Inference Approach to Enhance a Portuguese Open Information Extraction

Cleiton Fernando Lima Sena, Rafael Glauber and Daniela Barreiro Claro
FORMAS – Semantic Formalisms and Applications Research Group,
LaSiD/DCC/IME – Federal University of Bahia (UFBA),
Av. Adhemar de Barros, s/n, Campus de Ondina, Salvador, Bahia, Brazil

Keywords: Open Information Extraction, Inference, Transitivity, Symmetry, Portuguese.

Abstract: Open Information Extraction (Open IE) enables the extraction of facts in large quantities of texts written in natural language. Despite the fact that almost research has been doing in English texts, methods and techniques for other languages have been less frequent. However, those languages other than English correspond to 48% of content available on websites around the world. In this work, we propose a method for extracting facts in Portuguese without pre-determining the types of the facts. Additionally, we increased the quantity of those extracted facts by the use of an inference approach. Our inference method is composed of two issues: a transitive and a symmetric mechanism. To the best of our knowledge, this is the first time that inference approach is used to extract facts in Portuguese texts. Our proposal allowed an increase of 36% in quantity of valid facts extracted in a Portuguese Open IE system, and it is compatible in the quality of facts with English approaches.

1 INTRODUCTION

The volume of data available on the Web grows every day, and much of this data is accessible in natural language. The task of extracting relevant information from the Web has become difficult for humans due to the huge volume of data. Considering text written in natural language, this is even harder to be automated. In this direction, the Information Extraction (IE) emerged as a research area to identify relevant patterns in large quantities of textual documents (Soderland, 1999). The tasks employed by IE were carried out in specific, homogeneous and previously established domains. As a consequence, a first challenge was to scale traditional IE to the Web (Banko et al., 2007). However, some drawbacks were considered, such as low coverage of relations and human intervention for new relations. To overcome such limitations, a new paradigm called *Open Information Extraction* (Open IE) comes up which extracts information freely from texts and scales for the Web (Schmitz et al., 2012).

About 52% of the content available on websites are written in English[1], while other languages complement the list. Approaches such as: TextRunner (Etzioni et al., 2008), WOE (Wu and Weld, 2010), Reverb (Fader et al., 2011), OLLIE (Schmitz et al., 2012), ClausIE (Del Corro and Gemulla, 2013) and Stanford Open IE (Angeli et al., 2015) are examples of solutions that only act in English texts. Few initiatives for other languages are proposed, such as for Chinese, with the CORE (Tseng et al., 2014) and the ZORE (Qiu and Zhang, 2014) systems and for French, with (Gotti and Langlais, 2016) system. There are also initiatives for systems that perform their tasks in different languages such as ArgOE (Gamallo and Garcia, 2015). However, this type of approach does not deliver impressive results for languages other than English. Thus, there is an immeasurable amount of knowledge out of English written world that is still unexplored.

The low informativeness of the facts extracted from Open IE approaches has been emerging as a challenge. In the sentence, i.e. *Itabuna is a city of Bahia located in the Northeast* current methods can extract information such as: (*Itabuna, is a city of, Bahia*) and/or (*a city of Bahia, located in, Northeast*). However, the fact *Itabuna, located in, Northeast* could be extracted from a rule of transitive inference. In the work of (Bast and Haussmann, 2014) this

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

Sena, C., Glauber, R. and Claro, D.
Inference Approach to Enhance a Portuguese Open Information Extraction.
DOI: 10.5220/0006338204420451
In *Proceedings of the 19th International Conference on Enterprise Information Systems (ICEIS 2017) - Volume 1*, pages 442-451
ISBN: 978-989-758-247-9

problem is addressed. However, they employed only for English language. Moreover, their work is limited to the transitive approach to infer. On the other hand, our approach presents a method for Open IE that i) extracts facts in Portuguese texts, ii) extracts a large quantity of relevant facts and iii) proposes an inference by both transitive and symmetric.

The remainder of this work is organized as follows: Section 2 presents our proposal in detail and Section 3 describes how our method is constructed. In Section 4 we present our experimental method that aims to define the configurations of our experiments. Then, Section 5 describes our results comprising our approach with some related works. We discusses some of our extractions in Section 6 and we list some envision trends of this research domain in Section 7.

