DptOIE: A Portuguese Open Information Extraction system based on Dependency Analysis

Leandro de Oliveira
FORMAS*- DCC - LaSiD - IME
Federal University of Bahia
leo.053993@gmail.com

Daniela Barreiro Claro
FORMAS - DCC - LaSiD - IME
Federal University of Bahia
dclaro@ufba.br

Abstract

It is estimated that more than 80% of the information on the Web is stored in textual form. For humans, the task of extracting useful information from data that comes up daily is difficult. In order to automate the process, techniques of Open Information Extraction (OIE) methods, which are capable of extracting facts from large textual bases, have been proposed. At first, most OIE methods were developed for the English language. However, other languages, such as Portuguese, have tackled special attention, since it covers approximately 2.5% of all content available on websites. For English languages, methods based on hand-crafted rules and dependency analysis have gained good results. Nevertheless, methods based on similar approaches, in Portuguese, have not presented equivalent performance. We believe that the rules defined are generic and do not cover specific aspects of the language. For this reason, our DptOIE method defined a new set of hand-craft rules and explore sentences through a dependency analysis by a depth-first search (DFS) approach. DptOIE was compared against two other OIE methods which extract facts in Portuguese: PragmaticOIE and ArgOE. DptOIE outstands the other works, obtaining a greater area under the precision-yield curve. Precision was superior as well as the number of coherent facts extracts. As far as we know, this is the most outperforming method to extract fact on OIE for the Portuguese language.

Keywords

Open Information Extraction, Dependency Analysis, Depth-first search

1 Introduction

Technological advances and the popularity of the Internet have been contributing to the growth of textual databases which are daily generated. More than 80% of the Web (Barion & Lago, 2015) is stored in textual form. Most of them is heterogeneous. Analyzing all information published manually is a hard and a time consuming task. To minimize the effort of analyzing these type of data, Information Extraction (IE) techniques extracts and synthesizes information in an automated way. Traditional IE uses predominantly supervised approaches, which requires that relationships are specified through many examples for training (Bassa et al., 2018). This require a lot of manual effort every time the domain change. Moreover, new corpus needs to be labeled to deal with the new relationships. For this reason, traditional IE is not scalable to large corpus, such as the Web. To overcome this limitation, Banko et al. (2007) introduced the Open Information Extraction (OIE) paradigm.

OIE methods extract information without the need to pre-determine the set of target relationships, enabling greater scalability, more extractions and domain independence. OIE methods can be applied in question-and-answer systems, opinion mining, forensic computing and others. Schmitz et al. (2012) presented a scenario for OIE: " *A terrorist's computer has been seized, and intelligence analysts urgently need to find out information to avoid possible catastrophes.*". Certainly, these analysts have no knowledge of these data, so the important relationships to be discovered will probably not be pre-specified. OIE is an approach that could be considered in this scenario.

At the beginning, most OIE methods were focused on the English language. By 2019, it was estimated that English[1] had approximately 52% of all Web content, and it was stipulated that it covered approximately 25.5% of Web users[2]. However, other languages, such as Portuguese, are also important. After all, 169.0 million users

*http://formas.ufba.br/

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

[2] https://www.statista.com/statistics/262946/share-of-the-most-common-languages-on-the-internet/

on the Internet have Portuguese as their native language[3]. In addition, approximately 2.5% of website content is written in this language.

From the state-of-the-art analysis, methods which perform dependency analysis and use hand-crafted rules, such as ClausIE (Del Corro & Gemulla, 2013), have performed well in the English language (Rodríguez et al., 2016). In the Portuguese language, there are methods like DepOE (Gamallo et al., 2012), ArgOE (Gamallo & Garcia, 2015) and DependentIE (Oliveira et al., 2017), which use similar approaches. However, two of them (ArgOE and DepOE), are multilingual systems and their results are not always so good when compared to other methods that deal with specific languages. This may be due to the generic rules to extract the facts to cover all languages, imposing limitations on their methods. DependentIE is close to our approach, however the experiments showed that the Dependency Parser used, i.e.,MaltParser, did not perform good results and their rules were not able to generalize enough, since many sentences had no facts extracted. We believe that the use of dependency analysis and specific rules to handle a single language enable particular aspects of a language, favoring the increase of extractions and maintaining high precision. Thus, in this work we propose the DptOIE, an OIE method for the Portuguese language. We can summarize our main contributions: (I) we train models for POS *tagger* and Dependency Parser (DP) for Portuguese; (II) we implemented an OIE method for Portuguese based on dependency analysis; (III) we propose hand-crafted rules and adapt an Depth-first search (DFS) to explore the dependency tree; and (IV) we implemented modules to perform treatments from sentences with Coordinated Conjunctions (CC), Subordinate Clauses and Appositive.

This paper is organized as follows: section 2 describes our related work. Section **??** presents

2 Background

OIE methods extract facts from plain texts (Banko et al., 2007). Unlike traditional IE, OIE methods do not require prior relationships to be specified. OIE systems usually extract facts in the form of triples $t = (arg1, rel, arg2)$, where *arg1* and *arg2* are two arguments and *rel* establishes a semantic relationship between *arg1* and *arg2*. For example, in the sentence "The painkiller X causes nausea.", the triple ("The painkiller X", "causes" , "nausea") could be extracted without any prior specification. The main advantages of OIE are: domain independence, unsupervised extraction, and scalability for large amounts of text (Del Corro & Gemulla, 2013).

Sentences received by OIE methods go through a pre-processing step, which normally uses Natural Language Processing tools (NLP). Gamallo & Garcia (2015) organized the OIE methods into two broader categories: the systems that require automatically generated training data to learn a classifier and methods based on hand-crafted rules or heuristics. Moreover, they divided each category into two subtypes: systems that making use of shallow syntactic analysis, generally through POS tagging or chunking, and systems based on dependency parsing. Methods that perform shallow analysis typically achieve high precision, but suffers from low recall. Approaches that use dependency analysis usually have a high processing cost when compared to methods that perform shallow analysis. On the other hand, they trade efficiently for an improvement in precision and recall (Del Corro & Gemulla, 2013).

Our research focused on OIE systems both in English and Portuguese. We investigate NLP techniques and tools used by these methods and whether they are available in Portuguese. The Table 1 shows an overview of the analyzed methods. The works presented in this table were selected from the systematic mapping conducted by Glauber & Claro (2018) and literature review performed by Glauber et al. (2018).

2.1 English OIE Systems

The first Open Information Extraction method was TextRunner (Banko et al., 2007), which uses a self-supervised approach to label their own training data. Through machine learning techniques, authors trained a model based on Naive Bayes classifier, which is responsible for recognizing patterns and extracting facts. After TextRunner, other OIE methods have emerged, such as WOE (Wu & Weld, 2010), which is also self-supervised. The difference, when compared against to its predecessor, is that WOE uses heuristics and its classifier is trained from a corpus obtained from Wikipedia. WOE operates in two modes, WOE^{pos} which uses POS tagger and WOE^{dep} which parsers by a DP.

The use of machine learning in the second generation of OIE methods was quickly replaced by rule-based methods, such as ReVerb (Fader et al., 2011). This system receives sentences tagged

[3] http://www.internetworldstats.com/stats7.htm (13/08/2019)

Table 1: Overview of OIE methods. ML stands for Machine Learning.

System	Year	NLP	Techniques used	ML	Language
TextRunner	2007	POS, Chunk	Naive Bayes classifier	✓	English
WOE^{pos}	2010	POS, Chunk	Pattern Learner	✓	English
WOE^{parse}	2010	DP	CRF classifier	✓	English
ReVerb	2011	POS, Chunk	Syntatic and lexical constraints + logistic regression classifier	✓	English
Kraken	2012	DP	Hand-crafted rules		English
DepOE	2012	DP	Hand-crafted rules		Multilingual
OLLIE	2012	DP	Open Pattern Learning	✓	English
ClausIE	2013	DP	Hand-crafted rules		English
LSOE	2013	POS	Qualia Based Patterns		English
CSD-IE	2013	Constituency Parser	Hand-crafted rules		English
DepOE+	2014	DP	Hand-crafted rules + correference		Multilingual
ArgOE	2015	DP	Hand-crafted rules		Multilingual
RePort	2015	POS, Chunk	Syntatic constraints		Portuguese - BR
	2016	POS, Chunk, NER	CRF classifier	✓	Portuguese - BR
	2017	POS, Chunk	Syntatic constraints	✓	Portuguese - BR
DependentIE	2017	DP	Hand-crafted rules + DFS		Portuguese - BR
Neural OIE	2018		Encoder-decoder framework	✓	English
PragmaticOIE	2018	POS, Chunk	Syntatic consraints + Inference + Context + Intention		Portuguese - BR
DptOIE	2019	DP	Hand-crafted rules		Portuguese - BR

with POS tagger and chunker. Its extractions are based on syntactic and lexical constraints and the relations are mediated by verbs.

Akbik & Löser (2012) developed Kraken using their previous work, Wanderlust (Akbik & Broß, 2009), and they assumed that a limited number of patterns may be sufficient to analyze sentences in an in-depth way. For this, Kraken uses a dependency analyzer and generates n-ary facts.

Realizing that many previous methods normally perform their extractions only through verbs, Schmitz et al. (2012) proposed OLLIE (*Open Language Learning for Information Extraction*), which is also based on dependency analysis. In addition to identifying facts with relationships based on verbal phrases, OLLIE also checks relationships that are mediated by nouns or adjectives. Furthermore, it also analyzes the context of the fact. For example, in the sentence *" If he wins five key states, Romney will be elected President "*, previous systems would extract the following triple *(Romney; will be elected; President)*. But this fact is incoherent, which it claims that "*Romney will be elected president*", when in fact there is a condition for it, which is "*if he wins five key states*". For this reason, OLLIE add this new information and perform the following extraction: *((Romney; will be elected; President) ClausalModifier if; he wins five key states)*.

Shortly thereafter, LSOE *(Lexical-Syntactic patterns based Open Extractor)* (Xavier et al., 2013) was published. LSOE aims to be an efficient and simple method, without the need to use machine learning. It was the first method that uses hand-crafted rules in texts labeled with POS tagger, using the Qualia structure (Cimiano & Wenderoth, 2005), which provides new information about the role of words in a sentence.

Another method based on dependency analysis is ClausIE (Del Corro & Gemulla, 2013). This method separate "useful" segments of information in the sentences, called clauses, from the handcrafted rules. The clause constituent can be: Subject (S), Object(O), Verb (V), Adverbial (A),

Complements (C), others. From them, ClausIE can make combinations based on English grammar to extract facts.

