DependentIE: An Open Information Extraction system on Portuguese by a Dependence Analysis

DependentIE: An Open Information Extraction system on Portuguese by a Dependence Analysis

Leandro Souza de Oliveira, Rafael Glauber, Daniela Barreiro Claro

[1]Formalisms and Semantic Applications Research Group (FORMAS)
LASiD/DCC/IME – Federal University of Bahia (UFBA)
Av. Adhemar de Barros, s/n, Ondina, Salvador, Bahia, Brazil

leo.053993@gmail.com, rglauber@dcc.ufba.br, dclaro@ufba.br

Abstract. *The amount of textual data available on the Web grows every day. For humans, it is hard to manual extract useful information from such quantity of data. Traditional approaches extract information in a limited manner within a set of pre-determined features. Open information extraction (Open IE) extracts facts from plain texts without establishing the type of a relationship previously. In this work, we describe our DependentIE, an Open IE system for texts written in Portuguese that uses dependency analysis to identify clauses in a sentence. Our approach differs from the other methods such as ClausIE and ArgOE because we do not use handcrafted rules to extract the facts. Our results confirm that our approach is more precise than the current state of the art for Portuguese texts.*

1. Introduction

Nowadays the Web publishes a lot of data as texts written in a natural language format. The number of those texts are growing every day. Nevertheless, it is hard for humans to manual extract useful information from such a large data repository. Information extraction (IE) is a research area that automates the extraction of facts from textual documents [de Abreu et al. 2013]. According to Fader *et al.* [Fader et al. 2011] and Xavier *et al.* [Xavier et al. 2015], traditional IE is based on training an extractor with a target relationships previously defined. The main drawback of traditional IE approaches is their low coverage and their overfitting when applicable in a particular domain. Tackling with different domains a human intervention might be required. To overcome this problem, Open Information Extraction (Open IE) has emerged to extract facts without determining the set of relationships previously. Open IE extracts facts from sentences in triple forms, such as:

$$t = (arg1, rel, arg2) \tag{1}$$

where *arg1* and *arg2* are nominal phrases in a sentence and *rel* establishes a relationship between *arg1* and *arg2* through a verbal phrase. Estimates[1] show that 155 million of users on the Internet have Portuguese as their mother language. Considering the websites content, only 2.5% are written in Portuguese[2]. Holding the importance of all other languages over the World, the first works in Open IE have taken their attentions to texts

[1] http://www.internetworldstats.com/stats7.htm

[2] https://w3techs.com/technologies/overview/content_language/all

written in English. To the best of our knowledge, the Open IE work which presents the best results for English texts is called ClausIE [Del Corro and Gemulla 2013]. ClausIE is based on a set of manual rules and a dependency parser (DP) [Rodríguez et al. 2016]. For Portuguese languages, this kind of approach is carried by systems such as DepOE [Gamallo et al. 2012] and ArgOE [Gamallo and Garcia 2015]. Those systems define generic manual rules for different languages (multilingual systems). We do believe that compatible rules for different languages impose some limitations due to not dealing with particular aspects of each language. Different from them, our method extracts facts by a depth-first search (DFS) approach to identify the arguments. We can summarize our main contributions:

- We have created an Open IE system based on DP for Portuguese.
- We propose a new way of identifying relationship arguments using depth-first search.
- We create models for *tokenizer* tasks, *POS tagger* and *dependency parser* based on the *Universal Dependencies version 1.4* to Portuguese[3].
- We have created two datasets with sentences written in Portuguese and from different domains to evaluate our approach.

The remainder of this work is organized as follows: Section 2 presents our approach called DependentIE. Section 3 describes our experimental methodology and the materials used in our evaluation. In Section 4 we explain our results and discussed some insights in Section 5. Section 6 discusses the set of our closest works. Finally, Section 7 concludes and presents some envisioning work.

2. Our DependentIE Approach

DependentIE is an Open IE system for texts written in Portuguese language. As well as ArgOE [Gamallo and Garcia 2015] and ClausIE [Del Corro and Gemulla 2013] we use a Dependence Parser (DP) to identify clauses[4] (useful parts of a sentence). In this work, a clause is one of the following parts of a sentence: subject (S), direct and indirect objects (O), verb (V), adverb (A), complement (C) and modifier (M). Our method extracts facts using clauses based on the standard SV (Subject - Verb). The arguments are detected through a deep-search in the sentence dependency.

