NER and Open Information Extraction for Portuguese

Notebook for IberLEF 2019 Portuguese Named Entity Recognition and Relation Extraction Tasks

Pablo Gamallo[1], Marcos Garcia[2], and Patricia Martín-Rodilla[1]

[1] Centro de Investigación en Tecnoloxías Intelixentes (CiTIUS)
University of Santiago de Compostela, Galiza
{pablo.gamallo,patricia.martin.rodilla}@usc.es
[2] Universidade da Coruña, CITIC, Grupo LyS, Departamento de Letras, Galiza
marcos.garcia.gonzalez@udc.gal

Abstract This article describes the different systems we have developed to participate at the IberLEF 2019 Portuguese Named Entity Recognition and Relation Extraction Tasks (NerReIberLEF2019). Our objective is to compare rule-based and neural-based approaches. For this purpose, we applied our systems to two specific subtasks: Named Entity Recognition (Task 1) and General Open Information Extraction (Task 3) in Portuguese texts.

1 Introduction

The use of neural networks in tasks related to language technology and natural language processing (NLP) is currently rising very rapidly to the point that non-neural methods, including rule-based strategies, suffer at this momment a very large decline in popularity. However, it is important to know in which specific NLP tasks neural-based methods outperform other strategies and in which they do not. In a recent work [18], the authors assessed whether certain grammatical phenomena are more challenging for neural networks to learn than others. It is also important to take into account which are the characteristics of the target language, given that it is not the same to perform experiments on a Germanic language such as English, or a Latin one such as Portuguese, or even an Uralic language like Finnish with a very rich morphological base. In a recent work by [17] focused on comparing parsing methods for Finnish using neural and rule-based strategies, rule-based methods still outperform neural networks at a considerable distance.

In this article, we directly compare a neural-based tool for Named Entity Recognition (NER) with a rule-base system using the same test dataset. In addition, we also tested a rule-based strategy for Open Information Extraction (OIE), which is a complex task traditionally addressed through unsupervised or rule-based approaches. To the best of our knowledge, the recent work reported in [5] is the first time that the OIE task is addressed using a neural approach with promissing results. However, that system is still

dependent on traditional strategies as the neural OIE model described in the paper was trained with highly confident binary extractions bootstrapped from a state-of-the-art OIE system [14].

To evaluate the proposed systems and make the corresponding comparisons, we participated at the IberLEF 2019 Portuguese Named Entity Recognition and Relation Extraction Tasks (NerReIberLEF2019) [4].[1] The main goal is to allow participants to apply their systems to several tasks, including NER and OIE in Portuguese texts. These shared tasks are part of IberLEF 2019.

The article is organized as follows. In Section 2, we describe the two systems submitted for the NER task, while Section 3 describes the properties of our OIE approach. Experiments and evaluation are reported in Section 4, and conclusions are addressed in Section 5.

2 Named Entity Recognition for Portuguese

Two very different NER strategies have been developed, rule-based and neural-based, which are described in the following subsections.

2.1 A Rule-Based Approach with External Resources

We have adapted the NER module integrated in LinguaKit [7] and described in [8].[2] The NER is constituted by two kinds of rules: first, identification heuristics to select named entities from texts, and second, classification rules applied on previously identified named entities in order to classify them as Location, Person, Organization, or Miscellaneous. All rules require external resources to be applied.

Identification heuristics make use of lexicographic resources such as a lexicon of tokens and lemmas. Rules take into account letter capitalization of tokens, their position in the sentence, and lexicon membership. Considering these elements, a basic identification rule is the following example:

> *If a token with initial uppercase letter starts a sentence and it is not part of the regular lexicon, then it is a named entity candidate.*

Classification rules take identified named entities as input and assign them a semantic class. Two external resources are required: both a list of gazetteers for locations, persons and organizations, as well as a list of trigger words for the same three classes. These resources were automatically generated from Wikipedia. Given an identified named entity (NE), the classification algorithm works as follows. First, it verifies if the NE is an unambiguous expression appearing in just one gazetteer. If this is the case, it is assigned the class of the gazetteer. Second, if the NE appears in various gazetteers (ambiguity) or it is unknown (missing in gazetteers), then a disambiguation process is activated by searching relevant trigger words within its linguistic context. For instance, "Santiago" is an ambiguous NE that can be either a person or a location. In the following expression:

"Santiago é uma cidade galega" (*Santiago is a Galician town*)

It refers to a town and, therefore, should be classified as a location. In order to disambiguate it, the common noun "cidade" (*town*) is a trigger word in the list of locations that is used to select the appropriate class of the NE. If there are several trigger words of different classes in the context of the target NE, we give preference to the closest one. If there are two triggers at the same distance, the preference is given to the left position. If the NE remains ambiguous as cannot be disambiguated by using contextual triggers, then we check if its constituent expressions belong to the gazetteers or trigger words and apply the previous rules. If no rule is applied, then the NE is classified as *miscellaneous*.

