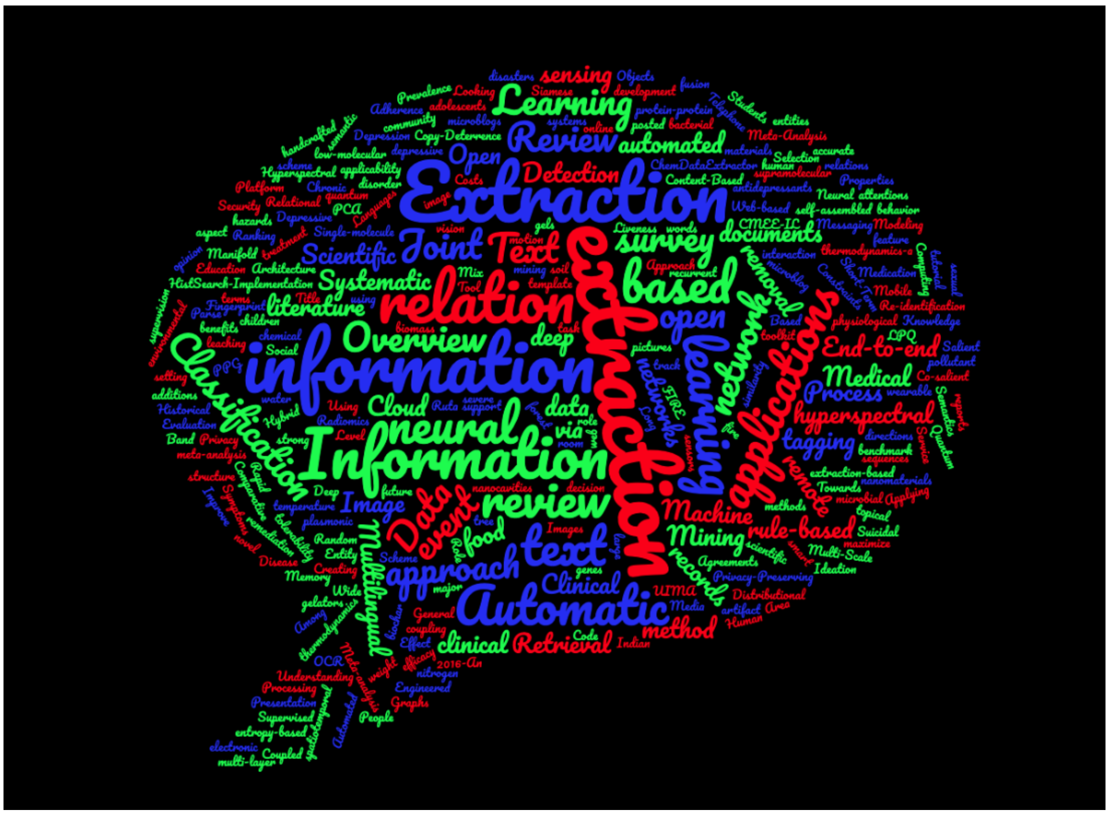
A Critical Analysis of the State-of-the-Art in Information Extraction

Emanuel Matos - Student PhD Informatics

2020-09-22



Supervisor:

- PhD António Teixeira - Universidade de Aveiro
- PhD Mario Rodrigues - Universidade de Aveiro

Contents

List of Figures

List of Tables

Abstract

"Information Extraction (IE) is a research area that can be used to automate the extraction of useful information from textual documents (ABREU; BONAMIGO; VIEIRA, 2013), through the extraction of Semantic Relationships (ER), that can be found in the text"(Oliveira, 2018).

"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 work was designed to evaluate in a brief, critical and historical path some of the main open information systems and their possible consequences. In this analysis of Information Extraction Systems we will have as chronology and the its evolution mapped from the 70s until the beginning of 2020, mainly in the English Language.

The main and most recent Open Language Information Systems in Portuguese language will have its own analysis. We have the challenge to obtain the Systems developed and suitable for treatment in the Portuguese Language.

1 Introduction

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".

2 Brief history of 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 1 which lists some of the historical systems and processes.

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

Table 1: History

2.1 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 1 of the ELISA concept.