The execution of our DependentIE needs some preprocessing tasks, such as a Tokenizer, a Part-Of-Speech Tagger and a Dependence Parser for Portuguese (Figure 1). Those three analyzers mark each sentence.

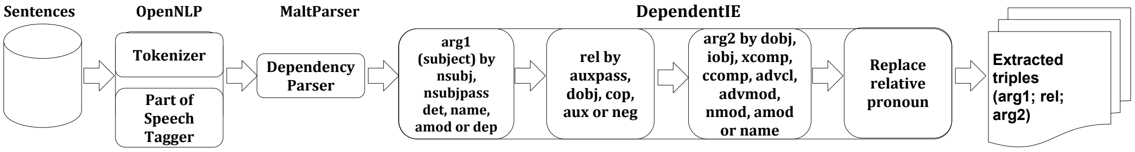

Sentences
OpenNLP
Tokenizer
Part of Speech Tagger
MaltParser
Dependency Parser
DependentIE
arg1 (subject) by nsubj, nsubjpass det, name, amod or dep
rel by auxpass, dobj, cop, aux or neg
arg2 by dobj, iobj, xcomp, ccomp, advcl, advmod, nmod, amod or name
Replace relative pronoun
Extracted triples (arg1; rel; arg2)

Figure 1. Overview of the process made by our DependentIE.

[3] http://universaldependencies.org/

[4] The clauses consist of a subject and a verb and their constituents, such as objects (direct and indirect), adverbs and others.

Figure 2 depicts an example of a preprocessed output sentence. For instance, consider the sentence *A espada, que era uma peça de decoração, foi apreendida para ser periciada em 1967.* (The sword, which was a decorative piece, was seized for inspection in 1967.). For each token, our approach finds all dependents. The other subject-dependent token (*nsubjpass*) "espada" (sword) is defined as an article (*det*) "A" (the) and a noun (*acl:relcl*) "peça" (piece).

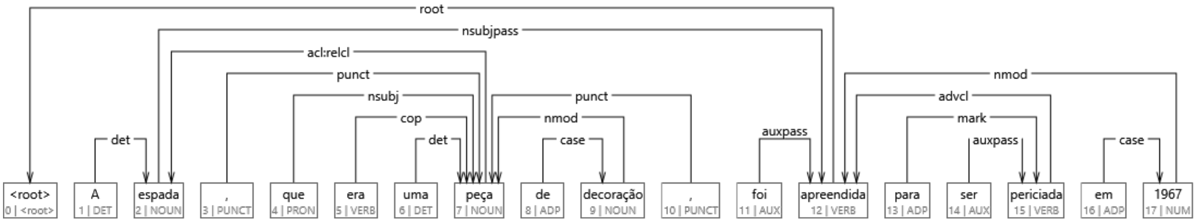


Figure 2. Example of a sentence with tokens and annotators of the POS tagger and the DP. Both Tokenizer and POS tagger were undertaken by Apache OpenNLP and the DP by MaltParser.

Taking our DependentIE (Figure 1), a sentence might contain a subject (arg1), a verbal phrase (rel) (SV) and one or more arguments (arg2). The identification of *arg1* starts locating the token identified by *nsubj* or *nsubjpass* in the dependency tree. In the next step, the following dependents are found: *det, name, amod, nmod or dep*. Considering our sentence "A espada" (The sword) and "which". The identification of *rel* is performed by the subject HEAD concatenated with the dependents such as: *auxpass, dobj, cop, aux or neg*. In the end, we identify the *arg2* starting again from the head of the subject following by different paths such as:

- $SV + path + O$ by *dobj, iobj*.
- $SV + path + C$ by *xcomp and ccomp*.
- $SV + path + A$ by *advcl and advmod*.
- $SV + path + PROPER\_NAMES$ by *name*.
- $SV + path + M$ by *nmod and amod*.

