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. "Information Extraction (EI) is the branch of the information retrieval area that uses techniques and algorithms to identify and collect desired information from documents, whether structured or not, storing them in a format appropriate field for future consultation "(Cabral, 2009).
This first part of the work was designed to briefly review the state of the art of some of the main open information systems and their possible consequences.
In this view of Information Extraction Systems, we will have its evolution from the 70's until the beginning of 2020, mainly in the English language. The main and most recent Open Language Information Systems in Portuguese will be described.

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

The history of extracting information from the records found, refers to the end 1960s, with the system called ELISA. In the early 1970s, with the article "GRAMMAR, MEANING AND THE MACHINE ANALYSIS OF LANGUAGE"by Yorick Wilks (Wilks, 1972) where he reported his work on Computable Semantic Derivations (CSD), focusing on Semantic disambiguation, based on the ELISA system.

The article "Methods for Domain-Independent Information Extraction from the Web: An Experimental Comparison" (Etzioni et al., 2004a), describes the classifier "KnowItAll". For this article we can believe that "KnowItAll" was a precursor to "TextRunner"and "ELISA" was one of the first Classifiers even though he did not pass the Turing.

TextRunner is considered to be a second generation of Open IE systems. In the article "An Overview of Open Information Extraction" (Gamallo, 2014), says "...: The first OIE system, TextRunner (Yates et al., 2007), belongs to this category.",but we can say that "KnowItAll" (Etzioni et al., 2004b) was the 1st. OpenIE process systems.

Below is table 2.1 which lists some of the historical systems and processes.

Table 2.1: Historic Develops from IE

System	Reference	Year	Language	Technology
ELISA	Weizenbaum (1966)	1966	EN	Semantic disambiguation
CSD	Wilks (1972)	1972	EN	Semantic disambiguation
OLLIE	Tablan et al. (2003)	2003	EN	ML
KnowItAll	Etzioni et al. (2004b)	2002	EN	Naive Bayes Classificator(NBC)
TextRunner	Yates et al. (2007)	2007	EN	Naive Bayes Classificator

2.3.1 ELISA

ELISA was the first so-called open information system, it worked with decomposition of rules through the keyboard, the system was "trained" in the decomposition and terms, so while the user was writing, the system grouped the terms and joined them and presented after specific commands.

The era of Open Information Systems was beginning, even without having realized. Below is a figure 2.1 of the ELISA concept.

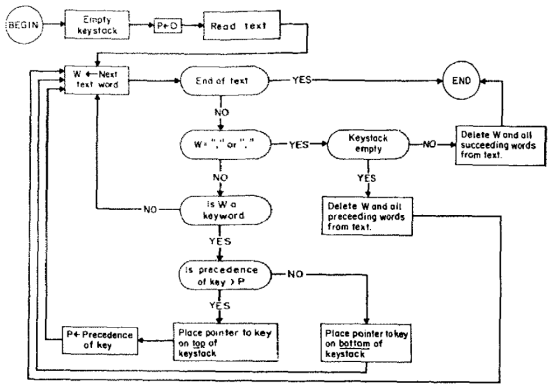


Figure 2.1: Keyboard Detection Flow / ELISA by (Weizenbaum, 1966)

2.3.2 CSD - Computable Semantic Derivations

CSD which is a process designed exclusively to disengage the senses of words. Unlike ELISA(Weizenbaum, 1966) ,does not produce an analysis tree of sentences, although it produces a small amount, which could be considered an analysis syntactic.

The semantics process is a list structure whose atomic elements are selected from a set of 53 primitive semantic classifiers(Katz and Fodor, 1963). Another feature of the CSD is the concept of expansion, which is comparable to our ability to recognize and understand words used in a new sense, in a metaphorical concept.

This one process has not been worked out on a large scale, just for a few examples. Like ELISA, this is yet another evolution towards the Information Systems Open formations.

2.3.3 OLLIE

OLLIE is a process for developing a framework environment for learning open and distributed. We can say that it is also an online application for annotation corpus that harnesses the power of Machine Learning (ML) and Information Extraction (IE) to facilitate and make the annotator's task more efficient.

We can characterize OLLIE as a process, made in JAVA, which is a facilitator information collection. The primary capacity for learning and distributing students data facilitates the process as a whole. Figure 2.2 shows the flow of the OLLIE process.

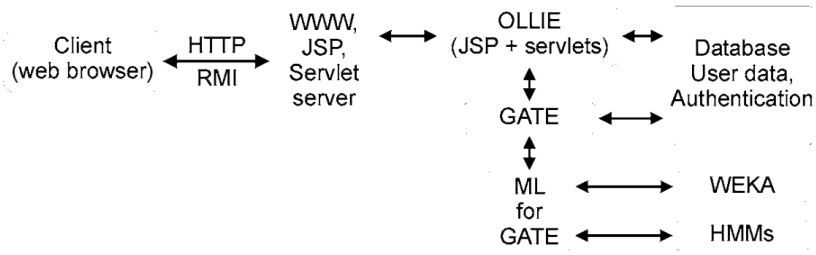


Figure 2.2: Architecture / OLLIE by (Tablan et al., 2003)

2.3.4 KnowItAll

KnowItAll is a system that aims to automate the process of extracting large volumes of data from the Web in an autonomous, scalable and independent of domain. There are seeds of Ontology that are inserted in addition to a small number of rules.

It used Naive-Bayes with Bootstrapping in his Extractor, due to the difficulty in obtaining the extraction through the WEB is quite different. At the time evaluated it was very early, the its use thus needed a larger volume of data for a better evaluation.

2.3.5 TextRunner

TextRunner is an Information System called Open Information Extraction (OIE), in which the system makes a single data-driven pass across the corpus and extracts a large set of relational tuples, without the need for any contribution human (Yates

et al., 2007).

In a single pass through all documents, marking phrases with tags from part of the part of speech and parts of substantive sentences. For each pair of nominal phrases that are not very distant and subject to several other restrictions, the concept of triples t=(*arg1*,*rel*,*arg2*).

From this extractor a good part of future Open Information Systems is used used this concept. Comparing with KnowItAll and brought a significant gain in correct extraction of sentences.

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.

Below we have in table 2.2 which lists some of the most recent Information Systems Open with several technologies.

Table 2.2: Recent Developments

System	Reference	Year	Language	Technology
ATP-OIE	(Rodríguez et al., 2020)	2020	EN	Rules
MCTS	(Liu et al., 2020)	2020	EN	ML:Markov
MinIE & MinScIE	(Gashteovski et al., 2017) (Lauscher et al., 2019)	2017/2019	EN	Rules/ ReverB/ ML:SVM
TruePIE	(Li et al., 2018)	2018	EN	ML:KNN
CoNEREL	(Phan and Sun, 2018)	2018	EN	ML:GRAPH/PAIR-LINK
Triplex-ST	(Mirrezaei et al., 2016)	2016	EN	ML:Bootstrapping
Sequence2Sequence	(Wiseman and Rush, 2016)	2016	EN	Generate sequence-labeling / ML:NEURAL
ReVerb	(Fader et al., 2011)	2011	EN	Rules + Analyze Syntactic

2.4.1 ATP-OIE

ATP-OIE or "Autonomous Open Information Extraction Method" is a System that uses semantic relations generated automatically from examples as a pattern of extraction. These relationships are generated from examples, so the more examples the greater autonomy, this difference from the methodology based on fixed rules. We can assess that this System "learns" based on examples.