2 BACKGROUND

Open IE is a paradigm that enables the discovery of facts in a large set of textual documents adjusting to their size and diversity (Banko et al., 2007). In this paradigm there is no need for a prior specification of the facts to be extracted (Fader et al., 2011). For example, considering the sentence *The table is in the center of the room*. The fact (The table, is in the center of the, room) would be extracted without the prior specification of the relation *in the center of the* or even the arguments *The table* and *room*. The main features of Open IE are: i) domain independence, ii) unsupervised extraction and iii) scalability for a large and varied number of texts (Del Corro and Gemulla, 2013). A fact extracted through Open IE is composed by the attribute of a relation between a pair of entities (Faruqui and Kumar, 2015) defined in the form of a triple $t = (e_1, rel_1, e_2, rel_2, e_3, ..., e_{n-1}, rel_n, e_n)$. Where $rel_1, rel_2, ..rel_{n-1}$ corresponds to the relations between the $(e_1, e_2), (e_2, e_3), ..., (e_{n-1}, e_n)$.

2.1 Related Work

The first Open IE system was TextRunner (Banko et al., 2007) (Etzioni et al., 2008) (Banko, 2009) which uses a self-supervised approach to train its own examples of how relationships are expressed in English. After TextRunner, new other systems emerged, such as WOE(Wu and Weld, 2010). It also uses self-supervised learning from heuristic combinations of attribute values from infoboxes of Wikipedia and their respective sentences. WOE operates in two modes: WOE*pos* which uses Part-of-Speech tagger (POS tagger) and WOE*parse* parsers using Dependence Parser (DP). Although WOE*parse* was more accurate in the task, the computational cost was 30x higher compared to WOE*pos* due to the use of a dependency analyzer. The second generation of Open IE systems left the stage of learning the patterns that express relationships. ReVerb (Fader et al., 2011) is the main representative of this approach that uses syntactic and lexical constraints to extract arguments and relations expressed by verbs in English sentences. One of the objectives of the method used by ReVerb was to reduce the number of incoherent extractions (relations without meaningful interpretation, incomprehensible) in comparison to its predecessors. According to Tab. 1, we see some examples of incoherent extractions possibly generated by first generation systems. This problem occurs because the extractors trained in the TextRunner and WOE methods take a set of decisions considering each word of the sentence to form the relational phrase.

Table 1: Examples of incoherent extractions (Fader et al., 2011).

Sentence	Incoherent Relation
The guide *contains* dead links and *omits* sites.	contains omits
The Mark 14 *was central* to the *torpedo* scandal of the fleet.	was central torpedo
They *recalled* that Nungesser *began* his career as a precinct leader.	recalled began

In Tab. 2 we have examples of extractions with the omission of useful information. The lexical constraint proposed by the ReVerb method has the aim of reducing this type of error. This type of problem is occasioned by the improper handling of relations that are expressed using a combination of a verb with a noun.

Table 2: Examples of uninformative extractions (Fader et al., 2011).

is	is an album by, is the author of, is a city in
has	has a population of, has a Ph.D. in, has a cameo in
made	made a deal with, made a promise to
took	took place in, took control over, took advantage of
gave	gave birth to, gave a talk at, gave new meaning to
got	got tickets to, got a deal on, got funding from

After Reverb, a new generation of methods began to use DP between the morphological classes of words and a set of rules for detecting useful parts in sentences (clauses). An example of this approach is DepOE (Gamallo et al., 2012) which supports English, Spanish, Portuguese and Galician languages and performs its task through three main steps. The first is dependency analysis, in which each input statement is analyzed by the multilingual DP tool *DepPattern* (Gamallo et al., 2012). Next, the method proposes the identification of verbal clauses and the determination of the function of each of them. Ultimately, a set of rules is applied in the constituent clauses in order to extract the triples $t = (arg_1, rel, arg_2)$.

OLLIE (Schmitz et al., 2012) is another Open IE system that uses dependency analysis for verbal extractions, but also proposes extractions mediated by nouns or adjectives. Clausie (Del Corro and Gemulla, 2013) and CSD-IE (Bast and Haussmann, 2013) use dependency analysis for the decomposition of sentences into clauses and contextual clusters, respectively. Despite the fact that Open IE still produces a considerable number of uninformative facts (Bast and Haussmann, 2014), the inference approach can enhance those approaches by generating new useful facts.

2.2 Inference

Inference is a mechanism used to deduce the veracity of a proposition based on its connection with other propositions already assumed to be true (Schoenmackers et al., 2008). The validity and veracity of inference are conditioned to its form and the truth value of its premises, respectively. Transitivity and symmetric are two of the possibilities to infer something.

2.2.1 Transitivity

A sentence has a transitive relation if it follows the pattern: If *A relation B* and *B relation C* then *A relation C*. Considering the same example: *Itabuna is a city of Bahia localized in the Northeast*. This sentence is classified as transitive because it follows the pattern: *Itabuna* (*A*) *is a city of* (relation) *Bahia* (*B*) *localized in* (relation) Northeast (*C*). As this sentence follows the pattern and it has a set of true premises it is possible to infer a new true fact: *Itabuna* (*A*) *localized in* (relation) *Northeast* (*C*). With the application of transitivity it is possible to minimize the problem of uninformative facts and to increase the amount of facts extracted in texts.