To adapt the NER module to the shared task requirements, we have added specific rules for dates, currencies and measures. The new rules are applied on external lists of currency names and measures as well as their usual abbreviations, e.g., *cm* for centimeters, or *min* for minutes.

2.2 A Neural-Based Approach with Cross-View Training

Our neural network approach to NER in Portuguese was based in Cross-View Training (CVT), which performs semi-supervised learning by combining supervised and unsupervised methods [2][3]. CVT improves the representation of a bidirectional long short-term memory encoder (Bi-LSTM) by adding, together with the annotated data, unlabeled representations to the input. In a NER scenario, CVT uses the unlabeled data to learn the different contexts in which a named entity occurs, apart from different properties (e.g., sequences of characters) of each entity type.

On the one hand, a CVT model needs annotated data for the desired task to be trained on. On the other hand, it also requires a large unlabeled corpus for the unsupervised learning process.

We obtained our supervised training data from the following resources:

- The corpus used to train the FreeLing NER modules for Portuguese [8,15,11]. As it had been labeled only with 'enamex' entities (Person, Place, and Organization), Value (VAL) and Time (TME) tags were automatically added with LinguaKit. Then, it was carried out a brief revision to correct the most frequent errors.
- LeNER [1]. This dataset, of Brazilian legal texts, was preprocessed by removing all the NE tags different from the ones used in the shared task.
- HAREM [20]. We used the NLTK-format corpus provided by [16][4].

It is worth noting that annotation guidelines used in these three corpora differ, namely the ones used in HAREM. For instance, the initial prepositions of prepositional phrases containing temporal expressions are labeled as 'TEMPO' by HAREM ("Durante$_{\text{B-TEMPO}}$ os$_{\text{I-TEMPO}}$ desolados$_{\text{I-TEMPO}}$ anos$_{\text{I-TEMPO}}$ Reagan$_{\text{I-TEMPO}}$"), while the other datasets consider they do not belong to the named entity [8,1]. In this respect, we automatically removed some differences by harmonizing the HAREM annotation with the one used in [8]. Apart from that, there are several other differences concerning the annotation of each resource, such as the representation of contractions, which some datasets keep in a single token while others split them in two elements. Obviously, training machine learning models in mixed resources from several datasets have an impact on the training process.

After the automatic processing of the corpora, they were merged into a single file, which was randomly splited in two sets (for training, and development). The size (in number of tokens) of the *train* set is of 898,157, while the *dev* has 50,120 tokens.

As unlabeled corpora, we combined resources from different varieties of Portuguese, totaling about 600 milion tokens: Wikipedia (300M), Jornal Público (215M), Jornal do Brasil (60M), and Europarl (31M). Additionally, we initialized the CVT model with the pre-trained GloVe embeddings described in [13]. The word embeddings have 300 dimensions, and the LSTM 1024 hidden layers.

Figure 1 shows the performance of our model in the *dev* set depending on the training steps.[5] As it can be seen, the improvement after 200k steps is very small, so we stopped at 250k with the following results: 88.89 precision; 93.11 recall, and 90.95 f1.[6]