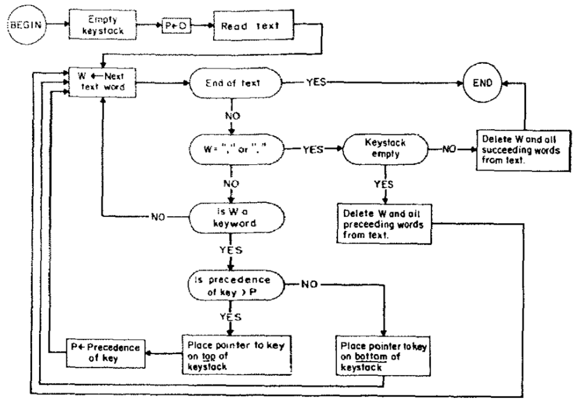


Figure 1: Keyboard Detection Flow / ELISA by (Weizenbaum, 1966)

2.2 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 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 shows the flow of the OLLIE process.

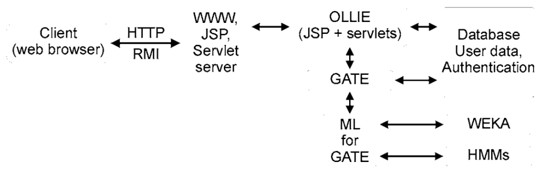


Figure 2: Architecture / OLLIE by (Tablan et al., 2003)

2.4 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.

2.5 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.

3 Recent Developments

Below we have in table 2 which lists some of the most recent Information Systems Open with several technologies.

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

Table 2: Recent Developments

3.1 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 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 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.

3.2 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 4 we have the MCTS Framework.

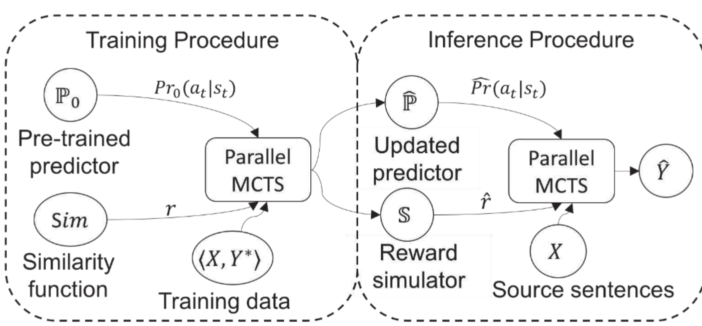


Figure 4: Framework MCTS for (Liu et al., 2020)

3.3 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 5 of the MinScIE Pipeline.

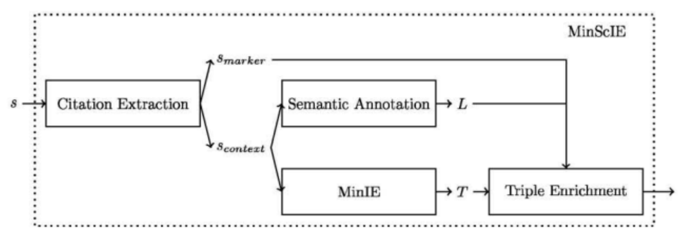


Figure 5: MinScIE pipeline from (Lauscher et al., 2019)

3.4 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 6 the TruePIE Framework.

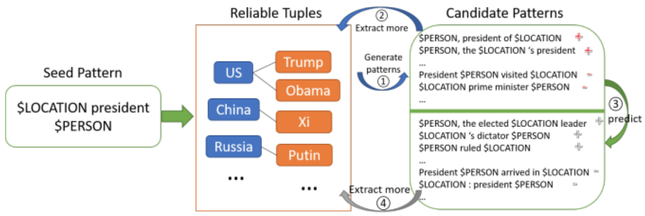


Figure 6: Framework TruePIE from (Li et al., 2018)

3.5 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 7 shows the basic architecture of CoNEREL.

