Explainable OpenIE Classifier with Morpho-syntactic Rules

Bruno Cabral and **Marlo Souza** and **Daniela Barreiro Claro**[1]

Abstract. Open information extraction (OpenIE) is a task of extracting structured information from unstructured texts independently of the domain. Recent advances have applied Deep Learning for Natural Language tasks improving the state-of-the-art, even though those methods usually require a large and high-quality corpus. The construction of an OpenIE dataset is a tedious and error-prone task, and one technique employed concerns the extractions from rule-based techniques and manual validation of those extraction triples. As low-resource languages usually lack available datasets for the application of high-performance Deep Learning techniques, our intuition is that a low-resource model based-on multilingual information can learn generalizations across languages and benefits from cross-lingual data. Moreover, we would like to interpret the set of generalized information gathered from multilingual learning to increase the Open IE classification task. In this paper, we introduce TabOIEC, a multilingual classifier based on generic morpho-syntactic features. Our classifier carries a glass-box method which can provide interpretation about some of the classifier decisions. We evaluate our approach through a small corpus of Open IE extractions for the English, Spanish, and Portuguese languages. Our results consider that for all languages our approach improves F1 measures, particularly for monolinguality. Experiments on Zero-shot learning provide evidence that our TabOIEC generalizes the classifier on other languages than that trained, although there is a shy transfer learning among them. Experiments on multilinguality do reduce the cost of training, however, in our experiments were difficult to provide appropriate generalizations.

1 Introduction

Every day we have a greater volume of data, and we need tools that help us to extract relevant information from this growing set. Much of this information is composed of texts created in an unstructured way, such as books, news and conversations. Open Information Extraction (OpenIE), as introduced by Banko et al. [2], is a useful tool in this context, because it is capable of extracting knowledge from large collections of textual documents independently of the domain [5]. By extracting information, we mean that these systems generate structured representation of information in the original documents, usually in the form of relational tuples, such as (arg_1, rel, arg_2), where arg_1 and arg_2 are the arguments of the relation, usually described by noun phrases, and rel a relation descriptor that describes the semantic relation between arg_1 and arg_2 [24]. For example, consider the sentence:

"I could only see the ball came in the goal, because it fell next to where I was."

An Open IE system can generate valid extractions, such as:

(the ball, came in, the goal).

Or the following invalid tuple:

(the ball, came in, it)

Since 2007, with the TEXTRUNNER [2], multiple OpenIE systems have been designed and proposed for the many different languages. These systems have had different types of approaches, from rule-based systems to deep neural networks. A continued number of innovations in Deep Learning have been pushing multiple Natural Language Processing (NLP) tasks to achieve a better performance, thanks in part to large-scale annotated datasets. Recently, OpenIE neural networks have been used for supervised learning in Open IE [57, 16, 58, 61], achieving state-of-the-art results for English.

As noted by Glauber and Claro [28], major advances in Open IE, have mainly focused on the English language. Although the focus on the English language may be due its origin and the usage language over the world, it has been recognized by the scientific community that the focus on the English language with its particular characteristics may introduce some bias to the area [7, 6].

While a constant number of innovations in Natural Language Processing (NLP) research enable models to achieve impressive performance, such developments are not available to all languages since only a handful of them have the labelled data necessary for training deep neural nets [12]. In fact, for Open IE, the availability of such datasets [56, 37] has led to the development of methods [57, 16, 58, 61] achieving the Open IE state-of-the-art results.

We believe one reason for this focus on the English language is the lack of available resources for the area in other languages. Unfortunately, manual creation of annotated corpora for Open IE is a difficult task, as noted by [30, 37], due to vague notion of semantic relation advocated in the area [60, 37] and the multiplicity of possible interpretations for the same sentence.

As Brants and Plaehn [8] observe, the use of automatic tools for assisting annotation of a corpus facilitates rapid semi-automatic corpus annotation in an interactive process. As noisy candidate extractions can be easily generated from a corpus based on simple morphosyntactic patterns [3, 21, 59] and parsing technology [26, 19, 29], an important bottleneck in an Open IE annotation process is deciding whether a given candidate extraction corresponds to a valid relation on the corpus. Hence, in this work we aim to construct a tool for assessing the quality/correctness of Open IE extractions, aiming to assist on semi-automatic construction of corpora for the area for different languages.

