Dependency Parser on Open Information Extraction for Portuguese Texts - DptOIE and DependentIE on IberLEF

Rafael Glauber, Daniela Barreiro Claro ✉[0000−0001−8586−1042], and Leandro Souza de Oliveira

Formalisms and Semantic Applications Research Group (FORMAS)
LASiD/DCC/IME
Federal University of Bahia, Salvador, Bahia, Brazil
http://formas.ufba.br/
rglauber@dcc.ufba.br, dclaro@ufba.br, leo.053993@gmail.com

Abstract. This paper describes the participation of the DependentIE and DptOIE systems in the Iberian Languages Evaluation Forum 2019. Our activities have focused on the "General Open Relation Extraction" task of relation extraction for Portuguese texts. We describe the choices adopted during the challenge, as well as the systems performed and their results.

Keywords: Shared Task · Open Information Extraction · Relation Extraction.

1 Introduction

Extract information from large repositories of texts within different domains is a hard task for humans. While the quantity and diversity of textual content grow on the Web, the traditional IE tools have low coverage in this scenario [4]. In the study conducted by [1] the authors proposed a new approach called Open Information Extraction (Open IE) that extracts facts from a sentence in the following triple format:

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

where *arg1* and *arg2* are nominal phrases in a sentence and *rel* establishes a relationship between *arg1* and *arg2* through a verb phrase. Open IE systems are useful in web-scale issues such as question answering and document filtering systems [6]. The Iberian Languages Evaluation Forum (IberLEF 2019) organized a Portuguese named entity recognition (NER) and relation extraction (RE) tasks

which included Open IE task [2]. Participants to this task should apply their systems/methods in activities related to NER or RE in Portuguese sentences. We applied two different Open IE systems in one task for the RE problem:

– Task 3: General Open Relation Extraction

In this work, we describe our Open IE systems and their results, as well as the choices and problems faced to perform this task. Our systems were based on dependency analysis and handcrafted rules to extract facts from Portuguese sentences. We participate with two of our systems: DependentIE[1] and DptOIE[2].

This paper is organized as follows: section 2 describes the problem statement; Section 3 presents our methods DependentIE and DptOIE; Section 4 describes our setup and section 5 presents our evaluation. Section 6 presents our results and we conclude in Section 7.

2 Problem Statement

The organization of the IberLEF (Iberian Languages Evaluation Forum) forum proposes a task that involves the automatic extraction of any relation descriptor expressing any semantic relation between a pair of entities or concepts mentioned in Portuguese sentences. In this task, the coordinators consider a relation description as a text chunk that describes the explicit semantic relation, occurring between two entities or noun phrases in a sentence.

The task was divided into two different tests. The first one, the participants must extract the relation descriptors between NP pairs from data provided by the coordinators. This data was annotated with NP information, and as a consequence, do not need to employ a NER system by participants. The second one, the data provided was not annotated with NP information. The goal of the task was to extract and classify the NPs from the test sentences, and then extract the relation descriptors between pairs of the NPs. We submitted our methods to both Test 1 and Test 2 of Task 3.

3 Our methods

We participate in the IberLEF 2019 with two Open IE systems. The first of them, the DependentIE is an Open IE system for Portuguese sentences [8]. As well as ArgOE [5] and ClausIE [3] we use a Dependence Parser (DP) to identify clauses[3] (useful parts of a sentence). In this work, a clause is one of the following parts of a sentence: subject (S), direct and indirect objects (O), verb (V), adverb (A), complement (C) and modifier (M). Our method extracts facts using clauses based on the standard SV (Subject - Verb). The arguments are detected through

[1] `http://formas.ufba.br/dclaro/tools.html#dependentie`

[2] `http://formas.ufba.br/dclaro/tools.html#dptoie`

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

a deep-search in the sentence dependency. It uses Malt Parser as the Dependence Parser.

