IberLEF 2019 Portuguese Named Entity Recognition and Relation Extraction Tasks

Sandra Collovini[1[0000−0003−1257−4100]], Joaquim Santos[1[0000−0002−0581−4092]], Bernardo Consoli[1[0000−0003−0656−511X]], Juliano Terra[1[0000−0003−3066−1531]], Renata Vieira[1[0000−0003−2449−5477]], Paulo Quaresma[2[0000−0002−5086−059X]], Marlo Souza[3[0000−0002−5373−7271]], Daniela Barreiro Claro[3[0000−0001−8586−1042]], and Rafael Glauber[3]

[1] Pontifical Catholic University of Rio Grande do Sul (PUCRS), Porto Alegre, Brazil {sandra.abreu,joaquim.santos,bernardo.consoli,juliano.terra}@acad.pucrs.br, renata.vieira@pucrs.br

[2] University of Évora, Évora, Portugal pq@di.uevora.pt

[3] Federal University of Bahia (UFBA), Bahia, Brazil {msouza1,dclaro}@ufba.br, rglauber@dcc.ufba.br

Abstract. This work provides and overview of the Named Entity Recognition (NER) and Relation Extraction (RE) in Portuguese shared tasks in IberLEF 2019, its participant systems and results. These tasks sought to challenge Portuguese NER and RE systems by offering five new datasets for testing, two for RE and three for NER. Of these new datasets, two in particular offered a novel challenge for NER: the first composed of official police documents and the second composed of hospital's clinical notes. These cannot be published due to their sensitive nature, but the other three have been released for public use.

Keywords: Natural Language Processing · NLP · Named Entity Recognition · NER · Relation Extraction · RE · Portuguese Language · IberLEF 2019.

1 Introduction

Information Extraction (IE), a task in the field of Natural Language Processing (NLP), consists of obtaining relevant information from texts and representing it in a structured way. This representation can, for example, take the form of a list, table or graph, all of which can easily be used for storage, indexing, and query processing by standard database management systems.

Some examples of IE applications are Named Entity Recognition (NER) and Relation Extraction (RE). NER aims to identify and classify a given text's Named Entities (NEs) and their categories (Organization, Place, Person, among

others) [17]; RE aims to identify relations that occur between said entities [14]. For instance, the "affiliation" relation between Person-type and Organization-type NEs is one of those relations sought by RE systems. According to [22], the identification of NEs is the first step towards the semantic analysis of a text, being crucial to relation extraction systems. In the literature, we find several works that consider NER to be an integral part of RE systems [2, 12, 16], given that NER can help with the identification of NEs that may possess some kind of relation between themselves.

In order to explore and contribute to the state of RE and NER for Portuguese, we proposed three workshop tasks that were part of the IberLEF 2019 (Iberian Languages Evaluation Forum). The first involved annotating Portuguese texts with NER systems, while the second and third focused on the extraction of open relations in Portuguese texts. Participants were free to apply for any combination of activities, be it only one, two or all of them.

This paper is organized as follows: Section 2 describes Task 1 and its participant's results; Section 3 describes Task 2 and its participant's results; Section 4 describes Task 3 and its participant's results; and Section 5 presents the concluding remarks.

2 Task 1: Named Entity Recognition

The first task we proposed was NER. As explained previously, this is the task of identifying NEs within a given text and classifying them into one of several relevant categories or to a default category known as Miscellaneous. Our objective with this task was to evaluate the participant system's performance for three test datasets composed of texts in different genres. The first test dataset, composed of assorted news stories, memorandums, e-mails, interviews and magazine articles, was annotated for the following categories: Person (tagged PER), Place (tagged PLC), Organization (tagged ORG), Value (tagged VAL) and Time (tagged TME). The second and third datasets, the former composed of a hospital's clinical notes on patients and the latter of police documents, were only annotated for the Person (tagged PER) category.

The task consisted of the following steps:

- Development Phase: For this phase, participants were required to develop a computational approach to NER. This approach, hereby referred to as system, must be capable of solving NER tasks for the proposed textual genres. Participants were free to develop their solution however they saw fit, so long as they complied with the requirements described in the training and test phases;
- Training Phase: The objective of this phase was that participants choose their training datasets. Participants were free to choose any datasets they so desired for training their systems to solve the task in the proposed textual genres;
- Test Phase: In this phase the coordinators evaluated the capacity and reproductibility of the submitted systems:

- Reproduction Stage: For this stage, the participants' proposed systems were reproduced and executed by the coordinators. Should the coordinators have been unable to reproduce or execute a system, said system would have been disqualified;
- Evaluation Stage: The proposed corpora were inputted into all systems that passed the Reproduction Stage. The expected output was to be in the ".txt" format, so that it could be evaluated via the CoNLL-2002 script[21].

2.1 Test data

This task makes use of three datasets composed of texts in different textual genres in order to evaluate how the submitted systems behave when exposed to each set's particularities.

General Dataset: this dataset was built from two existing corpora, SIGARRA [20] and Second HAREM [6]. SIGARRA was chosen since it is a more recently published dataset, and was not used as trainning data by any of the participants. We added the Second HAREM dataset only to complete with examples from the Value category which was not given in SIGARRA.

SIGARRA, a dataset of news stories collected by the SIGARRA system created by the University of Porto, is composed of 1000 news stories taken from 17 Organic Units within the University. SIGARRA is annotated for the following NE categories: Person, Place, Organization, Date, Time, Event, Organic Unit and Course. However in our corpus we only considered the first five categories (Person, Place, Organization, Date and Time). In regard to both Date and Time categories we mapped them to a single category (Time), as in the HAREM golden collection.

Second HAREM, a Dataset that comprises 129 Brazilian and European Portuguese texts with 7255 named entities manually annotated. Despite being annotated for 10 categories, we only used sentences annotated for the Value category.

In total, 5055 sentences made up of 179892 tokens were extracted.

