This division will allow you to weigh the data of the various segments differently. For example, in the machine learning task, the title text can be considered more relevant than the body text.

3.3 Golden Collection

The Golden Collection is a collection of annotated documents that allows you to evaluate the performance of the NLP chain. This annotation is performed manually, for this purpose 210 documents were randomly chosen, about 3% of the *Corpus*.

Due to the variation in the distribution of documents in the different units over time, as well as the reduction in the number of stamps, we conclude that there is a significant evolution of documents over time. To have a more representative collection of reality, we valued the most recent documents. Therefore, we have distributed the documents equally over the years.

In the Collection, there are 2,304 different words, of which 1,977 (86%) belong to the Portuguese language dictionary and the rest are numbers, marks, abbreviations or words in English. 128 misspelt words were also corrected and 70 words were removed as a result of OCR errors.

4 Procedures

Before writing down the documents of the Golden Collection, these documents had a selection and cleaning phase. The selection phase consisted of choosing, at random, the documents according to the distribution described in the Section 3.3. Then, they were submitted to a cleaning phase, the incorrect words were corrected and the OCR errors were erased. Even at this stage, when the end of a sentence did not have a punctuation character and the beginning of the next was a lowercase letter or a number, these two sentences were concatenated.

The Golden Collection annotation task was performed according to the following methodology. First, 10 documents were recorded, chosen at random. Then, these documents were submitted to validation by a linguist with knowledge both in the textual domain and in the classification guidelines used by STRING. Finally, the remaining 200 documents were noted based on the linguist's feedback.

This methodology was chosen to minimize the problem of lack of resources, both for people and time. The methodology we consider to be the most correct, but also the most expensive, is as follows: First, the Golden Collection would be noted by 3 people who understood the textual domain and the classification guidelines; Then, each named-entity would be discussed among the 3 people, the majority classification prevailing. In this way, the annotation would be less biased, since the decision of each one would be discussed together.

5 Evaluation and Results

This Section describes the procedures used in the evaluation of the Named-Entities recognition and classification system (Subsection 5.1) and the results (Subsection 5.2).

5.1 Evaluation

To assess the task of Named-Entities Recognition and Classifion, we were used an adaptation of the methodology used by HAREM [6]. This forum allows you to evaluate the correctness of the results through the use of a golden collection, that is, a reference document, usually annotated by hand, which presents the ideal output intended for the task to be evaluated.

The labeling of the original text, according to STRING's labeling rules, must contain each NE labeled by an opening and closing tag, similar to the tags used in XML. The opening label has the category and type assigned, optionally, it can also have the subtype.

5.1.1 Measures

This subsection presents the measures used in the task of Named-Entities Recognition and Classification. Concerning the identification task, it aims to measure the efficiency of the system and define the entities correctly, in comparison with the previously named-entities existing in the golden collection.

The evaluation of the semantic classification aims to measure the capacity of the system to be able to classify a Named-Entity taking into account the hierarchy of categories and types defined by STRING.

The measures used in this project are: precision, measures the quality of the system's response which measures the proportion of correct responses to all the responses given by the system; the recall, measures the percentage of correct responses that the system was able to identify; The F-measure combines the measures of precision and recall for each task; over-generation, measures the excess of spurious results that a system produces, that is, how many times it produces wrong results; subgeneration is a measure of how much the system has lacked, given the known solution, e.g. the golden collection.

5.2 Results

In this section of the document, we present the results obtained by the Natural Language processing chain during the evaluation of the Named-Entities Recognition and Classification task.

The evaluation was carried out by comparing the annotated documents of the Golden Collection and the output of the STRING NERC task. The general results are shown in the Table 1.

As previously mentioned, precision measures the relevance of the result. The accuracy of the Identified Entities task was 41.39%, so we concluded that the relevance of the result is low.

The scope measures the number of relevant results that have been returned. This value was higher than the precision, however, the value remains low, 55.41%.

To measure the balance between accuracy and comprehensiveness, we use F-Measure. This measure is calculated only on the basis of precision and scope. Thus, the F-measure describes the reliability of the result, which in this case is low.

The precision for the Semantic Classification by Category and Combined is high, however, when we take into account the Plain Classification, the precision is low. As in the Identification task, the recall of the results is low. the over-generation and the subgeneration have too high values. This is due to problems with STRING's RCEM task.

6 Conclusion and Future Work

During the processing of the *Corpus*, we noticed that the main difficulty comes from the bad reading of the OCR. Future work could be the development of an OCR tool that can identify unnecessary information, such as images and stamps, and the different sections of the document.

Another future work may be to vary the specific domain of *Corpus*. In this case, the domain is the correspondence of the Navy, however, there are similar organizations, such as a branch of the Armed Forces (e.g. Air Force) or a department of the Government of Portugal (e.g. Ministry of Defense), which may be of interest to apply the strategy used by this project.

Starting from the *Corpus* processed by this project, it would be interesting to develop two classification solu-

	Precision	Recall	F-Measure	Over-generation	Subgeneration
Identification	41,39%	55,41%	47,39%	25,30%	26,80%
Categories Classif.	86,42%	55,41%	67,53%	52,81%	55,94%
Combined Classif.	83,72%	53,68%	65,42%	-	-
Plain Classif.	67,96%	43,58%	53,10%	33,87%	26,80%

Table 1: General results for identification and classification.

tions, one using the process number and the other using the distribution table. The case number has information about the type of document, on the other hand, the distribution table references the recipients of the document. However, the two classifiers complement each other to complete the task of automatically distributing documents between departments. This task aims to optimize the flow of information and reduce human error and effort.