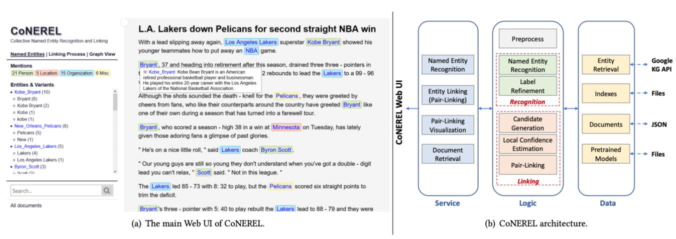
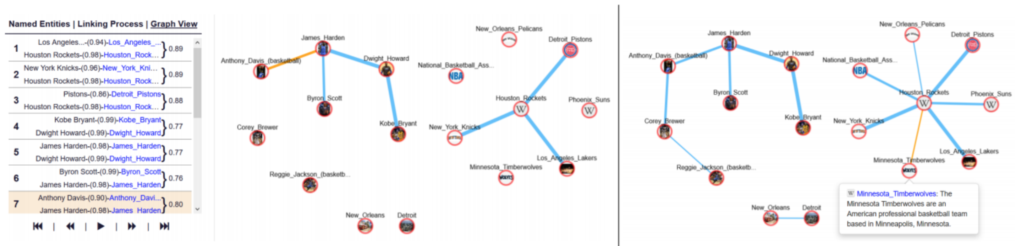


Figure 7: Architecture CoNEREL by (Phan and Sun, 2018)

As an example in the figure 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 8: Example CoNEREL from (Phan and Sun, 2018)

3.6 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 9 we have the example of TRIPLEX-ST as it is processed.

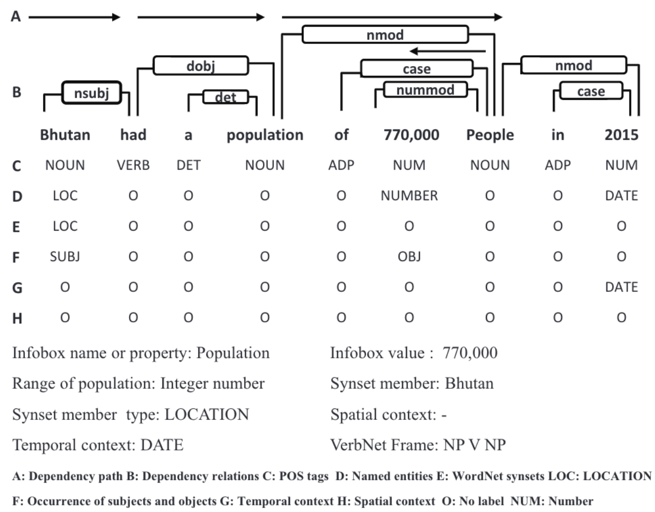


Figure 9: Example TRIPLE-ST from (Mirrezaei et al., 2016)

3.7 Sequence2Sequence

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 10 is an example of pipeline from Seq2Seq tool.

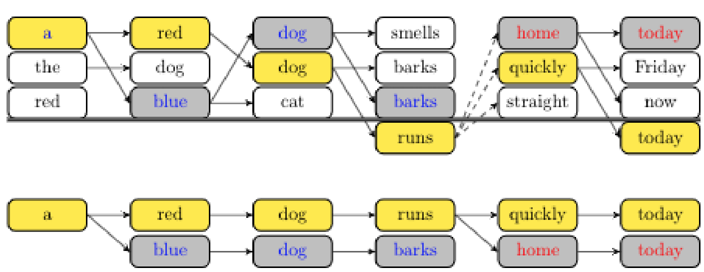


Figure 10: Example Seq2Seq of (Wiseman and Rush, 2016)

3.8 ReVerb

ReVerb based on TextRunner (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 11.

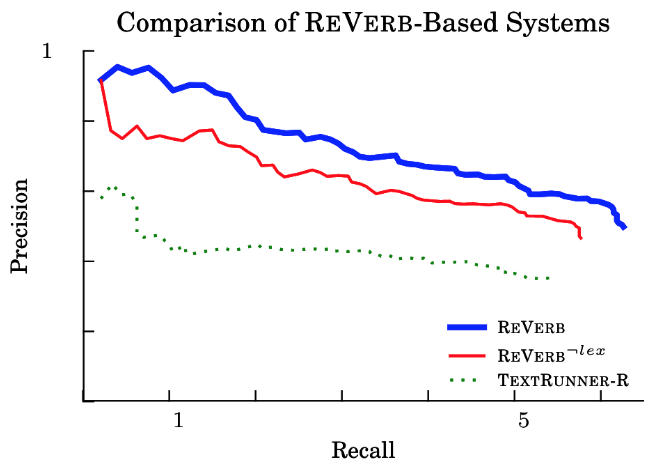


Figure 11: Results ReVerb from (Fader et al., 2011)

4 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 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 3: Table Portuguese

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

4.1 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 preprocessing ". Example in figure 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 12: Tabela com exemplos do Sistema RelP, de (Collovini et al., 2020)

4.2 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 13, on page 18 shows the process flow.

