FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Entity and relation extraction from web content

André Ricardo Oliveira Pires

DISSERTATION PLANNING

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Sérgio Sobral Nunes

Co-Supervisor: José Luís da Silva Devezas

February 1, 2017

Entity and relation extraction from web content

André Ricardo Oliveira Pires

Mestrado Integrado em Engenharia Informática e Computação

February 1, 2017

Abstract

The Named Entity Recognition (NER) task focuses on extracting relevant information from free text, such as news, which usually has a particular phrasing structure. Identified entities are then annotated with previously defined categories, being the most common defined in the MUC-6, namely Person, Location, Organization, Date, Time, Money and Percent. Entity detection supports more complex tasks, such as Relation Extraction or Entity-Oriented Search.

The NER task has been largely studied in recent years. There have been multiple solutions presented either for specific or generic domains and also for single or multiple languages. The first machine learning solutions were developed using supervised learning, such as Hidden Markov Models or Conditional Random Fields. However, in recent years, the focus has been on semi-supervised learning, in particular using bootstrapping, which involves a set of seeds to start the learning process, and unsupervised learning, where clustering is the typical approach. This shift in focus was due to the amount of work effort and the time consumed to obtain a significant training set. There have been some NER tools focused on the Portuguese language, such as Palavras or NERP-CRF, but their f-measure was still below the f-measure obtained by the available tools, for instance based on an annotated English corpus, trained with a Stanford Named Entity Recognizer or with OpenNLP.

ANT is an entity-oriented search engine for the University of Porto. This search system is limited to the information available in SIGARRA, the information system of the University of Porto. Currently it uses hand-crafted selectors based on XPath or CSS, which are dependant on the structure of the page. A machine learning method would allow the automation of the extraction tasks, making it scalable, structure independent and able to diminish the required work effort and consumed time.

This dissertation aims to evaluate existing NER tools in order to decide the best approach to use regarding the Portuguese language, particularly in the domain of SIGARRA news. Expanding the existing knowledge base will help index SIGARRA pages by providing a richer entity-oriented search experience with new information, as well as a better ranking scheme based on the additional context made available to the search engine. The scientific community will then have better tools to do research in this domain, particularly for the Portuguese language. Later, the developed work will be integrated with the ANT platform, developed at InfoLab, within the Faculty of Engineering of the University of Porto.

Resumo

A tarefa de Reconhecimento de Entidades Mencionadas (REM) foca-se na extração de informação relevante do texto livre, como notícias, que geralmente têm uma estrutura de frases particular. As entidades identificadas são então anotadas com categorias previamente definidas, sendo as mais comuns definidas na MUC-6, nomeadamente, Pessoa, Local, Organização, Data, Hora, Dinheiro e Percentagem. A deteção de entidades suporta tarefas mais complexas, como Extração de Relações ou Pesquisa Orientada a Entidades.

A tarefa de REM tem sido amplamente estudada nos últimos anos. Houve múltiplas soluções apresentadas para domínios específicos ou genéricos e também para idiomas únicos ou múltiplos. As primeiras soluções de aprendizagem computacional foram desenvolvidas usando aprendizagem supervisionada, como Modelos Ocultos de Markov ou Campos Aleatórios Condicionais. No entanto, nos últimos anos, o foco tem sido na aprendizagem semi-supervisionada, em particular com recurso a *bootstrapping*, que envolve um conjunto de sementes para iniciar o processo de aprendizagem, e aprendizagem não supervisionada, onde *clustering* é a abordagem típica. Esta mudança de foco deveu-se ao esforço e tempo significativo envolvido na obtenção de um conjunto de treino significativo. Houve algumas ferramentas de REM focadas na língua portuguesa, tais como o Palavras ou o NERP-CRF, mas o seu f-measure ainda estava abaixo do obtido usando as ferramentas disponíveis, por exemplo com base num *corpus* inglês anotado, treinado com o *Stanford Named Entity Recognizer* ou com o *OpenNLP*.

O ANT é um motor de busca orientado a entidades da Universidade do Porto. Este sistema de pesquisa está limitado às informações disponíveis no SIGARRA. Atualmente usa seletores construídos manualmente baseados em XPath ou CSS, que são dependentes da estrutura da página. Um método baseado em aprendizagem computacional permitiria a automatização das tarefas de extração, tornando-a escalável, independente da estrutura e capaz de diminuir o esforço de trabalho exigido e o tempo consumido.

Esta dissertação tem como objetivo avaliar as ferramentas de REM existentes para decidir a melhor abordagem a utilizar em relação à língua portuguesa, particularmente no domínio das notícias do SIGARRA. A expansão da base de conhecimento existente ajudará a indexar as páginas do SIGARRA proporcionando uma experiência de pesquisa orientada a entidades mais rica e com nova informação, bem como um melhor esquema de classificação baseado no contexto adicional disponibilizado ao motor de busca. A comunidade científica terá então melhores ferramentas para fazer pesquisas neste domínio, particularmente para a língua portuguesa. Posteriormente, o trabalho desenvolvido será integrado na plataforma ANT, desenvolvida no InfoLab, na Faculdade de Engenharia da Universidade do Porto.

Contents

CONTENTS

List of Figures

LIST OF FIGURES

List of Tables

LIST OF TABLES

Abbreviations

NE	Named Entity
NER	Named Entity Recognition
NLP	Natural Language Processing
IE	Information Extraction
KB	Knowledge Base
POS	Part Of Speech
CRF	Conditional Random Fields
HMM	Hidden Markov Model
MEMM	Maximum Entropy Markov Model
ME	Maximum Entropy
SVM	Support Vector Machine
OBIE	Ontology-Based Information Extraction
MUC	Message Understanding Conference
CoNLL	Conference on Natural Language Learning
ACE	Automatic Content Extraction
HAREM	HAREM Avaliação de sistemas de Reconhecimento de Entidades Mencionadas

Chapter 1

Introduction

This chapter contextualises Named Entity Recognition (NER) regarding scientific work, commercial applications and the applications in the ANT project. The main motivation for this work consists in expanding beyond structure dependent information extraction in the web. We close the chapter with the structure for the remaining document.

1.1 Context

With the vastness of information made available in the Web, there is a need for a method of filtering the relevant data and presenting it to the readers. Most of the Internet's information is not available in structured form. Natural Language Processing (NLP) is a field of Artificial Intelligence concerned with making human language understandable to computers. This enables structuring information in a way that it can be indexed and used by a machine for question answering, for example.