The second system DptOIE is an evolution of DependentIE system [7]. It uses Stanford's Dependency Parser, specific rules for extracting facts in Portuguese sentences, it adapts the depth-first search to explore the dependency tree, and it handles particular cases in sentences with coordinate conjunctions, subordinate clauses, and appositives. Furthermore, DptOIE is open to other dependency parsers, since sentences are in CoNLL-U and Universal Dependencies v2.1 Brazilian treebank format.

4 Setup

Both systems, DependentIE[4] and DptOIE[5], used to perform this task are available for download on FORMAS website. Our systems generate an output file in comma-separated values (CSV) format. For Test 1, each system extracted the facts contained in the test sentences. Then, each pair of NP contained in the test file is compared with the arguments of the facts extracted by both systems. For the comparison between the arguments of the extracted facts and the NPs of the test file, the following characters were ignored: " , . ( ) [ ] ? !. Moreover, to avoid minor divergences in the comparison of strings we removed a set of stopwords[6]. When identifying a pair of arguments in the output file of systems similar to an NPs pair of the test file, the text fragment corresponding to the relationship is selected as a result of Test 1.

Test 2 follows the free form suggested by the Open IE task. After running both systems for the set of test sentences, the next step is to convert our output format from CSV to the required format of IberLEF 2019.

5 Evaluation

Two scores were considered for the evaluation of Task 3: a completely correct relations score and a partially correct relations score [2]. Completely Correct Relations (CCR) occurs when all terms that make up the relation descriptors in the key are equal to the relations descriptors of the system's output. The score for each completely correct relation is 1, which represents a full hit. Partially correct relationships (PCR) occurs when at least one of the terms in the relation descriptors of the system's output corresponds to a term in the relation descriptors of the key.

[4] http://formas.ufba.br/dclaro/tools.html#dependentie
[5] http://formas.ufba.br/dclaro/tools.html#dptoie
[6] List of stopwords at https://github.com/stopwords-iso/stopwords-pt/blob/master/stopwords-pt.txt

5.1 Test 1 Evaluation

The systems extractions were matched against the relationship in Test 1 golden dataset, and the metrics of exact Precision, exact Recall, partial Precision, and partial Recall were computed.

5.2 Test 2 Evaluation

Since Open Relation Extraction identifies all possible information, and the sentences adopted in the evaluation of Test 2 are the same as Test 1 and training datasets, we did four different evaluations to provide a full panorama of the performance of our systems:

- Considering only the relationships in Test 2 golden dataset;
- Considering the relationships in Test 2 golden dataset and disregarding the relationship in the training dataset;
- Considering the relationships in Test 2 golden and Test 1 golden dataset and disregarding the relationship in the training dataset;
- Considering the relationships in all three datasets;

All datasets used are available at `http://www.inf.pucrs.br/linatural/wordpress/iberlef-2019/`. The details of the performed measures and datasets are described in [2].

6 Results

We organized the results of Task 3, considering the values obtained in Tests 1 and 2. Table 1 presents the results obtained by both systems performing the exact measures in Test 1. The values for all measures are not very expressive. Still, DptOIE has a slight advantage in comparing both systems. Next, we present the results for the partial measures in Table 2. Although the values obtained with the partial measures are better for both systems, Test 1 proved to be challenging to solve. The values obtained were very low for both systems in any of the experimental setup. The activity of identifying entities that are part of the arguments of a fact extracted by an Open IE system was the cause of part of the errors introduced. The arguments of the facts are NPs that contains other fragments of the sentence. Even when removing stopwords, other filters should be considered.

Another critical aspect in Test 1 is that the attempt to improve the measures, with a partial score, generated a little impact on the outcome. The increase in the values of Precision, Recall, and F-measures was small by the scale of the values presented.

Figures 1 and 2 present the results obtained for Test 2 in the four setups proposed by the coordinators. In this test, we were able to identify the best performance of DptOIE in performing the task. The difference is more significant when comparing the values for the partial measures between both systems.