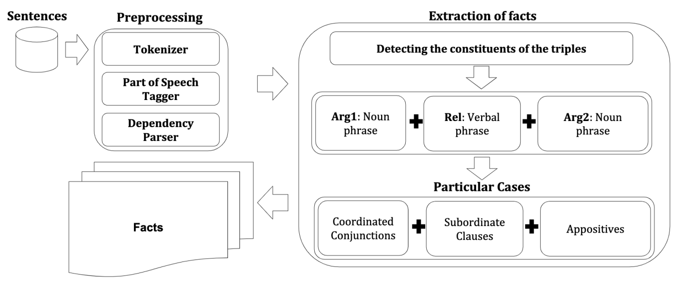


Figure 13: Process Flow of DeptOIE from (Oliveira and Claro, 2019)

4.3 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 14 is a flow from PragmaticOIE

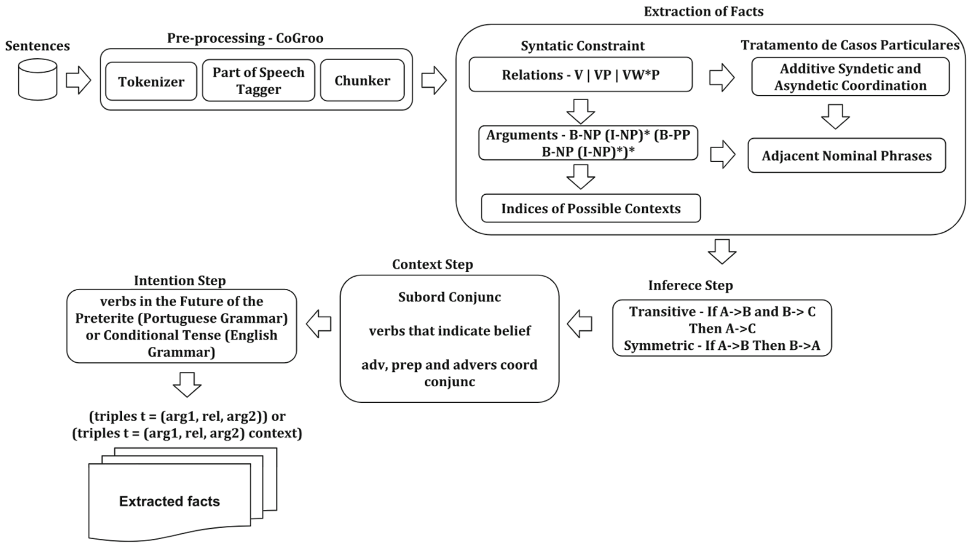


Figure 14: Flow PragmaticOIE by (Sena and Claro, 2018)

4.4 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 15 shows a pipeline from DependentIE.

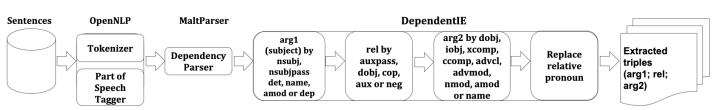


Figure 15: Pipeline DependetIE by (de Oliveira et al., 2017)

4.5 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 16 with the flow.

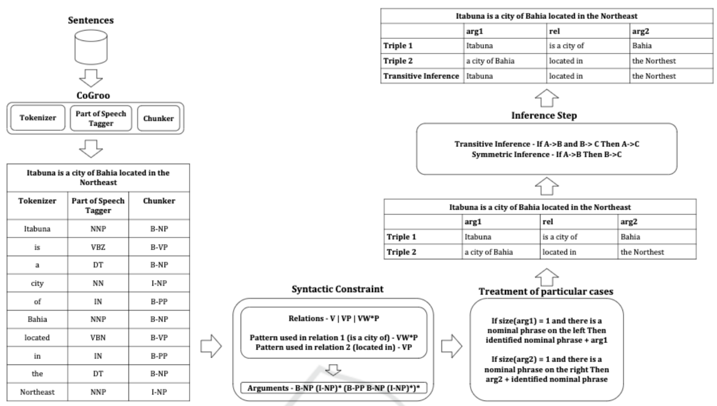


Figure 16: Flow InferReVerPt from (Sena et al., 2017)