Problems can arise if the examples are too random or too concentrated. ATP-OIE can use other methods like ReVer(Fader et al., 2011) and ClausIE(Del Corro and Gemulla, 2013), f not find semantic relations. At ATP-OIE there is an implementation that helps to avoid common mistakes in extracting Information. Following a comparative table 2.3 of metrics.

Methods	Precision	Recall	F1-Measure
ClausIE	0.467	0.519	0.492
OLLIE	0.456	0.416	0.435
ReVerb	0.633	0.319	0.424
MinIE-C	0.612	**0.593**	**0.6022**
ATP-OIE Standalone	0,650	0,294	0,401
ATP-OIE+R+C	**0,680**	0,401	0,504
ATP-OIE Online	0,670	0,390	0,493

Figure 2.3: Table with comparative metrics of the ATP-OIE of (Rodríguez et al., 2020)

ATP-OIE has been compared with other leading methods in a well-known database of texts: "Reuters-21578", obtaining a higher precision than with other methods.

2.4.2 MCTS

MCTS which stands for "Monte-Carlo Tree Search" is an Information Extraction system Open training, based on the Markov Chain(Levin et al., 1998). This process provides, based on a simulator, to learn the reward signs of a Reinforced Learning, with the Seq2Seq predictor (Wiseman and Rush, 2016) pre-trained who generates samples, explores candidate words during training.

The samples are feedback in order to improve the forecast. This technique in the evaluation empirical study demonstrated that the MCTS inference improves forecast accuracy (more than 10%) and achieves a leading performance in relation to other models of comparison of this generation. In figure 2.4 we have the MCTS Framework.

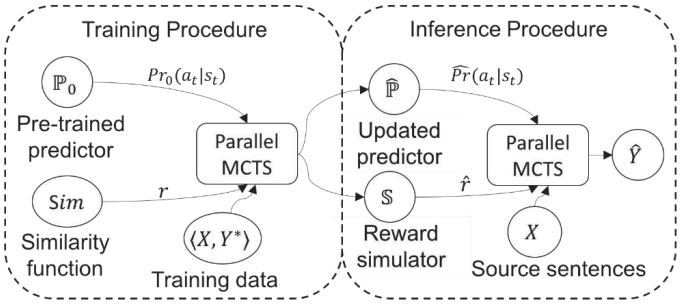


Figure 2.4: Framework MCTS for (Liu et al., 2020)

2.4.3 MinIE e MinScIE

MinIE is an Open Information System that addresses information, modality, assignment and quantities of semantic annotations instead of real extraction. Identifies and removes very specific parts. This system proposes useful, compact extractions of precision and recall. MinIE's semantic annotations represent information about polar- age, modality, assignment and quantity.

MinScIE is an optimized version of MinIE and has 3 percentage points of improvement model, according to the report. This model is adapted for the Scientific domain. Considering the occurrence of quotation marks and text, the system offers a more precise higher than its non-adapted core, MinIE. Its importance is that it allows the connect factual knowledge with references to scientific discourse. Below is Figure 2.5 of the MinScIE Pipeline.

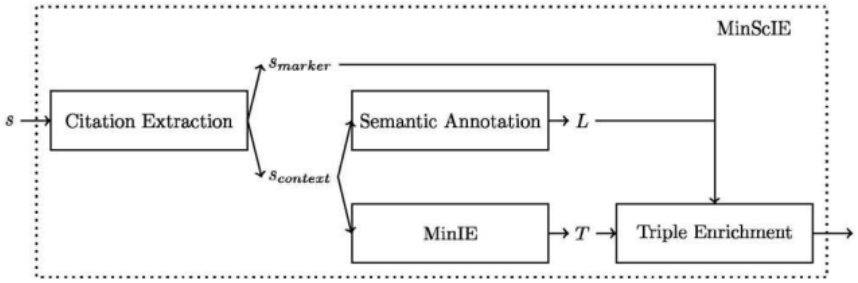


Figure 2.5: MinScIE pipeline from (Lauscher et al., 2019)

2.4.4 TruePIE

TruePIE is an NLP model that finds reliable standards where it can be extracted not only related information, but also correct information. TruePIE works with learning and repeats the feedback process for reliable standards, or Reinforcement Learning.

However, in the evaluation of this System it was found that one of the main reasons that cause errors in TruePIE is that devices are not able to distinguish enough to classify positive and negative patterns negative. Especially for standards with sparse or ambiguous named entities and low frequency and low coverage patterns. Next in figure 2.6 the TruePIE Framework.

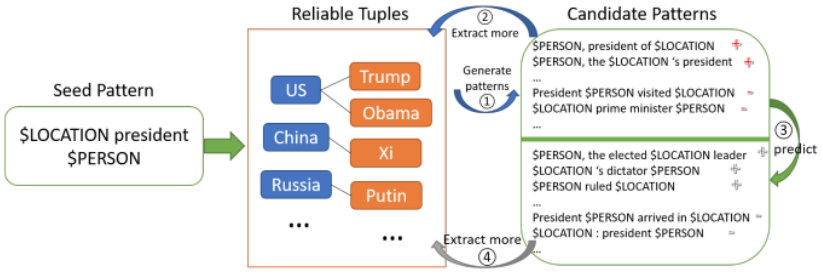


Figure 2.6: Framework TruePIE from (Li et al., 2018)

2.4.5 CoNEREL

CoNEREL is a Collective Recognition System, in batch mode, where it processes articles and comments in batch mode. It also uses comments and complex contexts shared. Basically it uses an article, its comments for recognition of the entities.

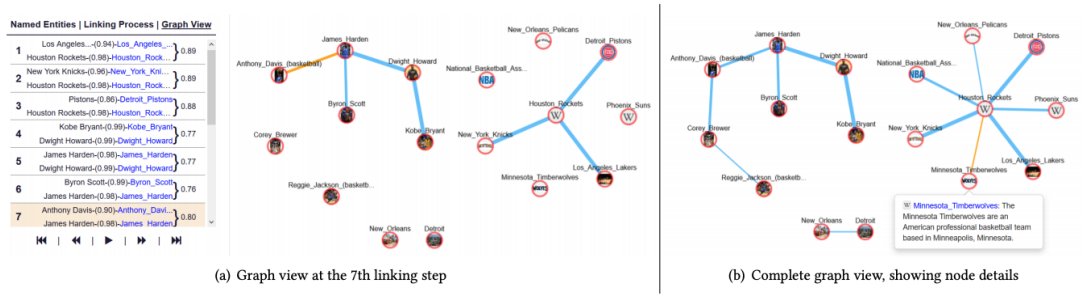
This systems uses co-reference of mentions to refine its class labels (e.g., person, location). Provides an interactive view of the linking process of pairs. Due to its implementation, it becomes fast and efficient in the study of the the text and comments. Figure 2.7 shows the basic architecture of CoNEREL.