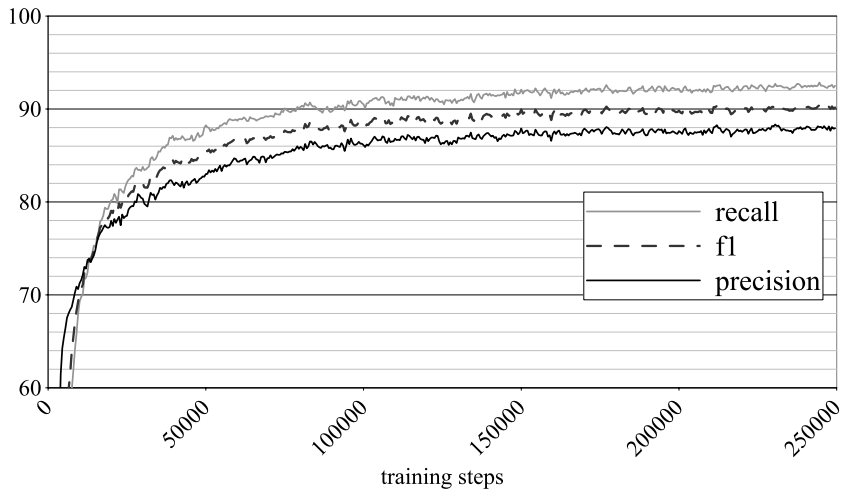

Figure 1. Precision, recall, and f-score of CVT model in the *dev* set *versus* training steps.

3 Open Information Extraction for Portuguese

The OIE system we have used for the shared task is an adapted version of the corresponding module installed in LinguaKit and described in [9] and [10]. The OIE module consists of two tasks: identification of argument structures and generation of relations (triples).

3.1 Argument Structure

Each clause has an argument structure which relies on a verb. To identify argument structures, the system takes a parsed sentence as input represented by means of the

dependency-based ConLL-X format. For each verb (V), the system selects all dependents whose syntactic function can be part of its argument structure. The functions considered to build an argument structure are the following: *subject* (S), *direct object* (O), *attribute* (A), and all complements headed by a preposition (C). So, there is no distinction between obligatory *vs.* optional arguments. Five types of argument structures were defined: SVO, SVC+, SVOC+, SVA, SVAC+, where "C+" means one or more complements.

Within a sentence, it is possible to find several argument structures corresponding to different clauses. Let us see an example. Table 1 shows three argument structures extracted from one of the input sentences of the test dataset provided organizers of NerReIberLEF2019. This example is quite complex as it includes a relative clause whose antecedent is not just a nominal phrase but a whole clause, namely the antecedent of "o que" (*which*) is the clause "Erlynne se envolve com Robert" (*Erlynne gets involved with Robert*). Our system wrongly substitutes the relative pronoun by "Robert" since the dependency parser identified this proper noun, and not the verb "se envolve" (*gets involved*), as the antecedent. While the two first argument structures in Table 1 are correct, the third one is wrong because of that odd dependency concerning the relative clause and its antecedent.

	Type	**Constituents**
1	SVC	**S**="Erlynne", **V**="se envolve", **C**="com Robert" *Erlynne, gets involved, with Robert*
2	SV0C	**S**="ele", **V**="está traindo", **O**="Meg" **C**="com a visitante" *he, is betraying, Meg, with the visitor*
3	SVO	**S**="Robert", **V**="gera", **O**="rumores" *Robert, generates, rumors*

Table 1. Three argument structures extracted by our system from the sentence "Erlynne se envolve com Robert, o que gera rumores de que ele está traindo Meg com a visitante" (*Erlynne gets involved with Robert, which generates rumors that he is betraying Meg with the visitor.*, which is part of the testing dataset of Task 3 at NerReIberLEF2019.

3.2 Generation of Triples

Once the argument structures have been detected in the previous task, the OIE system builds a set of verbal relations (triples) with two arguments. These *Arg1-Verb-Arg2* relations represent basic propositions or facts standing for minimal units of coherent, meaningful, and non over-specified information. For example, from the second argument structure in Table 1, two triples are generated and showed in Table 2.

To adapt the LinguaKit module to the requirements of task 3 at NerReIberLEF2019, we made some little adjustments of the OIE system by taken into account the annotation criteria found in the training/development dataset provided by the organizers. In particular, the shared task criteria include the fact that any relation between two noun phrases is to be considered. So, the main adjustment we made was to prevent from generating

Argument_1	Relation	Argument_2
ele	está traindo	Meg
he	*is betraying*	*Meg*
ele	está traindo Meg	com a visitante
he	*is betraying Meg*	*with the visitor*

Table 2. Two triples extracted from the second argument structure in Table 1.

arguments headed by verbs. To do that, the subordinated verb was placed within the verb relation after the main verb by giving rise to a composite verbal phrase. This way, the nominal argument of the subordinated verb was considered to be the argument of the verbal phrase and, thus, it was converted into the second argument of the corresponding triple. For instance, if we apply the official LinguaKit module on a sentence like "Mohsen Makhmalbaf decide realizar uma chamada" (*Mohsen Makhmalbaf decides to make a call*), it results in the following triple:

Argument_1	Relation	Argument_2
Mohsen Makhmalbaf	decide	realizar uma chamada
Mohsen Makhmalbaf	*decides*	*to make a call*

However, in the version adapted to the criteria of NerReIberLEF2019, the output is a slightly different triple:

Argument_1	Relation	Argument_2
Mohsen Makhmalbaf	decide realizar	uma chamada
Mohsen Makhmalbaf	*decides to make*	*a call*

4 Evaluation

4.1 NER Task

Tables 3, 4, and 5 show the results of the two systems submitted to the NER task (Task 1). As the organizers presented the results of each dataset individually, each Table depicts the results of each of the three datasets. As can be seen, the neural system based on Cross-View Training clearly outperforms the rule-based module of LinguaKit. If we compare these results with the rest of systems involved in the shared task (6 submissions), we must emphasize that our neural-based method was the one that obtained the best results with the first two datasets (legal and clinical), and the third best in the last one (general). Next, we analyze the results dataset to dataset.

Police Dataset is the result of manually annotating texts from Brazil's Federal Police for just the Person category. It consists of 30 texts containing 1,388 sentences with

37,706 tokens. In total, the annotators extracted 916 named entities of the Person category. As shown in Table 3, There is a big difference between the two strategies, CVT and LinguaKit, which also happens with regard to the other participants: there are three systems with low F1 scores between 30 and 40% (like LinguaKit), and three with very high scores between 88 and 90% (like CVT), being the highest F1 value obtained by CVT. It will be necessary to analyze the test dataset to explain these so important differences among systems.

System	Class	Prec	Rec	F1
CVT	PER	92.20%	89.73%	90.95%
Linguakit	PER	40.83%	25.92%	31.71%

Table 3. Police Dataset

Clinical Dataset consists of clinical notes which were annotated for the Person category. Clinical notes present particular challenges such as names with codes inside; for example, the annotators must understand "AnaR1" or "####Paulo" refer to Person entities. The corpus size is small: it consists of 50 notes with 50 sentences and 9,523 tokens. The total number of Person entities is 77. The performance of the neural system CVT is clearly better than LinguaKit, even though the F1 value remains discrete. CVT achieves the best score among all participants, which, therefore, gives also discrete values (ranging between 10 and 41%).

System	Class	Prec	Rec	F1
CVT	PER	36.36%	49.12%	41.79%
Linguakit	PER	22.08%	6.88%	10.49%

Table 4. Clinical Dataset

The evaluation with the General Dataset takes into account 5 categories: Person, Place, Organization, Time and Value. It was built from two different annotated corpora: SIGARRA [16] and Second HAREM (Relation Version) [6]. The total dataset contains 5,054 sentences with 179,892 tokens. The named entities were classified following this distribution: $2,159$ Person (PER), $1,593$ Place (PLC), $2,320$ Organization (ORG), $3,826$ Time itens (TME), and 106 Values of quantities (VAL). Table 5 shows the results of our two systems, CVT and LinguaKit, including micro and macro-average. In this corpus, the distance between the two systems is not so important. Even though the neural-based approach outperforms the rule-based in both micro and macro-average, the latter performs better on TME entities, which are, in fact, the most frequent class of named entities in this dataset.

System:	*CVT*		
Class	**Prec**	**Rec**	**F1**
PER	75.64%	58.83%	66.18%
ORG	54.24%	28.04%	39.27%
PLC	55.93%	42.47%	48.28%
TME	58.68%	58.57%	58.62%
VAL	96.23%	96.23%	96.23%
Micro-AV	61.27%	46.07%	52.60%
Macro-AV	68.14%	56.82%	61.71%

System:	*Linguakit*		
PER	56.79%	27.59.83%	37.14%
ORG	38.40%	19.99%	26.29%
PLC	39.61%	23.09%	29.17%
TME	44.59%	89.79%	59.59%
VAL	34.91%	42.05%	38.14%
Micro-AV	44.89%	32.97%	38.01%
Macro-AV	42,86%	40,50%	41.64%

Table 5. General Dataset (SIGARRA + SecHAREM)

4.2 Open Relation Extraction Task

The pure OIE task correspond, in fact, with Test 2 of Task 3 at NerReIberLEF2019. The objective is generate verbal relations with two nominal arguments, that is, triples referring to basic propositions.

