Universidade de Aveiro
2020

Departamento de Electrónica,
Telecomunicações e Informática
Programa de Doutoramento em Informática
das Universidades de Aveiro, Minho e Porto

Emanuel Matos

[Extração e Acesso a Informação em Linguagem Natural] para não desenvolvedores - Democratizando a Informação

Pré-tese

Extraction and Access to Information in Natural Language for Non-Developers - Democratizing Information

Pre-Thesis

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Automatic extraction of information from these natural language sources has many applications, including Business Intelligence, Forensics, and question answering systems. Despite their potential, creation of such systems for new domains is not simple and, in general, development has been limited to developers with in-depth knowledge of the area. As extraction methods improve, the amount of information also augments, making more and more difficult effective access by humans.

1.2 What is IE?

The basic concept of Information Extraction is that we do not need to determine to undermine the structure of relationships in advance, who the actor is and / or his action, allowing greater flexibility and scalability, in theory more extractions of relationships and independence of the domain.

Thus, we will have the possibility of discoveries that do not are directly evidenced. Some characteristics of an Open Information System: running a single execution in the corpus, guarantee scalability, independence of the corpus size and the domain. Have a single input, a corpus and an output that must be a set of extracted relations. Be unsupervised.

Information extraction will be useful in finding answers where we have some difficulties to assess the text structure, where we will have an untabbed volume of

text and the need to identify a certain type of response / information that does not have a structure formal evidence of content. The Open Information Extraction has the disadvantage to be less consistent than the Extraction of Traditional Information (Banko et al., 2008)

In general terms, the OIE is still developing, needs much more studies for improvement of technique and therefore theoretical and praxis improvement. This report intends to make a contribution in the historical, technical and works view with its authors providing what we call the OIE's "backbone".

1.3 IE Applications

The Information Extraction Technology (IE) is used to transform data in a structured way, suitable for automatic processing by machines in Information and then Intelligence to solve problems of the most varied forms and contents. OpenIE's goal is to recognize mentions to the specified entity and discover relational structures of unstructured data (ie text). The OpenIE system generally consists of two subtasks: (i) named entity recognition (NER) and (ii) relationship extraction (RE).

1.4 Open Problems

Currently, we need to stay updated and attentive to changes. This need makes us seek more information and discernment through various paths, through the WEB, through physical newspapers, through radios, in short, through an increasing diversity of data outputs. This data comes mainly in the form of Natural Language, in texts, speech, videos, etc. The approach of this work will try to explore the aspect of natural language that we have in texts on the Internet. Pages and / or sections that place texts where we should extract relevant content, for people who are not experts in the area of Open Information Extraction.

Chapter 2

Methods Review / Background and Related Work

2.1 Overview

The goals of study NER and OpenIE softwares are direct relationship with understand about the enviromnent develop , straightforward and easy to understand – but they aren't always easy to meet. This is because there are so many different ways to approach software engineering and so many outcomes that are possible. While we do have best practices and there are standards in place, every software engineer has a different approach and sometimes they don't always mesh well with other members of an IT team.

2.2 NER

The Named Entity Recognition (NER) is within the natural processing of language, being a process that from a sentence or part of it, this text is captured and verified its possibilities of finding entities that can be of the most varied categories such as names, organizations , places, quantities, monetary values, percentages, etc. NER algorithms can be trained on the basis of ML to extract the entities mentioned above. The NER for having a simple but effective approach, is also used as a kind of "Pre-Processing" for Open IE.

2.3 IE

Currently, IE is used to transform data in a structured way with automatic processing by machines in Information and Intelligence. We will use ML concepts that come from IE to structure OpenIE, so we will create and evolve on the basis.

2.4 OPenIE

The Open Information Extraction (Open IE) is a way to extract information based on obtaining non-predefined and independent domain relations of a text, this field of study for not having rigid search rules makes this process rich in obtaining results where not expected, this is your greatest strength.

2.5 Datasets

In this chapter, we list which possible data sets we can work with, these data sets must be easily accessible and available to the general public. What we realized is that when searching these data sets, only the data sets that are part of the guides for each software were available and easy to process. Below is the list of possible data sets and their characteristics.

Drop AllenNLP 2019 Dataset AllenNLP https://allenai.org/data/drop

CONLL 2003 – Since 1999, CoNLL (Conference on Natural Language Learning), is an annual SIGNLL meeting. CoNLL 2003 stood out because 1,393 news items were made available in English and 909 in German. The English corpus is free, but the German corpus is not. You will have access a few days after sending the organizational and individual contract for free. The entities are noted with LOC (local), ORG (organization), PER (person) and MISC (miscellaneous).

https://www.clips.uantwerpen.be/conll2003/ner/

https://github.com/davidsbatista/NER-datasets/tree/master/CONLL2003

Ontonote-5.0 – Entities:Organization, Art Work, Numbers in word, Numbers, Quantity, Person, Location, Geopolitical Entity, Time, Date, Facility, Event, Law,

Nationalities or religious or political groups, Language, Currency, Percentage, Product.