Information Extraction (IE) and NLP are intertwined, and IE's main task is to extract relevant data from documents. Documents can be either unstructured or structured texts. One of the main sub-tasks of information extraction is Named Entity Recognition. The concept of "named entity" was first introduced in 1996, at the Message Understanding Conference - 6 (MUC-6), by Grishman and Sundheim [GS96]. At that time, the concept of named entity referred to names of people, locations, organisations and number expressions such as money and dates. Over the years, there have been multiple redefinitions of named entities, mainly because there was a need to include other entities for specific purposes, for example DNA sequences for biomedical purposes.

NER provides a means for further, more complex tasks in information extraction. One of those tasks is Relationship Extraction, whose main objective is to identify semantic links between the entities identified in a sentence.

The launch of Google's Knowledge Graph, in May 2012, or even Facebook's Open Graph, boosted the focus on entity-oriented search. The ANT[1] project represents a similar effort. This project is being developed at InfoLab, at the Faculty of Engineering of the University of Porto. Its main objective is to index and make available entities in the domain of the University of Porto, such as students, staff or departments. This dissertation's main focus will be on information indexed by ANT and based on SIGARRA, the information system for the University of Porto.

1.2 Motivation and goals

Currently, entity extraction in the ANT project is made using hand-crafted rules, with selectors such as XPath and CSS. This means that they have to look into all of SIGARRA's web pages DOM to find out which selector rule to use for the extraction. This is extremely time consuming and requires a huge work effort. Furthermore, given their strong structure dependence, the extraction rules only work for specific pages, making it non-scalable. In addition, if the page's structure changes, the extraction rules have to change accordingly, which means the process of looking into the pages' DOM has to begin again. This also leads to not being able to extract entities from free text, such as SIGARRA's news.

The main goal of this dissertation is to automate this extraction process, making it less time consuming and greatly diminish the amount of work required. Moreover, the new extraction method will be less structure dependent, which makes it scalable to use in other circumstances.

I will also study approaches for Relation Extraction, in order to provide context to ANT's search engine, enabling it to improve entity ranking as well as to provide contextual information about the entities and their connections. Given SIGARRA is a Portuguese information system, the extraction methods will focus on this language.

There is a lot of Information Extraction software available, which already has significant performance in this field. A study of some of the state of the art approaches will be performed in order to decide on the best approaches to process the textual content available through ANT.

1.3 Document structure

This document contains three more chapters. In Chapter 2, the state of the art for NER and RE is presented, covering the extraction techniques and evaluation methods. In Chapter 3, the problem is further detailed and I present a possible solution for it. In Chapter 4, the work plan for future work is presented.

[1] See ant.fe.up.pt

Chapter 2

Named Entity Recognition and Relation Extraction

Named Entity Recognition (NER) and Relation Extraction (RE) can be performed using many different approaches. This chapter sums up the methods of extraction and the current evaluation techniques.

I will start briefly explaining what are NER and RE, along with its main extraction implementations. NER methods can be divided in two distinct approaches. The first approach hinges on creating a set of hand-coded rules to extract entities and the second approach falls in the category of machine learning systems, which is where a computer estimates the parameters using a set of algorithms.

The machine learning methods can be divided into three categories based on the data needed as input for the training algorithm. The learning algorithms can either be supervised, where the system needs an already annotated corpus, semi-supervised, where the main technique used is bootstrapping, that is to say it only requires a small set of marked examples and can extrapolate further for unmarked examples, and unsupervised which does not require annotated examples. An ontology-based extraction method is similar to the other methods, but takes into account an ontology to guide the extraction, mainly substituting the usual entity types with a pre-defined ontology.

2.1 Named Entity Recognition

NER is a sub-field of Information Extraction (IE). Its main purpose is to identify entities from unstructured text, such as news articles, or semi-structured text, such as Wikipedia articles. This extraction is the first part of a typical information extraction pipeline, supporting further information extraction methods, such as semantic analysis or relation extraction.

The concept of entity varies from approach to approach. It mainly depends on what is considered relevant to extract in each case. As already stated, the first categories of named entities emerged in the 6th MUC, where the entities were categorised as persons, locations, organisations, time expressions and numeral expressions. These have been the most common categories extracted in this field. However, some other categories are used in other, more specific, domains. Some examples are biomedical entities, such as drugs or DNA sequences, and other used for military purposes, like vehicles or weapons.

2.2 Relation Extraction

RE is also a sub-field of IE. Its objective is to recognise relations between the entities in a text. The type of relations extracted are usually highly dependent on the domain. Some of the most common relations are in the linguistic domain: hyponyms — a word with a more specific meaning —, for example, *car* and *vehicle*; antonyms — a word with contrary meaning —, for example, *happy* and *unhappy*; and meronyms — a part-whole relationship —, for example, *chapter* and *book*. Other relation types can be used in the film industry such as stars-in, director or film genres, or the food industry such as type of food, ingredients, quantities, or the social domain, such as parent-of, child-of, and geographical domain such as capital-of.

The main methods of extraction are similar to the NER methods. That being said, the extraction can be done using hand-coded patterns or machine learning methods. While NER extraction methods only focus on extracting entities, RE focuses on extracting the relationships between the entities.

Relation extraction often requires previous text analysis, which involves POS tagging, syntactic parsing, NER and choice of features to use.

2.3 Extraction methods

In this section I will present and explain the main methods for extracting entities and relations, which can be divided in hand-coded techniques, machine learning techniques and ontology-based techniques.

2.3.1 Hand-coded techniques

There are two approaches which can be considered hand-coded techniques, namely, rule-based, where patterns are used to extract entities, and dictionary-based, where tokens are matched with a gazetteer to recognise entities.

2.3.1.1 Rule-based

The first approaches to NER systems were based on the extraction of Named Entities (NEs) using grammar rules. These approaches focus on matching words using patterns, such as regular expressions.

One of the first works in this field was made by Lisa Rau [Rau91], in which she extracts company names from text using a set of manually created rules and heuristics such as capitalisation of words or detection of company suffixes, like "Inc." or "Corp.", and also generating acronyms for already existing company names.

An example of a pattern could be a title match, such as "Mr.", followed by one or more capitalised-tokens, indicating that those tokens are probably a NE with the person type. Mikheev et al. [MMG99] named these rules as "sure-fire", as they would most likely perform as expected and be successful, and used it in their system, where the match would only classify the words as *likely* candidates which, later, uses a Maximum Entropy model to decide the definitive NE classification, which is a Machine Learning (ML) technique.

This kind of pattern-matching approach can also be used for relation extraction, with patterns like "*person* lives in *location*". For instance, Barrière [Bar16] used a set of regular expressions for lexico-syntactic patterns. Hearst [Hea92] also used lexico-syntactic patterns to extract hyponym relations. Three starting patterns were defined and later, using a list of terms for which a specific relation is known to hold, more patterns can be created, either by hand-coding or by bootstrapping from an existing lexicon or Knowledge Base (KB), using ML techniques. This resulted in 66% accuracy.