In the evaluation, two scores metrics were considered: a completely correct relations score and a partially correct relations score. Completely correct relations (exact matching) stands when all terms that make up the relation descriptors in the key are equal to the relations descriptors of the system's output. Partially correct relations (partial matching) stands when at least one of the terms in the relation descriptors of the systems output corresponds to a term in the relation descriptors of the key.

Test 2 consists of a set of golden triples extracted from 25 sentences. A description of the constraints for extractions of relations and arguments is reported in [12].

Table 6 shows the results obtained by the 6 systems involved in this task. Most of the them have been described in previous work, for instance, DEPENDENTIE [22], INFERPOROIE [3], and ICEIS [21]. The OIE of LinguaKit is a more recent version of DepOIE [10] and ArgOIE [9]. It clearly outperforms the other systems in terms of Precision, both in exact and partial matching. However, in F1, LinguaKit is the first system only in exact matching. In partial matching, DPTOIE system performs better than LinguaKit as its Recall is higher. It is worth noting that all systems have very low recall, which shows the difficulty of the task.

System	Prec_Exact	Rec_Exact	F1_Exact	Prec_Part	Rec_Part	F1_Part
Linguakit	37.25%	4.29%	**7.70%**	55.34%	6.26%	11.25%
DPTOIE	10.45%	3.61%	5.37%	36.44%	13.98%	**20.20%**
ICEIS	9.23%	1.35%	2.36%	34.73%	5.30%	9.20%
INFERPOROIE	7.93%	1.13%	1.98%	32.10%	4.59%	8.03%
PRAGMATICOIE	7.35%	1.31%	1.96%	31.80%	4.85%	8.42%
DEPENDENTIE	5.00%	0.45%	0.82%	34.93%	3.13%	5.75%

Table 6. Evaluation of all systems in Task3, Test2 at NerReIberLEF2019 (Evaluation 4 - considering the relations in all datasets).

5 Conclusions

In this article, we compared a neural-based tool for NER with a rule-base system using the datasets of NerReIberLEF2019 Task 1. Moreover, we also compared a rule-based strategy with the rest of systems participating to the OIE shared task in NerReIberLEF2019 (Task 3 - Test 2).

For the NER task, the neural-based system, trained on a corpus of about $900k$ tokens and provided with pre-trained word embeddings, clearly outperformed the rule-based strategy in all datasets: legal, medical, and general. Concerning the OIE task, we could not make the same kind of comparison as there is no training corpus for this specific task, which has not been modelled so far using neural classifiers due to its excessive complexity. In this case, the precision of our rule-based tool clearly outperformed that of the other systems in the competition. However, it will be necessary to analyze the test dataset in order to know how to improve the recall, which remains still very low.

In future work, we will explore the possibility of developing a hybrid strategy mixing rules and neural networks, such as the recent study on sentiment analyzer described in [19], where the proposed technique mixes a deep learning approach (namely, Convolutional Neural Networks) and a rule-based method to improve aspect level sentiment analysis.

6 Acknowledgments

This work has received financial support from DOMINO project (PGC2018-102041-B-I00, MCIU/AEI/FEDER, UE), eRisk project (RTI2018-093336-B-C21), the Consellería de Cultura, Educación e Ordenación Universitaria (accreditation 2016-2019, ED431G/08), the Spanish Ministry of Economy, Industry and Competitiveness under its Competitive Juan de la Cierva Postdoctoral Research Programme (FJCI-2016-28032 and IJCI-2016-29598) and the European Regional Development Fund (ERDF). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