2.2.2 Symmetry

A sentence has a symetric relation if it follows the pattern: If *A relation B* then *B relation A*. For example, the sentence: *Barack Obama is married to Michele Obama* would be classified as symmetric, since it follows the previous pattern: *Barack Obama* (*A*) *is married to* (relation) *Michele Obama* (*B*). And *Michele Obama* (*B*) *is married to* (relation) *Barack Obama* (*A*). As stated that both facts are true, then the symmetry is valid. Similar to transitive approach, the application of symmetric allows an increase in the extraction of informative facts.

3 OUR PROPOSAL

We propose a method for Open IE for texts written in Portuguese. We have modified the approach described in (Fader et al., 2011) and refined our method through the inference approach. We are interested in new facts arising from inference, especially the identification of transitive and symmetric issues. We divided our method into four-folds: Syntactic Constraint, Inference Classifier, Transitivity Constraint and Symmetric Constraint. In those subsections, we depicted our workflow by an example and describe the materials used for the construction of the prototype.

3.1 Syntactic Constraint

Most of the relationships expressed in sentences written in Portuguese are through verbs as stated by the authors in (Fader et al., 2011). In Tab. 3 we present their syntactic constraint. In this way, a relation (*rel*) can be composed of a verb; or a verb followed by a preposition; or a verb followed by a sequence of words (nouns, adjectives, adverbs, pronouns) finalized by a preposition.

Table 3: Syntactic constraint for extract relations in Portuguese texts based in (Fader et al., 2011).

Pattern	V|VP|VW*P
V	verb particle? adv?
W	(noun | adj | adv | pron | det)
P	(prep | particle | inf. marker)

After identifying a possible relation in the sentence, the next step is to search for the arguments to the left of the relation (arg_1) and then to the right of the relation (arg_2). The identification of the arguments follows the new syntactic constraint pre-

sented in Tab. 4. These arguments are formed by noun phrases (NP) that can be composed of a noun, pronoun or adjective, as well as can be a combination of noun phrases connected through prepositions. Finally, our method extracts a triple as: $t = (arg_1, rel, arg_2)$ with the relation and both arguments.

Table 4: The constraints to identify the arguments proposed in this work.

Pattern	B-NP (I-NP)*(B-PP B-NP (I-NP)*)*
B-NP	Begin of Noun Phrase
I-NP	Middle of Noun Phrase
B-PP	Begin of Prepositional Phrase
I-PP	Middle of Prepositional Phrase

It is important to take into account the noun phrases due to more than one noun phrase, even on the right and on the left of a relation, can exist in a sentence. For example, in: *The movie characters refer to the Jules Jim movie* an extracted fact could be (The movie characters, refer to, Jules Jim). In this case we consider that *Noun Phrase chunker* (NP chunker) classified *Jules Jim* and *movie* as two NP. To address these cases we add a rule in our algorithm to form a single argument to the left (arg_1) or to the right of the relation (arg_2) through two NP, bound or not by a preposition. So, our method starts to extract a consistent fact, i.e. (The movie characters, refer to, Jules Jim movie).

3.2 Inference Classifier

The inference process proposed in our method aims to establish whether there is symmetry or transitivity in a set of sentences. This step is done by identifying at least one of the restrictions presented in Tab. 5. Then a classifier determines whether the sentence belongs to the class *transitive* or *symmetric*. We manually label 200 sentences with 100 belonging to the *transitive* class and 100 belonging to the *symmetric* class. We train a Support Vector Machine (SVM) algorithm within these set of sentences with 10-fold cross validation. Our model presented an accuracy of 83%, sensitivity of 90%, while its specificity was 75%.

3.3 Transitivity Constraint

When a sentence is detected as transitive, it is necessary to determine which type of transitivity to infer new facts. Table 6 indicates our proposed patterns for transitivity in a sentence. We emphasize that those patterns can also be considered in reverse. For example, the pattern 1 can also be recognized as arg_1 PART-OF arg_2 LOC arg_3. To simplify the presentation, we show only one of the transitivity orders for each type.

3.4 Symmetric Constraint

The symmetric is a simple constraint presented in Tab. 7. Both relational phrases considered symmetrical have a verb *to be*, a complement, and a preposition. Both inferences (transitive and symmetric) depend on the Syntactic Constraint step to gather the facts which will infer the new ones. Taking the transitive approach, it is necessary, at least, to extract two facts in a transitive sentence. If this does not occur, our method discards the inference. This problem does not happen in inference by symmetry. If a sentence is classified as symmetric and the pattern is detected, a single fact is enough to be extracted.