Berland and Charniak [BC99] attempted extracting meronym relations from text, using patterns. They began to identify two words *building* and *basement* with close proximity from a corpus. From that they extracted patterns, for example, using the possessive as in "building's basement", and other patterns. Although they managed to extract some correct relationships, their overall accuracy was low, at 55%.

Another example, using the relation *(author,title)* for books as use case, Brin [Bri99] developed DIPRE, whose algorithm worked as follows: given a small seed set of (author, title) pairs, find its occurrences in the web and recognise patterns where they appear; then, using the extracted patterns, new (author,title) pairs can be extracted; this steps can continue until some criteria is met. This approach can be extended for other relationships' types by providing other relation pairs as a seed, so that the algorithm can find meaningful patterns.

Although hand-coded patterns are not ideal, they can provide acceptable results. Using patterns requires building them for each NE and each relation, which is hard to write and hard to maintain. Furthermore, it is infeasible to write all the required patterns since there can be multiple distinct ways of expressing entities and their relations. Finally, hand-coded patterns are usually domain-dependent.

2.3.1.2 Dictionary-based

Many approaches rely on an already existing KB to extract entities from other texts. This KB is called *gazetteer*, which is a dictionary of a collection of entities. The main algorithm can be matching the words in a text with the gazetteer and, if a match occurs, the word is annotated as an entity.

Wikipedia can be used as a KB, because it provides an enormous amount of entities. Gattani et al. [GDL+13] developed a Wikipedia-based approach for NER in social media, where the relevant words from the text were linked to a Wikipedia page. This approach was used to classify and tag tweets.

SIEMÊS [Sar06], a participating system in HAREM [Car06], used similarity rules to obtain soft matching between the entities and a gazetteer. The gazetteer used was REPENTINO [SPC06], a publicly available gazetteer for Portuguese. After identifying possible candidates as NE, it uses similarity rules to formulate judgements about the possible classes of a NE. So instead of multiple hard-coded rules over the gazetteer it only has a small set of rules, such as exact matches, partial matches either on the beginning, the end or subsets of the NE candidate, and a check for frequent words in certain subclasses, to score a match in REPENTINO.

Using gazetteers proves to be a simple method for NER. However, the NEs recognised are dependent on whether a NE exists in the gazetteer, which means that even very large gazetteer only contain a portion of all used NEs. The performance of the NER system may be affected by the introduction of new NE.

2.3.2 Machine learning techniques

There are multiple machine learning techniques applied in NER. The most used are probabilistic techniques, such as Hidden Markov Models, Maximum Entropy Markov Models and Conditional Random Fields,

Hidden Markov Model (HMM) A HMM is a statistical Markov Model, in which the state is not directly visible. In NER, HMM states are usually a name of a category of NE, with an extra state for the current word not being a NE. Each state transition is dependent only on the current state, and represents the probability for the next word be of a specific category. This probability is calculated using word features, like capitalisation of the word or if it contains numbers. This model assumes that features are independent and the feature weights are set independently.

As Ponomareva et al. [PRPM07] explain, let $o = \{o_1, o_2, ..., o_n\}$ be a sequence of words from a text with length n. Let S be a set of states in a finite state machine, each associated with a label (categories for entities). Let $s = \{s_1, s_2, ..., s_n\}$ be a sequence of states that correspond to the labels assigned to words in the input sequence o. HMM defines the joint probability of a state given an input sequence to be:

$$P(s,o) = \prod_{i=1}^{n} P(o_i|s_i)P(s_i|s_{i-1}) \tag{2.1}$$

So, in order to train the HMM, the following probabilities have to be set:

1. Initial probabilities $P_0(s_i) = P(s_i|s_0)$ to begin from a state i;
2. Transition probabilities $P(s_i|s_{i-1})$ to pass from a state s_{i-1} to a state s_i;
3. Observation probabilities $P(o_i|s_i)$ of an appearance of a word o_i in a position s_i.

These probabilities are calculated using a training corpus.

An example of the use of this approach is Nymble, by Bikel et al. [BMSW97], achieving a f-measure of 93% for English and 90% for Spanish. And other example by Zhou and Su [ZS02], in which the HMM is based on the mutual information independence assumption instead of the conditional probability independence assumption after Bayes' rule, achieving a f-measure of 96.6%. Both used the MUC-6 dataset, and the latter also used the MUC-7 dataset.

Maximum Entropy Markov Model (MEMM) MEMM works in the same way as HMM. However, it no longer assumes that features are independent, which means there can be correlated features. While HMM is a generative model, MEMM is a discriminative model. In other words, HMM learns the joint probability distribution $p(s,o)$, while MEMM learns the conditional probability distribution $p(s|o)$.

The probability of the state sequence given the observation can be computed as:

$$P(s|o) = \prod_{i=1}^{n} P(s_i|s_{i-1}, o_i) \tag{2.2}$$

MENE [BSAG98] is an example of a maximum entropy framework for NER. It participated in the Message Understanding Conference 7 (MUC-7), resulting in a f-measure of 92.20%.

Also, inspired by MENE's results, Carvalho [Car07] developed a maximum entropy framework for NER for the portuguese language.

Conditional Random Fields (CRFs) CRFs [LMP01] work similarly to HMMs but are not constrained with local features. This means that CRFs are able to deal with a much larger set of features. Furthermore, while HMM's probabilities must satisfy certain constraints, in CRFs there are no restrictions.

As Teixeira et al. [TSO11] states, according to Lafferty et al. [LMP01] and McCallum and Li [ML03], let $o = \{o_1, o_2, ..., o_n\}$ be a sequence of words from a text with length n. Let S be a set of states in a finite state machine, each associated with a label (categories for entities). Let $s = \{s_1, s_2, ..., s_n\}$ be a sequence of states that correspond to the labels assigned to words in the input sequence o. CRFs define the conditional probability of a state given an input sequence to be:

$$P(s|o) = \frac{1}{Z_o} \exp\left(\sum_{i=1}^{n}\sum_{j=1}^{m} \lambda_j f_j(s_{i-1}, s_i, o, i)\right) \tag{2.3}$$

where Z_o is a normalisation factor of all state sequences, $f_j(s_{i-1}, s_i, o, i)$ is one of the m functions that describes a feature, and λ_j is a learnt weight for each such feature function.

Kazama and Torisawa [KT07] developed a CRF NER, using Wikipedia as external knowledge, to help classify entities and thus improve the accuracy of their NER model. They performed entity linking to a Wikipedia article, and extracted a category label from the first sentence of a Wikipedia article to use it a feature.