1. 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)
2. Clark, K., Luong, M.T., Manning, C.D., Le, Q.V.: Semi-supervised sequence modeling with cross-view training. In: EMNLP (2018)
3. Claro, D., Sena, C.: Inferportoie: A portuguese open information extraction system with inferences. Natural Language Engineering **25** (12 2018). https://doi.org/10.1017/S135132491800044X
4. 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. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). CEUR Workshop Proceedings, CEUR-WS.org (2019)
5. Cui, L., Wei, F., Zhou, M.: Neural open information extraction. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 407–413. Association for Computational Linguistics, Melbourne, Australia (Jul 2018), https://www.aclweb.org/anthology/P18-2065
6. Freitas, C., Mota, C., Santos, D., Oliveira, H.G., Carvalho, P.: Second HAREM: Advancing the state of the art of named entity recognition in Portuguese. In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10). European Languages Resources Association (ELRA), Valletta, Malta (May 2010)
7. Gamallo, P., Garcia, M., Piñeiro, C., Martinez-Castaño, R., Pichel, J.C.: LinguaKit: A Big Data-Based Multilingual Tool for Linguistic Analysis and Information Extraction. In: 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS). pp. 239–244 (2018). https://doi.org/10.1109/SNAMS.2018.8554689
8. Gamallo, P., Garcia, M.: A resource-based method for named entity extraction and classification. In: Antunes, L., Pinto, H.S. (eds.) Progress in Artificial Intelligence, 15th Portuguese Conference on Artificial Intelligence, EPIA 2011, Lisbon, Portugal, October 10-13, 2011. Proceedings. Lecture Notes in Computer Science, vol. 7026. Springer (2011). https://doi.org/http://dx.doi.org/10.1007/978-3-642-24769-9
9. Gamallo, P., García, M.: Multilingual Open Information Extraction. In: Progress in Artificial Intelligence - 17th Portuguese Conference on Artificial Intelligence, EPIA 2015, Coimbra, Portugal, September 8-11. pp. 711–722. Springer (2015)
10. Gamallo, P., Garcia, M., Fernández-Lanza, S.: Dependency-based open information extraction. In: ROBUS-UNSUP 2012: Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012). pp. 10–18. Avignon, France (2012)
11. Garcia, M.: Extracção de relações semânticas. Recursos, ferramentas e estratégias. Ph.D. thesis, University of Santiago de Compostela (2014)
12. 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: Computational Processing of the Portuguese Language - 13th International Conference, PROPOR 2018, Canela, Brazil, September 24-26, 2018, Proceedings. pp. 66–76 (2018). https://doi.org/10.1007/978-3-319-99722-3_7, https://doi.org/10.1007/978-3-319-99722-3_7
13. Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Silva, J., Aluísio, S.: Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In: Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology. pp. 122–131. Sociedade Brasileira de Computação, Uberlândia, Brazil (Oct 2017), https://www.aclweb.org/anthology/W17-6615

14. Mausam, M.: Open information extraction systems and downstream applications. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. pp. 4074–4077. IJCAI'16, AAAI Press (2016), http://dl.acm.org/citation.cfm?id=3061053.3061220
15. Padró, L.: Analizadores multilingües en freeling. Linguamática **3**(2), 13–20 (2011)
16. Pires, A.: Named entity extraction from Portuguese web text. Master's thesis, Universidade do Porto (2017)
17. Pirinen, T.A.: Neural and rule-based Finnish NLP models—expectations, experiments and experiences. In: Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages. pp. 104–114. Association for Computational Linguistics, Tartu, Estonia (Jan 2019), https://www.aclweb.org/anthology/W19-0309
18. Ravfogel, S., Goldberg, Y., Linzen, T.: Studying the inductive biases of RNNs with synthetic variations of natural languages. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 3532–3542. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019), https://www.aclweb.org/anthology/N19-1356
19. Ray, P., Chakrabarti, A.: A mixed approach of deep learning method and rule-based method to improve aspect level sentiment analysis. Applied Computing and Informatics (2019). https://doi.org/https://doi.org/10.1016/j.aci.2019.02.002, http://www.sciencedirect.com/science/article/pii/S2210832718303156
20. Santos, D., Seco, N., Cardoso, N., Vilel, R.: HAREM: An advanced NER evaluation contest for portuguese. In: 5th International Conference on Language Resources and Evaluation - LREC-2006. pp. 1986–1981. Genova, Italy (2006)
21. 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
22. Souza De Oliveira, L., Glauber, R., Claro, D.: Dependentie: An open information extraction system on portuguese by a dependence analysis. In: ENIAC-2017 XIV Encontro Nacional de Inteligência Artificial e Computacional (10 2017)

Notes

[1]http://www.inf.pucrs.br/linatural/wordpress/iberlef-2019/

[2]LinguaKit is freely available at: https://github.com/citiususc/Linguakit

[3]https://github.com/tensorflow/models/tree/master/research/cvt_text

[4]https://github.com/arop/ner-re-pt/tree/master/datasets/harem/nltk

[5]These values were obtained with the *tagging_scorer* script provided by CVT (see footnote 3).

[6]We achieve $F1 > 95\%$ using different corpora combinations (with more harmonized annotations) in preliminary experiments. However, we decided to submit this model as the training corpora were actually more balanced with regard to the different sources.