3.5 Our Workflow

Our method starts by processing the sentences through the POS tagger and NP chunker analyzers. With the words of each sentence labeled, we apply the syntactic constraint to extract facts in the sentences. Based on the sentence in Fig. 1 when applying the restrictions the following relations are obtained: 1 (*is a city of*) and 2 (*localized in*). The relation 1 is gathered from the pattern *VW*P*, which the relationship begins with a verb, followed by an article, a noun and ending with a preposition. While the 2 relation is obtained from the standard *VP*, which is a verb followed by a preposition. After identifying those relations, we use our new syntactic constraint to identify the arguments of a relation, i.e. we have the following arguments: for the relation 1 (Itabuna, Bahia) and the relation 2 (a city of Bahia, Northeast). The 1 argument is derived from the *B-NP* pattern, which indicates that the argument is a begin nominal phrase. Since no more elements are belonging to this syntactic group, it finishes the execution on catching the only item of the argument.

The argument (*Bahia*) of the relation 1 is also derived from the pattern *B-NP* and, for the same reasons as above it is the sole element of the argument. The argument (*a city of Bahia*) of the relation 2 is derived from the standard *B-NP I-NP B-PP B-NP I-NP*. This pattern indicates that the argument consists of a nominal phrase initial, followed by a nominal phrase medium, followed by a prepositional phrase initial, followed by a nominal phrase initial and ending in a nominal sentence medium. Finally, the *Northeastern* argument of the relation 2 is obtained from the standard *B-NP*. After identifying all the relations and

Table 5: Our patterns for transitive and symmetric features.

	Abbreviation	Type	Pattern
1	IS-A	Hyponymy	*verb_to_be art_a|an*
2	SIN	Synonymous	*verb_to_be nicknamed|called|known prep_of * |adv_as**
3	PART-OF	Meronymy	*verb_to_be|verb_to_do part_of*
4	LOC	Location	*verb_to_be art_the complement* | verb_to_be prep_in located|situated prep_in*
5	SYM	Symmetrical verbs	*verb_to_be complement* prep_as|of* art_the**

Table 6: Our patterns for transitive sentences.

1	arg_1	LOC	arg_2	PART-OF	arg_3
2	arg_1	LOC	arg_2	IS-A	arg_3
3	arg_1	LOC	arg_2	SYN	arg_3
4	arg_1	IS-A	arg_2	SYN	arg_3
5	arg_1	IS-A	arg_2	PART-OF	arg_3
6	arg_1	PART-OF	arg_2	SYN	arg_3

Table 7: Our pattern for symmetric sentences.

1	arg_1	SYM	arg_2

arguments in the sentence, the extractions are as follows: triple 1 (*Itabuna, is a city of, Bahia*) and triple 2 (*a city of Bahia, localized in, Northeast*). With all the extractions, the triples are loaded by the inference step. As this sentence was classified by our SVM as transitive, we apply our features and the argument 1 of the triple 1, the relation of the triple 2 and the argument 2 of the triple 2, formed a new extraction by the new fact: (*Itabuna, localized in, Northeast*).

3.6 Materials

We used CoGrOO (Moura Silva, 2013) system as POS tagger and NP chunker for Portuguese language. The CoGrOO project is a Portuguese-language spell checker that has analyzers trained with Brazilian and European Portuguese. For our classification model, which determines if the inference is transitive or symmetric, we used the *caret* (Kuhn, 2008) package for the *R* language.

4 EXPERIMENTAL SETUP

As we argued before, evaluating Open IE systems for languages other than English is a challenge. There is a need to create new datasets to establish a baseline, and it is hard work to identify other systems that can serve as a benchmark for comparing the results obtained. Therefore, we propose a comparison against two state-of-the-art methods: ReVerb (Fader et al., 2011) and DepOE (Gamallo et al., 2012). Considering that both systems made their comparison on English language, it was necessary to translate our set of sentences into English to allow the use of those works. We compare our proposal against ReVerb and DepOE into two aspects: a) the number of facts extracted and b) the precision. Our precision measure was calculated based on the ratio by the total # of valid facts and the total # of extracted facts (Equation 1). For each extraction, we classified if the fact was valid or invalid. This process was carried out by two experts who evaluated whether the extracted fact was consistent with the sentence.

$$Precision = \frac{\#(valid\,fact)}{\#(extracted\,fact)} \quad (1)$$

To evaluate our method, we propose the construction of two sets of sentences written in Portuguese. The first dataset was constructed from sentences retrieved from Wikipedia [2]. Sentences are retrieved randomly, but checking for any pattern (see Section 3) that indicates transitivity or symmetry. We retrieve 200 sentences with transitivity or symmetry in this dataset from now onwards called INFER-200. The second dataset was constructed from the Corpus of Electronic Texts Extracts NILCS/*Folha de São Paulo* newspaper (CETENFolha) [3] version 2008. This corpus was made up of texts about the most different contents (sports, politics, cooking, *etc.*). We randomly retrieved 200 sentences, and this dataset from now was called CETEN-200.