Besides the methods based on dependency analysis, there are methods that use constituency parser, such as the CSD-IE (Bast & Haussmann, 2013). A constituency parse breaks a text into sub-phrases. CSD-IE aimed at decompose a sentence into smaller pieces that semantically belong together, to generate minimal facts.

Recently, a novel OIE system was proposed, NeuralOIE (Cui et al., 2018). It uses machine learning techniques from an encoder-decoder framework. One of the main advantages reported by the authors is that their approach is able of generating facts with a high degree of confidence, without using hand-crafted patterns from other NLP tools.

Methods based on dependency analysis and hand-crafted rules has been presenting good performance in English. ClausIE, for example, despite being published in 2013, still presents results comparable to current approaches. For example, from Precision-Recall curve presented by Cui et al. (2018), ClausIE obtained, at the end of the experiment, a slightly higher precision and recall than NeuralOIE. However, the same experiment revealed that Neural OIE achieves the best AUC (Area Under the Curve).

2.2 Portuguese OIE systems

OIE approaches to the Portuguese language, such as DepOE (Gamallo et al., 2012) appeared some years after TextRunner. DepOE is based on dependency parser and hand-crafted rules. It is a multilingual system, i.e., besides the Portuguese language, it also performs extractions on sentences in Spanish, English and Galician. DepOE was also used from the output of sentences processed with LinkPeople, a Coreference Resolution system (Garcia & Gamallo, 2014). Soon after, DepOE was improved and emerged the ArgOE (Gamallo & Garcia, 2015), which is also a multilingual system. Both use similar approaches to perform the extractions. ArgOE attempted to be more open to other dependency parsers by using CoNLL-X format. Furthermore, it has been adapted to perform extractions in the French language.

Following the trend of using dependency parse features, Oliveira et al. (2017) developed DependentIE to extract facts from Portuguese texts. Their method is based on hand-crafted rules. Nevertheless, authors show in the experiments that the DP used, MaltParser (Nivre et al., 2006), did not achieve good results. Moreover, specified rules were not able to generalize enough, since many sentences had no facts to extract.

Unlike previous systems, Pereira & Pinheiro (2015) proposed Report, an OIE method for the Portuguese language based on shallow analysis. It is an adaptation of ReVerb with syntactic and lexical constraints. Sena et al. (2017) also developed an adapted method of ReVerb, using the syntactic constraints. However, their differential is that their method uses an inferential approach to extract new facts, using the binary SVM classifier between the transitive and symmetric classes. The down side reported by Sena et al. (2017) is that their trained model presented a high error rate (17%). More recently, Sena et al. (2017) improved their approach and permuted the SVM classification to rule-based approach into InferPortOIE (Sena & Claro, 2019). Moreover, they either improved their inferential approach and included a contextual and intentional aspect to reach a first pragmatic level within the PragmaticOIE system (Sena & Claro, 2018). Thus, Sena & Claro (2018) extract implicit facts from the text and, consequently, increase the quantity and variety of facts extracted.

Collovini et al. (2016) presented a method that uses CRF to find relationships and extract facts. However, their method is limited to extract relationships only between Named Entities.

Most of OIE methods for the Portuguese language performs shallow analysis, often involving a POS tagger and a chunker. Among them, PragmaticOIE has been the most outstand approach. It has overcome even the approaches that uses dependency analysis and hand-crafted rules, such as ArgOE and DependentIE.

2.3 Evaluation OpenIE Systems

Evaluation of OIE methods is based on Information Retrieval strategies. According to de Abreu et al. (2013), the most common metrics used in the evaluation process of relations extract methods are: Precision (P) (Equation 1), Recall (R) (Equation 2) e F1-Measure (F1) (Equation 3).

$$P = \frac{\#(coherent\ facts\ extracted)}{\#(facts\ extracted)} \quad (1)$$

$$R = \frac{\#(coherent\ facts\ extracted)}{\#(coherent\ facts)} \quad (2)$$

$$F1 = \frac{2 * P * R}{P + R} \quad (3)$$

Evaluating OIE methods is difficult, because, normally, people are required to judge each extraction manually. Also, calculating the recall is not a simple task. Due to OIE methods perform their extractions in an open domain, it is difficult to estimate all false negatives for recall calculation. A sentence can produce many combinations and have many interpretations of facts (Oliveira et al., 2017). For this reason, Schmitz et al. (2012) proposed the Yield (Y) metric, which refers to the amount of coherent facts. They claim that Yield is proportional to recall and can be calculated by multiplying the total of extracted triple by precision. In addition to the presented metrics, OIE methods can be evaluated through the Area Under Precision-Recall Curve. However, due to the difficulty of calculating the recall, Schmitz et al. (2012) evaluated their method by applying the Area Under the Precision-Yield Curve. To overcome this limitation, Stanovsky & Dagan (2016) proposed a benchmark to evaluate OIE methods in English, which enables to evaluate precision, recall and area under the curve in an automate way. But, to the best of our knowledge, there is no similar approach to Portuguese.

2.4 Portuguese vs English

There are significant differences between English and Portuguese OIE approaches. Such differences prevent the usage of techniques or rules from OIE English systems to be applied directly to Portuguese texts. We enumerate some relevant differences between them.

1. **Origin:** Portuguese is a Latin language, while English is derived from Germanic languages.

2. **Alphabet:** There are 26 letters in the Portuguese alphabet and 11 letters with diacritics. The English alphabet has 26 letters without diacritical marks.

3. **Hidden Subjects:** Hidden subject in the Portuguese language is common, which does not occur in English. For example, in the sentence "*Corremos a maratona.*", the pronoun "nós" is hidden. The same sentence in English must be "*We run the marathon.*". Hidden subject directly interferes within OIE methods, because facts are usually subject-verb based. For this reason, sentences with hidden subject often do not have relations extracted by the current methods of OIE in Portuguese.

4. **Adjectives:** In English, adjective is usually used before the noun, while in Portuguese it is more common to be placed after the noun. However, there are cases in Portuguese that same adjective can come after the noun and modify the meaning of a phrase, as is the case with the adjective "poor ". For example, the sentence "Poor boy." refers to a resource-poor or miserable kid. If adjective "poor" and noun "boy" alternate their positions in Portuguese, the semantics of the sentence is altered and may mean that boy inspires pity (not necessarily misery). Another observation is concerned the adjectives which can be inflected in number and gender, which does not occur in English. Thus, the position of the adjective when extracting the facts needs careful.

5. **Word order:** The basic structure of the most sentences in English and Portuguese adhere to subject-verb-object order. Moreover, words in both languages have high flexibility on how word could be shuffled within sentence (Bassa et al., 2018), which can be a challenge for the task of dependency analysis.

6. **Tools, resources and potential for OIE:** NLP tools for Portuguese are not always accurate than English NLP tools, for example. Datasets in Portuguese have not been annotated or reviewed as English has. Some times, it is even not available for downloading.

3 DptOIE

Thus, we developed a method to extract facts based on dependency analysis and hand-crafted rules. DptOIE follows an execution flow to extract new facts (Figure 1). Initially, DptOIE receives the sentences. Each sentence is preprocessed through a tokenizer, a POS tagger and Dependency Parser. A dependency tree is returned. Afterwards, the method starts the extraction and, if necessary, DptOIE uses three modules to handle particular cases: (I) Coordinate Conjunctions (CC), (II) Subordinate clauses and (III) Appositives.

3.1 Preprocessing

The preprocessing consists of using Information Extraction techniques inherited from NLP. Sentences were preprocessed following the same pattern of Brazilian Portuguese treebank V2.1 of

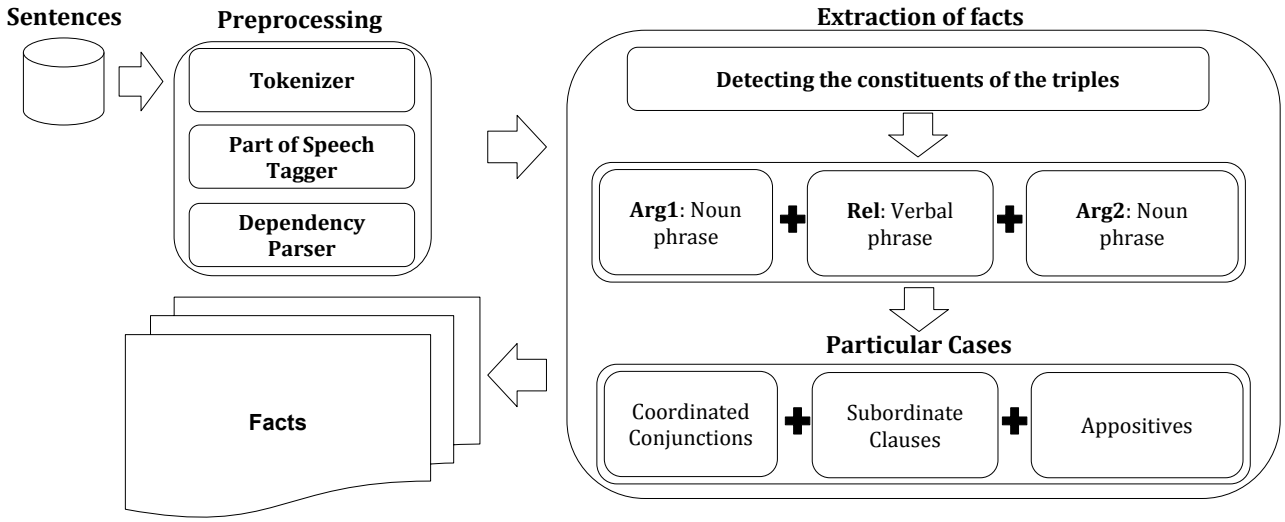


Figure 1: DptOIE workflow

Universal Dependencies (UD)[4], which is in the CoNLL-U format [5]. NLP techniques are: Tokenizer, Part of Speech Tagger and Dependency Parser.

After applying the tokenizer in each sentence, the compound words and contractions were separated to keep the same pattern as the treebank. For example, "guarda-roupa" becomes [guarda] [-] [roupa] and "da" becomes [de] [a]. Then, tokens were labeled with POS tagger in universal standard following the CoNLL-U format for DP. Figure 2 shows a fully preprocessed sentence.

3.2 Extraction of facts

The Extraction of facts receives as input a dependency tree in CONLL-U format. Initially, a mapping of each token in the sentence is done to know the dependent child tokens. For example, in Figure 2, only one dependent of the subject (*nsubj*), "Arsenal", is the determinant (*det*) "O" and the parent node of "Arsenal" is the verb "busca" (seeks).