Applying our DependentIE to our example we extract: t_1 = (The sword/a espada, was seized/foi apreendida, for inspection/para ser periciada), t_2 = (The sword/a espada, was seized/foi apreendida, in 1967/em 1967) and t_3 = (which/que, was a piece/foi uma peça, decoration/de decoração). However, the t_3 triple does not represent a coherent information. To solve this issue, we replace the relative pronoun "que" (which) by the subject "A espada" (The Sword). This is done by identifying the dependency *acl:relcl* between the head of "que" and the subject "A espada". When this type of fact is identified, our method always replaces the head with the subject as the example described above to extract facts more informative. In Table 1 we present all the facts extracted by our method from our sample sentence.

3. Experimental Setup

The performance of *DependentIE* was measured against two other Open IE systems on Portuguese: the *ArgOE* [Gamallo and Garcia 2015] and the proposal presented by Sena *et al.* (2017) in [Sena et al. 2017] (We call it InferReVerbPT). Each system received as

Table 1. Extracted triples

A espada, que era uma peça de decoração, foi apreendida para ser periciada em 1967 / The sword, which was a decorative piece, was seized for inspection in 1967.			
	arg1	rel	arg2
Triple 1	A espada / The sword	foi apreendida / was seized	para ser periciada / for inspection
Triple 2	A espada / The sword	foi apreendida / was seized	em 1967 / in 1967
Triple 3	A espada / The sword	era uma peça / was a piece	de decoração / decoration

input two datasets composed by sentences. The first dataset was built using the Corpus of Electronic Texts Extracts NILCS/*Folha de São Paulo* newspaper (CETENFolha) [5] version 2008. The second dataset was built from the digital encyclopedia *Wikipedia* [6]. Both datasets were composed of 200 sentences which were retrieved randomly. Those datasets were called CETEN200 and WIKI200 respectively. The selection of the sentences follows this set of criteria:

- A sentence must contain three or more tokens.
- A sentence should end with the dot (.) character.
- A sentence must be initiated with a capital letter.
- A sentence must have at least a noun or a pronoun.
- A sentence should not be initiated by a conjunction.

To evaluate the extraction by our DependentIE, two experts labeled the facts extracted from all three systems mentioned earlier. We consider two aspects of each extracted fact: (i) coherent or incoherent and (ii) minimal or not. A fact is considered coherent (i) if it keeps the same meaning as in the original sentence. A coherent fact may contain another fact in its arguments[Bast and Haussmann 2013]. In the sentence *Dener morreu no dia 19 de abril num acidente de carro, no Rio.* (Dener died on April 19 in a car accident in Rio.) a system can extract (Dener; morreu no; dia 19 de abril num acidente de carro no Rio) (Dener; died; on 19 April in a car accident in Rio). The fact is coherent, but we can still extract new coherent facts such as: (Dener; morreu; no dia 19 de Abril) / (Dener; died; on 19 April), (Dener; morreu; num acidente de carro) /(Dener; died; in a car accident) e (Dener; morreu; no Rio) / (Dener; died; in Rio). A fact which cannot be decomposed into new facts is considered as a Minimum fact (ii). Minimum facts can be more precise in Question Answering (QA) systems and either in the construction of ontologies [Bast and Haussmann 2013]. We measure the precision of our system by the ratio between all the extracted facts and their coherence. We use this ratio to measure the overall precision in *precision-c* Eq. 2 and by minimality *precision-m* Eq. 3. Considering the metric "recall", in OpenIE area, it is a difficult metric to calculate. Due to OpenIE

[5] http://www.linguateca.pt/cetenfolha/

[6] https://pt.wikipedia.org/

methods perform their extractions in an open domain, it is hard to estimate all the false negatives of a system to calculate the ”recall”. Only one sentence may have other combinations and interpretations of facts that can result to generate new facts. It depends on the human interpretation/analysis of each sentence. Thus, in this work, we did not calculate the ”recall” measurement because of its imprecise value in OpenIE domain.

$$precision - c = \frac{\#(coherent\_triples)}{\#(extracted\_triples)} \tag{2}$$

$$precision - m = \frac{\#(minimum\_triples)}{\#(coherent\_triples)} \tag{3}$$

An extracted fact is considered coherent or minimal if both evaluators agreed with each other. The agreement of the experts was verified by Cohen’s kappa coefficient [Carletta 1996]. In Table 2 we present the degree of agreement between those two experts.