5 RESULTS

We organize our results into two folds. The first group presents the results obtained for INFER-200 dataset and the second group for CETEN-200 dataset. The first part of our results aims to analyze our proposal into a maximum scenario, only with transitive

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

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

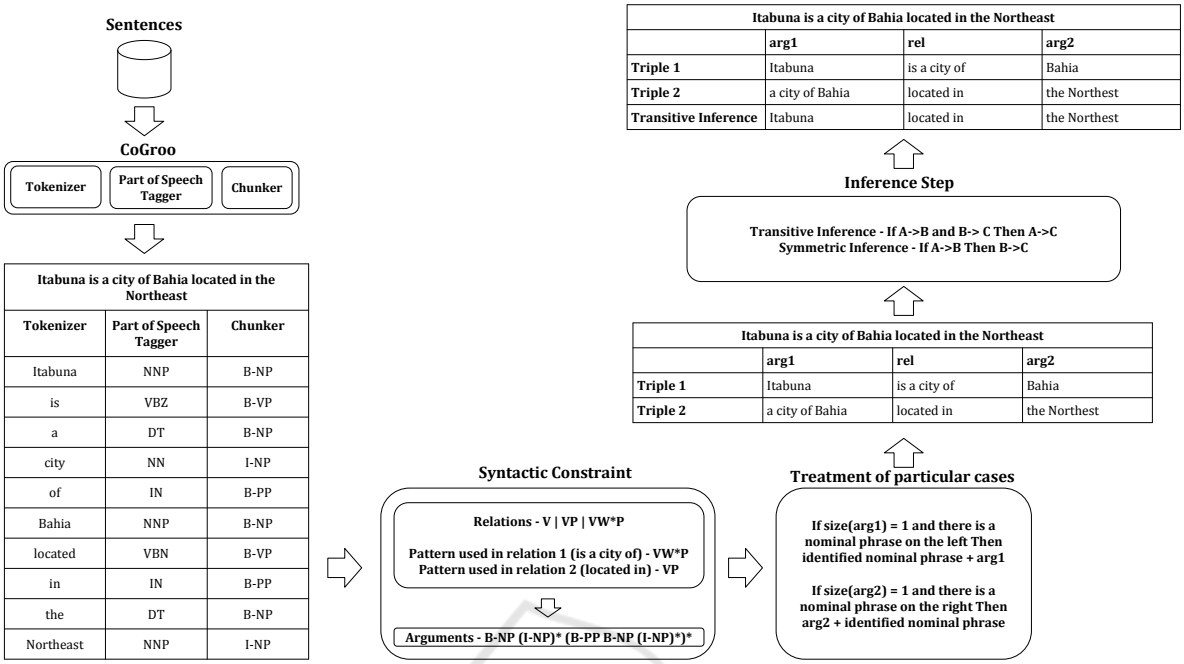

Figure 1: Our workflow.

and symmetric sentences. Within the second part of our experiments, we evaluate our proposal against the state-of-the-art work in a fair comparison. To evaluate the impact of our inference proposal we consider our method in two ways: OurMethod which is an adapted version of the syntactic restriction of Reverb for Portuguese and OurMethod+INFER which adds extracted facts by transitivity and symmetry.

5.1 Evaluation on INFER-200

In the evaluation with only transitive and symmetric sentences, our algorithm extracted 499 facts with 411 considered valid. In Fig. 2, we observed that OurMethod+INFER is able of extracting more than double valid facts about ReVerb demonstrating a significant superiority in this evaluated dimension. In Tab. 8 it is possible to observe the precision results of the evaluated methods. ReVerb has a slight advantage over our proposal. This trade-off between a high number of extraction and precision is expected (Buckland and Gey, 1994). However, the precision of 0.82 gives our method a satisfactory result.

In Tab. 8 it is possible to observe the precision results of the evaluated methods. ReVerb has a slight advantage over our proposal. This trade-off was already expected due to its adjustments for English written. However, our precision of 0.82 shows the effectiveness of our approach.

Figure 2: Results of the number of facts extracted by the evaluated methods in INFER-200. "+INFER" = our method with inference.

Table 8: Results of Precision by the evaluated methods in INFER-200.