Ontonote 5.0 is a Dataset provide by LDC "The Linguistic Data Consortium is an open consortium of universities, libraries, corporations and government research laboratories. LDC was formed in 1992 to address the critical data shortage then facing language technology research and development. The Advanced Research Projects Agency provided seed funding for the Consortium and the National Science Foundation provided additional support via Grant IRI-9528587 from the Information and Intelligent Systems division. " - https://www.ldc.upenn.edu/about

At the beginning of July I signed up and waited for the email to be validated, it didn't arrive, I did the same procedure in the second week with my email from the University of Aveiro, and I also didn't get a response. At the beginning of the third week, I requested DataSets, LDC 2013T19 - Ontonotes 5.0, LDC99T42 TreeBank, LDC2008T19 The new York Times and LDC2006T06 ACE 2005 Multilingual with a Non-Member request.

GMB(Groningen Meaning Bank) – Entities: Natural Phenomenon, Person, Geographical, Organization, Art Work, Event, Time, Geopolitical.

NAACL 2019 – Entities: Organization, Person, Location, Geopolitical, Facility, Vehicles.

Wnut2017 – Entities: Location, Person, Product, Groups, Corporations, Creative.

2.6 Tools

In the table 2.1 we have a some new,base and relations of tools from OpenIE and NER.

Table 2.1: Systems in General

System	HTTP	Access date
AllenNP	https://docs.allennlp.org/master/	07/Jun/2020
Linguakit	https://linguakit.com/en/full-analysis	10/Jun/2020
DptOIE	https://github.com/FORMAS/DptOIE	01/Ago/2020
BERT (Devlin et al., 2018) Relation Extraction (Soares et al., 2019)	https://tinyurl.com/yaen9y52	08/Jun/2020
CrossOIE	https://github.com/FORMAS/CrossOIE	01/Ago/2020
Stanford CoreNLP – Natural language software	https://tinyurl.com/yxq9sysp	07/Jun/2020
KnowItAll	https://github.com/knowitall	07/Jun/2020
Several Projects NLP	https://tinyurl.com/ybvv43jt	08/Jun/2020
The Top 40 Information Extraction Open Source Projects	https://tinyurl.com/yc76ewz7	08/Jun/2020
Open Information Extraction - Softwares	http://www.cse.iitd.ernet.in/~mausam/software.html	08/Jun/2020
Top 26 Free Software Text analysis-Text Mining	https://tinyurl.com/yb8sgjew	12/Jun/2020
NTLK - Python	https://www.nltk.org/book/ch07.html	07/Jun/2020

Chapter 3

Proposal

3.1 Objectives

The main objectives are:

- study and develop processes to simplify creation of information extraction pipelines for new domains ;
- study and develop methods to simplify access and exploration of extracted information, exploring recent developments in question answering and dialog systems;
- contribute to democratization of usage of such systems in non specialist.

Research Questions

There are many problems to be solved when it comes to Information Extraction. In the matter in question, it must have a sensitive and very complex part, because we can and must take into account, if possible, attitudes and desires that the texts can express without the main speaker and / or the listener being aware of what he is talking about and what are the main information that we can obtain at the end of Natural Language Processing. The main problem of this work is to create an interface or pipeline that is "friendly" and has as a work engine a new concept of Natural Language processing or better, a Pipeline that must be based on the most advanced OpenIE's and structured with automatic learning in each step of this pipeline. We propose the Tourism area as a study and work environment. This field

is an important aspect for the economy of several countries, therefore, this work will have aspects of dissemination of knowledge, construction and technological development and also with an economic leverage bias.

Initial research questions include:

- how to combine multiple methods and existing pipelines to improve performance and simplify development;
- can Open-IE be used to effectively bootstrap creation of relation extractors;
- how to use the extracted information to further improve extraction;
- how to couple extraction with users' interface to access information.

3.2 Approach

3.2.1 Engineering Design Process

In the site https://www.sciencebuddies.org/science-fair-projects/engineering-design-process/engineering-design-process-steps, we find a pipeline basic to develop our principal steps and design.