Another use of CRF for NER was made by Amaral and Vieira [dAV13], called NERP-CRF. This algorithm was trained using the HAREM corpora. First, sentence segmentation and POS tagging was performed, so that the complexity in applying the CRF method was lessened. Multiple features for the CRF algorithm were used, such as words around the NE and the capitalisation of the NE. Using the HAREM corpora to evaluate, it scored 57.92% and 48.43% for f-measure, respectively.

In Teixeira et al. [TSO11], they present a bootstrapping method using CRF. Firstly, using a dictionary-based approach, a set of non-annotated news items is annotated. In this phase, only entities with two or more words are considered. The sentences, where all the capitalised words are annotated, are used as a seed corpus to infer a CRF model. This model, is used to annotate the same corpus used in the first stage, resulting in an increase of annotated sentences. These are then used to infer a new CRF model. This cycle is repeated until the model stabilises.

Support Vector Machines (SVMs) SVMs, known as a large margin technique, defines an optimal hyperplane which separates categories. The SVMs algorithm is based on finding the best hyperplane which gives the maximum margin between the decision border and the closest objects from the classes. Although originally SVMs can only deal with linearly-separable tasks, by using kernel functions it can transform non-linearly separable data, in its original space, into a higher dimension space, where the data becomes linearly separable. SVMs can only deal with binary classification. However, it can be used in non-binary classification tasks (such as NER), by using methods like *one-against-one* approach, in which multiple classifiers are constructed and each one deals with two different classes. Afterwards, using a voting strategy, it chooses the best category for the present object.

Mididiú et al. [MD07] is an example of SVMs use for the Portuguese language, Ekbal and Bandyopadhyay [EB10] for Bengali and Hindi, and Asahara and Matsumoto [AM03] for the Japanese language.

2.3.3 Ontology-based

The concept of Ontology-Based Information Extraction (OBIE) is relatively new. Wimalasuriya and Dou [WD10] defined OBIE, in 2010, as:

> "An ontology-based information extraction system: a system that processes unstructured or semi-structured natural language text through a mechanism guided by ontologies to extract certain types of information and presents the output using ontologies."

OBIE systems' main goal is to identify concepts, properties or relations expressed in ontologies. Being NER a sub-task of IE, some techniques may also fall in this category. For instance, Pandolfo et al. [PPA16] developed a framework for automatic population of ontology-based digital libraries. They used some of OpenNLP's [Ope] modules to perform NER using an ontology as input so that the entities have the same name of the ontology classes considered for the automatic ontology population. Their triple extractor module can extract triples from text and add them to their knowledge base. This triple a relation between the entities extracted, using a gazetteer of verbs.

This approach is particularly useful when dealing with specific domains, mainly because the most common entity categories may not be sufficient in such cases. Yasavur et al. [YALR13] defined an ontology-based approach for the domain of behaviour and lifestyle change, creating a behavioural health ontology to model world knowledge. They also used WordNet to extend the ontology for NER purposes.

Dictionary-based approaches to OBIE consist in having a gazetteer whose entities follow a particular ontology. An example of this approach is Saggion et al. [SFMB07], which adapted GATE's ANNIE module using its own gazetteer containing countries and regions gathered from multiple sources.

2.4 Evaluation and datasets

Evaluating NER systems allows us to know if the new systems are evolving in a positive way, getting higher precision and recall. There is a need for having systematic evaluation, so that all NER systems have the same standards when evaluating their performance.

There are multiple techniques proposed to rank NER systems based on their ability to annotate text correctly. These techniques were defined and used in conferences, such as MUC, CONLL, ACE and HAREM. These conferences not only differ in their evaluation techniques but also on what it is considered an entity, so they have different entity categories and types. This makes it hard to compare different tools which participated in different conferences.

The evaluation task's main objective is to compare the output of the NER system to an actual correct output by human linguistics (Golden Standard). The most common metrics used to rate classification tasks are:

- *Precision*: the ratio of correct answers (True Positives) among the answers produced (Positives). This means checking if the answers marked as positive are truly positive.

$$precision = \frac{TP}{TP+FP} \tag{2.4}$$

- *Recall*: the ratio of correct answers (True Positives) among the total possible correct answers (True Positives and False Negatives). This means checking if all the positives are marked.

$$recall = \frac{TP}{TP+FN} \tag{2.5}$$

- *F-Measure*: the harmonic mean of precision and recall.

$$f\text{-}measure = \frac{2 * precision * recall}{precision * recall} \tag{2.6}$$

The four different classes of classification results are [Kon12]:

- *True Positive (TP)*: predicted value was positive and the actual value was positive.
- *True Negative (TN)*: predicted value was negative and the actual value was negative.
- *False Positive (FP)*: predicted value was negative and the actual value was positive.
- *False Negative (FN)*: predicted value was positive and the actual value was negative.

As an example, let's use the following correctly annotated text, marked up according to MUC guidelines:

```
<ENAMEX TYPE="LOCATION">Porto</ENAMEX> received multiple personalities, such as <
    ENAMEX TYPE="PERSON">James Jr</ENAMEX>, CEO of <ENAMEX TYPE="ORGANIZATION">Acme
    </ENAMEX>, and <ENAMEX TYPE="PERSON">Richard Doe</ENAMEX>, CTO of <ENAMEX TYPE=
    "ORGANIZATION">Software Inc</ENAMEX>.
```

Now, imagine the following output produced by a NER system:

```
<ENAMEX TYPE="PERSON">Porto</ENAMEX> received multiple personalities, such as <
    ENAMEX TYPE="PERSON">James Jr</ENAMEX>, CEO of Acme, and <ENAMEX TYPE="PERSON">
    Richard</ENAMEX> Doe, <ENAMEX TYPE="ORGANIZATION">CTO</ENAMEX> of <ENAMEX TYPE=
    "ORGANIZATION">Software Inc</ENAMEX>.
```

The system outputted four different errors, explained in Table 2.1, and three different correct answers. The main objective now is to figure out what score to give to this output.

The NER conferences have different ways of dealing with this, which will be detailed in the following sections.

2.4.1 Message Understanding Conference

The Message Understanding Conference 6 (MUC-6) [GS96] was the first conference to introduce the NER task. Consequently this corresponded to the first evaluation technique definition for this task. Its evaluation metrics were chosen based on other information retrieval tasks.

In MUC events, a system is scored according to two separate axes. One for its ability to find the correct type (category) regardless of its boundaries, and another to check whether the entity boundaries are correct, regardless of its type. For both of these axes three measures were kept: COR (correct answers), POS (number of original entities) and ACT (number of guesses). Both

Table 2.1: NER errors in example.