Our Method	0.82
Reverb	0.92
DepOE	0.70

5.2 Evaluation on CETEN-200

From the CETEN-200 OurMethod+INFER algorithm extracted 473 from which 339 were valid against 373 and 289 of OurMethod. Our method with inference is extracting the greater number of facts when comparing with other methods. In Fig. 3 it is

possible to verify that our method extracts a relevant number of valid facts when compared to the other.

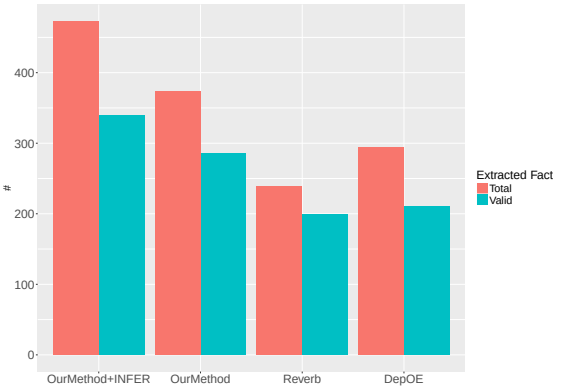

Figure 3: Results of the number of facts extracted by the evaluated methods in CETEN-200. "+INFER" = our method with inference.

In this scenario, our method and ReVerb had a slight drop on precision as presented in Tab. 9. Only the DepOE method remained with the same precision in both datasets evaluated.

Table 9: Results of Precision by the evaluated methods in CETEN-200. "+INFER" = our method with inference.

Our Method+INFER	0.71
Our Method	0.77
Reverb	0.83
DepOE	0.71

Despite the fact that our method had this slight drop regarding precision measure, it remains accurately compatible with other states-of-the-art works in this scenario.

6 ANALYSIS OF THE FACTS

The methods developed in Open IE approach extract facts without previously determining the type of relation. This freedom rises a problem: incoherent extractions. In this section, we discuss some of the extractions obtained by our method and our related works. We organize this discussion in two parts. In the first part, we present some sentences and the facts extracted by our method. In the second part, we compared some facts extracted by our method and other related works.

6.1 Analysis Concerning our Method

In the following sentence *A Igreja Paroquial de Santo Agostinho a Marvila é uma antiga obra religiosa localizada em Lisboa* (The Parish Church of Santo Agostinho the Marvila is an ancient religious work located in Lisbon) the proposed method extracted the facts presented in Tab. 10. All the extracted facts were classified as valid, and the triple 3 was inferred by transitivity.

Table 10: Example of valid extractions of transitive by our method.

Triple 1	(A Igreja_Paroquial_de_Santo_Agostinho a Marvila, é, uma antiga obra religiosa)
Triple 2	(uma antiga obra religiosa, localizada em, Lisboa)
Triple 3	(A Igreja_Paroquial_de_Santo_Agostinho a Marvila, localizada em, Lisboa)

Considering the symmetrical sentence *A Zona das Américas é uma das 3 zonas regionais da Copa Davis* (The Zone of the Americas is one of the three Davis Cup regional zones) our method extracts the triples presented in Tab. 11. The two facts extracted were considered valid. The 2 fact was inferred by symmetry.

Table 11: Example of valid extractions of symmetric by our method.

Triple 1	(A Zona das Américas, é uma das, 3 zonas regionais da Copa_Davis)
Triple 2	(uma das 3 zonas regionais da Copa_Davis, é, A Zona das Américas)

Our evaluation shows that our method extracts a higher number of valid facts than the others. However, it is observable that part of the extracted facts remains not coherent. Considering the sentence *Chilton (South Hylton, 16 de setembro de 1918 - 15 de junho de 1996) foi um futebolista e treinador inglês, que atuava como defensor* (Chilton (South Hylton, September 16, 1918 - June 15, 1996) was an English footballer and trainer, who acted as a defender) our method extracts the facts presented in Tab. 12. We consider the fact 2 as a valid extraction (although it is uninformative), and the facts 1 and 3 as invalid. Our method classified this sentence as transitive and performed a new invalid extraction. In this example, our method failed due to coreference issue(Van Deemter and Kibble, 1999) (the 1 argument is replaced by *Chilton*).

Table 12: Example of invalid extraction with transitive inference performed by our method.

Triple 1	(15 de junho de 1996, foi, um futebolista)
Triple 2	(treinador inglês que, atuava, defensor)
Triple 3	(15 de junho de 1996, atuava, defensor)

6.2 Analysis of ReVerb against our Method

Considering the sentence *Antonio Nocerino (Nápoles, 9 de abril de 1985) é um jogador de futebol italiano que joga como um meio-campista defensivo* (Antonio Nocerino (Naples, April 9, 1985) is an Italian footballer who plays as a defensive midfielder) in Tab. 13, it shows a comparison between the facts extracted by two methods: Reverb and our proposal. In the extracted triples it is noticed that the ReVerb method extracted a coherent information different from the proposed method. Actually, our method was unable to identify coherent arguments for the relationship. In this case, we verified that the construction of the sentence in Portuguese made impossible to extract a coherent fact. The English and Portuguese languages present a different set of grammatical rules and in some cases the syntactic restriction proposed in (Fader et al., 2011) is not efficient for Portuguese.