One of the objectives of this project was to adapt the NLP chain to a specific textual domain. This objective was not achieved because there was a need to improve the *Corpus* for the machine learning task. For, the two projects that proceed with this project were developed at the same time as this one. Although the Recognition and Classification of Named-Entities have not been improved, the developed *Corpus* allowed us to evaluate STRING for this specific domain. Thus, for future work to improve STRING, there is already a prepared *Corpus* and the identification of the main problems in the NLP chain.

Some improvements to this project would be to add to the STRING's vocabulary the unknown words, as well as the abbreviations. Then, based on the assessment carried out, identify the NEs that STRING was unable to recognize or partially recognized as correct. Finally, add the necessary rules for STRING to recognize and correctly classify the Named-Entities. After applying these improvements, it would be interesting to re-evaluate STRING's NERC task and compare it with this project. Thus, we conclude the improvements made.

Corpus processing became an arduous and challenging task, as the documents had a lot of unnecessary information that had to be removed; the documents were very diverse, both in terms of the type of document and the language used; the fact that this *Corpus* is used for two different classifications made different treatment necessary; finally, and very significant, the fact that OCR is not specialized for documents in Portuguese made it difficult to recognize words correctly, especially words with accents.

This project is the first part of two to carry out the task of machine learning in Navy documents, so the documents were subjected to a rigorous treatment to ensure that the *Corpus* data was valid and authentic. In this way, the next two projects will have the raw material necessary for success in the machine learning task.

References

[1] Babych, B., and Hartley, T. Improving machine translation quality with automatic named entity recognition. In *Proceedings of the 7th International EAMT workshop on MT and other language technology tools, Improving MT through other language technology tools, Resource and tools for building MT at EACL 2003* (01 2003), pp. 1–8.

[2] Baptista, J., Mamede, N., and Gomes, F. Auxiliary verbs and verbal chains in European Portuguese. In *Computational Processing of the Portuguese Language* (Berlin, Heidelberg, 2010), Springer Berlin Heidelberg, pp. 110–119.

[3] Diniz, C. RuDriCo2 - Um Conversor Baseado em Regras de Transformação Declarativas. Master's thesis, Instituto Superior Técnico, Lisboa, October 2010.

[4] Diniz, C., Mamede, N., and Pereira, J. RuDriCo2 - A Faster Disambiguator and Segmentation Modifier. In *INFORUM II* (September 2010), pp. 573–584.

[5] Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., and 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'04)* (Lisbon, Portugal, May 2004), European Language Resources Association (ELRA).

[6] Equipa da Linguateca. HAREM: Reconhecimento de entidades mencionadas em português. https://www.linguateca.pt/HAREM/, 2019. Último acesso 2019-10-24.

[7] Grishman, R., and Sundheim, B. Message Understanding Conference-6: A brief history. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 1* (Stroudsburg, PA, USA, 1996), COLING '96, Association for Computational Linguistics, pp. 466–471.

[8] Mamede, N., Baptista, J., Diniz, C., and Cabarrão, V. STRING: An Hybrid Statistical and Rule-Based Natural Language Processing Chain for Portuguese. In *PROPOR 2012* (April 2012), vol. Demo Session.

[9] Mandl, T., and Womser-Hacker, C. The effect of named entities on effectiveness in cross-language information retrieval evaluation. *Proceedings of the ACM Symposium on Applied Computing 2* (01 2005), 1059–1064.

[10] Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., and Gómez-Berbís, J. M. Named entity recognition: Fallacies, challenges and opportunities. *Computer Standards & Interfaces 35*, 5 (2013), 482 – 489.

[11] NetOwl. When 80% of the world's data is unstructured, entity extraction is a must. https://www.netowl.com/2017/08/11/80-worlds-data-unstructured-entity-extraction-must, 2017. Último acesso 2019/03/26.

[12] Nobata, C., Sekine, S., Isahara, H., and Grishman, R. Summarization system integrated with named entity tagging and ie pattern discovery. 1742–1745.

[13] Pizzato, L. A., Molla, D., and Paris, C. Pseudo relevance feedback using named entities for question answering. In *Proceedings of the Australasian Language Technology Workshop 2006* (Sydney, Australia, Nov. 2006), pp. 83–90.

[14] Reinsel, D., Gantz, J., and Rydning, J. The Digitization of the World - From Edge to Core. Tech. rep., Seagate, United States, November 2008.

[15] Ribeiro, R. Anotação morfossintática desambiguada do português. Master's thesis, Instituto Superior Técnico, Lisboa., March 2003.

[16] Ribeiro, R., Oliveira, L., and Trancoso, I. Using morphossyntactic information in TTS Systems: Comparing strategies for European Portuguese. In *PROPOR 2003 - Computational Processing of the Portuguese Language: 6th International Workshop* (01 2003), pp. 143–150.

[17] Santos, D., Seco, N., Cardoso, N., and Vilela, R. HAREM: An Advanced NER Evaluation Contest for Portuguese. 1986–1991.

[18] Sekine, S., and Isahara, H. IREX: IR and IE Evaluation project in Japanese. In *Proceedings of International Conference on Language Resources Evaluation (LREC 2000)* (04 2000).

[19] Shilakes, C., and Tylman, J. Enterprise Information Portals. Tech. rep., Merrill Lynch, United States, November 1998.

[20] Tjong Kim Sang, E. F. Introduction to the CoNLL-2002 Shared Task: Language-independent

Named Entity Recognition. In *Proceedings of the 6th Conference on Natural Language Learning - Volume 20* (Stroudsburg, PA, USA, 2002), COLING-02, Association for Computational Linguistics, pp. 1–4.

[21] Tjong Kim Sang, E. F., and De Meulder, F. Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition. In *Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4* (Stroudsburg, PA, USA, 2003), CONLL '03, Association for Computational Linguistics, pp. 142–147.