Clinical Dataset: clinical notes are textual data related by the hospital workers (nurse technicians, nurses, medical doctors...) about each of the hospital's patients, past and present. This kind of text contains names of patients, doctors and residents, results of medical exams and other assorted medical information. The Person category was manually annotated in a subset of the clinical notes. The manual annotation was made by 4 annotators on which each one annotated the clinical notes and after it was realized a discussion for the cases in which we did not have a consensus. The tool used for this annotation was WebAnno (https://webanno.github.io/webanno/), a web-based tool desined for usage on annotation tasks, we chose this tool due to the fact that it has a feature that allows direct export to the CoNLL-2002 format.

Clinical notes present particular challenges when it comes to their textual structure: words that should be separated by a space are not (for example "AnaR1")and several medical abbreviations.The pre-processing before annotation included only tokenization, so as to preserve the original formatting of this

data. In cases such as the above example, we understand "AnaR1" to be a Person, and we understand "####Paulo" to also be a Person.

In total, from 50 notes, we have 9523 tokens, summing up a total of 77 named entities of the Person category. As this date is of a sensitive nature, we cannot distribute it.

Police Dataset: a textual dataset from Brazil's Federal Police was manually annotated for the Person category. The data is divided in ten Testimony texts, ten Statement texts, and ten Interrogatory texts. These include names of deputies, scriveners and witnesses. This corpus contains well-structured, as well as grammatically correct texts, as they are all official documents. As with the clinical notes, the only pre-processing technique used on the text prior to annotation was tokenization.

In total, from 30 texts, we had 1,388 sentences, 37,706 tokens and a total of 916 named entities of the Person category. As for the clinical notes, we cannot distribute this dataset.

Table 1 shows the number of sentences and tokens by dataset. The quantity of named entities per category are shown in the Table 2.

Table 1: Number of tokens by dataset

Dataset	Number of Sentences	Number of Tokens
General Dataset	5,055	179,892
Clinical Dataset	50	9,523
Police Dataset	1,388	37,706

Table 2: Number of named entities by category

Categories	General Dataset	Clinical Dataset	Police Dataset
Person	2,159	77	916
Place	1,593	-	-
Organization	2,320	-	-
Time (*Date + Time*)	3,826	-	-
Value	106	-	-
Overall	10,004	77	916

2.2 Evaluation

Our evaluation process was divided in three stages, as shown in Figure 1 and described below:

i The participant's system are executed using one of the three proposed datasets as the input. Each system's expected output should have two columns: the

first containing the dataset's tokens (one per line of the column) and the second their predicted tags (as per the CoNLL-2002 format);

ii A third column, containing the expected tags, is then aligned with the other two;

iii The file generated in *ii* is used as an input for the CoNLL-2002 evaluation script, which calculates the final metrics.

In stage *ii*, the algorithm checks whether or not each of the output's tokens is the same as the expected token. Should the tokens be different, the algorithm returns the expected token and stops the alignment. This ensures that each system's output preserves the original dataset's integrity, for better and more accurate evaluation.

That said, none of the five systems evaluated output the expected sequence of tokens. As already mentioned, the Clinical Dataset has repeated sequences of the "#" character, and words joined by "_". All of these particularities are part of the text and of the language developed in the medical context. We then identified the instances of sequence breaking, and found that the systems were ignoring specific tokens, or capturing only part of them. An example of a partially captured token is in: "seg_sex" (medical abbreviation indicating the passage of time), where one system discarded everything that came after the "_".

Having alerted the participants of this, we asked them to resubmit their systems after altering them in such a way that they preserved in its totality the structure of the text. The results shown in the next section come from this second submission of the systems.

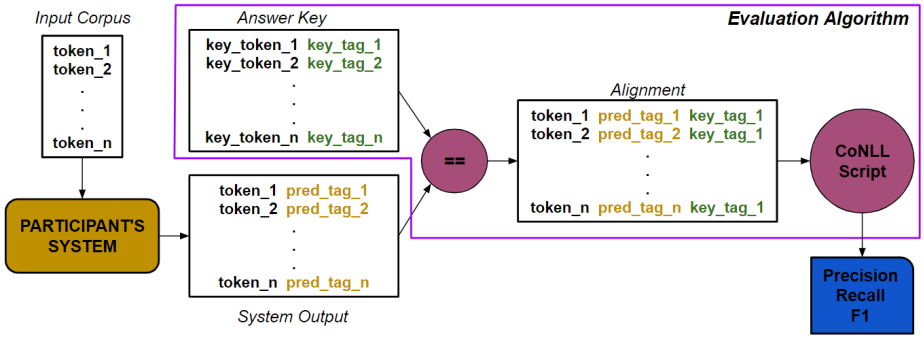


Fig. 1: Evaluation process

2.3 Results

We had five participants in total, one of whom submitted two systems.

1. BiLSTM-CRF-ELMo

2. CRF-LG
3. NLPyPort
4. CVT
5. BiLSTM-CRF-FlairBBP
6. Linguakit

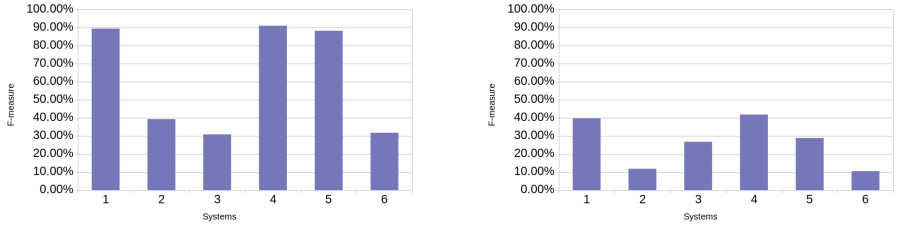
The evaluated systems used different training sets. The datasets used in the training process were: First HAREM[22], Second HAREM [6], MiniHAREM[23], WikiNER[18], Paramopama[13], LeNER-Br[4], FreeLing Corpus[5] and Datalawyer. Table 3 shows which datasets each participant used for training. *Linguakit* is not mentioned in Table 3 since it is based on rules and heuristics.