(a) The main Web UI of CoNEREL.

(b) CoNEREL architecture.

Figure 2.7: Architecture CoNEREL by (Phan and Sun, 2018)

As an example in the figure 2.8 of these systems, a processing of 500 articles was used of news collected from Yahoo!



(a) Graph view at the 7th linking step

(b) Complete graph view, showing node details

Figure 2.8: Example CoNEREL from (Phan and Sun, 2018)

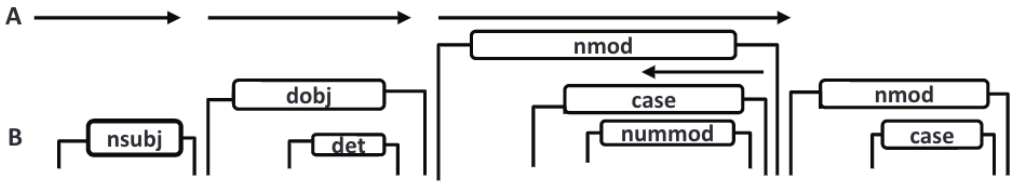
2.4.6 Triplex-ST

Triplex-ST Triplex-ST is an Extraction System aimed at extracting spatial-space information timing of texts. Triplex-ST has a supervised approach taking advantage of databases existing knowledge. Based on this procedure, models that capture facts from unpublished sentences, that is, we have an enrichment of information where there were no direct relations. Uses the YAGO knowledge base (Mahdisoltani et al., 2013) to create the models.

TRIPLEX and its TRIPLEX-ST extension involve an offline stage of data collection training instances (ie phrases that match triples), followed by the inference of extraction models of these cases. The models can then be used to extract new triplets of the text and these trebles are finally validated by a classifier.

TRIPLEX-ST extracts spatio-temporal information that involves dynamic or static information about entities and their properties. Therefore, it extends the general model of triples, considering the information related to the temporal and / or spatial context that qualify the facts expressed in triples, in the case of relationships that involve dynamic information and if this information is inserted in the text, validation is given for when and where they are define the triples.

Thus for an instant or period of time and / or for the region geospatial when and where they are valid. The evaluation of the TRIPLEX-ST was made by comparing the F1 between the TRIPLEX static model and the TRIPLEX-ST dynamic model where the dynamic showed better performance and still compared to OLLIE (Tablan et al., 2003), in static or dynamic facts. In figure 2.9 we have the example of TRIPLEX-ST as it is processed.



	Bhutan	had	a	population	of	770,000	People	in	2015
C	NOUN	VERB	DET	NOUN	ADP	NUM	NOUN	ADP	NUM
D	LOC	O	O	O	O	NUMBER	O	O	DATE
E	LOC	O	O	O	O	O	O	O	O
F	SUBJ	O	O	O	O	OBJ	O	O	O
G	O	O	O	O	O	O	O	O	DATE
H	O	O	O	O	O	O	O	O	O

Infobox name or property: Population — Infobox value : 770,000

Range of population: Integer number — Synset member: Bhutan

Synset member type: LOCATION — Spatial context: -

Temporal context: DATE — VerbNet Frame: NP V NP

A: Dependency path B: Dependency relations C: POS tags D: Named entities E: WordNet synsets LOC: LOCATION

F: Occurrence of subjects and objects G: Temporal context H: Spatial context O: No label NUM: Number

Figure 2.9: Example TRIPLE-ST from (Mirrezaei et al., 2016)

2.4.7 Sequence2Sequence

Sequence2Sequence is a general-purpose NLP tool that has proven to be effective for many tasks of text generation and sequence labeling. Seq2seq is based on the model deep neural language and inherits its accuracy in estimating local distributions of next word. The sequencer was based on the work of Daumé III and Marcu (Daumé III and Marcu, 2005). Figure 2.10 is an example of pipeline from Seq2Seq tool.

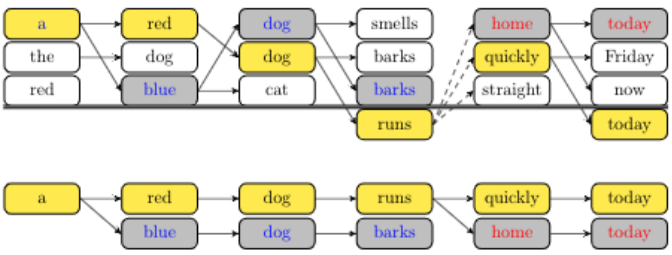


Figure 2.10: Example Seq2Seq of (Wiseman and Rush, 2016)

2.4.8 ReVerb

ReVerb based on TextRunnner (Yates et al., 2007), is an open extraction system with the possibility of reducing errors found in TextRunner, as it checks and validates the concept of holistic extractions, instead of word for word, potential phrases are filtered based on the statistics of a large corpus (the constraint implementation lexica).

ReVerb is "relationship first" instead of "arguments first place", which makes it possible to avoid a common mistake made by previous systems - confusing a noun in the relation phrase for an argument. The evaluation showed ReVerb higher than 30% in AUC than TextRunner, as shown in figure 2.11.