[1] Federal University of Bahia, FORMAS Research Group, Computer Science Department, Salvador - Bahia - Brazil, email: dclaro@ufba.br

While similar classifiers have been proposed before as post-processing tools in Second generation Open IE systems, e.g [21, 48, 18, 22], these classifiers are usually constructed in language-dependent manner, for which the generalization to other languages has not been investigated, and/or generate models which are not easily interpretable [4, 10].

An important characteristic of our method relies on the fact that we explore the use of machine learning methods which generate interpretable models. Since in Open IE manual annotation, as observed by [30], agreement among annotators can be very low and annotations have to be discussed. Our focus on interpretable models allow for the generation of explanations for the predictions, which can be exploited in this process, as well as to generate underlying non-documented rules/hypothesis in the annotation process - as explored by [8].

Interpretable or explainable models are decision models for which predictions can be traced back to explicit relationships in the data. Recently, the application of neural methods in natural language processing has led to a profound advances in the area. These advances, however, are hard to understand and evaluate, due to opaqueness of the new models developed in the area. Indeed, several recent researches [33, 40, 42] show that the predictions made by the systems in the area may be based on spurious or unclear reasons, thus subject to adversarial attacks, and that their reported performance may be explained by unrelated artifacts and regularities on the used datasets, not on the inherent quality of the model. In fact, adversarial examples seem to be an unavoidable characteristic of such methods, a rising from their foundation geometric principles [32].

In this work, we propose a classification method to asses the quality of Open IE system extractions aiming to assist on the semi-automatic annotation of data. This method is based on the use of tabular learning methods, i.e. methods specific to deal with tabular data and which generate interpretable models. By the use of generic features and multilingual pre-processing tools, our method can be directly trained on data from different languages without the need of engineering any pre-processing tools. To conduct our experiments, we investigate the application of several different explainable learning architectures on data from three different languages. This tool enables the classification of generated extractions of any previously developed OpenIE tool, independently of the language or type of implementation. In Portuguese, this model can trade recall performance for up to 65% improvement in F1 score.

This article is organized as follows: Section 2 presents some related work. Section 3 describes our approach and our methodology. Section 4 shows our experiments, results and discussions. Finally, Section 5 concludes our paper.

2 Related Work

Recently, new machine learning-based approaches for Open IE [57, 16, 58, 61] have been proposed, leading to a new generation of Open IE systems. While these systems represent the state-of-the-art in the area, their focus on the English language and need of annotated data make it hard to generalize their results to other languages. For the Portuguese language, new data-based methods have been proposed as a cross-lingual approach due to the lack of resources for this task [10]. Early methods use linguistically-inspired patterns for extraction, such as ArgOE [25], or adaptation of methods for the English language, such as SGS[18], SGC_2017 [55] and RePort [48]. Recently, new pattern-based methods have risen as the new state-of-the-art for the language [14] such as InferPORToie [54], PragmaticOIE [53] and DptOIE [29].

Classification-based tools to asses quality of extractions has been employed by different systems [48, 22], mainly following the success of the ReVerb [21]. These works are based on the manual construction of language-specific features to assess the quality of extractions, based on morphosyntactic patterns and grammatical rules for each language, which seldom generalize to other non-typologically related languages.

Language-independent classification methods have been proposed before [4, 10, 13]. The work of Barbosa and Claro [4] is the closest to ours, proposing a set of feature which the authors claim to be language-independent for the task of open IE extraction quality assessment. The authors' empirical evaluation of their proposed feature set on multilingual data and their proposed method is based on Support Vector Machine classifiers which are not easily interpreted. The work of Cabral et al. [11], on the other hand, proposes the use of multilingual language models, as M-BERT [20] and XLM [36] to perform quality assessment and classification of Open IE extractions. The authors evaluate their method on multilingual data, but due to the use of opaque language models and classification techniques, their predictions are not explainable and, thus, cannot be easily integrated within a semi-automatic annotation process.

3 TabOIEC

In this work, our goal is to have an explainable OpenIE triple classifier capable of supporting multiple languages, by changing the training dataset. In this Section, we briefly revisit the formulation of OpenIE, and the components used in our model.

3.1 Problem Definition

Let $X = \langle x_1, x_2, \cdots, x_n \rangle$ be a sentence composed of tokens x_i, an Open IE extractor is a function that maps X into a set $Y = \langle y_1, y_2, \cdots, y_j \rangle$ as a set of tuples $y\_i = \langle rel_i, arg1_i, arg2_i, \cdots, argn_i \rangle$, which describe the information expressed in sentence X. In this work, we consider that the tuples are always in the format of $y = (arg_1, rel, arg_2)$, where $arg1$ and $arg2$ are noun phrases, not necessarily formed from tokens present in X, and rel is a descriptor of a relation holding between arg_1 and arg_2. We do not consider extractions formed by n-nary extractions.

