Towards a Pragmatic Open Information Extraction for Portuguese Text - ICEIS17, InferPortOIE and PragmaticOIE on IberLEF

Rafael Glauber, Daniela Barreiro Claro ✉[0000−0001−8586−1042], and Cleiton Fernando Lima Sena

FORMAS Research Group, LaSiD/DCC/UFBA
Federal University of Bahia, Brazil
rglauber@dcc.ufba.br, dclaro@ufba.br,cflsena2@gmail.com
http://formas.ufba.br

Abstract. This paper describes the participation of the FORMAS research group with the systems ICEIS17, InferPortOIE, and PragmaticOIE 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 present our choices on this challenge, as well as the performance of our systems and their results.

Keywords: Shared Task · Open Information Extraction · Relation Extraction · Pragmatic Open IE · Inference Extraction.

1 Introduction

Information Extraction (IE) emerged as a research area to identify relevant patterns in large quantities of textual documents [10]. 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 [1]. However, some drawbacks were considered, such as low coverage of relations and human intervention for new relations. *Open Information Extraction* (Open IE) comes up to extract information freely from texts and scales for the Web [6]. While the quantity and diversity of textual content grow on the Web, the traditional IE tools have low coverage in this scenario [3]. 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 [4]. The Iberian Languages Evaluation Forum (IberLEF 2019) organized a Portuguese named entity recognition (NER) and relation extraction (RE) task which included Open IE task [2]. Participants should apply their systems/methods to this task related to NER or RE in Portuguese sentences. We applied three different Open IE systems to RE problem:

– Task 3: General Open Relation Extraction

We describe our Open IE systems and their results, as well as the choices and problems faced to perform this task. Our systems were implemented based on machine learning, inference, and handcrafted rules to extract facts from Portuguese sentences. We participate with three of our systems: ICEIS17, InferPortOIE, and PragmaicOIE.

This paper is organized as follows: section 2 describes the problem statement; Section 3 presents our methods ICEIS17, InferPortOIE, and PragmaticOIE; Section 4 describes our Setup, and section 5 presents our evaluation. Section 6 presents the 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. In the first test, participants 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. In the second test, 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 Task 3 with three systems:

3.1 ICEIS17

Our method called ICEIS17 [9] modified the approach described in [3] and refined through the inference approach. Within ICEIS17 method, 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[9].

3.2 InferPortOIE

InferPortOIE [8] takes into advance the structure of writing, especially asyndetic coordination sentences. In addition, the methodology of Reverb [3] was adapted to the Portuguese language [9]. InferPortOIE proposes two new rules that generalize both the inference by transitivity and by symmetry, thus increasing the number of extractions in a sentence. A new specific rule for symmetric reasoning is proposed based on a list of symmetric verbs reported in [5]. We divided InferPortOIE into six-folds: Pre-processing, Syntactic Constraint, Treatment of Particular Cases, Inference Detection, Transitivity Constraint and Symmetric Constraint [8].

3.3 PragmaticOIE

Our PragmaticOIE method achieves a first pragmatic level. Our first pragmatic level copes with inferential, contextual, and intentional aspects. The inferential module in PragmaticOIE has inherited from our previous work [8] and guarantees a semantic interpretation [7]. The new contextual layer of our PragmaticOIE system enhanced the method proposed by [6] and broadened it by the use of subordinate conjunctions, adverbs, prepositions, and adversative coordination sentences. Finally, the new intentional approach incorporated into our PragmaticOIE can extract implicit facts from a sentence, through verbs in Conditional Tense.

4 Setup

All systems, ICEIS17[1], InferPortOIE[2] and PragmaticOIE[3], employed to perform the Task 3 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 all 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[4]. The text fragment corresponding to the relationship is chosen

[1] `http://formas.ufba.br/dclaro/tools.html\#sgs\_iceis`
[2] `http://formas.ufba.br/dclaro/tools.html\#inferportoie`
[3] `http://formas.ufba.br/dclaro/tools.html\#pragmaticoie`
[4] List of stopwords at `https://github.com/stopwords-iso/stopwords-pt/blob/master/stopwords-pt.txt`

as a result of Test 1 when the pair of arguments in the output file is similar to those NP pair of the test file.

Test 2 follows the free form suggested by the Open IE task. After running all systems through the test sentences, the next step was 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 relation 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.

5.1 Test 1 Evaluation

The extractions of the systems were matched against the relationship in Test 1 golden dataset, and metrics of exact Precision (EP), exact Recall (ER), partial Precision (PP), and partial Recall (PR) were calculated. Exact and partial F-measure are identified by (EF) and (PF).