Correct solution	System output	Error
<Location> Porto </Location>	<Person> Porto </Person>	The system recognised an entity but assigned it the wrong label.
<Organisation> Acme </Organisation>	Acme	The system did not recognise the entity.
<Person> Richard Doe </Person>	<Person> Richard </Person>	The system recognised an entity but assigned it the wrong boundaries.
CTO	<Organisation> CTO </Organisation>	The system hypothesised an entity where there is none.

precision, recall and f-measure are calculated using the sum of these two axes. This measure gives partial credit to errors occurring only on one axis, full credit for having both axes correct and zero credit for errors in both axes.

MUC considered as NEs personal names, organisations, locations and, at a later stage, temporal entities, such as date and time, and numeral measurements, such as currency and percentage expressions. MUC focused only on the English language.

While the MUC-6 only centered on NER, the MUC-7 [CM98] provided one extra task regarding the identification of relations among categories (Template Relation). The main relationships were of *employee_of*, *product_of* and *location_of*.

2.4.2 Conference on Natural Language Learning

The Conference on Natural Language Learning (CoNLL) [TD03], provided evaluation for systems either in English or German. Contrary to the MUC evaluation, CoNLL only gives credit to exact matches. That is to say, it only gives credit when both the boundaries and the type of the guessed entity are correct, corresponding to an exact match of the corresponding solution. Consequently, having only one of these axes correct results in zero credit.

This is a simple evaluation system, giving a lower estimate of a system score. This technique is often too restrictive, giving zero credit to an "almost good" answer, which sometimes is enough for the task at hand.

This conference concentrates on four types of named entities, namely persons, locations, organisations and miscellaneous.

2.4.3 Automatic Content Extraction

The Automatic Content Extraction (ACE) [DMP+04] succeeds MUC and has a different view of the NER task. This program relates to English, Arabic and Chinese texts and considers multiple

types of NEs, such as persons, organisations, locations, facilities, weapons, vehicles and geo-political entities. It also considers sub-types for these types.

The ACE program not only deals with NER, but also relation detection and event extraction, so they have different evaluation techniques for each of these tasks. ACE's 2003 relation types are *ROLE*, corresponding to a role a person plays in an organisation, *PART*, part-whole relationships, *AT*, location relationships, *NEAR*, relative locations, and *SOCIAL*, such as parent.

Its evaluation in based on a complex algorithm where each distinct NE type and each type of error have a different weight. Partial matches are allowed to a certain extent. The final score is 100% minus the sum of the penalties from these weights. The relation detection and event extraction are evaluated using the same algorithm.

This is a powerful method of evaluation due to its ability to customise the cost of error. However, given the complexity of this method with multiple parameters, it makes it difficult to compare different systems.

2.4.4 HAREM Avaliação de Reconhecimento de Entidades Mencionadas

HAREM[1] is an evaluation contest for NER in Portuguese. There were two main HAREM events, in 2005 and 2008. In the HAREM conference, there were types for entities, but also categories and subtypes. It evaluated the task of identification, the task of morphological classification and the task of semantic classification.

The first event [Car06, SSCV06] gave partial credit to producing a correct type identification of an entity and having wrong boundaries, and to producing wrong type identification and having the correct boundaries. This partial credit is given through the equation:

$$score = 0.5\frac{n_c}{n_d} \tag{2.7}$$

where n_c represent the number of common terms, and n_d the number of distinct terms between the output entities and the ones in the HAREM's golden collection, being 0.5 the maximum partial credit. The full credit is given when both type and boundaries are identified correctly.

Besides precision, recall and f-measure, the HAREM event also used other metrics, such as under and over-generation and combined error.

The second event [FMS+10] had a different approach. This was due to the introduction of a new system of classification. This new system allows the classification with multiple alternatives in each entity, using the *ALT* category, for example:

```
<ALT><Barcelona Olympic Games> | <Barcelona> <Olympic Games></ALT>
```

in which it can output the "Barcelona Olympic Games" as an event, or, "Barcelona" place and "Olympic Games" event.

[1] HAREM - HAREM Avaliação de sistemas de Reconhecimento de Entidades Mencionadas

This second event only has a single new measure, which is an extension of the combined measure of the first HAREM, taking into account the existence of subtypes and the optionality of all values.

The second event not only included the NER task but also the task of identifying the semantic relations between the NE - ReRelEM [FSM+09] track. The relations defined in this event are *Identity*, *Inclusion*, *Location* and *Other*. Relations were scored as correct, missing or incorrect. That being said, only correct identifications received one point and the remaining received none, where a correct identification only considered the triples which linked the correct NE and whose relation was well classified. The HAREM task is considerably more difficult and fine-grained than other classical NER tasks.

HAREM's golden collection [SC06], is a collection of portuguese texts from several genres, such as web pages and newspapers, in which NEs have been identified, semantically classified and morphologically tagged in context. There were 10 categories identified, namely Works of art (*Obra*), Event (*Acontecimento*), Organisation (*Organização*), Misc (*Variado*), Person (*Pessoa*), Abstraction (*Abstração*), Time (*Tempo*), Value (*Valor*), Local (*Local*) and Thing (*Coisa*).

2.4.4.1 Participants overview

In both HAREM conferences, almost all participants resorted to hand-coded techniques. For the first conference, out of nine participants, only two (NERUA and MALINCHE5) used machine learning techniques, and, for the second conference, only one out of ten (R3M). [FMS+10]

For the NER task, the top scoring participants for the first HAREM conference were PALAVRAS, using a rule-based approach and scored a f-measure of 58%, and SIEMÊS, which used similarity rules, scored 53%. For the second HAREM, Priberam, with a rule-based approach, scored 57% and REMBRANDT, also a rule-based approach, but using Wikipedia as a KB, scored 56%.

In the second HAREM, a RE task was added. Of the ten participants, only three participated in the RE task (REMBRANDT, SEI-Geo and SeRelEP), where the others only participated in the NER task. Since they chose to cover different relation types, it is not possible to compare them directly. However, taking into account all relations, REMBRANDT takes the lead with a f-measure of 45%.

2.4.5 Other

There are other attempts at evaluation apart from the shared-tasks conferences, mainly corresponding to specific cases. Furthermore, the above-mentioned conferences provide an evaluation only for participating members, being hard for non-participating programs to evaluate their own approaches. For instance, Marrero et al. [MSCMA09] evaluated multiple programs which had not participated in conferences, using mainly precision and recall, but also several features such as the typographical, lexical, semantic or heuristic factors used by each evaluated program. Other example is Konkol [Kon12], who states that attributing the correct span is hard, thus it gives more importance to the categorisation.