Table 2. Analysis of cohen’s kappa in both datasets for coherence and minimality

	CETEN200	WIKI200
Kappa (coherent)	0,807	0,846
Kappa (minimality)	0,778	0,860

Both experts had almost perfect agreement in evaluating the coherence in both datasets. In the evaluation of minimality, both experts had a full agreement in CETEN200 and almost a complete agreement in WIKI200.

3.1. Materials

The DP used by DependentIE was *MaltParser* [Nivre et al. 2006] with the library *LIBSVM* [Chang and Lin 2011]. To the best of our knowledge, there is no model for DP in Brazilian Portuguese language pattern *Universal Dependencies* version 1.4[7] (UDv1.4). We train the DP using the *Bosque* treebank in the CoNLL format available in UDv1.4 The input of a DP is a separate sentence into tokens within a POS tagger annotation. For both tasks, we use the Apache OpenNLP[8]. As we did not find a model for both tools based on UDv1.4 for Portuguese, we trained our model for tokenizer with a subset of 9047 sentences randomly chosen from the treebank *Bosque* following the annotation presented in the Table 3. Our tokenizer model presented a high precision as shown in Table 4.

To train the POS tagger, we also used the Bosque treebank. All sentences were pre-processed to stay in the format shown in Table 5. We used 10-fold cross-validation. Our POS Tagger model presented 95.84% of accuracy.

The model created for DP was evaluated using the following metrics: Labeled attachment score (LAS), unlabeled attachment score (UAS) and label accuracy (LA). LAS checks the accuracy of both the relation and the syntactic analysis. The UAS checks the accuracy of tokens’ relationships *tokens*. The LA checks the accuracy of the syntactic analysis of each *token*.

[7] `http://universaldependencies.org/#pt_br`

[8] https://opennlp.apache.org/

Table 3. Pre-processed sentences to train the Tokenizer

Sentences	Sentences Pre-processed to be trained
A espada, que era uma peça de decoração da casa, foi apreendida para ser periciada / The sword, which was a decorative piece of the house, was seized for inspection	A espada<SPLIT>, que era uma peça de decoração da casa<SPLIT>, foi apreendida para ser periciada / The sword<SPLIT>, which was a decorative piece of the house<SPLIT>, was seized for inspection
Morena (Nanda Costa) e Wanda (Totia Meirelles) declararam guerra uma à outra. / Morena (Nanda Costa) and Wanda (Totia Meirelles) declared war on each other.	Morena <SPLIT>(Nanda Costa <SPLIT>) e Wanda <SPLIT>(Totia Meirelles<SPLIT>) declararam guerra uma à outra<SPLIT>. / Morena <SPLIT>(Nanda Costa<SPLIT>) and Wanda <SPLIT>(Totia Meirelles<SPLIT>) declared war on each other<SPLIT>.

Table 4. Tokenizer results

Precision	**Recall**	**F1-measure**
99,99%	87,63%	93,40%

Table 5. Pre-processed sentences to train the POS tagger.