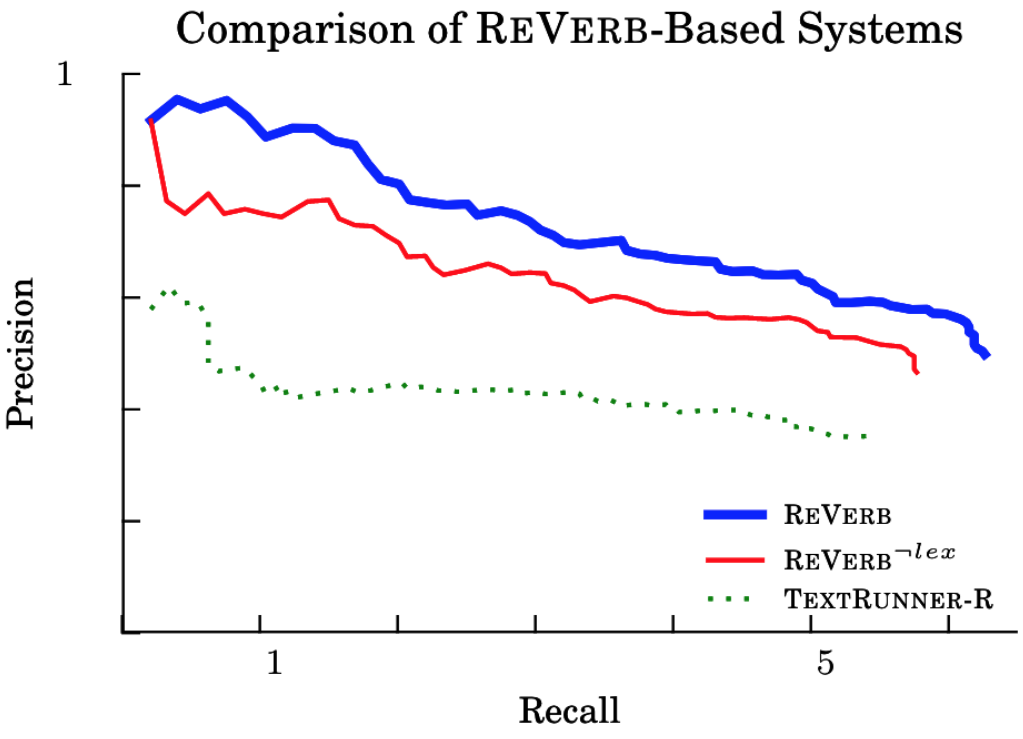


Figure 2.11: Results ReVerb from (Fader et al., 2011)

2.5 Recent developments for Portuguese

Below is a table that lists the main and most recent Information Systems Open training for texts in the Portuguese language. After table 2.3 we will make a brief contextualization of each of the Systems.

System	Reference	Year	Language	Technology
RelP	(Collovini et al., 2020)	2019	PT	Marking / Pre-processing, Probabilistic Model CRF
DptOIE	(Oliveira and Claro, 2019)	2018	PT	Rules
PragmaticOIE	(Sena and Claro, 2020)	2018	PT	Restrictions syntactic + inference+ context + intention
DependentIE	(de Oliveira et al., 2017)	2017	PT	Rules
InferReVerbPT	(Sena et al., 2017)	2017	PT	Restrictions syntactic, Classifier of Inference, Restrictions Transitivity and Symmetry
CRF-EN-pt	(Collovini et al., 2016)	2016	PT	Classifier CRF
RePort	(Victor Pereira, 2015)	2015	PT	Based Reverb /Rules for selecting the verbal relation and extracting the arguments
ArgOE	(Gamallo and Garcia, 2015)	2015	PT, EN, SP	Heuristic + Analyze Syntactic
DepOE	(Gamallo et al., 2012)	2012	PT, EN, SP, GA	Rules

Table 2.3: Table Portuguese

The systems are briefly described in the next subsections in chronological order

reverse.

2.5.1 RelP

RelP is a tool designed to try to extract any description of relationship explicitly between named entities in the **organization's domain**. The probabilistic model CRF - (Conditional Random Fields) is used to classify the relationship descriptor. It is tool is based on extracting the explicit relation that occurs between pairs of entities named in the figure of the triple t=(*arg1*,*rel*,*arg2*), where we seek the existence of the organization, Person or Location in the arguments and their relations.

There is a pre-processing with automatic text marking and NER. Classifies the correlation with the CRF Probabilistic model, considered the representation scheme and characteristics presented in Collovini's 2014 papers 2014(Collovini et al., 2014) and 2015(Collovini et al., 2015). This tool is geared towards the Portuguese language and the business and economic environment.

To work with this system we need to have a prior understanding of the environment environment and its context for "marking" and so despite this chosen environment, imagine the availability of applying this same technology in other environments provided that you have prior information on the environment in order to carry out a "marking and pre-processing ". Example in figure 2.12 below.

Configuration	Triples
(Config. 1) Context: NE Brasil	(Biblioteca_da_Real_Academia, seguir para, Brasil) (Serrambi, locação de automóvel em, Brasil) (Legião_da_Boa_Vontade, fundar em, Brasil) (Marfinite, abrir perspectiva em, Brasil) (FCI, em Brasil) (Creative_Commons, em, Brasil) (Brasil, manter sobre, Inglaterra)
(Config. 2) Context: NE Place Brasil	(Biblioteca_da_Real_Academia, seguir para, Brasil) (Serrambi, locação de automóvel em, Brasil) (Legião_da_Boa_Vontade, fundar em, Brasil) (Marfinite, abrir perspectiva em, Brasil) (FCI, em Brasil) (Creative_Commons, em, Brasil)
(Config. 3) Context: NE Person Santos_Ferreira	(Santos_Ferreira, saber de, Caixa) (Santos_Ferreira, ter sucesso em, BCP)
(Config. 4) Context: NE Organisation Legião_da_Boa_Vontade	(Legião_da_Boa_Vontade, implantação em, Portugal) (Legião_da_Boa_Vontade, fundar em, Brasil) (Legião_da_Boa_Vontade, em, Hora_da_Boa_Vontade) (Legião_da_Boa_Vontade, em, Rádio_Globo) (Legião_da_Boa_Vontade, fundar por, Alziro_Zarur)
(Config. 5) Context: Relation descriptor presidente de	(Rudy_Giuliani, presidente de, Câmera) (Almeida_Henriques, presidente de, Associação_do_Viseu) (Antônio_Nunes, presidente de, Autoridade_de_Segurança) (Fernando_Gomes, presidente de, Câmara_do_Porto) (Biblioteca_Nacional, presidente de, Pedro_Corrêa_do_Lago)

Figure 2.12: Tabela com exemplos do Sistema RelP, de (Collovini et al., 2020)

2.5.2 DeptOIE

DeptOIE is an OIE system or process that fundamentally is for Portuguese, as it is a language that, due to its different characteristics from English - which is more direct - there is a pre-processing that tokenizes the sentences. Uses a labeler POS and a dependency analysis. In this sense, the system also works with the triple t=(*arg1,rel,arg2*). There is a 3 part module that is prepared to work with special cases. In Figure 2.13, on page 17 shows the process flow.