3.2.1 Extracting the triples

Triples extraction requires a sentence with a Subject (*Arg1*), a verbal phrases to compose the relation (*Rel*) and one or more relationship-dependent arguments (*arg2*). To detect the constituents of the triple, an Depth-First Search (DFS) was adapted. Table 2 presents the handcrafted rules used by DptOIE to conduct the DFS. In it, words in italics represent the dependency relationships that are searched by DFS to find the constituents of the triples. Relationships have been grouped into Objects, Modifiers, Adverbial, Subjects, Complements, Auxiliary or Copular verbs and Others. The latter group contains dependency relationships, which are: *conj, appos, expl:pv, acl:part, acl:relcl and dep*, which respectively serve to represent: conjunctions, appositives, reflexive pronouns, a connection between sentences, the term referring to the relative pronoun (usually) and unspecified dependencies.

Thus, when the DptOIE receives a sentence as input, it applies the rules from Table 2. Initially, DFS searches subjects in the sentence. Then, relationships are found from another DFS, which starts from the parent node of the subject or child node of the subject labeled with "acl:part". The last step is to find Arg2. Another DFS is performed from the child nodes of the relation. Extractions of events occur when DFS encounters a leaf node. From the sentence "Yesterday, John traveled to Salvador with his car / Yesterday, John traveled to Salvador with his car" in Figure 3, DptOIE was applied. Extractions can be seen in Table 3. We can verify that 4 triples were extracted. The fourth extraction was derived from the combination of Arg2 from triples 2 and 3, in order to generate a fact with more information. Another example can be seen from the sentence "Arsenal desperately seeks to reinforce its offensive system" (Figure 2), which DptOIE extracts the fact (The Arsenal desperately reinforcing their offensive system) / (The Arsenal, desperately seeks to strengthen their offensive system). Although the adverb (desperately / desperately) be a leaf node, DptOIE did not extract the fact (The Arsenal; desperately) / (The

[4] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2515 (Acessado em 07/20/2018)

[5] http://universaldependencies.org/format.html

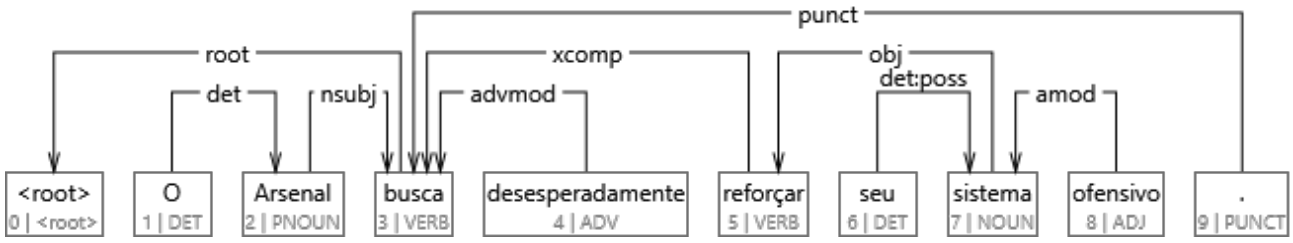


Figure 2: Sentence "O Arsenal busca desesperadamente reforçar seu sistema ofensivo / The Arsenal desperately seeks to strengthen their offensive system" parsed with DP

Table 2: Hand-crafted rules used by DptOIE to perform DFS.

Triple	**DFS - start**	**DFS - search**
Arg1	Subject (*nsubj / nsubj:pass*)	Modifiers{*nummod, advmod, nmod, amod*}, Others{*dep, obj, conj, appos* (must be labeled by POS tagger as "NUM")}
Rel	Node parent of the subject or child nodes labeled as "acl:part"	**Tokens before the parent node of the subject:** Auxiliar/Copular verbs {*aux, aux:pass, cop*}, Objects{*obj, iobj*}, Adverbial modifier{*advmod*}, Others{*expl:pv, mark*}; **Tokens after the parent node of the subject:** others{*expl:pv, acl:part*}
Arg2	relation root	Objects{*obj, iobj*}, Modifiers {*nmod, nummod, advmod, amod*}, Complement{*xcomp*}, Complement/Adverbial clause modifier{*ccomp and advcl* (as long as they don't have child nodes labeled as "*nsubj/nsubj:pass*")}, Others{*acl:relcl, conj, appos, acl:part, dep*}

Arsenal; seeks; desperately), as it omits critical information, which is " to strengthen their offensive system ". The same situation would occur if leaf node is an adjective.

3.3 Particular cases

This section presents the treatment of the particular cases performed by DptOIE, which are: Coordinated conjunctions, Subordinate clause and Appositive.

3.3.1 Coordinated conjunctions

A coordinate conjunction (CC) has the function of connecting sentences or words with the same syntactic level (Bechara, 2012; Sacconi, 2012). The dependency relations with conjunctions labeled as "conj". From conjunctions, DptOIE can perform particular treatments in the relation or in the second argument. Treatments of CC are applied only on conjunctions "e / and" and "ou / or". For instance, take the sentence "Eu compro e vendo banana, maçã e pera; / I buy and sell banana, apple and pear;" (Figure 5). Using the basic rules, DptOIE generates only the triple ("Eu"; "compro"; "banana , maçã e pera") / ("I"; "buy"; "banana, apple and pear"). By applying Coordinated conjunctions, DptOIE checks the token "vender", which is the child node of the subject's parent node ("compro"), was labeled as conjunction and if it is a verb. If this verification is valid, DptOIE can generate another triple: ("Eu"; "vendo"; "banana , maçã e pera") / ("I"; "sell"; "banana, apple and pear"). In addition, this module can perform another treatment on Arg2, because the same sentence contains an enumeration, "banana , maçã e pera / banana, apple and pear". From the dependency relations of these tokens, DptOIE can derive another six triples: ("Eu"; "compro"; "banana")/("I"; "buy"; "banana"), (Eu; compro; maçã)/("I"; "buy"; "apple"), ("Eu"; "compro"; "pera")/("I"; "buy"; "pear"), ("Eu"; "vendo"; "banana")/("I"; "sell"; "banana"), ("Eu"; "vendo"; "maçã")/("I"; "sell"; "apple"), ("Eu"; "vendo"; "pera")/("I"; "sell"; "pear").

3.3.2 Subordinate clause

A subordinate clause does not provide a complete thought. For this reason, it can not stand alone as a complete sentence. Subordinate clauses can be divided into: Adjective Clause, Adverb Clause and Noun Clause.

An adjective Clause is a dependent clause that act as an adjective, modifying the antecedent term. They are beginning with relative pronoun (Sacconi, 2012). Relative pronouns are usually

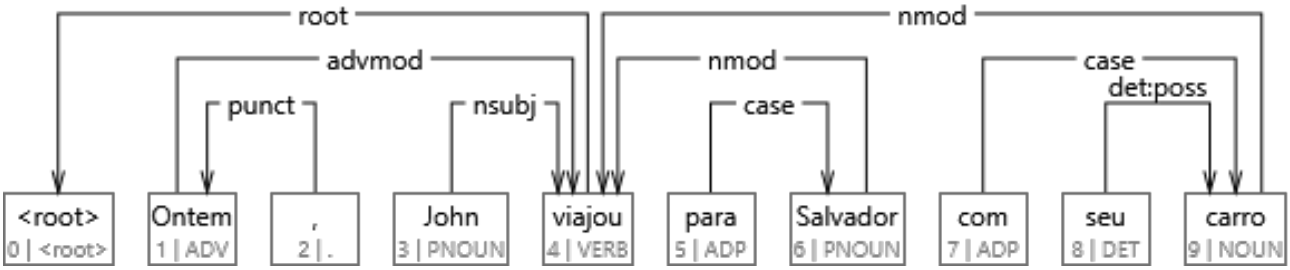


Figure 3: Sentence "Ontem, John viajou para Salvador com seu carro / Yesterday, John traveled to Salvador with his car" parser with the trained DP.

Table 3: Extracted facts by DptOIE: "Ontem, John viajou para Salvador com seu carro / Yesterday, John traveled to Salvador with his car"

Id	Arg1	Rel	Arg2
1	John	viajou/traveled	ontem/yesterday
2	John	viajou/traveled	para Salvador/to Salvador
3	John	viajou/traveled	com seu carro/with his car
4	John	viajou/traveled	para Salvador com seu carro/to Salvador with his car

labeled as subjects (*nsubj*), as was the case of the token "que" in Figure 6. By applying default rules, one of the triples that the DptOIE can extract is ("que"; "estava"; "em casa") / ("which"; "was"; "at home"). However, this extraction is not coherent. To solve this problem, the token "que / which" was replaced by "a espada / The sword", generating the following triple: (a espada; estava; em casa) / ("The sword"; "was"; "at home"). This substitution is performed every time the relative pronoun is labeled as a subject (nsubj) and its parent node is labeled with "acl:relcl", which in this case was the token "estava". Since the parent node of "estava" is the token "espada", then the DFS is started from it to find the argument that will replace the relative pronoun.

A noun clause is a dependent clause that acts as a noun. In most cases, the noun clause begins with the conjunction "que", as is the case of the sentence "Ele afirmou que o cidadão se comportou bem. / He said the citizen behaved well." (Figure 7). From this sentence, we can verify that it has two subjects, "Ele / He" and "cidadão / citizen". The first refers to the main clause and the second to the subordinate clause. In addition, it can be seen that the conjunction "que" was labeled with "mark" to mark it introduces a subordinate clause. Thus, by applying the default rules, DptOIE could extract only the triple: t_sub_clause = ("o cidadão"; "se comportou"; "bem") / ("the citizen"; "behaved"; "well"). However, this triple is a clause subordinated to the main clause "Ele afirmou / He said". Since it was not possible to obtain an Arg2 for the main clause, then the triple was not generated. DptOIE adds the conjunction "que / that" to the main clause and generates the triple t_main_clause = (Ele; afirmou; que) / (He; said; that). However, this triple is still not coherent. DptOIE goes futher and links the triple of the main clause to the triple of subordinate clause, generating the following fact composed of two triples: (Ele; afirmou; que) -> (o cidadão; se comportou; bem) / (He; said; that) -> ("the citizen"; "behaved"; "well"). It is important to note that the detection of the main and subordinate clauses was possible because there is a dependency relation, "ccomp", between the "afirmou / said" and "comportou / behaved" verbs of each clause. Furthermore, the linkage can be performed because both sentences have explicit subjects. If the sentence were "Ele afirmou que se comportaram bem. / He said they behaved well.", the method would only extract the triple ("Ele"; "afirmou"; "que se comportaram bem") / ("He"; "said"; "they behaved well").