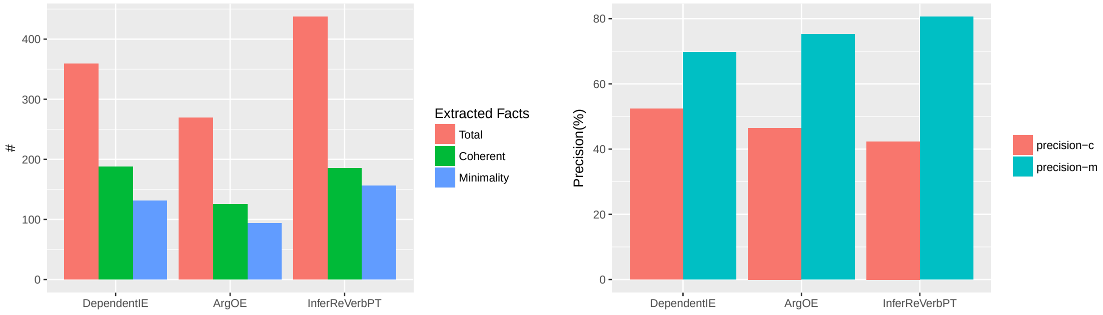
Sentences	Sentences Pre-processed to be trained
A espada, que era uma peça de decoração da casa, foi apreendida para ser periciada / The sword, which was a decorative piece of the house,was seized for inspection	A_DET espada_NOUN ,_PUNCT que_PRON era_VERB uma_DET peça_NOUN de_ADP decoração_NOUN da_ADP casa_NOUN ,_PUNCT foi_AUX apreendida_VERB para_ADP ser_AUX periciada_VERB / The_DET sword_NOUN,_PUNCT which_PRON was_VERB a_DET decorative_NOUN piece_NOUN of_ADP the_DET house_NOUN,_PUNCT was_AUX seized_VERB for_ADP inspection_VERB

Table 6. Dependency parser results

LAS	**UAS**	**LA**
81,7%	84,3%	91,4%

4. Results

We compared the *DependentIE* against *ArgOE* and *InferReVerbPT* in quantity of facts extracted and precision. We organized our results presenting initially the numbers obtained in CETEN200 and then in WIKI200. The amount of facts extracted in CETEN200 is shown in Figure 3(a) and the precision in the Figure 3(b). Our proposal extracted 359 facts of which 188 are coherent. We obtained 52.4% by *precision-c* and 69.7% by *precision-m*. With more coherent facts extracted and a higher precision, we have the highest gain among all three systems compared in CETEN200. In a direct comparison with *ArgOE* that also uses DP, we had 33.5% of more coherent facts.

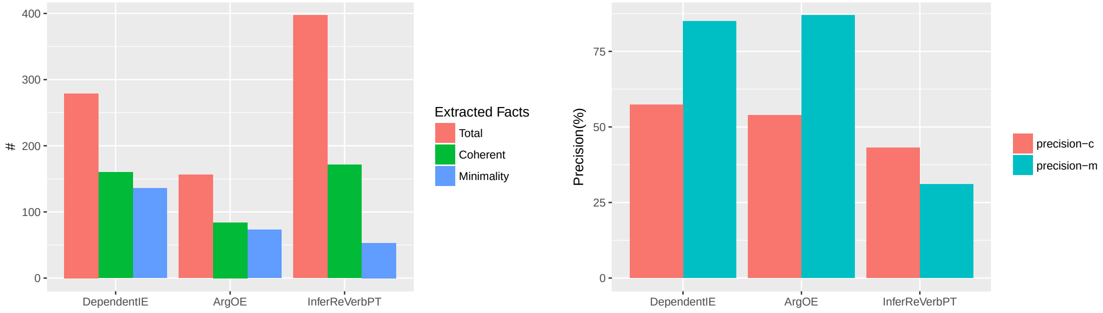


(a) Number of facts extracted in CETEN200.

(b) Precision-c and precision-m in CETEN200.

Figure 3. Result in CETEN200 dataset.

The results about the facts extracted in WIKI200 are presented in Figures 4(a) and 4(b). In WIKI200 the DependentIE was able to extract 279 triples of which 160 were coherent. Our method obtained 57.34% in *precision-c* and 85% of the coherent extracted facts were minimal. In WIKI200 the *ArgOE* performed better for *precision-m*. However, *DependentIE* continues to present the greatest gain among the approaches considering the number of coherent facts and *precision-c*. *InferReVerbPT* has a small advantage on the number of coherent facts. However, it has a low gain approach that is evident in WIKI200. In WIKI200 dataset our method can maintain a good proportion between the number of coherent facts and their minimality. These characteristics give our proposal an alternative application in Open IE for Portuguese texts in both scenarios.



(a) Number of facts extracted in WIKI200.

(b) Precision-c and precision-m in WIKI200.

Figure 4. Result in WIKI200 dataset.