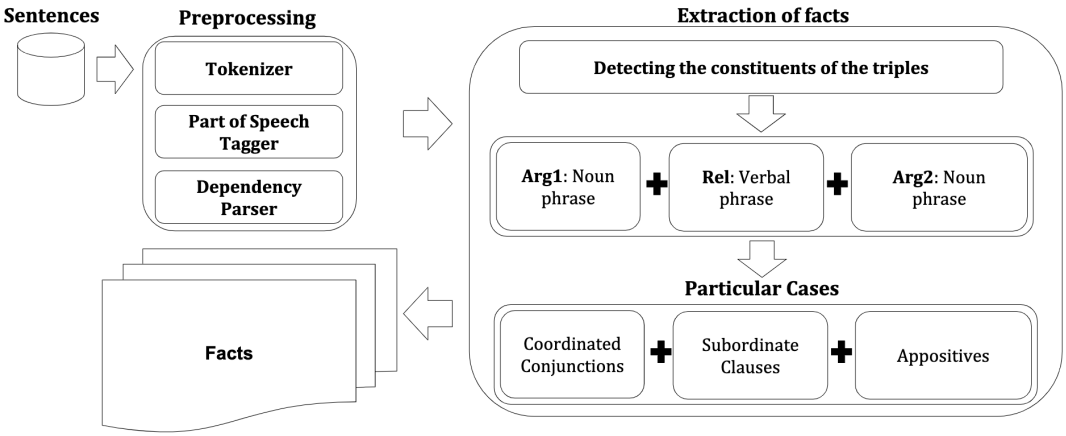


Figure 2.13: Process Flow of DeptOIE from (Oliveira and Claro, 2019)

2.5.3 PragmaticOIE

PragmaticOIE is a tool that tries a different aspect in the structure of the Open Information Extraction, this tool seeks this extraction based on the information intention, inference and context that the text tries to reveal. Even based on this new structuring, the concept on the triple t=(*arg1,rel,arg2*) continues to be used and evaluated comparatively. The intentional part was dealt with by evaluating implicit facts. For this fact, PragmaticOIE becomes a possibility of Extracting Intentional Information and this evaluation should require a more in-depth study as we will have to add not only the relationships, but the intentionality of the context. Thus the Environment as ”fourth” part. Figure 2.14 is a flow from PragmaticOIE.

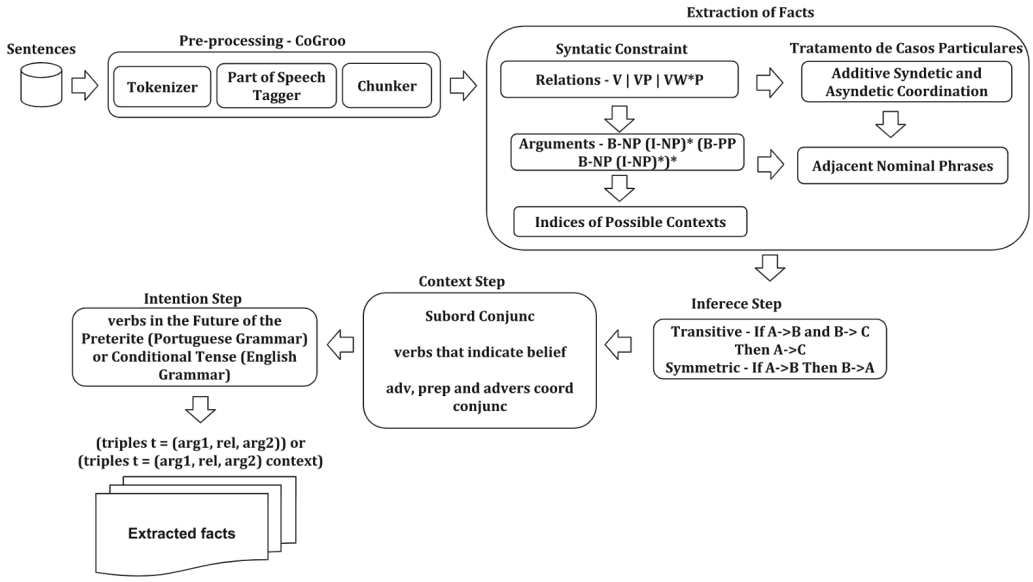


Figure 2.14: Flow PragmaticOIE by (Sena and Claro, 2018)

2.5.4 DependetIE

DependentIE system developed based on the triple t=(*arg1,rel,arg2*), for texts in the Portuguese language. It uses pre-processing with Tokenization and POS tag. The arguments are detected through sentence dependency searches. It is used parts of sentences after Tokenization.

The difference is that the rules for tokenization do not are fixed and neither is the creation of dependencies. As the author needs to improve the Precision and Recall. Figure 2.15 shows a pipeline from DependentIE.

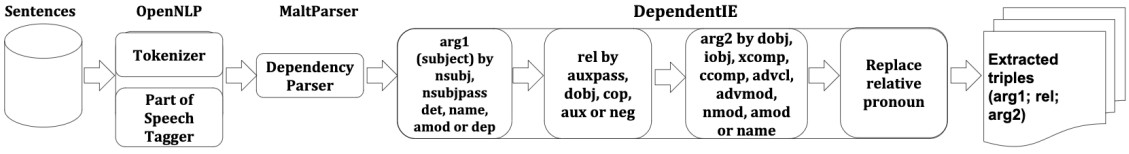


Figure 2.15: Pipeline DependetIE by (de Oliveira et al., 2017)

2.5.5 InferReVerbPt

InferReVerbPt this method was idealized for texts in the Portuguese language, for the inference approach. The issues of transitivity and symmetry are of interest for the creation of this method, which was divided into 4 parts:

- syntactic constraint
- inference classifier
- transitivity constraint
- symmetry constraint

Pre-processing was used which takes into account the triples t=(*arg1,rel,arg2*), to use of the model. Figure 2.16 with the flow.

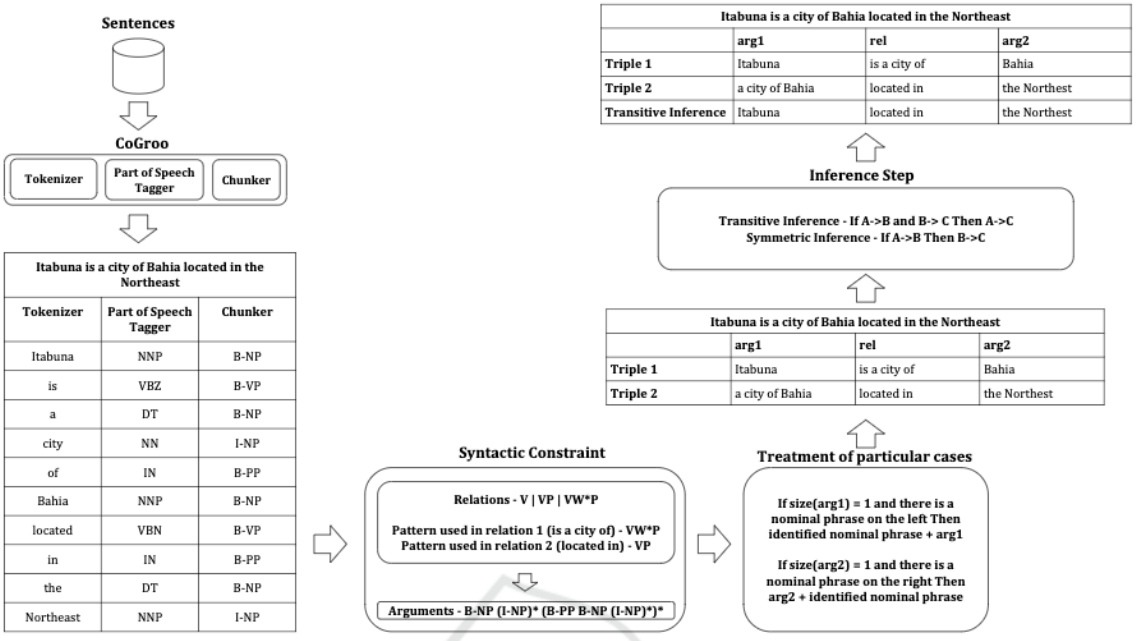


Figure 2.16: Flow InferReVerPt from (Sena et al., 2017)

2.5.6 CRF-EN-pt

CRF-EN-pt this model was applied to extract categories belonging to People, Location and Organization. The CRF metric was used in the structuring of relations between entities seeking to express explicit relations.