Given a sentence X as above, we are interested in determining for every extraction $y_i \in Y$ whether y_i is a valid extraction from X , the factors that the classifier made their decision well as the confidence score for such classification . An OpenIE extraction classifier can be expressed as a decision function that for every single sentence X and extractions Y, returns a pair $(Z, P) \in \{0,1\}^{|Y|} \times [0,1]^{|Y|}$, where $Z = \langle z_1, z_2, \cdots, z_n \rangle$ is a binary vector s.t. $z_i = 1$ denotes that y_i is a valid extraction, and $P = \langle p_1, p_2, \cdots, p_n \rangle$ is a probability vector, s.t. p_i denotes that extraction y_i has an associated probability p_i of being classified as z_i, given the input sentence X.

3.2 Fine-tuned Multilingual Contextual Embedding

In this work, our plan is to create an explainable language-agnostic classifier, and for that, we use a Multilingual Contextual Embeddings. Multilingual means that those models represent words of multiples languages into a shared semantic representation space. As such, these models are able to represent semantic similarities between words in different languages. Contextual Embeddings means that the meaning of the word is represented taking its context into consideration.

One such Multilingual Contextual Embedding is M-BERT [20], a 12-layer transformer trained on 104 languages from a Wikipedia with a shared word piece vocabulary. According to tests conducted by Pires et al. [49], M-BERT is able to transfer knowledge between languages with no lexical overlap, an indication that it captures multilingual representations. It is capable of generating across languages because common word pieces such as numbers are mapped to a shared space, spreading the effect to other word pieces, until similar words in different languages are close in the vector space [49].

The problem with using M-BERT directly is that it does not fulfill our requirement of an explainable classifier, due to its ability to represent tokens in a multidimensional vector of values. One alternative is the use of UDify model, a multilingual multi-task model capable of predicting universal part-of-speech, morphological features, lemmas, and dependency trees across 75 languages [34]. This model uses M-BERT and fine-tunes it on the Universal Dependencies (UD) dataset, as it provides syntactic annotations consistent across a large collection of languages [43]. UDify is able to represent of syntactic knowledge transfer across multiple languages including lemmas (LEMMAS), treebank-specific part-of-speech tags (XPOS), universal part-of-speech tags (UPOS), morphological features (UFEATS), and dependency edges and labels (DEPS) for each sentence [34].

Finally, for training our classifier, we use the final output of UDify to extract features of sentences' inputs and extractions. Those features are than tabulated in a specific format so that they can be used in classification algorithms that create rules on a set of predefined attributes. One example of such algorithm is a Decision Tree [9]. This type of classifier has the characteristic of creating high-interpretable models.

3.3 Architecture

Our general architecture and classifier are illustrated in Figure 1. It consists of three main steps. Firstly, we pre-process the input, then we generate the feature set, and finally we feed the computed features to a Classifier. Each step is detailed in the subsections below.

3.3.1 Pre-processing

In the pre-processing step our objective is to convert the textual output of the OpenIE Extractors to a structured format to be processed in the later steps. Relational triple data is textual and its contents cannot be used directly in the classification algorithms implemented in TabOIEC. This step is illustrated on Figure 2.

It first receives a sentence X and a list of extractions Y, each in the form $y_i = \langle arg_1, rel, arg_2 \rangle$. The first step is to split the sentence into tokens. For the tokenization step we utilize the Spacy [31] *xx_ent_wiki_sm* tokenizer, a Multi-lingual CNN trained on Nothman et al.[51] Wikipedia corpus. Afterwards we perform the contraction expansion. For example, the English contraction *I'm* could be tokenized as the two words *I am*, and *we've* could become *we have*. This is needed because in an extraction, different parts of a token could appear on different parts of an extraction.

3.3.2 Feature Extraction

As explainability is a requirement in our classifier, we chose to use classifiers that work on a fixed set of features. For that, we need to convert the Sentence and the extractions to a set of features. This process consists of running the feature function and saving the value obtained to a tabular structure. The process is depicted on Figure 1.