According to the results, we noticed that no single system had better F-measures for all datasets. The system with the best F-measure for the Police Dataset and Clinical Dataset was System 4. We also noted based on Figure 2 (a) and (b) that the best three systems for these two datasets used approaches based on Neural Network and Language Models. The results for the Police Dataset in particular showed a remarkable difference between approaches that were based on Neural Networks and those that were not.

System 5 achieved the best F-measure for the General Dataset. However, Figure 3 (a) shows that the results for the General Dataset had the least F-measure variance out of all test datasets. This can be explained by the fact that the General Dataset is structurally similar to the datasets used to train the systems. Figure 3 (b) shows the F-measure for the Organization category, where the highest score, achieved by System 5, was of over 45%. For the Person category, Figure 3 (c), System 1 had the best performance with an F-measure of over 80%. For the Place category, Figure 3 (d), the best performance was achieved by System 5 with an F-measure of over 58%. Figure 3 (e) shows results for the category, where the highest metric was the one achieved by System 5. For the Value category, Figure 3 (f), the best F-measure was achieved by System 4. Overall, the systems with higher F-measures used approaches based on Neural Networks. However, still on the General Dataset, the systems based on rules showed competitive results for the Organization category as seen in Figure 3 (b). Detailed results are presented in Appendix 1.

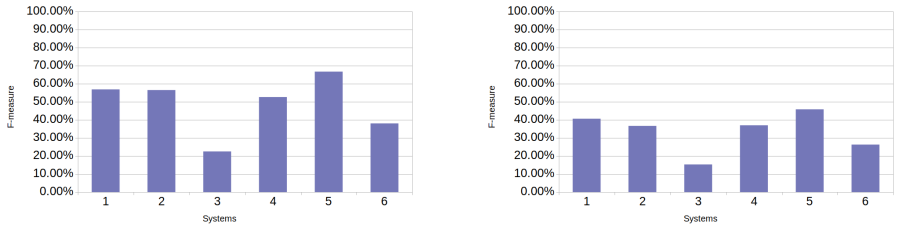
Table 3: Training datasets by System

System	BiLSTM CRF ELMo	CRF-LG	NLPyPort	CVT (Embeddings)	BiLSTM CRF FlairBBP
Training Corpora	WikiNER I HAREM MiniHAREM LeNER-Br Paramopama Datalawyer	I HAREM II HAREM MiniHAREM	II HAREM	LeNER-Br II HAREM FreeLing Corpus	I HAREM

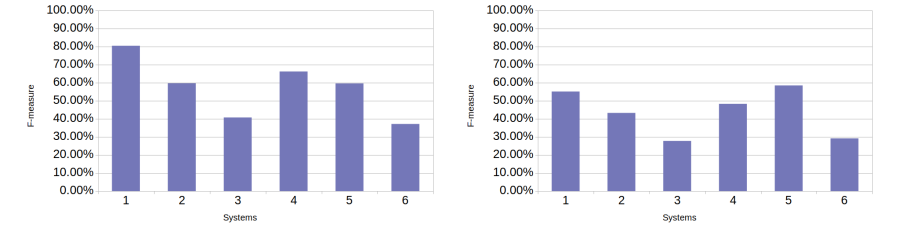


(a) Police Dataset - Person (b) Clinical Dataset - Person

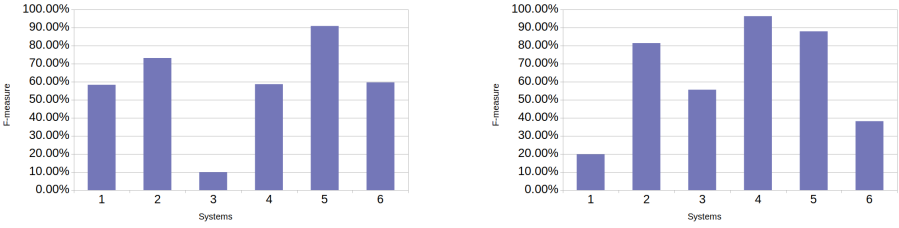
Fig. 2: Systems F-measure for the Police and Clinical Dataset



(a) Overall (b) Organization



(c) Person (d) Place



(e) Time (f) Value

Fig. 3: Systems F-measure for General Dataset

3 Task 2: Relation Extraction for Named Entities

The task of open relation extraction (RE) from texts faces many challenges, as it requires large amounts of linguistic knowledge and sophistication in the language processing techniques employed to solve it. We proposed a RE task that included the automatic extraction of a relation descriptor expressing any type of relation between a pair of NEs of the Person, Place and Organization categories, in Portuguese texts. The relation descriptor is defined as the text chunk that describes the explicit relation occurring between these entities in a sentence [7, 1].

For example, we have the relation descriptor "diretor de" (*director of*) that occurs between the NEs "Ronaldo Lemos" (PER) and "Creative Commons" (ORG) in the sentence below:

> "No próximo Sábado, Ronaldo Lemos, diretor da Creative Commons, irá participar de um debate [...]"
> *Next Saturday, Ronaldo Lemos, director of Creative Commons, will participate in a debate [...]*

The relation descriptor identified in the sentence is represented as a triple: (Ronaldo Lemos, diretor de, Creative Commons).

This RE task consisted of the following steps:

- Systems Development Phase: In this phase, the coordinators made a small annotated dataset available for the participants' use in developing their RE systems;
- Test Phase: The test phase included two options for participants:
  - Test 1: For this test, participants had to extract relation descriptors between NE pairs (all of which belonging to one of the following categories: PER, PLC or ORG) from data provided by the coordinators. This data was already annotated with NE information when provided, and as such did not need the application of a NER system by participants;
  - Test 2: For this test, the data provided was not annotated with NE information. As such, the objective of the task was to extract and classify (using only the following categories: PER, PLC or ORG) the NEs from the test sentences, and then they had also to extract the relation descriptors between pairs of the recognized NEs;
- Evaluation Phase: In this phase the participants sent their results from the Test Phase. Results were submited for Test 1 only. Afterwards, the analyzed results were sent back to the participants. The metrics used for evaluation phase were Precision, Recall and F-measure.