The CRF Classifier was considered for the exact and partial match of the data set. The organization of the triple t=(*arg1,rel,arg2*) was adequate as a subject, predicate and object, it was performed a POS tag in the data set that was not very large. Example results in figure 2.17:

Relation instance (reference)	Exact matching	Partial matching
Na *Biblioteca Nacional*, o **presidente da** instituição, *Pedro Corrêa do Lago* (...)	presidente<I-REL> de<I-REL>	presidente<I-REL> de<O>
(In *Biblioteca Nacional*, the **president of** the institution, *Pedro Corrêa do Lago* (...))	president<I-REL> of<I-REL>	president<I-REL> of<O>

Figure 2.17: Example CRF-EN-pt by (Collovini et al., 2016)

2.5.7 RePort

RePort is an Open Information Extraction model developed and adapted for the Portuguese Language based on ReVerb that was made for English Language. This model has them as the search for the Confidence metric of the triples t=(*arg1,rel,arg2*). There is a sentence detector, tokenization, expression identification, a POS tag thus classifying tokens. Applications of syntactic rules and identification of Nominal and sentence phrases.

The result of this model is similar to ReVerb, but it was evaluated with a small number of sample. Following figure 2.18 is the basic RePort process:

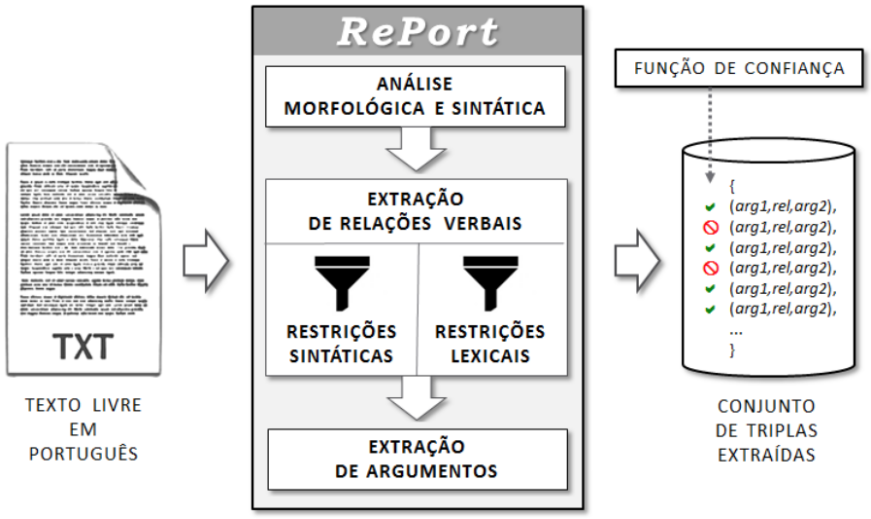


Figure 2.18: Flow RePort by (Victor Pereira, 2015)

2.5.8 ArgOE

ArgOE this Open Information Extraction system, according to its author (Gamallo and Garcia, 2015) is based on heuristics using syntactic analysis as base of work in the definition of the structure of the relations within the triples, t=(*arg1,rel,arg2*). The system seeks the broad structure of the arguments. The analysis includes: subject objects, attributes, locations, instruments, modes, etc. No distinction between arguments and adjuncts. The method is characterized by two stages: detection of arguments and generation of triples. The difference is that this system was applied to different languages: English, Spanish and Portuguese. With triples of different granularities and multilingual analysis. In the evaluation of this system, according to the author "be overcome by other methods similar rules based, it achieves better results than those strategies based on training data "(Gamallo and Garcia, 2015). What it differs from was the first to have worked in more than one language. In figure 2.19 below, we have the comparative data for ArgOE.

Systems	correct extractions	total extractions
textrunner	286	798
reverb	388	727
woe	447	1028
ollie	547	1242
argoe	582	1162
clausie	1706	2975

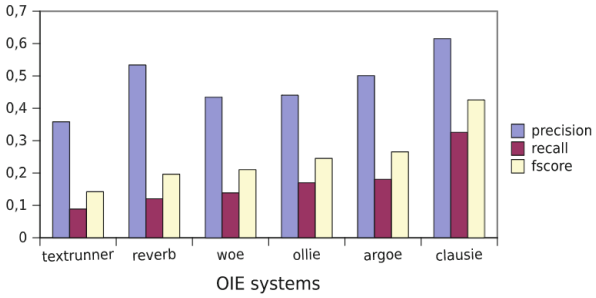


Figure 2.19: Comparative Results ArgOE from (Gamallo and Garcia, 2015)

2.5.9 DepOE

DepOE is an Open Information Extraction system consisting of three steps: Dependency Analysis, Structure Rules and Extraction of trebles,t=(*arg1,rel,arg2*). Binary relationships are not dealt with, but deeper dependency information is facilitating the construction of the relations between the arguments. This procedure helps to find relations that are not expressed by verbs.

This system tends to be multilingual and has been applied to English, Spanish, Galician and Portuguese. The method is based on deep syntactic information, such as dependency. It was possible to perform the extraction of open information, such as dependence based on rules and standards-based extraction rules, maintaining scalability. Figure 2.20 show the example.

patterns	**triples**
subj-vp-dobj	Arg1 = subj Rel= vp Arg2 = dobj
subj-vp-vprep	Arg1 = subj Rel= vp+prep (prep from vprep) Arg2 = np (from vprep)
subj-vp-dobj-vprep	Arg1 = subj Rel= vp+dobj+prep Arg2 = np (from vprep)
subj-vp-attr	Arg1 = subj Rel= vp Arg2 = attr
subj-vp-attr-vprep	Arg1 = subj Rel= vp+attr+prep (from vprep) Arg2 = np (from vprep)

Figure 2.20: Structure DepOE by (Gamallo et al., 2012)

2.6 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.

2.7 Tools

In the table 2.4 we have a some new,base and relations of tools from OpenIE and NER.