2.4.6 Conferences summary

Table 2.2 summarises the information for each conference.

Table 2.2: Conferences summary.

Conference	Relevant years	Entity types	Languages
MUC	1996, 1998	Person, Organisation, Location, Date, Time, Money, Percent	English
CoNLL	2002, 2003	Person, Location, Organisation, Miscellaneous	Spanish, Dutch, English, German
ACE	2003	Person, Organization, Location, Facility, Weapon, Vehicle and Geo-Political Entity	English, Arabic, Chinese
HAREM	2005, 2008	Pessoa, Organização, Local, Tempo, Obra, Acontecimento, Abstração, Coisa, Valor, Variado	Portuguese

2.5 Summary

NER and RE have been largely studied in recent years. The first approaches were mainly done through hand-coded techniques, either with patterns or with dictionary matching. However, throughout the years there was a shift in focus, where machine learning techniques started to gain more interest. This is due to the scalability of these techniques and the amount of work required. Machine learning supervised approaches continue to be the most used techniques, but recently, many semi-supervised approaches, involving bootstrapping, started to appear.

There were a lot of attempts to evaluate the current state of the art for NER and RE. Some examples, are the conferences, such as MUC, CoNLL, ACE or HAREM. They had different views on which type of entities would be considered and how to evaluate their recognition, for instance, while MUC allows partial NE recognition, CoNLL only considers exact-matches. HAREM provided a good state of the art for the Portuguese language but it is outdated since there were no more conferences focused on the Portuguese language in recent years. Nevertheless, the HAREM conferences showed that hand-coded techniques were still preferred over machine learning ones.

Although there are some attempts at NER and RE for the Portuguese language they still perform worse than for other languages, such as English.

Chapter 3

Optimisation of available tools

In this chapter, I will provide detailed information about the problem discussed in this dissertation, concerning a new method for extracting entities and relations for the Portuguese language, and propose a solution for it, mainly by exploring existing tools and adapting them for the Portuguese language.

3.1 Problem

The ANT project already has a knowledge base (KB) populated with entities. These entities were extracted using DOM selectors, such as XPath and CSS. This way of extraction requires a huge effort since, for every SIGARRA's page, the ANT team has to search for the specific place where a named entity is. Additionally to the amount of effort required, this poses a problem in two ways. Firstly, it is extremely dependent on the page structure, which means that, if the page structure changes, the current method does not work, and the process of extraction has to begin from scratch. Secondly, this method only works for specific pages, which does not support the extraction of named entities from free text, namely SIGARRA's news.

For the extraction of information from news, a new method has to be created. This new extraction approach has to be scalable, in order to work on every page and in free text. The main goal is to provide the means to link the extracted information to the already existing KB, leading to improved results in the search engine.

There are a lot of tools capable of Named Entity Recognition (NER) and Relation Extraction (RE). However, most of them only provide trained models for English and languages other than Portuguese. So there is a need for a trained model for the Portuguese language, which in turn requires a Portuguese corpus to be available to work as a training and test set. To my knowledge, the only significant publicly available dataset is the HAREM collection. This collection, although extensive, is outdated, and does not fit all the needs of an academic search engine, such as ANT.

3.2 Proposed solution

Some tools stood out in my research, namely OpenNLP, Stanford NER, NLTK and spaCy. They are all free tools and can all perform NER. However, only Stanford NER and NLTK can perform RE. I will evaluate their performance for the Portuguese language and decide, based on the results, the most appropriate tools to use.

Furthermore, I will manually annotate some SIGARRA's news, in order to have an updated dataset. This will be done using brat rapid annotation tool[1]. Apart from this dataset, I will also use the HAREM collection.

3.2.1 Tools

I have selected four tools to evaluate. While experimenting with the tools, I will produce documentation containing a set of guidelines for each of them so that later it will be easy for anyone to install and run the baseline configuration. The tools are the following:

OpenNLP[2] This toolkit performs some Natural Language Processing (NLP) tasks such as tokenization, sentence segmentation, part-of-speech (POS) tagging, NER, chunking, parsing, and coreference resolution. This tool uses Maximum Entropy (ME) and a neural network, perceptron based machine learning. OpenNLP is developed in Java.

Stanford CoreNLP[3] This tool supports multiple NLP tasks, including NER (Stanford Named Entity Recognizer) and RE (Stanford Relation Extractor). It uses Conditional Random Fields (CRF) models for training. Stanford CoreNLP is developed in Java.

NLTK[4] This tool supports the entire Information Extraction (IE) pipeline, including NER and RE. It uses ME for learning and it is developed in Python.

spaCy[5] spaCy can perform multiple NLP tasks, such as tokenization, sentence segmentation, POS tagging, NER. To my knowledge, it does not perform RE. spaCy is developed in Python.

After evaluating these tools with the default configuration, I will select two of them to further improve their extraction performance. This means, I will evaluate the importance of the features to use in the training process, the parameters for each tool and the effects of pre-processing the data.

[1] http://brat.nlplab.org/
[2] https://opennlp.apache.org/
[3] http://nlp.stanford.edu/
[4] http://www.nltk.org/
[5] https://spacy.io/

Table 3.1: Tools summary.

Tool	**NER**	**RE**	**Other tasks**	**License**	**Available models for NER**
OpenNLP	yes	no	tokenization, sentence segmentation, POS tagging, chunking, dependency parsing, lemmatization, coreference resolution	Apache License	English, Spanish, Dutch
Stanford CoreNLP	yes	yes	tokenization, sentence segmentation, POS tagging, lemmatization, dependency parsing, sentiment analysis, coreference resolution	GNU General Public License	Arabic, Chinese, English, French, German, Spanish
NLTK	yes	yes	tokenization, sentence segmentation, POS tagging, dependency parsing, stemming coreference resolution, sentiment analysis	Apache License	
spaCy	yes	no	tokenization, sentence segmentation, POS tagging, lemmatization, dependency parsing	MIT License	English, German

3.2.2 Datasets

The HAREM collection provides a good initial baseline to use as a training and testing dataset for the Portuguese language. Despite its length, it is has not been updated recently. So, in order to best evaluate the systems performances, I will manually annotate some SIGARRA's news. Furthermore, since I will not use all of HAREM's entity categories and sub-categories, I will manipulate the corpora to fit my needs.

3.2.3 Evaluation method

The ANT project requires exact matches for NER. For that purpose, I will use CoNLL's evaluation method, since it only scores the systems based on the amount of exact matches. Furthermore, it is the simplest evaluation method and the one that pessimistically evaluates the systems.