DptOIE also deals with adverbial clauses with the same treatment as substantive clauses. The only difference is that instead of "ccomp", the connection between the main and subordinate clause is through the dependency relation between two verbs through "advcl".

3.3.3 Appositive

DptOIE can also derive new triples from sentences with appositive labeled as proper names (PNOUN), creating a synthetic clause with the verb "é / is". For example, in the sentence "O dire-

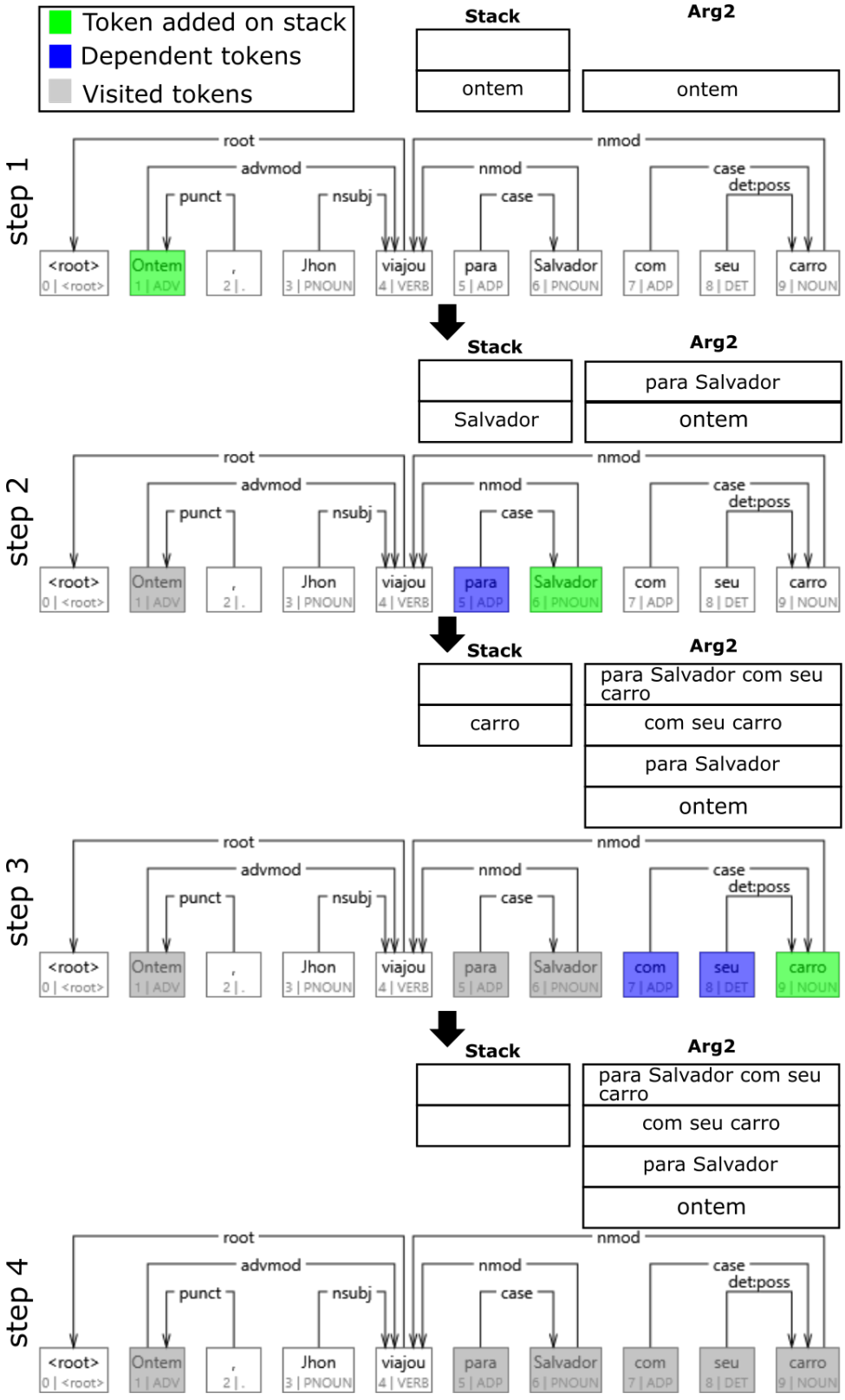


Figure 4: Detecting the *Arg2* of the sentence "Ontem, John viajou para Salvador com seu carro / Yesterday, John traveled to Salvador with his car" by means of an Depth-First Search, which start in the token "viajou".

tor do hospital , Júlio , vendeu sua fazenda. / The hospital's director, Júlio, sold his farm." (Figure 8), DptOIE extracts only triple t1 = ("O diretor de o hospital"; "vendeu"; "sua fazenda") / ("The hospital's director"; "sold"; "his farm"). However, in this sentence it is possible to verify that there is a relation between the appositive (appos) "Júlio" and the noun "director". Thus, DptOIE can gen-

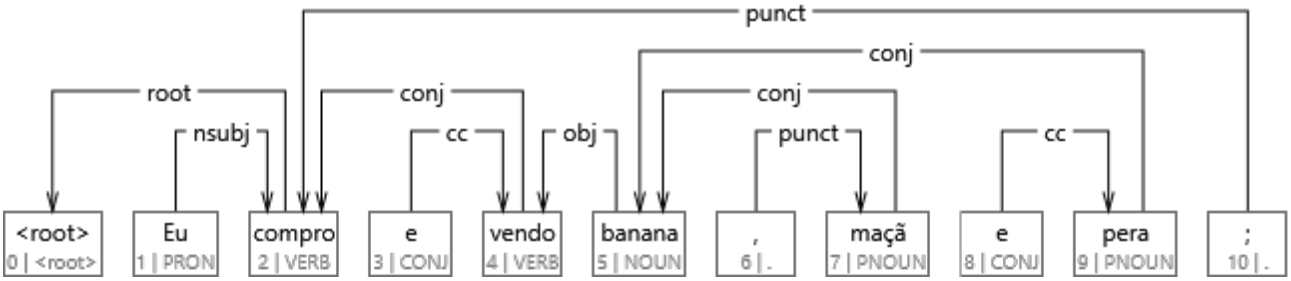


Figure 5: Sentence "Eu compro e vendo banana, maçã e pera; / I buy and sell banana, apple and pear" parsed with DP.

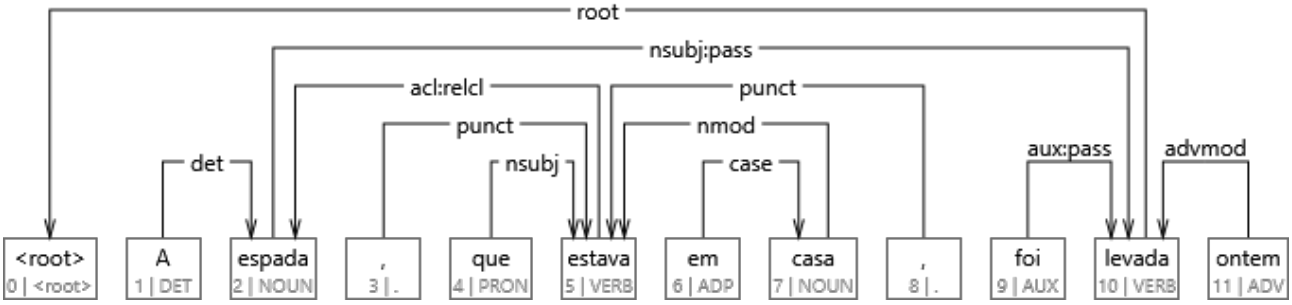


Figure 6: Sentence "A espada, que estava em casa, foi levada ontem / The sword, which was at home, was taken yesterday" parsed with DP.

erate a new triple, which is: t2 = ("O diretor do hospital"; "é"; "Júlio") / ("The hospital's director"; "is"; "Júlio"). In addition, DptOIE can create another triple through transitivity, since the appositive was considered equivalent to the element that it relates. Thus, a triple equivalent to t2 is: ("Júlio"; "é"; "O diretor de o hospital") / ("Júlio"; "is"; "The hospital's director"). Hence, if "Júlio é diretor do hospital / Julio is director of the hospital" and "O diretor do hospital vendeu sua fazenda / The director of the hospital sold his farm", through transitivity DptOIE generates triple t3 = ("Júlio"; "vendeu"; "sua fazenda") / ("Júlio"; "sold"; "his farm").

4 Experimental setup

In order to validate DptOIE, we compare DptOIE against the most recent version of ArgOE (Gamallo & Garcia, 2017)[6] and PragmaticOIE (Sena & Claro, 2018). Datasets to evaluate all methods were the same as those used in the PragmaticOIE evaluation. The first dataset was called CETEN200. It is a subset with 200 sentences randomly selected from the Corpus of Electronic Texts Extracts NILCS/Folha de Sao Paulo newspaper (CETENFolha) [7]. The second dataset has 200 randomly selected sentences from Wikipedia, called WIKI200. The comparison was based on the following criteria: (I) amount of extracted facts, (II) amount of coherent extracted facts (Yield), (III) amount of minimal extracted facts, (IV) precision of the methods, (V) precision x yield curve analysis and (VI) area under the curve.

Evaluation of facts in OIE is not simple since the idea of semantic relation is very broad. There is no standardization of how each method performs its extractions. For example, ArgOE can add objects in the relation, PragmaticOIE can create another argument to perform contextual treatments, DependentIE, ClausIE and DptOIE leave the preposition in Arg2, while other methods put it in the relation and so on.

To avoid harming the methods and having a fair evaluation, the concept of relationship was generalized. Thus, in this work, *a fact was considered coherent if the information expressed in it has a meaningful interpretation based on the sentence. In addition, relations must contain a verb and may be accompanied by nouns, pronouns, prepositions or adverbs.*

On the other hand, minimality was redefined as a fact that can not be decomposed into oth-

[6] https://github.com/citiususc/Linguakit. (05/29/2018).