3.2.1.1 Define the Problem

The engineering design process starts when you ask the following questions about problems that you observe:
What is the problem or need? simplified and automation, more domains
Who has the problem or need?General people, not programmers
Why is it important to solve?Democratic Knowledge

3.2.1.2 Background Research

Learn from the experiences of others — this can help you find out about existing solutions to similar problems, and avoid mistakes that were made in the past. So, for an engineering design project, do background research in two major areas:

- Users or customers

- Existing solutions

3.2.1.3 Specify Requirements

Design requirements state the important characteristics that your solution must meet to succeed. One of the best ways to identify the design requirements for your solution is to analyze the concrete example of a similar, existing product, noting each of its key features.

3.2.1.4 Brainstorm Solutions

There are always many good possibilities for solving design problems. If you focus on just one before looking at the alternatives, it is almost certain that you are overlooking a better solution. Good designers try to generate as many possible solutions as they can.

3.2.1.5 Choose the Best Solution

Look at whether each possible solution meets your design requirements. Some solutions probably meet more requirements than others. Reject solutions that do not meet the requirements.

3.2.1.6 Develop the Solution

Development involves the refinement and improvement of a solution, and it continues throughout the design process, often even after a product ships to customers.

3.2.1.7 Build a Prototype

A prototype is an operating version of a solution. Often it is made with different materials than the final version, and generally it is not as polished. Prototypes are a key step in the development of a final solution, allowing the designer to test how the solution will work.

3.2.1.8 Test and Redesign

The design process involves multiple iterations and redesigns of your final solution. You will likely test your solution, find new problems, make changes, and test new solutions before settling on a final design.

3.2.1.9 Communicate Results

In order to guide, deepen and consolidate the latest knowledge and technologies and be able to better address the research questions and the proposed problems, as well as reevaluate other possibilities, we will deal with the process of internalizing knowledge on three fronts:

- A bibliographic search with a view to new technologies and new developments for structuring a Pipeline that is also conceptually to structure the topic of automatic learning.

- A comparative and adaptive study of OpenIE Projects that are well positioned as a reference for new technologies.

- Conduct a field study with the new tool that we must develop throughout this work, proving its contribution to society and to the academy, since it is expected that a work of this level will bring this type of contribution.

3.3 Proposed solution

Proposing a solution is very difficult, what we are proposing is the construction of a Project / Process that is seen as a facilitator so that people can extract information from texts related to aspects of tourism in Portuguese, and with technological gain in Portuguese, the back-end that will serve as a pipeline for the development of this project, this is what we have as a goal today. This name will be CAIT , Computer App Information Tourism.

In figure 3.1 is a basic CAIT baseline

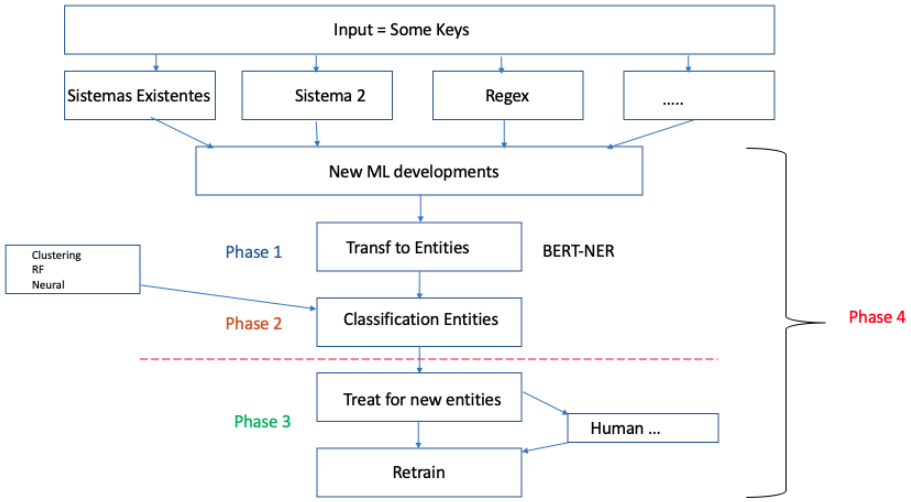

Figure 3.1: Pipeline

3.4 Main Tasks

During this period, in order to seek the development and delivery of the project, we have divided it into some parts, which may make sense and complementarity to use them.

Literature Review / State of Art ——

We will revisit the works that are currently considered cutting edge cases for OpenIE, NER and IE.

Tools search and selection ——

Based on tools that we will evaluate in the academic and business environment (provided they are free), we will be choosing which one or which should be within the initial pipeline. This search for better options must follow procedures and metrics to guide the best choice. These tools will be the starting point for creating the Project pipeline. What will result in the software that we already call Computer App Information Tourism (CAIT).