4.6 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 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 17: Example CRF-EN-pt by (Collovini et al., 2016)

4.7 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 18 is the basic RePort process:

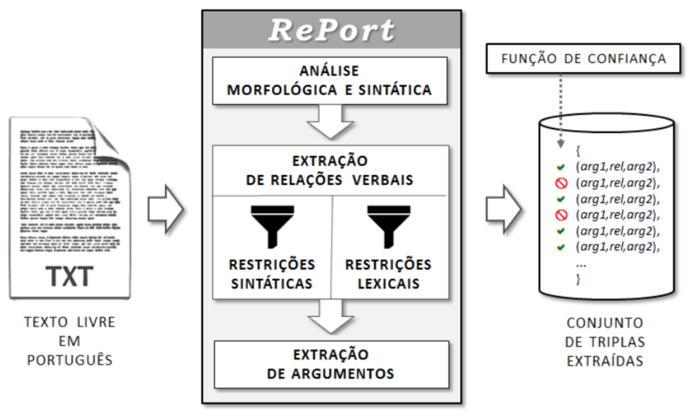


Figure 18: Flow RePort by (Victor Pereira, 2015)

4.8 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 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

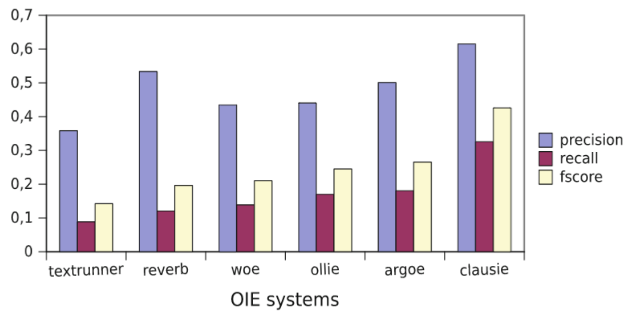


Figure 19: Comparative Results ArgOE from (Gamallo and Garcia, 2015)

4.9 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 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 20: Structure DepOE by (Gamallo et al., 2012)

5 Critical Analysis

The Open Information Systems briefly evaluated, tend in an evolutionary way to resolve some errors of the previous models, so we have within the historical concept a development as an ascendant in the structuring and reduction of errors of the Systems.

When looking for concepts or processes that have different strategies, we can endorse 3 systems:

- TextRunner(Yates et al., 2007), which was the basis for several other developments.
- CoNEREL(Phan and Sun, 2018) hat deals with relationships using concepts graphs.
- TRIPLEX-ST(Mirrezaei et al., 2016) that takes into account the temporal condition.

The Seq2Seq(Daumé III and Marcu, 2005) also called me to attention because it introduces an idea of extraction based on the holistic view, not the word word.

5.1 Limitations

Some more general limitations can be pointed out:

- Interrelate different types of objects, figures, texts, etc ...
- Create hybrid models, such as Random Forest and Neural Networks.
- Apply on a large scale.

5.2 Limitations for PT

Explaining the Systems that were conceived or adapted for the Portuguese Language, we can have besides the generic limitations others:

- Small volume of examples, much tested on the same things.
- Reduced applicability.
- a lot of POS tag based.
- has not been widely applied to "real" applications.

6 Major Challenges

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.

7 Conclusion

In this short time of research and creation of the report, I had the opportunity to verified a part of Open Information Extraction Systems, its history, its technical development as well as triple or sequential extraction processes. So, in this learning exercise, I was sure that there is a field too big to work on and look for new concepts that help people in making of decision or better understanding of a problem that may be local or of global proportion. The next steps that I believe to be of added value should be based on the processes the latest and most recent developments in Open Information Systems.

With this, you will see the prototyping, testing, complete revisiting of the model and its application in the environment original and in an uncontrolled environment. Based on this replication we will be able to evaluate and direct our future research as well as split all models presented between Extraction Tools and Systems.

The bibliographic references presented refer to each paper that was used in the construction of this report, the deal was to bring the state of the art to each Open Information Extraction presented.