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

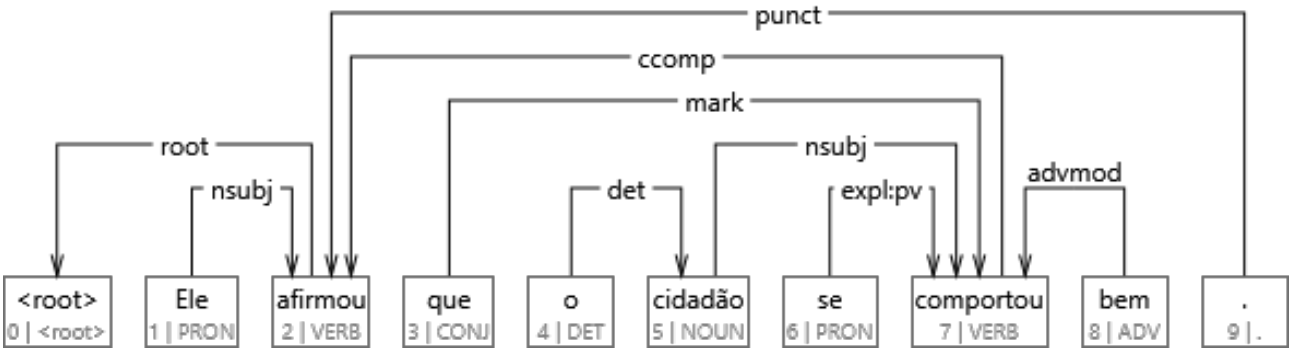


Figure 7: Sentence "Ele afirmou que o cidadão se comportou bem. / He said the citizen behaved well." parsed with DP.

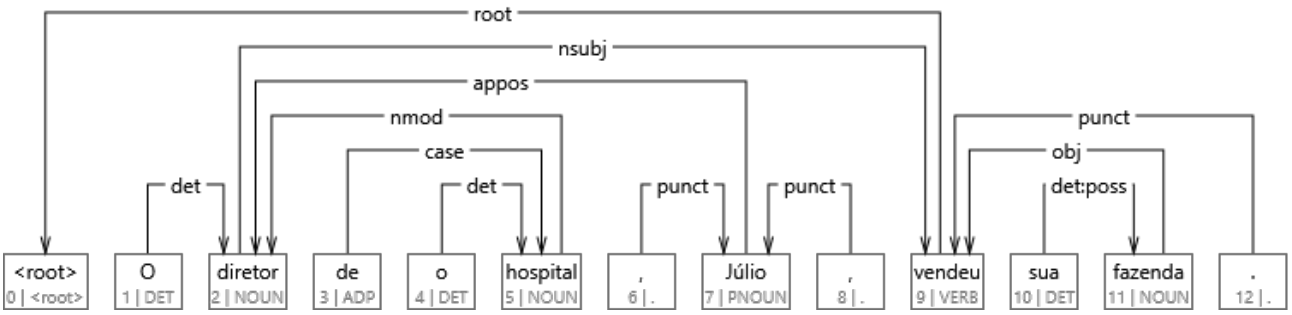


Figure 8: Sentence "O diretor do hospital, Júlio, vendeu sua fazenda. / The hospital's director, Júlio, sold his farm." parsed with DP.

ers from its arguments (Sena & Claro, 2018). For example, in the sentence *"Ele viajou para a Rússia no dia 19 de Junho. / He traveled to Russia on June 19."*, a method can extract ("Ele"; "viajou"; "para a Rússia no dia 19 de Junho") / ("He"; "traveled"; "to Russia on June 19"). This fact is coherent, but it could not be considered minimal, since it is still possible to derive other facts: ("Ele"; "viajou"; "para a Rússia") / ("He"; "traveled"; "to Russia") e ("Ele"; "viajou"; "on June 19") / ("He"; "traveled"; "para a Rússia"). Moreover, direct determinants and modifiers were not considered sufficient to make a fact not minimal in our evaluation. For example, in the sentence "os 4 rapazes unidos formam uma boa equipe. / the 4 boys together make a good team." the token "unidos / together" can be considered as an adverbial modifier and the following fact ("os 4 rapazes unidos"; "formam"; "uma boa equipe") / ("the 4 boys together"; "make"; "a good team") can be considered as minimal.

Facts extracted by each method were judged manually by two native Brazilian specialists and with knowledge in IE area. In this work, one fact was considered coherent or minimal if both experts agreed. It is worth mentioning that human beings may have different biases and interpretations. For this reason, Cohen's Kappa was used to evaluate the degree of agreement of the judges.

Furthermore, to identify the errors obtained by DptOIE, we propose to analyze the quantity of errors from each extracted fact in both datasets from the following points: (I) error in the basic module (without the treatment of particular cases), (II) error in subordinate clauses, (III) error in coordinate conjunctions, (IV) error in appositive, (V) DP error.

DptOIE modules were evaluated separately, as this allows us to analyze the impact they had on our results. Table 4 shows how this separation was made.

4.1 Materials

The tool used in pre-processing was Stanford CoreNLP v3.9.1 (Manning et al., 2014). CoreNLP does not have tools ready to be used to the Portuguese language. But, it offers an Application Programming Interface (API), which enable to train new tools for other languages. Thus, DP used was trained though the Brazilian Portuguese treebank V2.1 of Universal Dependencies, which comes from the dataset of the uni-

Table 4: Nomenclature used for the modules

Nomenclature	Modules evaluated
DptOIE_B	Default module
DptOIE_B_SC	DptOIE_B + Subordinate clause
DptOIE_B_SC_A	DptOIE_B_SC + appositive
DptOIE_B_SC_AT	DptOIE_B_SC_A + Transitivity
DptOIE_B_SC_AT_CC	DptOIE_B_SC_AT + Coordinated Conjunction

versal Google [8]. This treebank consists of texts from newspaper articles, blogs and consumer reviews. DP trained needs to annotate sentences with POS tagger, so we also trained a model to perform such task from the same treebank. It is worth mentioning that there are other POS Taggers available to Portuguese. Even so, we have chosen to train another to ensure that the output of it follows the universal format and CoreNLP, making it easier to manipulate the data. POS tagger receives separate sentences in tokens. Tokenizer used in this work was already trained to Spanish model of CoreNLP. It was chosen because it is a language of the same origin as Portuguese and the pattern followed is very close to the Brazilian treebank.

DP model was evaluated using the following metrics: *Labeled Attachment Accuracy* (LAS) e *Unlabeled Attachment Accuracy* (UAS). According to Jurafsky & Martin (2017), LAS and UAS is the most common way of evaluating dependency parsers. LAS represents the percentage of tokens with the parent node (HEAD) and the dependency relation correctly labeled. UAS represents the percent of tokens with only the parent node correctly labeled. Our model obtained LAS of 87.39% and UAS of 89.31%. The POS tagger obtained 96.89% of accuracy.

5 Results

We present in this section the performance of DptOIE separately as shown in Table 4, as well as the comparison against ArgOE and PragmaticOIE within the datasets: CETEN200 and WIKI200.

5.1 Evaluation on CETEN200

In this dataset, DptOIE_B extracted 557 facts of which 363 were considered coherent by the experts. DptOIE_B_SC extracted 727 facts, of which 467 were considered coherent. DptOIE_B_SC_A obtained 784 facts of which 502 were considered coherent. When adding the transitivity, DptOIE_B_SC_AT extracted 859 facts of which 541 were considered coherent. The full version of DptOIE, DptOIE_B_SC_AT_CC, obtained a total of 945 extracted facts of which 582 were considered coherent.

From Figure 9, we can verify that DptOIE was superior in almost all aspects, when compared against ArgOE and PragmaticOIE. The basic module the DptOIE performed approximately 1.93 times more coherent extractions than the other methods and our full version extracted about 3.09 times more facts. As for minimum facts, DptOIE presented results comparable to the other methods. Although most of the DptOIE facts are not minimal, this does not mean that the facts extracted by DptOIE are all too long. Many of the facts were not considered minimum by the judges because DptOIE added more information in the arguments.

Figure 10 shows Precision x Yield curve and AUC-PY of all methods. From it, we can observe that DptOIE outstands the other methods. The basic module (DptOIE_B) obtained 65.17% of precision. As new modules have been added, the precision has decreased because these modules have to deal with particular cases of Portuguese grammar and DP, which are more complex. Even so, the difference between precision of DptOIE_B and DptOIE_B_SC_AT_CC in this dataset was only 3.59%. ArgOE achieved a precision of 58,56% and PragmaticOIE 53,17%. In addition to these results, Figure 10 shows the area under the curve of the methods. DptOIE_B obtained approximately 2.2 times higher AUC-PY than the other methods, while the version with all modules obtained around 3.5 times.

This work also proposed the analysis of the failures of DptOIE. Table 5 presents the amount of coherent and incoherent facts that DptOIE modules produced. From it, it is verified that practically all modules presented more extractions coherent than incoherent ones. The exception was the CC module, which obtained 47 coherent and 51 incoherent facts.

In addition, the influence of DP was also an-

[8] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2515 (accessed November 19, 2018)

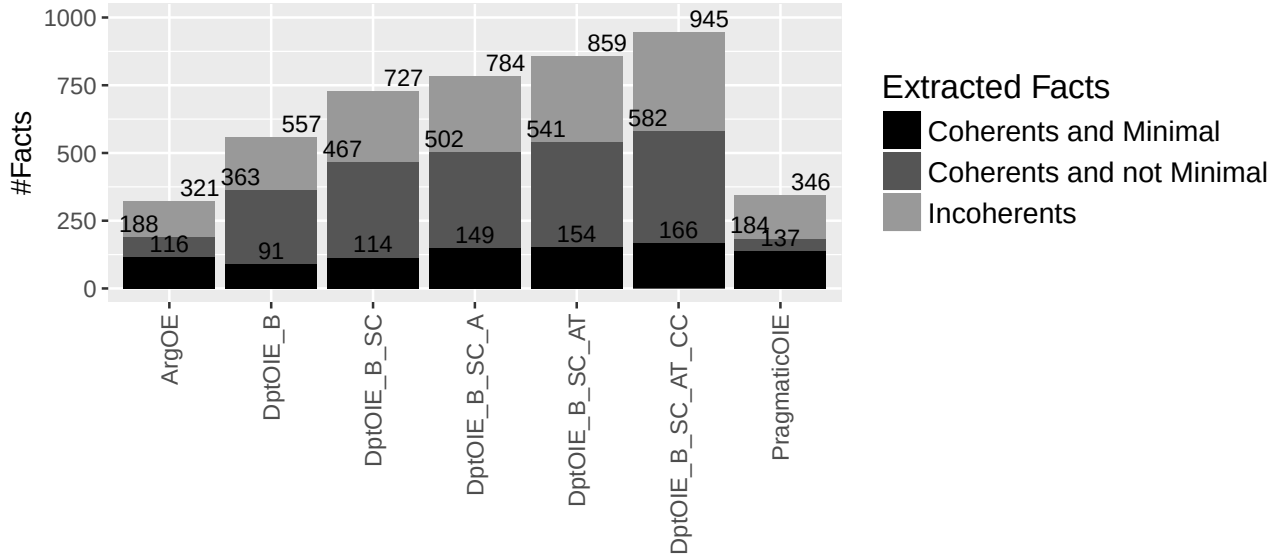
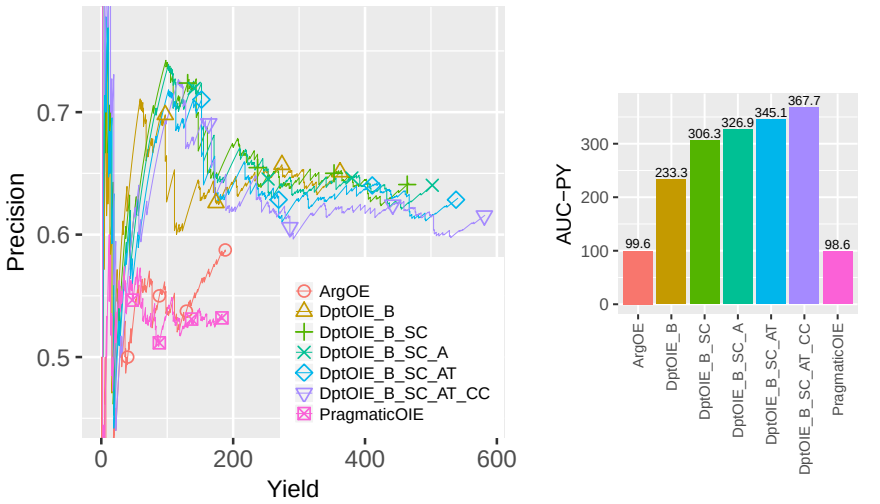