NER without annotaded data ——

Another important stage will be the construction of the NER's. These should be based on the combinations of NERs that will exist after evaluating the tools. We will try to create a metric that is a statistic based on the spontaneous measurements and discoveries of each base NER within the pipeline. We will also discuss the possibility of adding generic NERs to a personal

list. For entities, we will propose the technique of Bidirectional Coding of Representations of Transformers (BERT-NER).

Tool for Dummies / NER ——

In creating an APP that we will use as an intentional engine for knowledge extraction, process evaluation and usability. This APP must have some types of input, such as being able to take pictures of the text and insert a pdf file to collect the data that the user will use for after our processing has access to the NER, if you want you can "retrofeed" with the experience that he himself has. The main objective will be with the data collected through file, typed or by photo (ease of entry) the system will respond with a list of NER's where the output can be a csv file, a list on the screen and even a list where the user can or not to validate the NER found. This tool at this stage will be offered to a restricted audience for adjustments in the construction process of the CAIT Learning Machine. This audience will also provide feedback on the usability and perceived correctness of the APP.

Tool targeting Turism ——

The tool that we must develop will be based on the Tourism area, this environment was chosen for several reasons: There is a group in the AU that focuses on the Tourism area, so that we can be complementary and synergistic to the work in progress. Tourism is one of the areas where Portugal and Brazil, Portuguese-speaking countries, have a strong economic appeal. This area has always called my attention to discover the best options for Hotels, Tours, according to the availability of accesses. For at least the 3 circumstances indicated, we need good, correct, updated and reliable information, so developing a web page that is easy to access and simple to use will also be one of the bets of this work and behind this simplicity we will have an engine that will guarantee the execution within the characteristics sought, that is, correct, updated and reliable information of the content to be evaluated.

One-Shot or Few-shot ——

We will use a technique called One Shot and Few Shots, which will consist of using an existing system and without much programming or complex programming, we will bring Information based on this IE. We will have 4 phases:

- phase 1: we will validate the output by an appropriate ontology by specialists in the field;

- phase 2: with the material already filtered we will train and retrain iteratively; in this phase we will study a stop or convergence point that minimally satisfies the target audience, in addition to assuming performance metrics;
- phase 3: we will explore new paths where, for example, the verb is not present, but there is a relationship between the entities, in addition to analyzing what is failing and, thus, providing feedback to the system;
- phase 4: new ML models are to be incorporated into the system engine.

CAIT with Relations ——

For a more in-depth development of CAIT, we will use the predominant relationships, that is, in the composition of the items to be evaluated, they will have the greatest "strength" within the evaluation concepts that we will format.

Natural access and exploration of extracted information ——

We will explore other problem-solving capabilities, as, instead of texts, we will assess whether indexing will propose a reduction in processing time or an improvement in responses.

3.5 Expected results

The expected result is to have an App that is easy to access, easy to handle, with a response pattern where the result is superior to that found in approved systems and with greater difficulty in handling. In other words, to democratize quality information, for access by non-developer professionals in the area in question. Without neglecting the supply of cutting edge technology.