3.1 Resources

For the purpose of accomplishing this task, the coordinators provided subsets of Portuguese texts annotated with RE information as described in [7, 1]. The

authors presented a subset of the Golden Collections from the two HAREM conferences [22, 6], to which they added manual annotation of RE information expressed between NEs belonging to certain categories (ORG, PER and PLC). This resulted in a total of 516 RE annotated text instances, which were added to the 236 RE annotated texts from the Summ-it++ corpus [3, 8], for a total of 752 instances (positive and negative) of RE annotated texts. In Table 4 examples of positive relation instances are shown.

Table 4: Example of positive relation instances

Examples
(1) A **Marfinite** fica em o **Brasil**. Relation Descriptor: fica em Triple: (Marfinite, fica em, Brasil)
(2) **Hugo_Doménech**, professor de a **Universidade_Jaume_de_Castellón**. Relation Descriptor: professor de Triple: (Hugo_Doménech, professor de, Universidade_Jaume_de_Castellón)
(3) **António_Fontes** de a **AIPAN**. Relation Descriptor: de Triple: (António_Fontes, de, AIPAN)

The organizers selected 3 positive example subsets from the RE dataset for each step of Task 2. Table 5 shows the data distributed by NE pairs, for a total of 390 examples. For the Systems Development Phase, 90 positive examples annotated with relation descriptors (seeds) were made available for the participants. The available data for Test Phase (Test 1 and 2) was not annotated with the relation descriptors.

Table 5: Datasets

Phases	**NE Pairs**	**# Examples**	**# Total**
Systems Development Phase	ORG-ORG	29	90
	ORG-PER	31	
	ORG-PLC	30	
Test Phase - Test 1	ORG-ORG	47	149
	ORG-PER	56	
	ORG-PLC	46	
Test Phase - Test 2	ORG-ORG	47	151
	ORG-PER	56	
	ORG-PLC	48	

3.2 Evaluation

We considered two scores for Task 2 evaluation metrics: a completely correct relations score and a partially correct relations score. These were adapted from First HAREM's evaluation metrics for named entities [23].

- Completely Correct Relations (CCR): 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 (see Appendix 2);
- Partially Correct Relations (PCR): when at least one of the terms in the relation descriptors of the systems output corresponds to a term in the relarion descriptors of the key. The score for a partially correct relation is calculated as shown in the Appendix 2.

3.3 Results

For Task 2, the FactPyPort system participated on Test 1. Table 6 shows the results. Of the 149 examples in Test 1, FactPyPort system identified 144 examples and from those, 106 were Completely Correct Relations (CCR). There was no evaluation for Test 2 since there were no registered participants.

Table 6: Evaluation Results for Test 1

SYSTEM	P_{exact}	R_{exact}	F_{exact}	$P_{partial}$	$R_{partial}$	$F_{partial}$
FactpyPort	73.61%	71.14%	72.35%	76.62%	74.82%	75.71%

4 Task 3: General Open Relation Extraction

The task of general open relation extraction aims to identify structured representations of the information contained in unstructured sources, such as textual documents. This task faces many challenges, considering the generality of the problem, as well as the required linguistic knowledge to automatically perform such task.

This task involves the automatic extraction of any relation descriptor expressing any type of semantic relation between a pair of entities or concepts mentioned in Portuguese sentences. As before, a relation descriptor is defined as the text chunks that describe the explicit semantic relation, occurring between these entities in a sentence [7, 1]. This task is a generalization of Task 2 by removing the requirement of the entities being named in the text, meaning that any relation between two Noun Phrases (NP) is to be considered.

For example, the relation descriptor "diretor de" (*director of*) that occurs between noun phrases "Ronaldo Lemos" and "uma organização sem fins lucrativos" (*a non-profit organization*) in the sentence below:

> "No próximo Sábado, Ronaldo Lemos, diretor de uma organização sem fins lucrativos, irá participar de um debate [...]"
> (Next Saturday, Ronaldo Lemos, director of a non-profit organization, will participate in a debate [...])

The relation descriptor identified in the sentence is represented as a triple: (Ronaldo Lemos, diretor de, uma organização sem fins lucrativos).

The idea of this proposal is to request the participation of systems/solutions for the task of RE between NPs in Portuguese texts. The systems' results were evaluated using a set of annotated test data provided by the coordinators.

For the purpose of accomplishing this task, the coordinators will provide two sets of Portuguese texts: the first one, composing *Test 1*, is annotated with NPs information aims for the systems to identify the the relation descriptors, similar to what is provided in Task 2; the second, composing *Test 2*, is presented without any annotation, aiming to evaluate the system's capacity to identify relations and its arguments in texts. The authors present a set of 25 sentences annotated with NPs and RE information presented in [11].

This RE task consists of the following steps:

- Systems Development Phase: In this phase, the coordinators will make a small annotated dataset (seeds) available for the participants' use in developing their RE systems;
- Test Phase: The test phase includes two options for participants:
  - Test 1: For this test, participants must extract relation descriptors between NP pairs from data provided by the coordinators. This data will already be annotated with NP information when provided, and as such will not necessitate the application of a NER system by participants;
  - Test 2: For this test, the data provided will not be annotated with NP information. As such, the goal of the task will be to extract and classify the NPs from the test sentences, and then they must also extract the relation descriptors between pairs of the recognized NPs;
- Evaluation Phase: In this phase the participants will send their results from the Test Phase. They may submit results from Test 1, Test 2 or both to evaluation by the coordinators. Afterwards, the analyzed results will be sent back to the participants. The metrics used for evaluation phase will be Precision, Recall and F-measure.

4.1 Resources

Task 3 was evaluated using the Portuguese Open IE corpus proposed by Glauber et al [11]. This corpus is composed of 442 relation triples extracted from 25 setences obtained from sources such as the Portuguese section of Wikipedia, the CETENFolha corpus, movie reviews from Adoro Cinema and the Europarl corpus v7.0 [15]. The relations were manually extracted by 5 human annotators in two rounds.