Figure 9: Results of the number of extracted facts, coherent extracted facts and minimum facts in dataset CETEN200



(a) Precision-Yield curve

(b) Area under P-Y curve (AUC-PY)

Figure 10: Results in CETEN200 dataset.

Table 5: Analysis of errors and correctness of DptOIE in CETEN200

Modules	**#Coherent facts**	**#Incoherent facts**
Basic	363	194
Coordinated Conjunction	47	51
Subordinate Clause	106	63
Appositive	35	22
Transitivity	58	33

alyzed through the incoherent facts, from Figure 11. In it, we found that DP interfered negatively, in particular, in appositive, Subordinate clause, and Basic module. Other errors are related to DptOIE itself, since the specified rules do not cover all possible combinations that Portuguese grammar can offer in a sentence.

In this dataset, the Kappa coefficients for the coherent and minimum facts were 0.879 and 0.836, respectively. This indicates that the evalu-

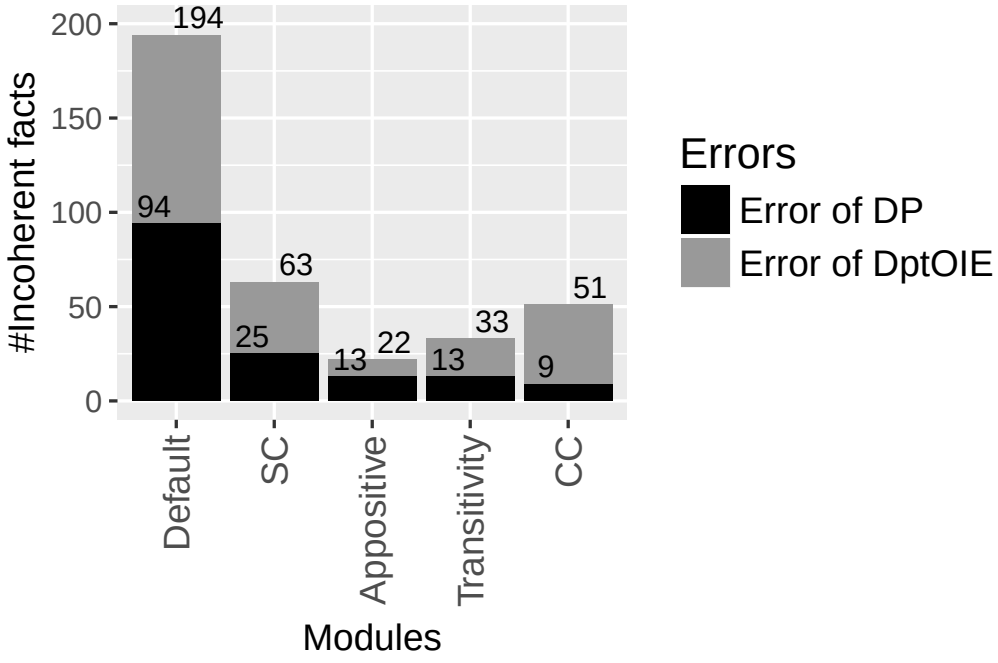


Figure 11: Influence of DP on incoherent facts of CETEN200

ation of the specialists in this dataset had a high degree of confidence, confirming the results obtained by the methods.

5.2 Evaluation on WIKI200

In this dataset, DptOIE_B extracted 720 facts, 522 of which were considered coherent by experts. DptOIE_B_SC extracted 789 facts of which 566 were considered coherent. DptOIE_B_SC extracted 789 facts, and 566 was coherent. When adding the transitivity DptOIE_B_SC_AT extracted 967 facts of which 625 were considered coherent. The full version of DptOIE, DptOIE_B_SC_AT_CC, obtained a total of 1063 facts extracted and 673 facts considered coherent.

From Figure 12, we can also verify that DptOIE was superior in practically all aspects. ArgOE had the worst performance. The basic version of DptOIE performed approximately 1,72 times more coherent extractions than PragmaticOIE, while the full version presented about 2,22 times more extractions. For minimal facts, PragmaticOIE had better performance, obtaining 230 minimal facts. The full version of DptOIE obtained a similar value, extracting 3 minimum facts less than PragmaticOIE.

From Precision x Yield curve and AUC-PY, it is observable that DptOIE obtained the best performance in the basic module (DptOIE_B), which obtained 72.50% precision at the end of the experiment. As new modules were added, precision decreased with DP being one of the factors that had the greatest negative influence on this dataset. PragmaticOIE achieve precision comparable to DptOIE_B, 73,72%. However, due to the fact that DptOIE_B extracted more facts, its AUC-PY was 394,47, whereas the AUC-PY of PragmaticOIE was 231,16. Despite having the least precision, AUC-PY of the full version of DptOIE was 457.88, as it presented 370 more coherent facts than PragmaticOIE.

Table 6 presents the amount of coherent and incoherent facts that DptOIE produced in WIKI200. From this, we can verify that the transitivity treatment module only extracted 15 coherent facts and 62 incoherent facts. This was due to the fact that DP had a high error rate in the appositive module (Figura 14). Of the 62 incoherent facts extracted by the transitivity, the DP negatively influenced in 56. The CC module in the WIKI200 presented a performance similar to CETEN200. It extracted 48 coherent facts and 55 incoherent facts, of which 21 was extracted wrongly because of DP. The other modules presented more coherent than incoherent extractions. However, the results could be better if the DP performance was higher.

In this dataset, Kappa coefficients for coherent and minimum facts were 0.914 and 0.892, respectively. This indicates that the evaluation of the experts in this dataset had a high degree of confidence, confirming, once again, the results obtained.

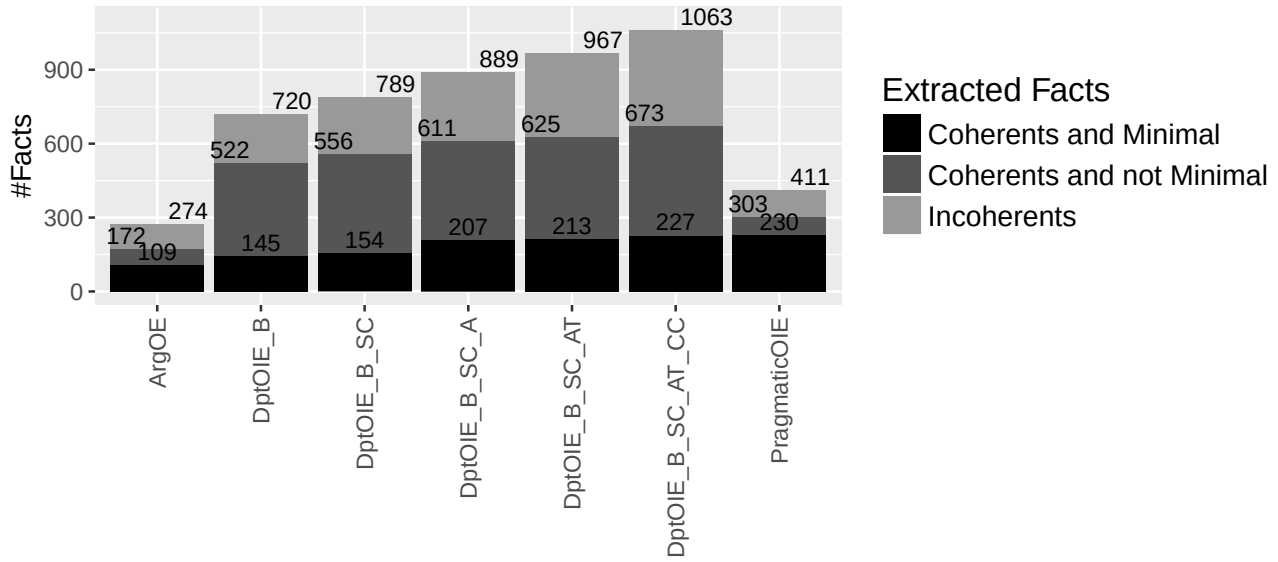
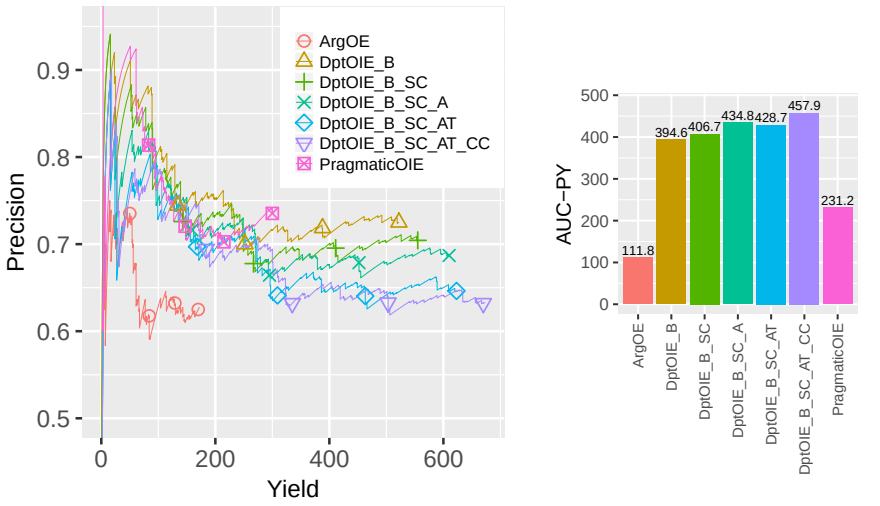


Figure 12: Results of the number of extracted facts, coherent extracted facts and minimum facts in dataset WIKI200



(a) Precision-Yield (P-Y) curve

(b) Area under P-Y curve (AUC-PY)

Figure 13: Results in WIKI200 dataset.