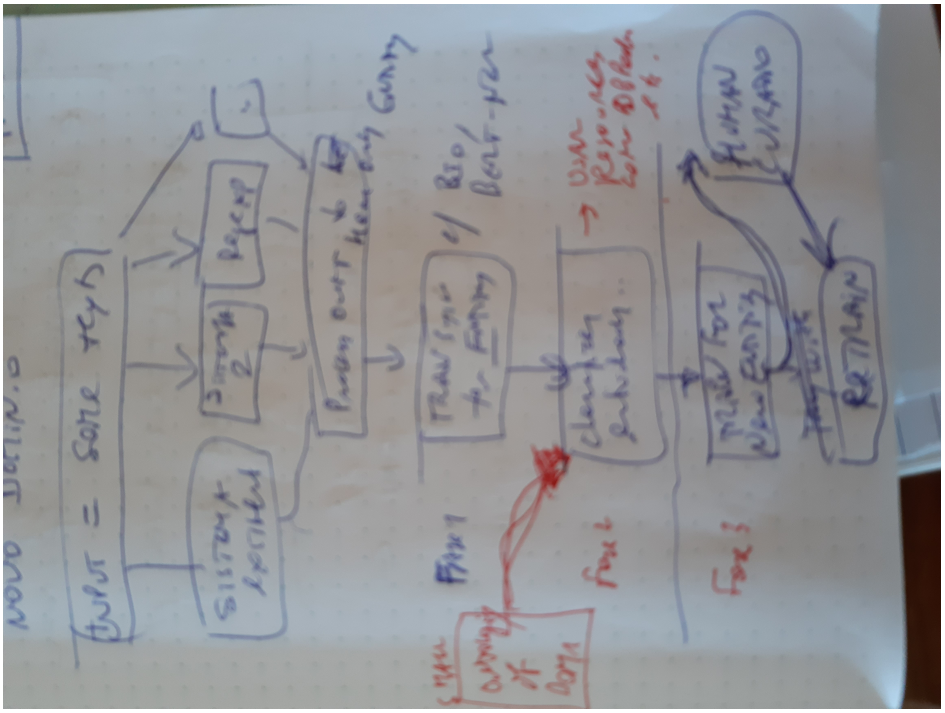

Figure 3.2: Caption

Chapter 4

Workplan

Can start by 20 september.
Before we need chapter 3!

4.1 Research Timeline

In the table 4.1 we have the proposal timetable.

Table 4.1: Thesis Schedule

	2020-2021					2021-2022				2022-2023			
	Now	1Q	2Q	3Q	4Q	1Q	2Q	3Q	4Q	1Q	2Q	3Q	4Q
State of Art Review													
Tools													
NER "without data" [M1] - Eval													
Tool for Dummies - Eval													
Tool App Turismo (CAIT) Included Info Exploration (Simple Version) [M2]													
One Shot/ Few Slot (OpenIE)													
CAIT w/Relations Extration / Exploration[M3]													
Publications													
Thesis[M4]													
Milestones					M1				M2		M3		M4

4.2 ?Workpackages

4.3 Milestones?

It is important to define around 2 or 3 per year...

- M1
- M2
- M3
- M4

4.4 Publication plan?

4.5 Initial work

During 1st year some preparatory work was already performed... Did

4.5.1 initial review of NER and OpenIE (Supervised Study)

short summary with info on number of systems, papers consulted etc

1/2 page is enough

4.5.2 first selection of NER and OpenIE tools (Scientific Activities)

short summary with info on number of systems, metrics, datasets, ...

1/2 page is enough

Chapter 5

Conclusion

This document presented and detailed the doctoral thesis proposal entitled "Extraction and Access to Information in Natural Language for Non-Developers - Democratizing Information". The thesis proposal was proposed with its motivation, objectives and expected contributions, the risks we have, we will be mitigating throughout the process and the uncertainties when they occur will be evaluated. We present the pipeline in order to provide a broad understanding of the main areas and tasks most relevant to this work. In conceptual creation, the emphasis was placed on creating the learning mechanism based on several fronts of OpenIE and NER. The revision of the work in the State of the Art that we proposed concerns the integration of planning and learning in the domain of OpenIE and NER. This served not only to detail the state of the art in this field of research, but also to point out problems and successes between the work of literature and the work proposed here. Finally, we present and detail the tasks identified to achieve the proposed objectives and contributions, as well as their chronology and the general pipeline that should result from the final product of this thesis. The presentation of the preliminary proposal, the milestones and the publication strategy finalize this work.

This document presented and discussed the PhD thesis proposal entitled “Planning and Learning for Games”. The thesis proposal was first contextualized by presenting its motivation, goals and expected contributions to the state of the art, as well as any expectable risks. The presentation of the technical background ensued, in order to provide a theoretical understanding of the main research areas relevant to this work. Special emphasis was given to Automated Planning, Reinforcement Learning and research concerning their integration. Next, followed the

review of the related work, particularly that concerning the integration of planning and learning in the context of games. This served not only to layout the state of the art in this research domain, but also to point out the differences between the work in the literature and the work proposed here. Finally, the tasks identified in order to carry out the goals and contributions proposed, as well as their timeline and the conceptual architecture of the integrating framework that should result as the final outcome of this thesis were presented and discussed. The presentation of the preliminary work and the publication strategy concluded the discussion.

Artificial Intelligence [adapt to Information Extraction] is a thriving research field, moving at a very fast pace, with innovative and groundbreaking work emerging every day. The range of application of AI grows continuously and is by no means limited to video games and entertainment. AI will undoubtedly revolutionize the way humans live and even the way humans think about themselves, as intelligent acting agents. This thesis intends to contribute to this exciting and innovative field of research.

Bibliography

Banko, Oren, Etzioni, M., Soderland, S., and Weld, D. S. (2008). Open information extraction from the web. *Communications of the ACM*, 51(12):68–74.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Soares, L. B., FitzGerald, N., Ling, J., and Kwiatkowski, T. (2019). Matching the blanks: Distributional similarity for relation learning. *arXiv preprint arXiv:1906.03158*.