Since the annotation of all possible extractions from a sentence is a hilghly subjective and, therefore, difficult task to perform systematically, the authors imposed some restrictions on the form of extractions that may appear in the corpus [11]:

C1 When there is a word chain through a preposition forming a noun phrase (NP), we first select the fragment that is composed of a noun, proper noun orpronoun, its respective determinants and direct modifiers (articles, numerals, adjectives and some pronouns).
C2 When a sentence has a transitive verb with preposition (indirect mode),the preposition will be attached to the realtion descriptor.
C3 We call minimal fact (minimal) any extracted fact having as arguments NPs composed only of a noun, proper noun or pronoun with its determinants and direct modifiers.
C4 If there are fragments with a noun function (preposition chain) that modify arguments in minimal facts, new facts (not minimal) must be added bythe annotator (see C3 second triple example).
C5 A fact must only be extracted from a sentence if it contains a propernoun or pronoun in, at least, one of the arguments.
C6 For n–ary facts, if there is no significant loss of information, the annotator must extract multiple binary facts.
C7 The coordinating conjunctions with additive function can generate multiple extracted facts and also a fact with the coordinated conjunction.
C8 Relations and arguments in the extracted facts must agree in number.

The corpus was divided into three fragments containing randomly selected relations: a training fragment (train dataset) composed of 90 realtion triples that were made available to the participants in the Systems Development Phase; and two test fragments each composed of 176 relation triples used to evaluate the systems submissions for Test 1 (test 1 dataset) and Test 2 (test 2 dataset). The three fragments are pair-wise disjoint and each fragment contains extractions from all the 25 sentences that compose the original annotatted corpus.

4.2 Evaluation

General Open Relation Extraction is concerned with indentifying all possible information contained within a sentence, without any a priori restriction on the kind of relation to be extracted. Since the train and test datasets were composed of relations extracted from the same 25 sentences, the evaluation for Task 3 needs to consider the possibility of a system correctly identifying a relation and this relation not being in the test dataset being considered, as such we proceed with different evaluation scenarios.

For Test 1, since the relations to be extracted were pre-defined a priori by setting the arguments, the participating systems needed only to identify the relation descriptor. As such, we performed one evaluation scenario using the test 1 dataset as golden resource and comparing it to the participants' systems outputs.

For Test 2, however, due to the fact that the training and test datasets were constructed by taking examples of extractions from the same set of 25 sentences, we performed the evaluations of the participanting systems in four different evaluation scenarios. The reason for this is that in the simplest scenario (Scenario 1) in which only the test 2 dataset is used to compute the evaluation metrics, the systems may suffer for identifying correct relations which are not in this dataset, i.e. the correct relation was selected for the train or test 1 datasets in the corpus fragmentation. As such, we consider the following scenarios:

Scenario 1) Test 2 dataset is the golden resource. Since the system may perform correct extractions that do not appear in test 2 dataset, e.g. relations contained in test 1 or train dataset, the evaluation of the system's precision may be affected.
Scenario 2) Test 2 dataset contains the target relations to be identified, but we matched the extractions performed by the participant systems against the relation in test 2 and train datasets. The metrics are then computed considering only those relations that were matched with some relation in the test 2 dataset, since the relations in the train dataset were known *a priori* by the systems and cannot be used in the evaluation. In this evaluation, the systems may suffer in precision for not considering those relations in the Test 1 dataset, but also in recall due to the fact that some relations that could be partially matched with relations in the test 2 dataset may have been matched with relations in the trainning dataset.
Scenario 3) in this scenario, we consider as target relations, those contained in Test 1 and Test 2 datasets and matched the extractions performed by the participant systems against them, as well the relations contained in the trainning dataset. The metrics are then computed considering only those relations that were matched with some target relation, disregarding those that were matched with relations in the trainning dataset. In this evaluation, the systems may suffer in recall due to the fact that some relations that could be partially matched with relations in the test 1 or test 2 dataset may have been matched with relations in the trainning dataset.
Scenario 4) in this scenario we consider the union of all three datasets as golden resource and compute the evaluation metrics for each system. In this evaluation, the systems may have gained in precision and recall, since the realtions in the train dataset have been provided in advance for the systems to train over.

As for Task 2, we adapted the First HAREM's evaluation metrics for named entity recognition [23] to the task of relation extraction considering both a completely correct relations score and a partially correct relations score (see Appendix 2). We also followed the same matching strategy as used in Task 2.

4.3 Results

For Test 1, we had two groups participating in the task: group 1 submitted two systems - DEPENDENTIE [19] and DPTOIE; group 2 submitted three systems - ICEIS [26], INFERPORTIE [25] and PRAGMATICOIE [24].

Table 7: Evaluation Results for Test 1

SYSTEM	P_{exact}	R_{exact}	F_{exact}	$P_{partial}$	$R_{partial}$	$F_{partial}$
DEPENDENTIE	1.14%	1.14%	1.14%	1.42%	1.35%	1.38%
DPTOIE	3.41%	3.41%	**3.41%**	4.40%	4.29%	**4.34%**
ICEIS	1.14%	1.14%	1.14%	1.28%	1.42%	1.35%
INFERPORTOIE	0.00%	0.00%	0.00%	0.36%	0.45%	0.40%
PRAGMATICOIE	0.00%	0.00%	0.00%	0.36%	0.45%	0.40%

Considering only the exact matches, the system DPTOIE had the best result with F_{exact} of 3.4% and the systems ICEIS and DEPENDENTIE achieved 11% of F_{exact}, while the systems INFERPORTIE and PRAGMATICOIE had negligible results. When considering partial matches, DPTOIE had the best results, achieving a $F_{partial}$ score of 4.3%.

For Test 2, we had three groups participating in the task: the two groups that participated in Test 1 and a third group which submitted the system Linguakit 2, an adpated version of the relation extraction module of Linguakit [10] described in [9] .