Table 1. Results for both systems in Task 3/Test 1 and Exact measures.

System	Exact Precision	Exact Recall	Exact F-measure
DependentIE	0.011364	0.011364	0.011364
DptOIE	0.034091	0.034091	0.034091

Table 2. Results for both systems in Task 3/Test 1 and Partial measures.

System	Partial Precision	Partial Recall	Partial F-measure
DependentIE	0.014205	0.013494	0.013840
DptOIE	0.044034	0.042850	0.043434

DptOIE presents higher values for precision and Recall, in addition to greater harmony between these measures, which generated higher values of F-measure.

For Test 2, the execution of partial scores generates a significant impact on the outcome. There's an improvement in the results for the DptOIE. In addition, the DependentIE precision result gets a large increase. Although this type of evaluation approach generates better results, one aspect should be considered: the absence of some terms in the arguments of the facts extracted by the Open IE systems may indicate invalid facts, and this cannot be discarded from a more fair evaluation.

7 Conclusions

This paper described the participation of the DependentIE and DptOIE systems in IberLEF 2019. Both systems were submitted to the "General Open Information Extraction" task through Test 1 and Test 2. In particular, the DptOIE system presented the best results. When the values for Test 2 are analyzed, it becomes more evident.

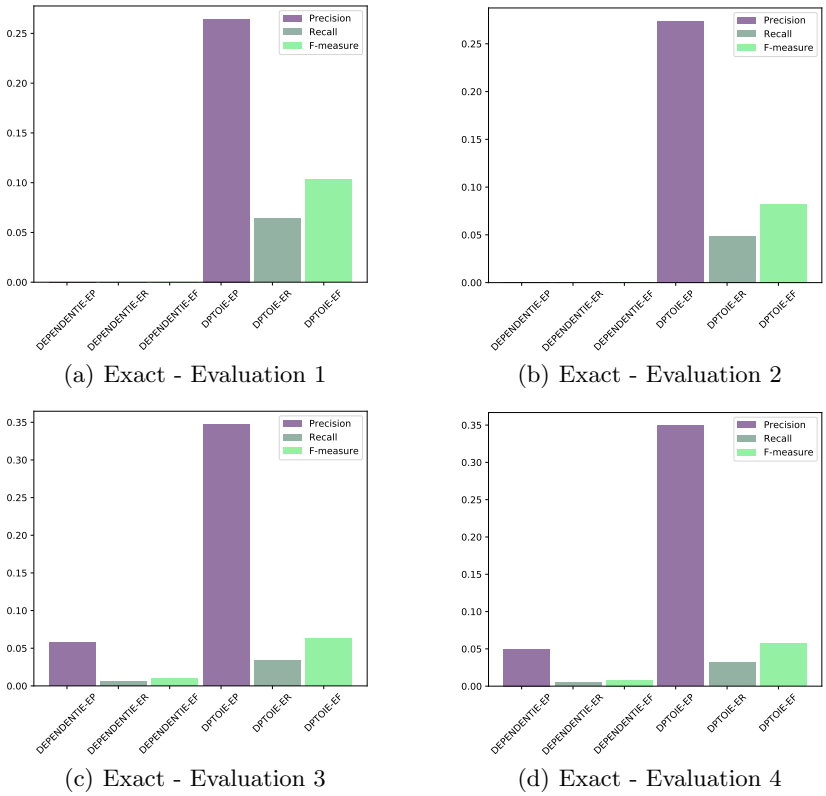
In general, the Open IE task presents the trade-off of approaches that prioritize greater coverage. While discovering a more significant number of relationships, the precision values obtained by Open IE systems are low. The results obtained in the participation of the DependentIE and DptOIE systems confirm this problem.

Acknowledgement

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior - Brasil (CAPES) - Finance Code 001.