Table 2.4: 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

2.8 Considerations

By having a view of some of the Open Information Systems and their process of development, it is clear that we have a lot to do both in a global context and specific to the Portuguese language. The challenges that we believe can make sense we list below:

- Develop methodology for Portuguese Language based on Machine Learning and less in Rules.
- Check new extraction processes such as Seq2Seq and redirect them to Portuguese Language.
- Work in a relational way, that is, with graphs to assess whether it makes sense.
- Create a system that can be "plugged" into a random text, which extracts information not directly related and easy to handle and understand.

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

We find and use an Engineering Design Process (Plan and Khandani, 2005) to develop our path, principal steps and design. We will work with Engineering Design Process for their interactivity in solving problems, you can find at any time in the process that you need to go back to a previous step, so we will have a virtuous cycle for the solution of the "problem" in question. Below the figure 3.1 that exemplifies.

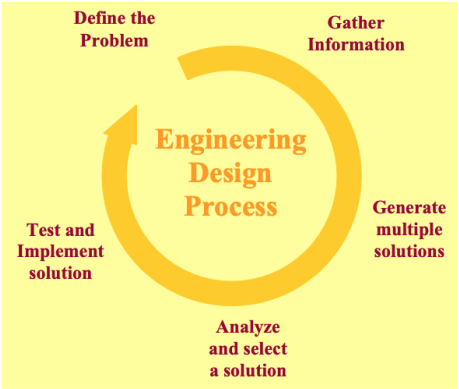


Figure 3.1: Continuous iterative process from Plan and Khandani (2005)

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 3.2 is a basic CAIT baseline

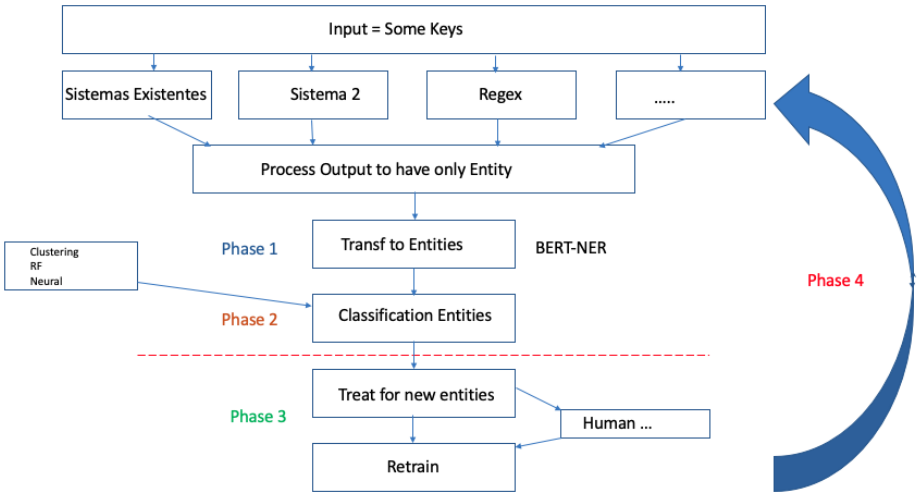


Figure 3.2: CAIT pipeline

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.

Chapter 4

Workplan

4.1 Research Timeline

In the figure 4.1 we have the proposal timetable.

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

Figure 4.1: Timetable

4.2 Milestones

Each Milestone refers to a delivery, product or report as follows:

- M1 - which is the 1st milestone is like the delivery of the “NER without data

registration", this delivery will take place at the end of the 1st year, that is, until 20/21 Sep.

- M2 - which is the second delivery will be the delivery of the CAIT App in a simpler version around 4Q on 21/22.

- M3 - Will already be the delivery of CAIT in a version with Exploration of other relationships within the domain in 2Q to 22/23.

- M4 - must be the delivery of the document with the details of how the App was developed, the technology shipped and the results obtained the delivery will be in September 2023.

4.3 Publication plan

For publications we will be directing efforts based on the availability of Conferences, international magazines, national magazines (Portugal and Brazil) and associations in the machine learning area.

In the table 4.1, a list of possible places of publication.

In figure 4.2 we have a random cloud based in bibliography references (place publications)

Figure 4.2: Publications

Table 4.1: Publication possibilities

Publication / Place
International Conference on Computational Processing of the Portuguese Language
International Conference on World Wide Web
International ACM SIGIR Conference on Research
Development in Information Retrieval
Ibero-American Conference on Artificial Intelligence
Conference on Empirical Methods in Natural Language Processing
The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)
Symposium in Information and Human Language Technology - Sociedade Brasileira de Computação
International Conference on Compute and Data Analysis
Symposium on Languages, Applications and Technologies
Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CONLL)
Association for Computational Linguistics (ACL) System Demonstrations
Communications of the ACM
Journal of Information Science
Encontro Nacional de Inteligência Artificial e Computacional
Springer
Association for Computational Linguistics
Association for Computing Machinery
Association for Computational Linguistics
PROPOR
Linguamatica

4.4 Initial work

The first six months of this year, from February to September 2020, were used to conduct a supervised study on the timing of NLP in the current context of technology. In parallel, work was carried out with some OpenIE and NER projects and systems, with a focus on the most recent systems and close to the state of the art, if possible in Portuguese, to assess the opportunities arising from a critical view of support and possible relationship between technologies and gains aimed at the democratization of knowledge.

4.4.1 Initial review of NER and OpenIE

A study and review was carried out with some scientists who wrote about OpenIE and NER, this work was supervised by professors Prf. Phd. Antonio Teixeira

and Prof. Phd. Mario Rodrigues. This oriented study was done to seek a better foundation and digitalization in the Natural Language Processing environment. This is a study that never ends, every day we need to research new technologies and new processes in this field, as it still requires a lot of work and attention.

We have some numbers to indicate the volume of work done, we read about 40 titles, we evaluate something around 20 different developments, which are listed and in some detail in Section 2.4 and Section 2.5 of this document.

4.4.2 First selection of NER and OpenIE tools

In section 2.7 we list the tools that can be useful in verifying, evaluating and exemplifying the construction of the CAIT knowledge extraction engine (name given to the product that will be the interface of our development and the user). We tried to evaluate 3 tools, AlleNLP, DeptOIE and Linguakit, due to the circumstances we went through, we only implemented and guided the processing of the tests. Other tools must be evaluated with these 3, in order to broaden the horizons of investigation. In total, we understand that we have about 20 tools that we can evaluate.

4.4.3 Dataset

In section 2.6 we detail some data sets that we have found that can be useful in the conceptual evaluation of the model that we will build. In addition, we will have to search for a new data set and / or build one for our work.

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.

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.

Cabral, L. d. S. (2009). Extração de informação usando integração de componentes de pln através do framework gate. Dissertação de mestrado.