Table 8: Evaluation Results of Test 2 (Scenario 1)

SYSTEM	P_{exact}	R_{exact}	F_{exact}	$P_{partial}$	$R_{partial}$	$F_{partial}$
DEPENDENTIE	0.00%	0.00%	0.00%	26.43%	6.42%	10.33%
DPTOIE	3.92%	3.41%	3.65%	27.20%	29.60%	**28.35%**
ICEIS	4.62%	1.70%	2.49%	27.75%	11.34%	16.10%
INFERPORTOIE	1.59%	0.57%	0.84%	21.42%	8.47%	12.14%
Linguakit 2	19.61%	5.68%	**8.81%**	38.77%	11.55%	17.79%
PRAGMATICOIE	1.47%	0.57%	0.82%	20.60%	8.67%	12.21%

In Scenario 1, considering only the exact matches, the system Linguakit 2 had the best results with F_{exact} of 8,8%. When considering partial matches, DPTOIE had the best results, achieving a $F_{partial}$ score of 28.3%.

Table 9: Evaluation Results of Test 2 (Scenario 2)

SYSTEM	P_{exact}	R_{exact}	F_{exact}	$P_{partial}$	$R_{partial}$	$F_{partial}$
DEPENDENTIE	0.00%	0.00%	0.00%	27.38%	4.85%	8.24%
DPTOIE	6.06%	3.41%	4.36%	29.48%	19.19%	**23.25%**
ICEIS	6.00%	1.70%	2.65%	30.72%	9.21%	14.17%
INFERPORTOIE	2.08%	0.57%	0.89%	22.39%	6.29%	9.82%
Linguakit 2	26.32%	5.68%	**9.35%**	44.82%	9.71%	15.96%
PRAGMATICOIE	1.96%	0.57%	0.88%	21.89%	6.43%	9.94%

In Scenario 2, considering only the exact matches, the system Linguakit 2 had the best results with F_{exact} of 9.35%. When considering partial matches, DPTOIE had the best results, achieving a $F_{partial}$ score of 23.25%. Notice that, as conjuctured, the systems' precisions were positivelly impacted in the evaluation Scenario 2, but due to the fact that some relations have been partially matched with relations in the train dataset, the $R_{partial}$ has been negativelly impacted in this scenario.

Table 10: Evaluation Results of Test 2 (Scenario 3)

SYSTEM	P_{exact}	R_{exact}	F_{exact}	$P_{partial}$	$R_{partial}$	$F_{partial}$
DEPENDENTIE	5.71%	0.57%	1.03%	34.73%	3.44%	6.26%
DPTOIE	11.11%	3.69%	5.54%	35.60%	12.78%	**18.81%**
ICEIS	8.93%	1.42%	2.45%	34.08%	5.63%	9.67%
INFERPORTOIE	1.92%	0.28%	0.50%	25.52%	3.86%	6.71%
Linguakit 2	33.33%	4.26%	**7.56%**	51.84%	6.53%	11.60%
PRAGMATICOIE	1.79%	0.28%	0.49%	25.48%	4.10%	7.06%

In Scenario 3, considering only the exact matches, the system Linguakit 2 had the best results with F_{exact} of 7.56%. When considering partial matches, DPTOIE had the best results, achieving a $F_{partial}$ score of 18+81%. Notice that since the number of target relations has increased greatly from the previous scenarios to Scenario 3, we can percieve a decrease in the systems' Recall, associated with an increase in their Precision.

Table 11: Evaluation Results of Test 2 (Scenario 4)

SYSTEM	P_{exact}	R_{exact}	F_{exact}	$P_{partial}$	$R_{partial}$	$F_{partial}$
DEPENDENTIE	5.00%	0.45%	0.83%	34.93%	3.13%	5.75%
DPTOIE	10.46%	3.62%	5.38%	36.44%	13.98%	**20.21%**
ICEIS	9.23%	1.36%	2.37%	34.74%	5.31%	9.21%
INFERPORTOIE	7.94%	1.13%	1.98%	32.11%	4.59%	8.03%
Linguakit 2	37.25%	4.30%	**7.71%**	55.35%	6.26%	11.25%
PRAGMATICOIE	7.35%	1.13%	1.96%	31.81%	4.85%	8.42%

In Scenario 4, considering only the exact matches, the system Linguakit 2 had the best results with F_{exact} of 7.71%. When considering partial matches, DPTOIE had the best results, achieving a $F_{partial}$ score of 20.21%. Given that no extraction of the system is excluded from evaluation, it is noticible that the systems' recall scores have improved, when compared to those achieved in Scenario 3, specially considering partial matches. Notice, however, that the performance of some systems, such as DEPENDENTIE and ICEIS, have decreased, due to the fact the the number of target relations to be extracted have increased - in-

dicating that these systems have partially macthed some of the relations in the train dataset, thus performing decreasing the overall performance of the systems when considering those relations in the evaluation.

Despite the fact that we performed 4 evaluation scenarios, the overall evaluation of the systems have remained consistent across the different scenarios. This indicates that our evaluation results are robust - considering the particularities of the dataset considered in this task. Overall, the systems DPTOIE and Linguakit 2 have performed best in all evaluation scenarios, with Linguakit 2 dominating the exact match evaluations and DPTOIE the partial matches evaluations. These facts indicate that, while both systems were able to extract a great deal of the manually identified relations in the corpus, Linguakit 2 is the most consistent in their extractions with the restrictions in the extractions imposed by the dataset, while DPTOIE is capable of extracting a great number of relations from the sentences.

5 Concluding Remarks

In this work, we presented three tasks involving annotating Portuguese texts with NER systems and open relation extraction in Portuguese texts. As a result, we had a total of eleven teams registered to participate on the proposed tasks. Seven of them sent their results, while four dropped out. In the ende, a total of thirteen submissions from 6 different institutions were evaluated, as presented in Table 12.