References

1. Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O.: Open information extraction from the web. In: Proceedings of IJCAI. vol. 7, pp. 2670–2676 (2007)

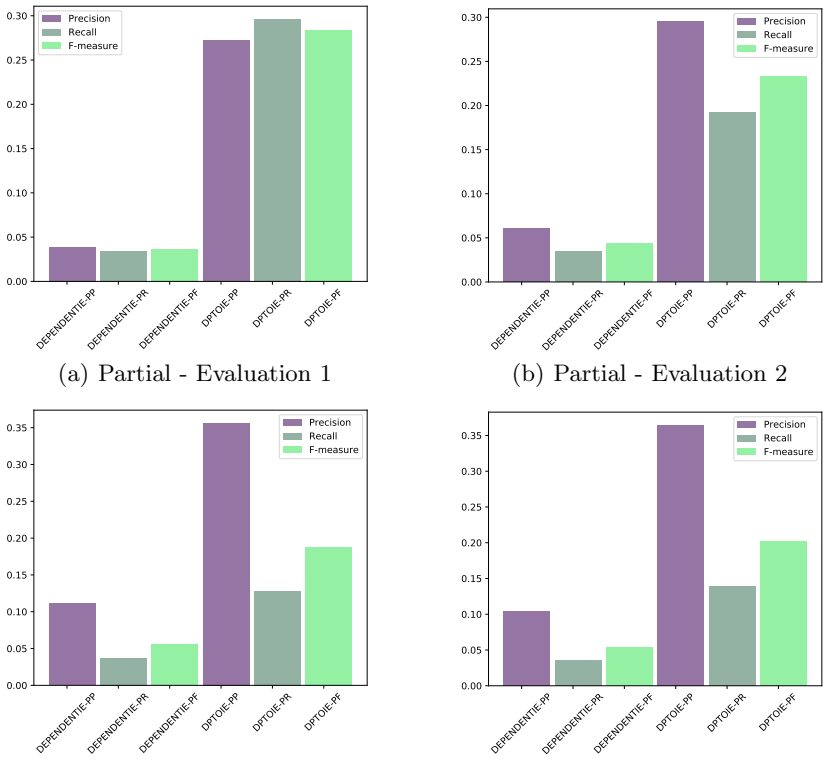


(a) Exact - Evaluation 1 (b) Exact - Evaluation 2

(c) Exact - Evaluation 3 (d) Exact - Evaluation 4

Fig. 1. Results for both systems in Task 3/Test 2 and Exact measures.

2. Collovini, S., Santos, J., Consoli, B., Terra, J., Vieira, R., Quaresma, P., Souza, M., Claro, D.B., Glauber, R., a Xavier, C.C.: Portuguese named entity recognition and relation extraction tasks at iberlef 2019 (2019)
3. Del Corro, L., Gemulla, R.: Clausie: clause-based open information extraction. In: Proceedings of WWW. pp. 355–366. ACM (2013)
4. Fader, A., Soderland, S., Etzioni, O.: Identifying relations for open information extraction. In: Proceedings of EMNLP. pp. 1535–1545. Association for Computational Linguistics (2011)
5. Gamallo, P., Garcia, M.: Multilingual open information extraction. In: Proceedings of EPIA. pp. 711–722. Springer (2015)
6. Glauber, R., Claro, D.B.: A systematic mapping study on open information extraction. Expert Systems with Applications (2018)
7. de Oliveira, L.S., Claro, D.B.: DPTOIE: Um Método para Extração de Informação Anerta na Língua Portuguesa baseado em Análise de Dependência. Master's thesis, Universidade Federal da Bahia (2018)
8. de Oliveira, L.S., Glauber, R., Claro, D.B.: Dependentie: An open information extraction system on portuguese by a dependence analysis. In: Proceedings of ENIAC. pp. 271–282. FC-UFU (2017)



(a) Partial - Evaluation 1

(b) Partial - Evaluation 2

(c) Partial - Evaluation 3

(d) Partial - Evaluation 4

Fig. 2. Results for both systems in Task 3/Test 2 and Partial measures.