Table 13: Example that ReVerb is better than our method.

Our method	(9 de abril de 1985, é um, jogador de futebol italiano)
ReVerb	(an Italian footballer, plays as, a defensive midfielder)
Best fact	**Reverb**

Considering now the sentence *Canne al vento é um romance de Grazia Deledda* (Reeds in the wind is a novel by Grazia Deledda) in Tab. 14, it shows a comparison between the triples extracted by the same two methods. According to the sentence in question, our method extracts a valid fact, while ReVerb does not. It is possible to notice that the process of identification of the NP made an error in the ReVerb method possibly by an incorrect classification of the morphological classes that composes the book name. We corrected this type of deficiency in our method by proposing new restrictions to identify extraction arguments.

Table 14: Example that our method is better than ReVerb.

Our method	(Canne al vento, é um romance de, Grazia_Deledda)
ReVerb	(the wind, is a novel by, Grazia Deledda)
Best fact	**Our method**

Finally, analyzing the sentence *O condado de Dawson é um dos 254 países do estado americano de Texas* (Dawson County is one of the 254 countries of the American state of Texas) in Tab. 15, it shows another comparison between the facts extracted. We consider both facts as valid. In this example, ReVerb extracts more details compared to our method. However, we believe that both methods were successful on carrying this kind of task.

Table 15: Example that ReVerb and our method are equivalent.

Our method	(O Condado de Dawson, é um dos, 254 condados do estado)
ReVerb	(Dawson County, is one of, the 254 countries of the American state of Texas)
Best fact	**Both**

6.3 Analysis of DepOE against our Method

Given the following sentence *End of Time é uma música da cantora americana Beyoncé gravada para seu quarto álbum de estúdio 4* (End of Time is a song by American singer Beyoncé recorded for her fourth studio album 4) in Tab. 16, it shows a comparison between the facts extracted by our method and DepOE. In the facts presented in Tab. 16 it is possible to observe that DepOE extracted a coherent information regarding the sentence. In a different way, our method did not obtain a coherent fact. This occurred because the POS tagger used failed to set *End of Time* ("of" was classified as a verb in Portuguese). As in English grammar, proper names are not translated into Portuguese, and the analyzer has made a mistake.

Table 16: Example that DepOE is better than our method.

Our method	(End, of, Time é uma canção da cantora norte-americana)
DepOE	(End of Time, is, a song)
Best fact	**DepOE**

Considering now the sentence *Ione é uma cidade no estado dos EU de Oregon, no condado de Morrow* (Ione is a city in the US state of Oregon, in Morrow County) in Tab. 17, it shows the facts extracted. In the presented facts, our method extracted a coherent information while DepOE did not. The information presented in the fact extracted by DepOE did not correspond to the information contained in the sentence.

Table 17: Example that our method is better than DepOE.

Our method	(Ione, é, uma cidade)
DepOE	(Ione, is, state of Oregon)
Best fact	**Our method**

Finally, analyzing the sentence *A Supercopa de Portugal é uma competição de rugby em Portugal*

(The Super Cup Portugal is a rugby competition in Portugal) in Tab. 18, it shows again a comparison between the facts extracted by the two methods analyzed. Both facts extracted were classified as valid.

Table 18: Example that DepOE and our method are equivalent.

Our method	(A Supertaça de Portugal, é a uma competição de, rugby em Portugal)
DepOE	(The super@cup@portugal, is a rugby competition in, Portugal)
Best fact	**Both**

7 CONCLUSION AND FUTURE WORK

This paper describes a method that aims to extract facts from texts written in the Portuguese language. To the best of our knowledge, this is the first time that inference approach, considering both transitive and symmetric, is being used to extract facts in Portuguese texts with a large volume of extractions. Our related works have presented few advances in Open IE in texts written other than English. The construction of our method resulted in the following contributions:

- We have improved the method proposed in (Fader et al., 2011) and applied it to Portuguese.
- We have created a way of identifying the NP (relations arguments) that increase the informativeness of the extracted facts. Our new constraints allow the union of different NP to get coherent arguments from the facts.
- We enhanced the method proposed by (Bast and Haussmann, 2014) with inference by transitivity in Portuguese. With our proposal, it was possible to acquire new useful facts thus improving the results obtained by our method.
- We proposed a new approach to Open IE: inference by symmetric. This new method of inference allows the extraction of new facts thus increasing the number of valid extractions.
- We propose a set of restrictions to identify transitive and symmetric sentences written in Portuguese. Our method uses machine learning with low supervision to determine whether or not a sentence has an inference feature.