Table 12: Participating teams by task

Task	**Teams**	**Systems**
Task 1	CISUC, University of Coimbra	NLPyPort
	CiTIUS, University of Santiago de Compostela	CVT
	CiTIUS, University of Santiago de Compostela	LinguaKit
	Pontifícia Universidade Católica do Rio Grande do Sul	BiLSTM-CRF-FlairBBP
	Universidade Federal do Espírito Santo	CRF-LG
	Universidade Federal de Goiás	BiLSTM-CRF-ELMo
Task 2	CISUC, University of Coimbra	FactpyPort
Task 3	CiTIUS, University of Santiago de Compostela	LinguaKit 2
	Universidade Federal da Bahia - Team 1	DEPENDENTIE
	Universidade Federal da Bahia - Team 1	DPTOIE
	Universidade Federal da Bahia - Team 2	ICEIS
	Universidade Federal da Bahia - Team 2	INFERPORTOIE
	Universidade Federal da Bahia - Team 2	PRAGMATICOIE

As a contribution of this work, we made available annotated datasets for RE in Portuguese; we evaluated different systems/solutions for NER and RE; and we had the oportunity to test the solutions/systems for NER on various textual genres. The resources to reproduce this work are available in our GitHub

(https://github.com/jneto04/iberlef-2019). As future work we would like to propose an evaluation of the systems where the same data sets are used for trainning the systems.

Appendix 1 - Detailed systems results

Table 13: BiLSTM-CRF-ELMo's results

System	Corpus	Category	Prec	Rec	F1
BiLSTM-CRF-ELMo	Police Dataset	PER	86.14%	92.82%	89.35%
	Clinical Dataset	PER	32.47%	51.02%	39.68%
	General Dataset	Overall	63.11%	51.69%	56.83%
		ORG	57.90%	31.26%	40.60%
		PER	83.93%	77.17%	80.41%
		PLC	54.61%	55.59%	55.10%
		TME	59.49%	57.10%	58.27%
		VAL	11.32%	80.00%	19.83%

Table 14: CRF-LG's results

System	Corpus	Category	Prec	Rec	F1
CRF-LG	Police Dataset	PER	29.59%	58.41%	39.28%
	Clinical Dataset	PER	14.29%	10.09%	11.83%
	General Dataset	Overall	56.26%	49.26%	52.53%
		ORG	42.27%	31.77%	36.28%
		PER	57.39%	60.62%	58.96%
		PLC	37.35%	49.34%	42.52%
		TME	71.33%	73.68%	72.48%
		VAL	80.19%	82.52%	81.34%

Table 15: NLPyPort's results

System	Corpus	Category	Prec	Rec	F1
NLPyPort	Police Dataset	PER	21.72%	53.07%	30.83%
	Clinical Dataset	PER	27.27%	26.25%	26.75%
	General Dataset	Overall	26.08%	19.78%	22.50%
		ORG	19.41%	12.62%	15.30%
		PER	50.07%	34.34%	40.74%
		PLC	42.31%	20.64%	27.75%
		TME	8.99%	11.11%	9.94%
		VAL	56.60%	54.55%	55.56%

Table 16: CVT's results

System	Corpus	Category	Prec	Rec	F1
CVT	Police Dataset	PER	92.20%	89.73%	90.95%
	Clinical Dataset	PER	36.36%	49.12%	41.79%
	General Dataset	Overall	61.27%	46.07%	52.60%
		ORG	54.24%	28.04%	36.97%
		PER	75.64%	58.83%	66.18%
		PLC	55.93%	42.47%	48.28%
		TME	58.68%	58.57%	58.62%
		VAL	96.23%	96.23%	96.23%

Table 17: BiLSTM-CRF-FlairBBP's results

System	Corpus	Category	Prec	Rec	F1
BiLSTM-CRF-FlairBBP	Police Dataset	PER	94.21%	82.82%	88.15%
	Clinical Dataset	PER	22.08%	41.46%	28.81%
	General Dataset	Overall	75.28%	59.82%	66.66%
		ORG	65.13%	35.32%	45.80%
		PER	65.96%	54.33%	59.58%
		PLC	55.81%	61.40%	58.47%
		TME	94.43%	87.44%	90.80%
		VAL	88.68%	87.04%	87.85%

Table 18: LinguaKit's results

System	Corpus	Category	Prec	Rec	F1
LinguaKit	Police Dataset	PER	40.83%	25.92%	31.71%
	Clinical Dataset	PER	22.08%	6.88%	10.49%
	General Dataset	Overall	44.89%	32.97%	38.01%
		ORG	38.40%	19.99%	26.29%
		PER	56.79%	27.59%	37.14%
		PLC	39.61%	23.09%	29.17%
		TME	44.59%	89.79%	59.59%
		VAL	34.91%	42.05%	38.14%

Appendix 2 - Metrics calculation details

The evaluation metrics are: Precision, Recall and F-measure, where:

CR = correct relations
IR = identified relations
TR = total relations

- Precision (P): measures the proportion of correct responses when compared to the sum of all responses given by the system;
- Recall (R): measures the percentage of answers the systems can give when compared to all answers available in the key;
- F-measure (F): combines the metrics of Precision and Recall.

Considering only the Completly Correct Relations (CCR):

$$P_{exact} = \frac{CR}{IR} \tag{1}$$

$$R_{exact} = \frac{CR}{TR} \tag{2}$$

$$F_{exact} = \frac{(2 * P_{exact} * R_{exact})}{(P_{exact} + R_{exact})} \tag{3}$$

The score for a partially correct relation is calculated as

$$PCR = 0.5 \cdot \frac{\#\text{correct terms in the annotation}}{\text{greatest value from terms in the key and the system's output}} \tag{4}$$

Considering both CCR and PCR:

$$P_{partial} = \frac{(CR + PCR)}{IR} \tag{5}$$

$$R_{partial} = \frac{(CR + PCR)}{TR} \tag{6}$$

$$F_{partial} = \frac{(2 * P_{partial} * R_{partial})}{(P_{partial} + R_{partial})} \tag{7}$$

To compute the partially correct score, we matched each of the systems outputs R_1 to a relation in the corpus (R_2) that maximize the following matching score, in with $match(R_1, R_2)$ denotes the number of terms in common between the two extractions and $len(R)$ denotes the number of terms in the extraction:

$$score(R_1, R_2) = (2 \cdot match(R_1, R_2) - (len(R_1) + len(R_2))) - |len(R_1) - len(R_2)| \tag{8}$$

Notice that the matching score minimizes mismatches between the relations R_1 R_2, i.e. it is maximal and when the relations R_1 ad R_2 are an exact matches, i.e. the same relation. When the match is only partial, the score privileges matchs with the fewer number of mismatched terms. Notice that the term $|len(R_1) - len(R_2)|$ is used to guarantee that the relations are the closest possible, i.e. it is used to rule out macth candidates with high number of matched tokens but differring too much from relation (R_1).