5.2 Test 2 Evaluation

Since Open Relation Extraction recognizes all possible information, and the sentences adopted in Test 2 are the same as Test 1, 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 both Tests 1 and 2. Table 1 exhibits the results achieved by all systems considering

the exact measure in Test 1. Values for all measures are not very expressive. It is noteworthy that ICEIS17 has a slight advantage when compared with the other systems. Both InferPortOIE and PragmaticOIE systems obtained null values for the exact score.

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

System	Exact Precision	Exact Recall	Exact F-measure
ICEIS17	0.011364	0.011364	0.011364
InferPortOIE	0.000000	0.000000	0.000000
PragmaticOIE	0.000000	0.000000	0.000000

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

System	Partial Precision	Partial Recall	Partial F-measure
ICEIS17	0.012784	0.014205	0.013457
InferPortOIE	0.003551	0.004545	0.003987
PragmaticOIE	0.003551	0.004545	0.003987

Table 2 presents the results for the partial measure. Although the values with the partial measure are better for all systems, Test 1 proved to be challenging to solve. The values obtained were very low for all systems in any of the experimental setup.

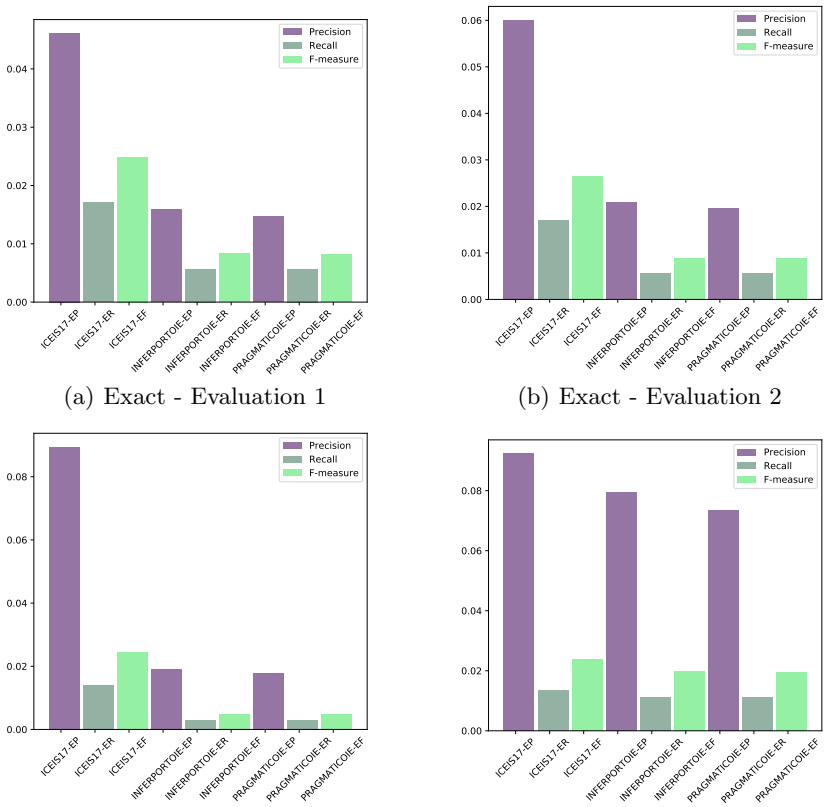
The activity of identifying entities that are part of the arguments of a fact extracted by an Open IE system was the cause of part of the errors introduced. The arguments of the facts are NPs that contains other fragments of the sentence. Even when removing stopwords, other filters should be considered.

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 presents the results for Test 2 for the four setups proposed by the coordinators. In this test, we were able to identify the best performance of ICEIS17 on executing this task. The difference is significant when comparing the values for the partial measures among the three systems. In this case, ICEIS17 presents a higher value on precision. However, the performance of ICEIS17 on recall is low. We can thus conclude that, for Test 2, the execution of partial scores generates a significant impact on the outcome.