7.1 Future Work

We analyze some of the facts extracted by our method against other methods of the state of the art. This analysis allows us to identify in which moments our method was superior to the evaluated methods. It also makes it possible to analyze deficiencies in our proposal. A question was identified regarding the problems generated by the used analyzers (POS tagger and NP chunker). The errors gathered by the parsers propagated in our method generating invalid facts. The improvements of such natural language processing tools are required to the advance of Open IE in Portuguese. Another open question in our method is the treatment for coreference. Some sentences are written with pronouns or writing styles that do not favor our approach. A new version of our method can ensure greater informativeness by replacing occurrences of coreference with the mentioned entity. Besides, it was possible to identify that also in Portuguese, a sentence writing style has a strong impact on the method used. In a case presented in Section 6 it was possible to verify that the syntactic constraint does not meet the pattern identified in the sentence. A review of Portuguese grammar can lead to improvements in the syntactic constraint proposed by (Fader et al., 2011) dealing with aspects of Portuguese.

Our method showed a large gain in the amount of extracted facts, and this naturally leads to a loss of precision. This trade-off was addressed in the work of (Fader et al., 2011) by adding the lexical constraint. Our work focused on the adaptation of the syntactic constraint due to the lack of appropriate resources for the Portuguese language. A new work is the proposal of a lexical restriction to verify the validity of the facts extracted to guarantee a greater precision to our method.

ACKNOWLEDGEMENTS

We would like to thank FAPESB and CAPES for their scholarships to support this work.

REFERENCES

Angeli, G., Premkumar, M. J., and Manning, C. D. (2015). Leveraging linguistic structure for open domain information extraction. *Linguistics*, (1/24).

Banko, M. (2009). *Open information extraction for the web*. PhD thesis, University of Washington.

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

Bast, H. and Haussmann, E. (2013). Open information extraction via contextual sentence decomposition. In *Semantic Computing (ICSC), 2013 IEEE Seventh International Conference on*, pages 154–159. IEEE.

Bast, H. and Haussmann, E. (2014). More informative open information extraction via simple inference. In *Proceedings of the 36th European Conference on IR Research on Advances in Information Retrieval - Volume 8416*, ECIR 2014, pages 585–590, New York, NY, USA. Springer-Verlag New York, Inc.

Buckland, M. and Gey, F. (1994). The relationship between recall and precision. *Journal of the American society for information science*, 45(1):12.

Del Corro, L. and Gemulla, R. (2013). Clausie: clause-based open information extraction. In *Proceedings of the 22nd international conference on World Wide Web*, pages 355–366. ACM.

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

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

Faruqui, M. and Kumar, S. (2015). Multilingual open relation extraction using cross-lingual projection. *arXiv preprint arXiv:1503.06450*.

Gamallo, P. and Garcia, M. (2015). Multilingual open information extraction. In *Portuguese Conference on Artificial Intelligence*, 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*, pages 10–18. Association for Computational Linguistics.

Gotti, F. and Langlais, P. (2016). Harnessing open information extraction for entity classification in a french corpus. In *Canadian Conference on Artificial Intelligence*, pages 150–161. Springer.

Kuhn, M. (2008). Caret package. *Journal of Statistical Software*, 28(5).

Moura Silva, W. D. C. d. (2013). *Improving the Corrector Gramatical CoGrOO*. PhD thesis, University of São Paulo.

Qiu, L. and Zhang, Y. (2014). Zore: A syntax-based system for chinese open relation extraction. In *EMNLP*, pages 1870–1880.

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.

Schoenmackers, S., Etzioni, O., and Weld, D. S. (2008). Scaling textual inference to the web. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 79–88. Association for Computational Linguistics.

Soderland, S. (1999). Learning information extraction rules for semi-structured and free text. *Machine learning*, 34(1-3):233–272.

Tseng, Y.-H., Lee, L.-H., Lin, S.-Y., Liao, B.-S., Liu, M.-J., Chen, H.-H., Etzioni, O., and Fader, A. (2014). Chinese open relation extraction for knowledge acquisition. In *EACL*, pages 12–16.

Van Deemter, K. and Kibble, R. (1999). What is coreference, and what should coreference annotation be? In *Proceedings of the Workshop on Coreference and its Applications*, pages 90–96. Association for Computational Linguistics.

Wu, F. and Weld, D. S. (2010). Open information extraction using wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 118–127, Stroudsburg, PA, USA. Association for Computational Linguistics.