5. Discussion

There was a subset of sentences for which no extractions were obtained. In CETEN200 *InferReVerbPT* left to extract facts from 7 sentences and *ArgOE* left 85 sentences without extractions. Our *DependentIE* did not get extractions in 42 sentences. However, our approach presented the best proportion of coherent facts and *precision-c*. In WIKI200 the number of sentences without extraction increased for our *DependentIE* up to 66 sentences.

Even so, our proposal presents the highest gain among the evaluated systems. In most cases, the DP did not correctly label the sentence preventing the extraction of some facts. Another important observation was the behavior of our method for *precision-m* in both datasets. We believe that their writing style influenced on our experiments. CETEN200 is formed of journalistic texts, and the WIKI200 dataset is formed with encyclopedia texts thus confirming the different styles. For a better understanding of this kind of problem, we consider observing some examples of sentences from both datasets. On many occasions, the systems make different extractions. Each approach has different characteristics, and this reflects in the results. For example, in Table 7 the *DependentIE* had better performance by extracting 2 coherent and minimal triples. This happens because our proposal does a search in the dependency tree and we can recover more arguments than the other systems. In case of *ArgOE*, there was no extraction. We believe that this fact is justified by some error in DP as the researchers themselves emphasize. The *InferReVerbPT* made only a coherent extraction, but that it is not minimal. Methods that use shallow parsers have some difficulties retrieving arguments away from the verbal phrase.

Table 7. Example where DependentIE is better than the other methods.

Sentence pt-br/en: Esse asteroide foi descoberto em 24 de Maio de 1931 por Cyril Jackson. / This asteroid was discovered on 24, May in 1931 by Cyril Jackson.

System	Triples (pt-br / en)	Coherent	Minimality
DependentIE	(Esse asteroide; foi descoberto; em 24 de Maio de 1931) / (This asteroid; was discovered; on May 24 in 1931)	yes	yes
	(Esse asteróide; foi descoberto; por Cyril Jackson) / (This asteroid; was discovered; by Cyril Jackson)	yes	yes
ArgOE	-	-	-
InferReVerbPT	(Esse asteroide; foi descoberto; em 24 de Maio de 1931 por Cyril Jackson) / (This asteroid; was discovered; on 24, May in 1931 by Cyril Jackson)	yes	no

We present another example in the Table 8. The method *DependentIE* did not extract facts because the DP did not find the subject of the sentence. Meanwhile, *ArgOE* had a better performance by making both two coherent and minimal extractions. The *InferReVerbPT* made two incoherent extractions because it was wrong about the definition of the arguments in both cases.

In the example on Table 9, both *DependentIE* and *ArgOE* did not perform extractions. Once again the DP did not find a subject. However, the *InferReVerbPT* made a coherent and minimal extraction. The low precision of more sophisticated language tools for Portuguese is still a barrier to Open IE. While analyzers such as POS tagger and Chunker present a precision close to 100%[9] the *MaltParser* present a precision near to 80% for the DP task. Thus, we believe that more effort is still needed to create efficient linguistic tools beyond the English language.

[9] InferReVerbPT uses the CoGrOO `http://ccsl.ime.usp.br/cogroo/`

Table 8. Example where ArgOE is better than the other methods.

Sentence pt-br/en: O primeiro serviço do MD-11 nos Estados Unidos ocorreu ainda em 1990 e foi,inaugurado pela Delta Airlines. / The first service of MD-11 in the United States occurred in 1990 and was inaugurated by Delta Airlines.

System	Triples (pt-br/en)	Coherent	Minimality
DependentIE	-	-	-
ArgOE	(O primeiro serviço do MD-11; ocorreu nos; Estados Unidos) / (The first service of MD-11; ocurred in; United States)	yes	yes
	(O primeiro serviço do MD-11; ocorreu em; 1990) / (The first service of MD-11; ocurred in; 1990)	yes	yes
InferReVerbPT	(O primeiro serviço do MD-11 nos Estados_Unidos; ocorreu ainda em; 1990 foi inaugurado pela Delta Airlines) / (The first service of MD-11; ocurred in; 1990 was inaugurated by Delta Airlines)	no	-
	(O primeiro serviço do MD-11 nos Estados Unidos ocorreu ainda em 1990; foi inaugurado pela; Delta Airlines) / (The first service of MD-11 ocurred in 1990; was inaugurated; by Delta Airlines)	no	-

Table 9. Example where InferReVerbPT is better than the other methods.