Chapter 4

Conclusions and future work

This chapter states the main conclusions for this stage of the dissertation process, regarding the state of the art for Named Entity Recognition (NER) and Relation Extraction (RE). Furthermore, it explains the future work plan, which is divided in four main tasks, namely analysing the datasets, evaluating the chosen tools, picking the best tools and optimising them for this domain.

4.1 Conclusions

After investigating the state of the art for NER and RE, the approach to use in this dissertation has become clearer. These areas have been extensively researched through recent years, so a lot of techniques were developed to deal with the inherent problems of these Information Extraction (IE) sub-tasks. The techniques developed can be divided in using hand-coded rules for matching entities and relations either with patterns or with gazetteers, and machine learning techniques, in which a machine is able to learn features from already annotated data to, later, be able to annotate unannotated text.

There were multiple conferences with the purpose of evaluating the different approaches for NER and RE, each having a different evaluation techniques. I will use CoNLL's evaluation technique because of its simplicity. Since the Portuguese language is the main focus for this dissertation, I will use portuguese corpora, namely the Golden Collection from HAREM, for training purposes. Furthermore, I will manually annotate a subset of SIGARRA's news, to serve as a comparison.

4.2 Work plan

For this dissertation, I estimate the following work plan:

1. Analyse the available Portuguese datasets

Determine the entity categories to use in the project

Manually annotate SIGARRA's news

Analyse the HAREM collection

2. Analyse and evaluate available tools and check their baseline performance; for each tool, provide documentation with guidelines for basic installation and usage

   OpenNLP

   Stanford NER

   NLTK

   spaCy

3. Pick (two) tools with the best performance

4. Optimise the chosen tools for this use case, namely for the Portuguese language and for SIGARRA's news; continuously evaluate results

Table 4.1: Detailed work plan.

Task Name	**Start Date**	**End Date**	**Duration**
Analyse datasets	**13-02-2017**	**03-03-2017**	**15d**
Determine category types	13-02-2017	20-02-2017	6d
HAREM corpora	21-02-2017	27-02-2017	5d
Manually annotate SIGARRA's news	21-02-2017	03-03-2017	9d
Analyse available tools	**06-03-2017**	**31-03-2017**	**20d**
OpenNLP	06-03-2017	10-03-2017	5d
Stanford NER	13-03-2017	17-03-2017	5d
NLTK	20-03-2017	24-03-2017	5d
spaCy	27-03-2017	31-03-2017	5d
Pick best two tools	**03-04-2017**	**05-04-2017**	**3d**
Optimise for SIGARRA's news	**06-04-2017**	**19-06-2017**	**53d**

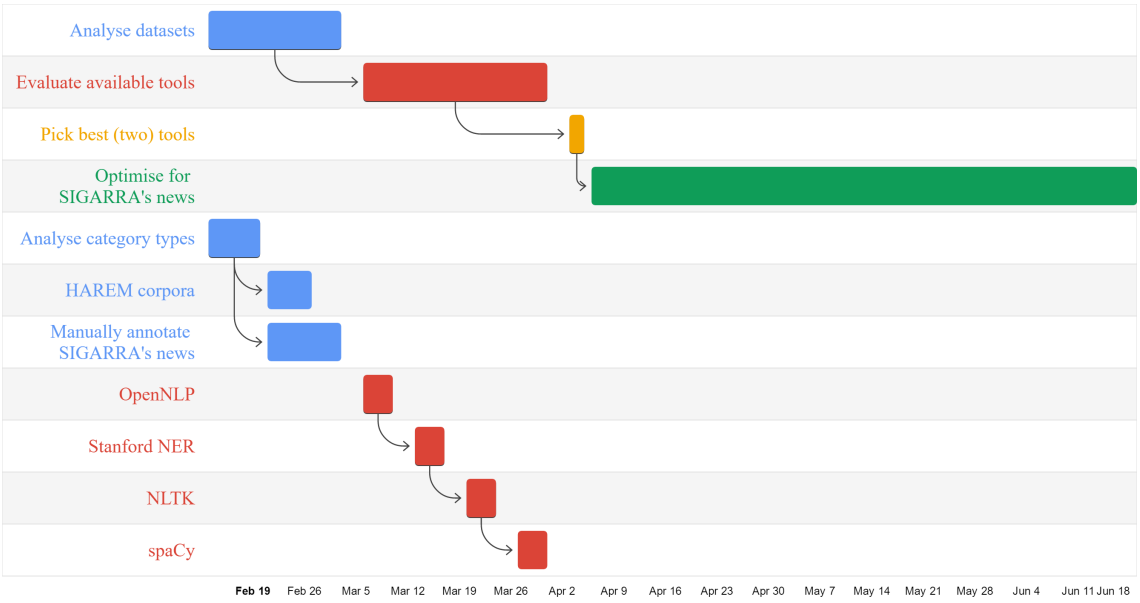

Figure 4.1: Gantt diagram for proposed work.

References

[AM03] Masayuki Asahara and Yuji Matsumoto. Japanese Named Entity extraction with redundant morphological analysis. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03*, volume 1, pages 8–15, Edmonton, Canada, 2003. Association for Computational Linguistics.

[Bar16] Caroline Barrière. Pattern-Based Relation Extraction. In *Natural Language Understanding in a Semantic Web Context*, chapter IV, pages 205–229. Springer International Publishing, Cham, 2016.

[BC99] Matthew Berland and Eugene Charniak. Finding parts in very large corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics -*, pages 57–64, Morristown, NJ, USA, 1999. Association for Computational Linguistics.

[BMSW97] Daniel M. Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. Nymble: a High-Performance Learning Name-finder. In *5th International Conference on Applied Natural Language Processing*, pages 194–201, Washington, DC, 1997. Association for Computational Linguistics.

[Bri99] Sergey Brin. Extracting Patterns and Relations from the World Wide Web. In *Selected Papers from the International Workshop on The World Wide Web and Databases*, pages 172–183, London, UK, 1999. Springer-Verlag.

[BSAG98] Andrew Borthwick, John Sterling, Eugene Agichtein, and Ralph Grishman. Exploiting Diverse Knowledge Sources via Maximum Entropy in Named Entity Recognition. In *Proceedings of the Sixth Workshop on Very Large Corpora*, pages 152–160, 1998.

[Car06] Nuno Francisco Pereira Freire Cardoso. Avaliação de Sistemas de Reconhecimento de Entidades Mencionadas. Master's thesis, University of Porto, 2006.

[Car07] Wesley Seidel Carvalho. Reconhecimento de entidades mencionadas em português. Master's thesis, Universidade de São Paulo, São Paulo, feb 2007.

[CM98] Nancy Chinchor and Elaine Marsh. MUC-7 Information Extraction Task Definition. In *Proceedings of a 7th Message Understanding Conference (MUC-7)*, pages 359–367, 1998.