Acknowledgments

We thank the CNPQ, CAPES and FAPERGS for their financial support.

References

1. Collovini de Abreu, S., Vieira, R.: Relp: Portuguese open relation extraction. Knowledge Organization **44**(3), 163–177 (2017)
2. Agichtein, E., Gravano, L.: Snowball: Extracting relations from large plain-text collections. In: 5th ACM International Conference on Digital Libraries. pp. 85–94 (2000)
3. Antonitsch, A., Figueira, A., Amaral, D., Fonseca, E., Vieira, R., Collovini, S.: Summ-it++: an enriched version of the summ-it corpus. In: of the Language Resources and Evaluation Conference (LREC). pp. 2047–2051 (2016)
4. de Araujo, P.H.L., de Campos, T.E., de Oliveira, R.R., Stauffer, M., Couto, S., Bermejo, P.: Lener-br: A dataset for named entity recognition in brazilian legal text. In: International Conference on Computational Processing of the Portuguese Language. pp. 313–323. Springer (2018)
5. Carreras, X., Chao, I., Padró, L., Padró, M.: Freeling: An open-source suite of language analyzers. In: LREC. pp. 239–242 (2004)
6. Carvalho, P., Oliveira, H.G., Mota, C., Santos, D., Freitas, C.: Segundo harem: Modelo geral, novidades e avaliação. In: Mota, C., Santos, D. (eds.) Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM (2008)
7. Collovini, S., Machado, G., Vieira, R.: Extracting and structuring open relations from portuguese text. In: Computational Processing of the Portuguese Language - 12th International Conference, PROPOR 2016. Lecture Notes in Computer Science, vol. 9727, pp. 153–164. Springer, Tomar, Portugal (2016)
8. Collovini, S., Pereira, B., dos Santos, H.D.P., Vieira, R.: Annotating relations between named entities with crowdsourcing. In: Natural Language Processing and Information Systems - 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018. pp. 290–297. Paris, France (2018)
9. Gamallo, P., García, M.: Multilingual open information extraction. In: Proceedings of Progress in Artificial Intelligence - 17th Portuguese Conference on Artificial Intelligence, EPIA 2015. pp. 711–722. Coimbra, Portugal (2015)
10. Gamallo, P., Garcia, M.: Linguakit: uma ferramenta multilingue para a análise linguística e a extração de informação. Linguamática **9**(1), 19–28 (2017)
11. Glauber, R., de Oliveira, L.S., Sena, C.F.L., Claro, D.B., Souza, M.: Challenges of an annotation task for open information extraction in portuguese. In: International Conference on Computational Processing of the Portuguese Language. pp. 66–76. Springer (2018)

12. Hasegawa, T., Sekine, S., Grishman, R.: Discovering relations among named entities from large corpora. In: ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. p. 415. Association for Computational Linguistics, Morristown, NJ, USA (2004)
13. Júnior, C.M., Macedo, H., Bispo, T., Santos, F., Silva, N., Barbosa, L.: Paramopama: a brazilian-portuguese corpus for named entity recognition. Encontro Nac. de Int. Artificial e Computacional (2015)
14. Jurafsky, D., Martin, J.H.: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Prentice Hall series in artificial intelligence, Pearson Education Ltd., London, 2 edn. (2009)
15. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: MT summit. vol. 5, pp. 79–86 (2005)
16. Li, H., Bollegala, D., Matsuo, Y., Ishizuka, M.: Using graph based method to improve bootstrapping relation extraction. In: CICLing (2). pp. 127–138 (2011)
17. Mota, C., Santos, D., Ranchhod, E.: Avaliação de reconhecimento de entidades mencionadas: Princípio de harem. In: Santos, D. (ed.) Avaliação Conjunta: Um novo paradigma no processamento computacional da língua portuguesa, chap. 14, pp. 161–176. IST Press (2007)
18. Nothman, J., Ringland, N., Radford, W., Murphy, T., Curran, J.R.: Learning multilingual named entity recognition from wikipedia. Artificial Intelligence **194**, 151–175 (2013)
19. de Oliveira, L.S., Glauber, R., Claro, D.B.: Dependentie: an open information extraction system on portuguese by a dependence analysis. Encontro Nacional de Inteligência Artificial e Computacional (2017)
20. Pires, A.R.O.: Named entity extraction from Portuguese web text. Master's thesis, Faculdade de Engenharia da Universidade de Porto, Porto, Portugal (2017)
21. Sang, T.K., Erik, F.: Introduction to the conll-2002 shared task: language-independent named entity recognition. In: Proceedings of CoNLL-2002. pp. 155–158 (2002)
22. Santos, D., Cardoso, N.: Breve introdução ao HAREM, chap. 1, pp. 1–16. Linguateca (2007)
23. Santos, D., Cardoso, N.: Reconhecimento de entidades mencionadas em português: Documentação e atas do HAREM, a primeira avaliação conjunta na área. Linguateca, Lisboa, PT (2007)
24. Sena, C.F.L., Claro, D.B.: Pragmatic information extraction in brazilian portuguese documents. In: International Conference on Computational Processing of the Portuguese Language. pp. 46–56. Springer (2018)
25. Sena, C.F.L., Claro, D.B.: Inferportoie: A portuguese open information extraction system with inferences. Natural Language Engineering **25**(2), 287–306 (2019)
26. Sena, C.F.L., Glauber, R., Claro, D.B.: Inference approach to enhance a portuguese open information extraction. In: ICEIS (1). pp. 442–451 (2017)