6 Discussion

In this section, we discuss some of the extracted facts by DptOIE, ArgOE and PragmaticOIE. Two sentences were manually selected. The extracted facts can be seen from Table 7. The first sentence (S1), PragmaticOIE extracted 3 facts, having one of them considered coherent, the fact A3. ArgOE also extracted 3 facts. All of them were considered coherent, but just the fact A4 was considered minimal. DptOIE extracted 4 coherent facts, of which just the fact A9 was considered minimal. Fact A8->A8.1 was derived from the module that treats subordinate clauses, A9 fact was derived from appositive and transitivity, DptOIE was able to extract the fact A10-> A10.1.

When analyzing S2 from Table 8, PragmaticOIE extracted 2 coherent facts that were considered not minimal. ArgOE performed no extraction in this sentence. DptOIE extracted 9 facts, of which 5 were considered coherent and 2 were minimal. Facts B3, B4 and B5 were extracted by the basic module and facts B6 and B7 were generated by the CC module. The fact B8 was derived from the appositive module. It was considered incoherent because the DP labeled "Williamsport" as an appositive, where it

Table 6: Analysis of errors and correctness of DptOIE in WIKI200

Modules	**#Coherent facts**	**#Incoherent facts**
Basic	522	199
Coordinated Conjunction	48	55
Subordinate Clause	34	28
Appositive	55	45
Transitivity	15	62

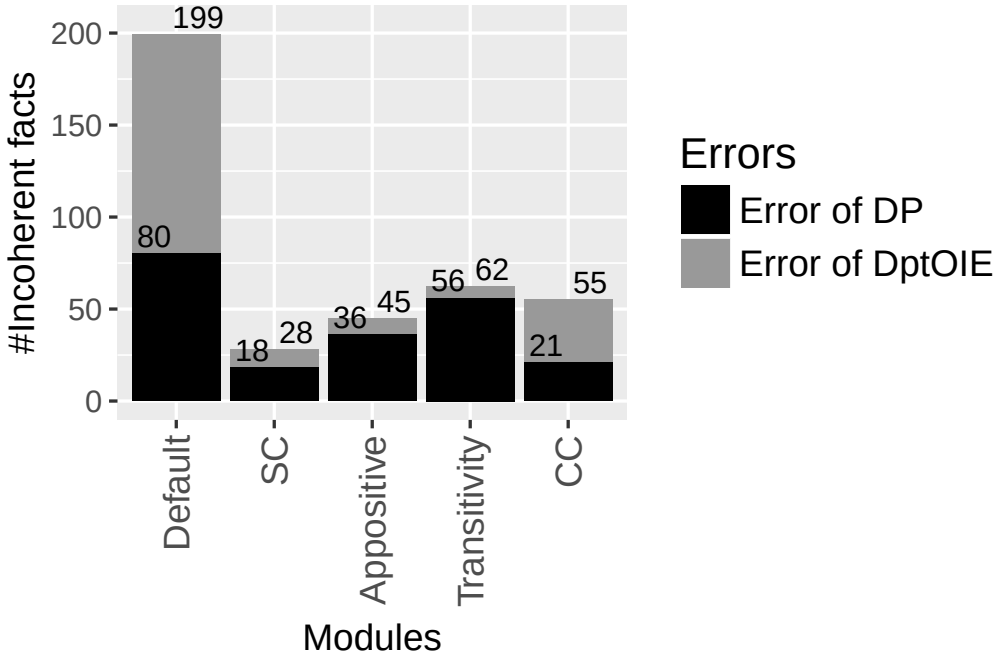


Figure 14: Influence of DP on incoherent facts of WIKI200

is associated with the birthplace of "Joanna Dove Hayes". Because of this, DptOIE extracted the incoherent facts B9, B10 and B11 by applying the transitivity. Many sentences had similar characteristics to the sentence S2 in WIKI200, which potentiated the amount of incoherent facts generated by the transitivity because of the DP error.

We have observed that the treatment of the coordinated conjunctions in this work was a challenge and this reflected in the results. Part of the problems were related to various situations and combinations that can occur in the sentences with the CCs. For example, in the sentence "Os integrantes dessa força, os primeiros de nacionalidade não americana a chegarem ao Haiti, são de Antígua, Barbados, Belize, Jamaica e Trinidad - Tobago. / Members of this force, the first non - American nationals to arrive in Haiti, are from Antigua, Barbados, Belize, Jamaica and Trinidad - Tobago." of CETEN200 an enumeration occurs and by applying the CC module, one of the facts that DptOIE extracts was ("Os integrantes dessa força"; "são"; "de Antígua") / ("Members of this force"; "are"; "from Antigua"). This fact was considered incoherent by the judges, because in this case it should not be decomposed. Other problems with CC can occur in sentences that present figures of speech, like *elipse*, in which there is an omission of a term that can be implied in the text. Part of the other problems are related to DP errors. Despite the difficulties presented, the number of coherent facts would be reduced by approximately 7% if the CC module was not used in both datasets.

DptOIE outstands the other extractors when analyzed the amount of coherent facts extracted and the AUC-PY in both datasets. In precision, DptOIE performed well in CETEN200 and was comparable to PragmaticOIE in WIKI200, despite the tradeoff between precision and yield. As for the quantity of minimal facts, results of DptOIE were similar to PragmaticOIE. Moreover, another factor to be analyzed is the variation of the behavior of the methods from the two datasets. The precision of PragmaticOIE varied much from CETEN200 to WIKI200 (20,55%), whereas the DptOIE, in the version that most os-

S1 - O gerente da empresa, Hilton Naetzke, disse que a empresa estava providenciando sua transferência para o parque industrial da cidade. / The Company manager, Hilton Naetzke, said that the company was arranging its transfer to the city's industrial park. (CETEN200)

System	Id	Facts (pt-br/en)	C	M
PragmaticOIE	A1	("O gerente da empresa"; "disse que"; "a empresa") / ("The Company manager"; "said that"; "the company")	×	-
	A2	("O gerente da Hilton_Naetzke"; "disse que"; "a empresa") / ("The Hilton Naetzke manager"; "said that"; "the company")	×	-
	A3	("a empresa"; "estava providenciando sua transferência para"; "o parque industrial da cidade") / ("the company"; "was arranging its transfer to"; "the city's industrial park")	✓	✓
ArgOE	A4	("a empresa"; "estava providenciando"; "sua transferência") / ("the company"; "was arranging"; "its transfer")	✓	✓
	A5	("a empresa"; "estava providenciando sua transferência para"; "o parque industrial de a cidade") / ("the company"; "was arranging its transfer to"; "the city's industrial park")	✓	✓
	A6	("O gerente de a empresa"; "disse"; "que a empresa estava providenciando sua transferência para o parque industrial de a cidade") / ("The Company manager"; "said that"; "the company was arranging its transfer to the city's industrial park")	✓	×
DptOIE	A7	("a empresa"; "estava providenciando"; "sua transferência para o parque industrial de a cidade") / ("the company"; "was arranging its transfer"; "to the city's industrial park")	✓	×
	A8 A8.1	("O gerente de a empresa"; "disse"; "que") − > ("a empresa"; − > "estava providenciando"; "sua transferência para o parque industrial de a cidade") / ("The Company manager"; "said"; "that") − > ("the company"; "was arranging its transfer"; "to the city's industrial park")	✓	×
	A9	("O gerente de a empresa"; "é"; "Hilton Naetzke") / ("The Company manager"; "is"; "Hilton Naetzke")	✓	✓
	A10 A10.1	("Hilton Naetzke"; "disse"; "que") − > ("a empresa"; "estava − > providenciando"; "sua transferência para o parque industrial de a cidade") / ("Hilton Naetzke"; "said"; "that") − > ("the company"; "was arranging its transfer"; "to the city's industrial park")	✓	×

Table 7: Extracted facts by PragmaticOIE, ArgOE and DptOIE from Sentence 1. Letter "C" stands for "coherence" and letter "M" for "minimality"

cillated (DptOIE_D), the variation of precision was 7,33%. Authors of PragmaticOIE have argued that the precision on CETEN200 was low because it is a dataset that presents sentences with more complex structures, with excessive use of punctuations such as commas and semicolons, that makes it the method with the worst performance in this dataset. In contrast, DptOIE did not suffer so much from the complexity of the sentences in CETEN200, since it was 8.41-12% more accurate than PragmaticOIE. This shows that our method has been less sensitive to writing style, which can give more confidence in multidomain applications with diverse writing styles.

As for ArgOE, DptOIE was superior to it in all respects, regardless of the dataset or module. On the other hand, we understand the importance of multilingual systems, since they cover a greater number of languages and can extract more information. However, many of the languages covered by ArgOE already have specific OIE methods that often have better results in the treated language. So, in a real application, it may be more interesting to use a set of methods that treat specific languages along with PLN techniques to detect languages. This shows that specific languages methods also has great relevance.