7 Conclusions

This paper described the participation of the ICEIS17, InferPortOIE, and PragmaticOIE systems in IberLEF 2019. All systems were submitted to the "General



(a) Exact - Evaluation 1

(b) Exact - Evaluation 2

(c) Exact - Evaluation 3

(d) Exact - Evaluation 4

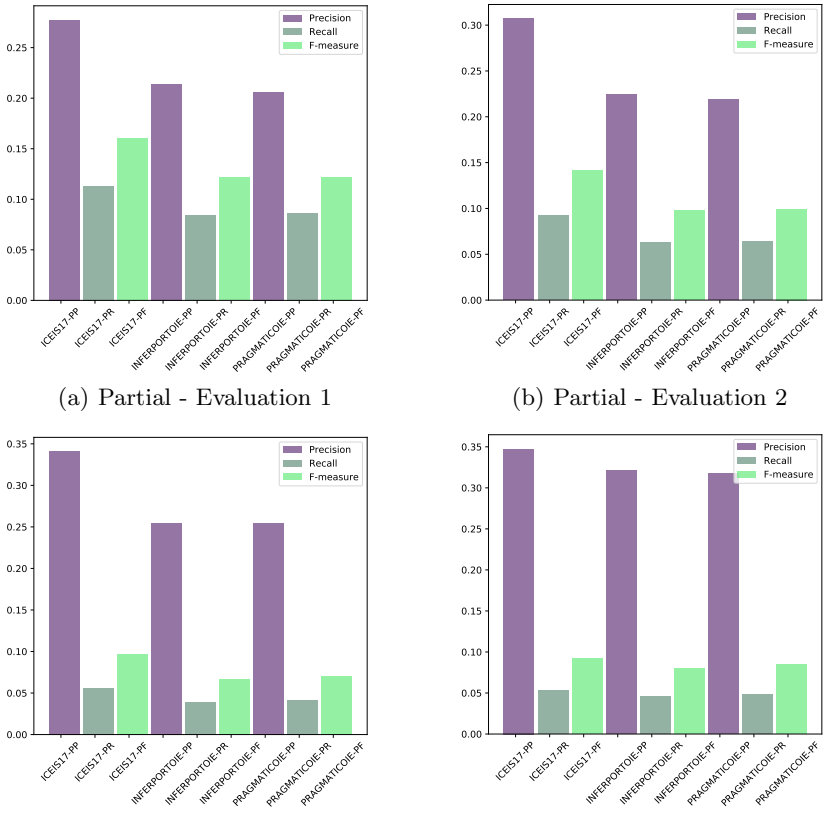
Fig. 1. Results for systems in Task 3/Test 2 and Exact measures.

Open Information Extraction" task through Test 1 and Test 2. In particular, the ICEIS17 system presented the best results (especially, when we isolate precision). When the values for Test 2 are analyzed, it becomes more evident.

The approach used in the evaluated systems demonstrated a low performance for the proposed task. While Open IE systems prioritize the identification of a high number of facts in sentences, our methods, that utilize shallow analyzers, have been little efficient.

Acknowledgement

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



(a) Partial - Evaluation 1

(b) Partial - Evaluation 2

(c) Partial - Evaluation 3

(d) Partial - Evaluation 4

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

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)
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. Fader, A., Soderland, S., Etzioni, O.: Identifying relations for open information extraction. In: Proceedings of EMNLP. pp. 1535–1545. Association for Computational Linguistics (2011)
4. Glauber, R., Claro, D.B.: A systematic mapping study on open information extraction. Expert Systems with Applications (2018)
5. GODOY, L.: Os verbos recíprocos no PB: interface sintaxe-semântica lexical. 2008. Ph.D. thesis, Dissertação (Mestrado em Estudos Linguísticos)-Faculdade de Letras, UFMG, Belo Horizonte (2008)

6. Mausam, Schmitz, M., Bart, R., Soderland, S., Etzioni, O.: Open language learning for information extraction. In: Proceedings of the EMNLP–CoNLL. pp. 523–534. ACL (2012)
7. Sena, C.F.L., Claro, D.B.: EXTRACAO DE RELAÇÕES PRAGMÁTICAS EM DOCUMENTOS DO PORTUGUÊS DO BRASIL. Master's thesis, Universidade Federal da Bahia (2017)
8. Sena, C.F.L., Claro, D.B.: Inferportoie: A portuguese open information extraction system with inferences. Natural Language Engineering **25**(2), 287–306 (2019). https://doi.org/10.1017/S135132491800044X
9. Sena., C.F.L., Glauber., R., Claro, D.B.: Inference approach to enhance a portuguese open information extraction. In: Proceedings of the 19th International Conference on Enterprise Information Systems - Volume 1: ICEIS,. pp. 442–451. INSTICC, SciTePress (2017). https://doi.org/10.5220/0006338204420451
10. Soderland, S.: Learning information extraction rules for semi-structured and free text. Machine learning **34**(1-3), 233–272 (1999)