Collovini, S., de Bairros P. Filho, M., and Vieira, R. (2015). Analysing the role of representation choices in portuguese relation extraction. In Mothe, J., Savoy, J., Kamps, J., Pinel-Sauvagnat, K., Jones, G., San Juan, E., Capellato, L., and Ferro, N., editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 105–116, Cham. Springer International Publishing.

Collovini, S., Gonçalves, P. N., Cavalheiro, G., Santos, J., and Vieira, R. (2020). Relation extraction for competitive intelligence. In *International Conference on Computational Processing of the Portuguese Language*, pages 249–258. Springer.

Collovini, S., Machado, G., and Vieira, R. (2016). Extracting and structuring open relations from portuguese text. In *International Conference on Computational Processing of the Portuguese Language*, pages 153–164. Springer.

Collovini, S., Pugens, L., Vanin, A. A., and Vieira, R. (2014). Extraction of relation descriptors for portuguese using conditional random fields. In *Ibero-American Conference on Artificial Intelligence*, pages 108–119. Springer.

Daumé III, H. and Marcu, D. (2005). Learning as search optimization: Approximate large margin methods for structured prediction. In *Proceedings of the 22nd international conference on Machine learning*, pages 169–176.

de Oliveira, L. S., Glauber, R., and Claro, D. B. (2017). DependentIE: An open information extraction system on portuguese by a dependence analysis. *Encontro Nacional de Inteligência Artificial e Computacional*.

Del Corro, L. and Gemulla, R. (2013). Clausie: Clause-based open information extraction. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, page 355–366, New York, NY, USA. Association for Computing Machinery.

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*.

Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., and Yates, A. (2004a). Web-scale information extraction in knowitall: (preliminary results). In *Proceedings of the 13th international conference on World Wide Web*, pages 100–110.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., and Yates, A. (2004b). Methods for domain-independent information extraction from the web: An experimental comparison. In *AAAI*, pages 391–398.

Fader, A., Soderland, S., and Etzioni, O. (2011). Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, page 1535–1545, USA. Association for Computational Linguistics.

Gamallo, P. (2014). An Overview of Open Information Extraction (Invited talk). In Pereira, M. J. V., Leal, J. P., and Alberto Simões, publisher = Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, a. . D. U. . h. U. . u. d. . O. a. . K., editors, *3rd Symposium on Languages, Applications and Technologies*, volume 38 of *OpenAccess Series in Informatics (OASIcs)*, pages 13–16.

Gamallo, P. and Garcia, M. (2015). Multilingual open information extraction. *EPIA 2015. LNCS (LNAI)*, 9273(711-722):22.

Gamallo, P., Garcia, M., and Fernández-Lanza, S. (2012). Dependency-based open information extraction. In *Proceedings of the joint workshop on unsupervised and semi-supervised learning in NLP*, pages 10–18. Association for Computational Linguistics.

Gashteovski, K., Gemulla, R., and Corro, L. d. (2017). Minie: minimizing facts in open information extraction. Association for Computational Linguistics.

Katz, J. J. and Fodor, J. A. (1963). The structure of a semantic theory. *language*, 39(2):170–210.

Lauscher, A., Song, Y., and Gashteovski, K. (2019). Minscie: Citation-centered open information extraction. In *Proceedings of the 18th Joint Conference on Digital Libraries*, JCDL '19, page 386–387. IEEE Press.

Levin, E., Pieraccini, R., and Eckert, W. (1998). Using markov decision process for learning dialogue strategies. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, volume 1, pages 201–204 vol.1.

Li, Q., Jiang, M., Zhang, X., Qu, M., Hanratty, T. P., Gao, J., and Han, J. (2018). Truepie: Discovering reliable patterns in pattern-based information extraction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, KDD '18, page 1675–1684, New York, NY, USA. Association for Computing Machinery.

Liu, G., Li, X., Wang, J., Sun, M., and Li, P. (2020). Extracting knowledge from web text with monte carlo tree search. In *Proceedings of The Web Conference 2020*, WWW '20, page 2585–2591, New York, NY, USA. Association for Computing Machinery.

Mahdisoltani, F., Biega, J., and Suchanek, F. M. (2013). YAGO3: A Knowledge Base from Multilingual Wikipedias. In *CIDR*, Asilomar, United States.

Mirrezaei, S. I., Martins, B., and Cruz, I. F. (2016). A distantly supervised method for extracting spatio-temporal information from text. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPACIAL '16, New York, NY, USA. Association for Computing Machinery.

Oliveira, L. d. and Claro, D. B. (2019). DptOIE: a portuguese open information extraction system based on dependency analysis.

Phan, M. C. and Sun, A. (2018). Conerel: Collective information extraction in news articles. In *The 41st International ACM SIGIR Conference on Research Development in Information Retrieval*, SIGIR '18, page 1273–1276, New York, NY, USA. Association for Computing Machinery.

Plan, E. and Khandani, S. (2005). *Engineering design process*. dphu.org. 504 cites: https://scholar.google.com/scholar?cites=15579240992028560648&as$_s dt$ $=$ 2005&$sciodt = 0,5$&$hl = en$.

Rodríguez, J. M., Merlino, H. D., and Pesado, P. (2020). Atp-oie: An autonomous open information extraction method. In *Proceedings of the 2020 the 4th International Conference on Compute and Data Analysis*, ICCDA 2020, page 197–202, New York, NY, USA. Association for Computing Machinery.

Sena, C. F. L. and Claro, D. B. (2018). Pragmatic information extraction in brazilian portuguese documents. In *International Conference on Computational Processing of the Portuguese Language*, pages 46–56. Springer.

Sena, C. F. L. and Claro, D. B. (2020). PragmaticOIE: a pragmatic open information extraction for portuguese language. *Knowledge and Information Systems*, pages 1–26.

Sena, C. F. L., Glauber, R., and Claro, D. B. (2017). Inference approach to enhance a portuguese open information extraction. In *ICEIS (1)*, pages 442–451.

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

Tablan, V., Bontcheva, K., Maynard, D., and Cunningham, H. (2003). Ollie: on-line learning for information extraction. In *Proceedings of the HLT-NAACL 2003 workshop on Software engineering and architecture of language technology systems-Volume 8*, pages 17–24. Association for Computational Linguistics.

Victor Pereira, V. P. (2015). RePort - um sistema de extração de informações aberta para língua portuguesa. In *Proceedings of Symposium in Information and Human Language Technology - Sociedade Brasileira de Computação*, pages 191–200.

Weizenbaum, J. (1966). ELIZA — a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Wilks, Y. A. (1972). *Grammar, meaning and the machine analysis of language*. Routledge & Kegan Paul London.

Wiseman, S. and Rush, A. M. (2016). Sequence-to-sequence learning as beam-search optimization. *arXiv preprint arXiv:1606.02960*.

Yates, A., Banko, M., Broadhead, M., Cafarella, M. J., Etzioni, O., and Soderland, S. (2007). Textrunner: open information extraction on the web. In *Proceedings of*

Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), pages 25–26.