Sentence pt-br/en: Põlde é localizado no município de Abja vald e Viljandimaa, no sul do país, e 150 km,ao sul da capital Tallinn. / Põlde is located in the municipality of Abja vald and Viljandimaa in the south of the country, and 150 km south of the capital Tallinn.

System	Triples (pt-br/en)	Coherent	Minimality
DependentIE	-	-	-
ArgOE	-	-	-
InferReVerbPT	(Põlde; é localizado no; município de Abja vald) / (Põlde; is located in the; municipality of Abja vald)	yes	yes

Overall, from a preliminary analysis, it was verified that in many cases the DependentIE made incoherent extractions due to the Dependency Parser produces a dependency tree with bad grammar structures. In the sentence ”Para o Bank of America a opção foi a de não converter nenhum tostão.” the DependentIE returned the subject ”a de não” because the DP labeled the article ”a” as ”nsubj”. As a consequence, DependentIE generates an incoherent triple (a de não; opção nenhum tostão; Para o Bank of America). We believe that the improvement of NLP tools can improve the quality of our DependentIE, allowing it to do more coherent extractions with both high precision-m and precision-c. Another problem detected concerns the Portuguese language. The extractions of triples are made based on a subject, but in Portuguese, there are implicit subjects, which does not exist in

English language. For example, in the sentence ”faremos o trabalho” DependentIE does not extract any triples. This problem could be solved by using some tool to verify that there is an implicit subject (the pronoun ”nós”), allowing the method to extract the triple (nós; faremos; o trabalho). Moreover, to increase the informativeness of this triple, it is possible to use some task of coreference. Thus the pronoun ”nós” would be associated with the real subject (the people who are going to do the work).

6. Related Works

The precursor systems in Open IE performed their task in four stages: (i) Labeling of sentences using heuristics or distant supervised; (ii) An extractor training using a sequential labelling model (*e.g.* conditional random field (CRF)); (iii) Identification of the noun phrases (arguments); and (iv) the extractor identifies the relational phrase (if there is) between the arguments. The first proposals such as *TextRunner* [Banko et al. 2007] and *WOE* [Wu and Weld 2010] used this kind of approach. The use of machine learning was quickly replaced by methods that use heuristics or handcrafted rules for identifying relational phrases. New systems such as *ReVerb* [Fader et al. 2011] substituted the extractor by regular expression and started the task by first identifying the relational phrase to then discover its arguments. After the *ReVerb* that gave great pulse to the area, new methods emerged as *OLLIE* [Schmitz et al. 2012] and *ClausIE* [Del Corro and Gemulla 2013], but all exclusively for English texts. To the best of our knowledge, the first Open IE system for Portuguese texts was the *DepOE* [Gamallo et al. 2012] published in 2012. Although presenting a multilingual proposal, including Spanish and Galician, the evaluation of the method is only made for facts extracted in English sentences. Only in 2015 is published a new version of *DepOE* called *ArgOE* [Gamallo and Garcia 2015] where their evaluation for Portuguese was completed. This method uses handcrafted rules on a dependency tree to identify useful parts in sentences (clauses) and then uses the clauses to compose the extracted facts. Open IE with DP approach and handcrafted rules has been receiving greater attention because of their higher *recall*. On the other hand, for texts in Portuguese, there are methods that use shallow analysis. In 2013, *LSOE* [Xavier et al. 2013] published a work that uses POS tagger and a set of regular expressions identifying facts in an unsupervised way. This approach is similar to the method proposed by Sena *et al.* (2017). In this case, the syntactic restrictions of the *ReVerb* were used in Portuguese sentences. Furthermore, new syntactic restrictions were proposed to identify transitivity and symmetry to generate new facts and increase the *recall*. In Collovini *et al.* (2016) a method that uses CRF is presented to Portuguese. However, their extraction is limited to a restricted set of named entities. We believe that this limitation restricts the use of this method in specific applications, as listed by the authors.