S2 - Joanna Dove Hayes (Williamsport, 23 de dezembro de 1976) é um atleta barreirista e campeã olímpica norte-americana. / Joanna Dove Hayes (Williamsport, December 23, 1976) is a hurdles athlete and Olympic champion North-American. (WIKI200)

System	Id	Facts (pt-br/en)	C	M
PragmaticOIE	B1	(“Joanna_Dove_Hayes”; “é um atleta barreirista”; “campeã olímpica”) / (“Joanna_Dove_Hayes”; “is a hurdles athlete”; “olympic champion”)	✓	✓
	B2	(“um atleta barreirista e campeã olímpica”; “é”; “Joanna_Dove_Hayes”) / (“a hurdles athlete and Olympic champion”; “is”; “Joanna_Dove_Hayes”)	✓	×
ArgOE	-	-	-	-
DptOIE	B3	(“Joanna Dove Hayes”; “é um atleta”; “barreirista e campeã olímpica norte - americana”) / (“Joanna Dove Hayes”; “is a athlete”; “hurdler and olympic champion North-American”)	✓	×
	B4	(“Joanna Dove Hayes”; “é”; “um atleta”) / (“Joanna Dove Hayes”; “is”; “a athlete”)	✓	✓
	B5	(“Joanna Dove Hayes”; “é”; “barreirista e campeã olímpica”) / (“Joanna Dove Hayes”; “is”; “hurdler and olympic champion North-American”)	✓	×
	B6	(“Joanna Dove Hayes”; “é um atleta”; “barreirista”) / (“Joanna Dove Hayes”; “is a athlete”; “hurdler”)	✓	✓
	B7	(“Joanna Dove Hayes”; “é um atleta”; “campeã olímpica norte - americana”) / (“Joanna Dove Hayes”; “is a athlete”; “olympic champion North-American”)	✓	×
	B8	(“Joanna Dove Hayes”; “é”; “(Williamsport , 23 de dezembro de 1976)”) / (“Joanna Dove Hayes”; “is”; “(Williamsport, December 23, 1976)”)	×	-
	B9	(“(Williamsport , 23 de dezembro de 1976)”; “é um atleta”; “barreirista e campeã olímpica norte - americana”) / (“(Williamsport, December 23, 1976)”; “is a athlete”; “hurdler and olympic champion North-American”)	×	-
	B10	(“(Williamsport , 23 de dezembro de 1976)”; “é um atleta”; “barreirista”) / (“(Williamsport, December 23, 1976)”)”; “is a athlete”; “hurdler”)	×	-
	B11	(“(Williamsport , 23 de dezembro de 1976)”; “é um atleta”; “campeã olímpica norte - americana”) / (“(Williamsport, December 23, 1976)”)”; “is a athlete”; “olympic champion North-American”)	×	-

Table 8: Extracted facts by PragmaticOIE, ArgOE and DptOIE from Sentence 2. Letter “C” stands for “coherence” and letter “M” for “minimality”

7 Conclusion and future work

In this work, we present DptOIE, an OIE system for Portuguese. Our method uses stanford’s DP, specific rules for extracting facts in the Portuguese language, an adaptation of a DFS to explore the dependency tree, and modules to handle particular cases in sentences with coordinate conjunctions, subordinate clauses, and appositives. Furthermore, DptOIE is open to other dependency parsers, ever since it is provided sentences in CoNLL-U and Universal Dependencies v2.1 Brazilian treebank format. Our method and trained models are available at FORMAS[9]. DptOIE was compared against ArgOE and PragmaticOIE from datasets with sentences from journalistic texts and encyclopedias. We believe that the use of dependency analysis and specific rules can increase the precision and quantity of coherent facts in the Portuguese language, the same way the OIE methods in English. Our results confirmed that DptOIE extracted more information, obtained high precision and high AUC-PY. Despite this, DP used still presented many errors. We believe that potential improvements in

[9] http://formas.ufba.br/page/downloads

the dependency parser and training datasets can lead to an increase in precision and yield. In future, we intend to adapt our approach to extract n-ary facts, in order to better segment the sentences, preserving all the textual information and avoiding that the arguments becomes excessively long.

Acknowledgements

Authors would like to thank FAPESB for financial support. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

References

de Abreu, Sandra Collovini, Tiago Luis Bonamigo & Renata Vieira. 2013. A review on relation extraction with an eye on portuguese. *Journal of the Brazilian Computer Society* 19(4). 553–571.

Akbik, Alan & Jügen Broß. 2009. Wanderlust: Extracting semantic relations from natural language text using dependency grammar patterns. Em *SemSearch Workshop day at World Wide Web Conference (WWW2009)*, vol. 48, .

Akbik, Alan & Alexander Löser. 2012. Kraken: N-ary facts in open information extraction. Em *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, 52–56. Association for Computational Linguistics.

Banko, Michele, Michael J Cafarella, Stephen Soderland, Matthew Broadhead & Oren Etzioni. 2007. Open information extraction from the web. Em *IJCAI*, vol. 7, 2670–2676.

Barion, Eliana Cristina Nogueira & Decio Lago. 2015. Mineração de textos. *Revista de Ciências Exatas e Tecnologia* 3(3). 123–140.

Bassa, Akim, Mark Kroll & Roman Kern. 2018. Gerie-an open information extraction system for the german language. *Journal of Universal Computer Science* 24(1). 2–24.

Bast, Hannah & Elmar Haussmann. 2013. Open information extraction via contextual sentence decomposition. Em *Semantic Computing (ICSC), 2013 IEEE Seventh International Conference on*, 154–159. IEEE.

Bechara, Evanildo. 2012. *Moderna gramática portuguesa*. Nova Fronteira.

Cimiano, Philipp & Johanna Wenderoth. 2005. Automatically learning qualia structures from the web. Em *Proceedings of the ACL-SIGLEX workshop on deep lexical acquisition*, 28–37. Association for Computational Linguistics.

Collovini, Sandra, Gabriel Machado & Renata Vieira. 2016. Extracting and structuring open relations from portuguese text. Em *International Conference on Computational Processing of the Portuguese Language*, 153–164. Springer.

Cui, Lei, Furu Wei & Ming Zhou. 2018. Neural open information extraction. *CoRR* abs/1805.04270. http://arxiv.org/abs/1805.04270.

Del Corro, Luciano & Rainer Gemulla. 2013. Clausie: clause-based open information extraction. Em *Proceedings of the 22nd international conference on World Wide Web*, 355–366. ACM.

Fader, Anthony, Stephen Soderland & Oren Etzioni. 2011. Identifying relations for open information extraction. Em *Proceedings of the conference on empirical methods in natural language processing*, 1535–1545. Association for Computational Linguistics.

Gamallo, Pablo & Marcos Garcia. 2015. Multilingual open information extraction. Em *Portuguese Conference on Artificial Intelligence*, 711–722. Springer.

Gamallo, Pablo & Marcos Garcia. 2017. Linguakit: uma ferramenta multilingue para a análise linguística e a extração de informação. *Linguamática* 9(1). 19–28.

Gamallo, Pablo, Marcos Garcia & Santiago Fernández-Lanza. 2012. Dependency-based open information extraction. Em *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*, 10–18. Association for Computational Linguistics.

Garcia, Marcos & Pablo Gamallo. 2014. Entity-centric coreference resolution of person entities for open information extraction. *Procesamiento del Lenguaje Natural* 53. 25–32.

Glauber, Rafael & Daniela Barreiro Claro. 2018. A systematic mapping study on open information extraction. *Expert Systems with Applications* 112. 372–387.

Glauber, Rafael, Leandro Souza de Oliveira, Cleiton Fernando Lima Sena, Daniela Barreiro Claro & Marlo Souza. 2018. Challenges of an annotation task for open information extraction in portuguese. Em *International Conference on Computational Processing of the Portuguese Language*, 66–76. Springer.

Jurafsky, Dan & James H. Martin. 2017. Vector semantics. Em Dan Jurafsky & James H. Martin (eds.), *Speech and Language Processing*, chap. 6, 101–130. https://web.stanford.edu/ jurafsky/slp3/ed3book.pdf: Prentice Hall 3rd edn. Draft of September 23, 2018.

Manning, Christopher D, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard & David McClosky. 2014. The stanford corenlp natural language processing toolkit. Em *ACL (System Demonstrations)*, 55–60.

Nivre, Joakim, Johan Hall & Jens Nilsson. 2006. Maltparser: A data-driven parser-generator for dependency parsing. Em *Proceedings of LREC*, vol. 6, 2216–2219.

Oliveira, Leandro, Rafael Glauber & Daniela Barreiro Claro. 2017. Dependentie: An open information extraction system on portuguese by a dependence analysis. Em *ENIAC - 2017 XIV Encontro Nacional de Inteligência Artificial e Computacional*, http://comissoes.sbc.org.br/ce-ia/pg/historico/?file=ENIAC-2017|Anais-ENIAC-2017.pdf.

Pereira, Victor & Vládia Pinheiro. 2015. Reportum sistema de extração de informações aberta para língua portuguesa (report-an open information extraction system for portuguese language). Em *Proceedings of the 10th Brazilian Symposium in Information and Human Language Technology*, 191–200.

Rodríguez, Juan M, Hernán D Merlino, Patricia Pesado & Ramón García-Martínez. 2016. Performance evaluation of knowledge extraction methods. Em *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, 16–22. Springer.

Sacconi, Luiz Antonio. 2012. *Gramática para todos os cursos e concursos -teoria e prática.* Nova Geração 5th edn.

Schmitz, Michael, Robert Bart, Stephen Soderland, Oren Etzioni et al. 2012. Open language learning for information extraction. Em *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 523–534. Association for Computational Linguistics.

Sena, Cleiton F. L. & Daniela Barreiro Claro. 2019. Inferportoie: A portuguese open information extraction system with inference. *Natural Language Engineering* 25. 287–306. doi:10.1017/S135132491800044X. https://doi.org/10.1017/S135132491800044X.

Sena, Cleiton Fernando Lima & Daniela Barreiro Claro. 2018. Pragmatic information extraction in brazilian portuguese documents. Em Aline Villavicencio, Viviane Moreira, Alberto Abad, Helena Caseli, Pablo Gamallo, Carlos Ramisch, Hugo Gonçalo Oliveira & Gustavo Henrique Paetzold (eds.), *Computational Processing of the Portuguese Language*, 46–56. Cham: Springer International Publishing.

Sena, Cleiton Fernando Lima, Rafael Glauber & Daniela Barreiro Claro. 2017. Inference approach to enhance a portuguese open information extraction. Em *Proceedings of the 19th International Conference on Enterprise Information Systems (ICEIS)*, vol. 1, 442–451. INSTICC ScitePress. doi:10.5220/0006338204420451.

Stanovsky, Gabriel & Ido Dagan. 2016. Creating a large benchmark for open information extraction. Em *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2300–2305.

Wu, Fei & Daniel S Weld. 2010. Open information extraction using wikipedia. Em *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 118–127. Association for Computational Linguistics.

Xavier, Clarissa Castellã, Vera Lúcia Strube de Lima & Marlo Souza. 2013. Open information extraction based on lexical-syntactic patterns. Em *Intelligent Systems (BRACIS), 2013 Brazilian Conference on*, 189–194. IEEE.