"I always thought something was fundamentally wrong with the universe" (Adams, 1995)

References

D. Adams. *The Hitchhiker's Guide to the Galaxy*. San Val, 1995. ISBN 9781417642595. URL `http://books.google.com/books?id=W-xMPgAACAAJ`.

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

L. d. S. Cabral. Extração de informação usando integração de componentes de pln através do framework gate. Dissertação de mestrado, 7 2009.

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

S. Collovini, M. de Bairros P. Filho, and R. Vieira. Analysing the role of representation choices in portuguese relation extraction. In J. Mothe, J. Savoy, J. Kamps, K. Pinel-Sauvagnat, G. Jones, E. San Juan, L. Capellato, and N. Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 105–116, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24027-5.

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

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

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

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

L. Del Corro and R. Gemulla. 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, 2013. Association for Computing Machinery. ISBN 9781450320351. doi: 10.1145/2488388.2488420. URL `https://doi.org/10.1145/2488388.2488420`.

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

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

A. Fader, S. Soderland, and O. Etzioni. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, page 1535–1545, USA, 2011. Association for Computational Linguistics. ISBN 9781937284114.

P. Gamallo. An Overview of Open Information Extraction (Invited talk). In M. J. V. Pereira, J. P. Leal, and a. . D. U. . h. U. . u. d. . O. a. . K. Alberto Simões, publisher = Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, editors, *3rd Symposium on Languages, Applications and Technologies*, volume 38 of *OpenAccess Series in Informatics (OASIcs)*, pages 13–16, 2014. ISBN 978-3-939897-68-2.

P. Gamallo and M. Garcia. Multilingual open information extraction. *EPIA 2015. LNCS (LNAI)*, 9273(711-722):22, 2015. URL `https://doi.org/10.1007/978-3-319-23485-472`.

P. Gamallo, M. Garcia, and S. Fernández-Lanza. 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, 2012.

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

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

A. Lauscher, Y. Song, and K. Gashteovski. Minscie: Citation-centered open information extraction. In *Proceedings of the 18th Joint Conference on Digital Libraries*, JCDL '19, page 386–387. IEEE Press, 2019. ISBN 9781728115474. doi: 10.1109/JCDL.2019.00083. URL `https://doi.org/10.1109/JCDL.2019.00083`.

E. Levin, R. Pieraccini, and W. Eckert. 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, 1998.

Q. Li, M. Jiang, X. Zhang, M. Qu, T. P. Hanratty, J. Gao, and J. Han. 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, 2018. Association for Computing Machinery. ISBN 9781450355520. doi: 10.1145/3219819.3220017. URL `https://doi.org/10.1145/3219819.3220017`.

G. Liu, X. Li, J. Wang, M. Sun, and P. Li. 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, 2020. Association for Computing Machinery. ISBN 9781450370233. doi: 10.1145/3366423.3380010. URL https://doi.org/10.1145/3366423.3380010.

F. Mahdisoltani, J. Biega, and F. M. Suchanek. YAGO3: A Knowledge Base from Multilingual Wikipedias. In *CIDR*, Asilomar, United States, Jan. 2013. URL https://hal-imt.archives-ouvertes.fr/hal-01699874.

S. I. Mirrezaei, B. Martins, and I. F. Cruz. 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, 2016. Association for Computing Machinery. ISBN 9781450345897. doi: 10.1145/2996913.2996967. URL https://doi.org/10.1145/2996913.2996967.

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

L. S. Oliveira. DptOIE: um sistema para extração de informação aberta na língua portuguesa baseado em análise de dependˆencia. Dissertação de mestrado, 7 2018.

M. C. Phan and A. Sun. 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, 2018. Association for Computing Machinery. ISBN 9781450356572. doi: 10.1145/3209978.3210165. URL https://doi.org/10.1145/3209978.3210165.

J. M. Rodríguez, H. D. Merlino, and P. Pesado. 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, 2020. Association for Computing Machinery. ISBN 9781450376440. doi: 10.1145/3388142.3388166. URL https://doi.org/10.1145/3388142.3388166.

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

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

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

V. Tablan, K. Bontcheva, D. Maynard, and H. Cunningham. 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, 2003.

V. P. Victor Pereira. 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, November 2015.

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

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

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

A. Yates, M. Banko, M. Broadhead, M. J. Cafarella, O. Etzioni, and S. Soderland. 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, 2007.