The application of Open IE for Portuguese texts is scarce. The *RAPPort* system [Rodrigues and Gomes 2015] have proposed a method that uses DP and a standard subject-predicate-object to extract facts. The facts extracted by the Open IE module are used to enhance a QA system. In the best of our knowledge, other applications such as ontology construction, classification or clustering texts have not been identified. Thus, our proposal is positioned in this scenario of evolution from the Open IE to the Portuguese language.

7. Conclusions and Future Work

In this work, we proposed the *DependentIE*, an open IE for Portuguese texts. Our proposal differs from the state of the art by not using fixed manual rules. This feature enhanced our method when compared with another system of the same approach. To go through the dependency tree, we adapt an in-depth search that identifies the arguments of a relationship. The *DependentIE* was compared to the Open IE *ArgOE* and *InferReVerbPT* systems, obtaining greater gain under two different datasets: journalistic and encyclopedia texts. We believe that linguistic tools to a target language can increase this type of task. Our results suggested that this hypothesis is correct, but also reveals that Open IE for Portuguese still requires some efforts. These efforts go beyond the adaptation of existing approaches for other languages and the construction and improvement of tools for natural language processing.

As future works we need to enhance the DP tool for a more accurate *precision* and *recall*. We observed that our method did not extract facts from a significant number of sentences due to not identifying the subject on these sentences. There are also particular issues of the Portuguese language as an implicit or indeterminate subject and the anaphoras. All these concerns are current open problems. We believe that minimizing these problems can further enhance our results.

References

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., and Etzioni, O. (2007). Open information extraction from the web. In *Proceedings of IJCAI*, volume 7, pages 2670–2676.

Bast, H. and Haussmann, E. (2013). Open information extraction via contextual sentence decomposition. In *Proceedings of ICSC*, pages 154–159. IEEE.

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, 22(2):249–254.

Chang, C.-C. and Lin, C.-J. (2011). Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

de Abreu, S. C., Bonamigo, T. L., and Vieira, R. (2013). A review on relation extraction with an eye on portuguese. *Journal of the Brazilian Computer Society*, 19(4):553–571.

Del Corro, L. and Gemulla, R. (2013). Clausie: clause-based open information extraction. In *Proceedings of WWW*, pages 355–366. ACM.

Fader, A., Soderland, S., and Etzioni, O. (2011). Identifying relations for open information extraction. In *Proceedings of EMNLP*, pages 1535–1545. Association for Computational Linguistics.

Gamallo, P. and Garcia, M. (2015). Multilingual open information extraction. In *Proceedings of EPIA*, pages 711–722. Springer.

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*, ROBUS-UNSUP '12, pages 10–18. Association for Computational Linguistics.

Nivre, J., Hall, J., and Nilsson, J. (2006). Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of LREC*, volume 6, pages 2216–2219.

Rodrigues, R. and Gomes, P. (2015). Rapport—a portuguese question-answering system. In *Portuguese Conference on Artificial Intelligence*, pages 771–782. Springer.

Rodríguez, J. M., Merlino, H. D., Pesado, P., and García-Martínez, R. (2016). Performance evaluation of knowledge extraction methods. In *Proceedings of IEA/AIE*, pages 16–22. Springer.

Schmitz, M., Bart, R., Soderland, S., Etzioni, O., et al. (2012). Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 523–534. Association for Computational Linguistics.

Sena, C. F. L., Glauber, R., and Claro, D. B. (2017). Inference approach to enhance a portuguese open information extraction. In *Proceedings of the 19th International Conference on Enterprise Information Systems - Volume 1: ICEIS,*, pages 442–451. INSTICC, ScitePress.

Wu, F. and Weld, D. S. (2010). Open information extraction using wikipedia. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, pages 118–127. ACL.

Xavier, C. C., de Lima, V. L. S., and Souza, M. (2013). Open information extraction based on lexical-syntactic patterns. In *Intelligent Systems (BRACIS), 2013 Brazilian Conference on*, pages 189–194. IEEE.

Xavier, C. C., de Lima, V. L. S., and Souza, M. (2015). Open information extraction based on lexical semantics. *Journal of the Brazilian Computer Society*, 21(1):1.