[dAV13] Daniela O. F. do Amaral and Renata Vieira. O Reconhecimento de Entidades Nomeadas por meio de Conditional Random Fields para a Língua Portuguesa. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, pages 59–68, 2013.

REFERENCES

[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 *4th International Conference on Language Resources and Evaluation*, pages 24–30, Lisbon, Portugal, 2004.

[EB10] Asif Ekbal and Sivaji Bandyopadhyay. Named Entity Recognition using Support Vector Machine : A Language Independent Approach. *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, 4(September):155–170, 2010.

[FMS+10] Cláudia Freitas, Cristina Mota, Diana Santos, Hugo Gonçalo Oliveira, and Paula Carvalho. Second HAREM : Advancing the State of the Art of Named Entity Recognition in Portuguese. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, number 3, pages 3630–3637, 2010.

[FSM+09] Cláudia Freitas, Diana Santos, Cristina Mota, Hugo Gonçalo Oliveira, and Paula Carvalho. Relation detection between named entities: report of a shared task. In *Proceedings of the NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 129–137, Boulder, Colorado, 2009. Association for Computational Linguistics.

[GDL+13] Abhishek Gattani, AnHai Doan, Digvijay S. Lamba, Nikesh Garera, Mitul Tiwari, Xiaoyong Chai, Sanjib Das, Sri Subramaniam, Anand Rajaraman, and Venky Harinarayan. Entity extraction, linking, classification, and tagging for social media. In *Proceedings of the VLDB Endowment*, volume 6, pages 1126–1137. VLDB Endowment, aug 2013.

[GS96] Ralph Grishman and Beth Sundheim. Message Understanding Conference-6. In *Proceedings of the 16th conference on Computational linguistics*, volume 1, page 466, Morristown, NJ, USA, 1996. Association for Computational Linguistics.

[Hea92] Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics*, volume 2, page 539, Morristown, NJ, USA, 1992. Association for Computational Linguistics.

[Kon12] Michal Konkol. Named Entity Recognition PhD Study Report. Technical report, University of West Bohemia in Pilsen, Pilsen, Czech Republic, 2012.

[KT07] Jun'ichi Kazama and Kentaro Torisawa. Exploiting Wikipedia as External Knowledge for Named Entity Recognition. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 698–707, Prague, Czech Republic, 2007. Association for Computational Linguistics.

[LMP01] John Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the eighteenth international conference on machine learning*, pages 282–289, 2001.

[MD07] Ruy Luiz Milidiú and Julio Cesar Duarte. Machine Learning Algorithms for Portuguese Named Entity Recognition. *Intel. Artif.*, 11(36), 2007.

REFERENCES

[ML03] Andrew McCallum and Wei Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, volume 4, pages 188–191, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[MMG99] Andrei Mikheev, Marc Moens, and Claire Grover. Named Entity recognition without gazetteers. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, page 1, Morristown, NJ, USA, 1999. Association for Computational Linguistics.

[MSCMA09] Mónica Marrero, Sonia Sánchez-Cuadrado, Jorge Morato, and Yorgos Andreadakis. Evaluation of Named Entity Extraction Systems. *Research In Computer Science*, 41:47–58, 2009.

[Ope] Apache OpenNLP. https://opennlp.apache.org/. Accessed: 2017-01-07.

[PPA16] Laura Pandolfo, Luca Pulina, and Giovanni Adorni. A Framework for Automatic Population of Ontology-Based Digital Libraries. In *AI*IA 2016 Advances in Artificial Intelligence*, pages 406–417. Springer International Publishing, 2016.

[PRPM07] Natalia Ponomareva, Paolo Rosso, Ferran Pla, and Antonio Molina. Conditional Random Fields vs. Hidden Markov Models in a biomedical Named Entity Recognition task. In *Proc. of Int. Conf. Recent Advances in Natural Language Processing, RANLP*, pages 479–483, 2007.

[Rau91] Lisa F. Rau. Extracting company names from text. In *[1991] Proceedings. The Seventh IEEE Conference on Artificial Intelligence Application*, volume i, pages 29–32. IEEE Comput. Soc. Press, 1991.

[Sar06] Luís Sarmento. SIEMÊS - A named-entity recognizer for Portuguese relying on similarity rules. In Renata Vieira, Paulo Quaresma, Maria das Graças Volpe Nunes, Nuno J. Mamede, Cláudia Oliveira, and Maria Carmelita Dias, editors, *Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006, Itatiaia, Brazil, May 13-17, 2006. Proceedings*, volume 3960 LNAI, pages 90–99. Springer Berlin Heidelberg, 2006.

[SC06] Diana Santos and Nuno Cardoso. A Golden Resource for Named Entity Recognition in Portuguese. In *Proceedings of the 7th International Workshop, PROPOR 2006*, pages 69–79. Springer Berlin Heidelberg, 2006.

[SFMB07] Horacio Saggion, Adam Funk, Diana Maynard, and Kalina Bontcheva. Ontology-based Information Extraction for Business Intelligence. In *Proceedings of the 6th International The Semantic Web and 2Nd Asian Conference on Asian Semantic Web Conference*, pages 843–856, Berlin, Heidelberg, 2007. Springer-Verlag.

[SPC06] Luís Sarmento, Ana Sofia Pinto, and Luís Cabral. REPENTINO – A Wide-Scope Gazetteer for Entity Recognition in Portuguese. *LNAI*, 3960:31–40, 2006.

[SSCV06] Diana Santos, Nuno Seco, Nuno Cardoso, and Rui Vilela. HAREM: An Advanced NER Evaluation Contest for Portuguese. In *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC'2006*, pages 1986–1991, 2006.

REFERENCES

[TD03] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, volume 4, pages 142–147, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[TSO11] Jorge Teixeira, Luís Sarmento, and Eugénio Oliveira. A Bootstrapping Approach for Training a NER with Conditional Random Fields. In *Progress in Artificial Intelligence*, pages 664–678. Springer Berlin Heidelberg, 2011.

[WD10] Daya C. Wimalasuriya and Dejing Dou. Ontology-based information extraction: An introduction and a survey of current approaches. *Journal of Information Science*, 36(3):306–323, 2010.

[YALR13] Ugan Yasavur, Reza Amini, Christine Lisetti, and Naphtali Rishe. Ontology-based Named Entity Recognizer for Behavioral Health. *In Proceedings of the 26th International FLAIRS Conference*, (3):249–254, 2013.

[ZS02] Guodong Zhou and Jian Su. Named Entity Recognition using an HMM-based Chunk Tagger. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 473–480, Philadelphia, 2002. Association for Computational Linguistics.