The process is the following: feed the sentence X and the list of extractions Y to the multilingual words embedding model (in our case, the UDify model) to compute the set of features of each token in the sentence. Afterwards, the indexing step goal is performed to identify the start and end positions of each relation inside the triple arguments through the rest of the sentence. The sentence X and the list of extractions Y are inputted to the Algorithm 1.

```
Input: Original sentence $S$ , $arg_1$, $rel$, $arg_2$
Output: $Feat\_arg_1$, $Feat_rel$, $Feat/arg_2$
$Feat\_sen \leftarrow GenerateUdifyFeatures(S)$
for part in [ $arg_1$, $rel$, $arg_2$] do
    // Check if the string is a substring
       of the original sentence
    if substring(part, S) then
        $Feat\_part \leftarrow$
          $GetSubsetFeatures(part, Feat\_sen)$;
        // Extract the features of this part
           from the already generated
           features from the whole sentence
    else
        // The relation is not a substring
           of the original sentence, thus
           generate new features isolated
        $Feat\_part \leftarrow GenerateUdifyFeatures(part)$
    end
end
```

Algorithm 1: Finding Features from a sentence

This algorithm first generates the features using the original sentence and then tries to match the constituent parts of each extracted triple to the original sentence, as shown visually in Figure 1. This is necessary due to the way that contextual embeddings work: a word will have a different set of features, depending on the full sentence, and we want the representation to be the same as the original sentence.

In some cases, the constituents are not a sub sequence of the original sentence, such as in implicit extractions. For example, in the sentence *"The covid-19 virus is very dangerous"*, the triple *(Covid-19, is a, virus)* is valid, however the tokens *"is a"* are not present directly in the original sentence. This makes it impossible to determine the start and end of the relation extraction in the original sentence.

In this case, we generate a new embedding as if the individual part is a sentence. The output of the algorithm is $Feat\_arg_1$, $Feat\_rel$, $Feat\_arg_2$, each is an array of features for each constituent of an extracted triple. Each array of features is then transformed into a fixed-length vector of a manually defined feature as can be seen in Table 1. All features are based on the Universal Dependencies (UD) version 2.3 tagset and each one is described below:

- **1-3 – Relative distance between parts**
  Those features represent the relative distance between each constituent part of the relation. The objective of this feature is to capture improbable distances. Analyzing the rules learned by the classifiers, we identified that this feature represents the location of the constituents, which together with the features below is a good indicator if those relationships happen in the correct order.
- **4-6 – UPOS features** These tags mark the core part-of-speech (POS) categories. There are in this version of UD, 17 Universal

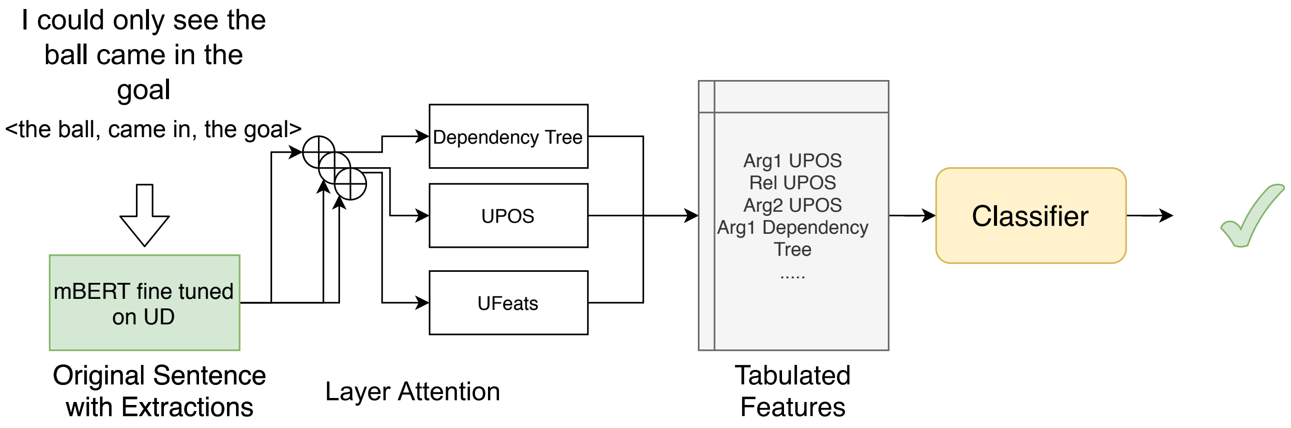


Figure 1. Architecture overview

Figure 2. TabOIE Pre-processing overview.

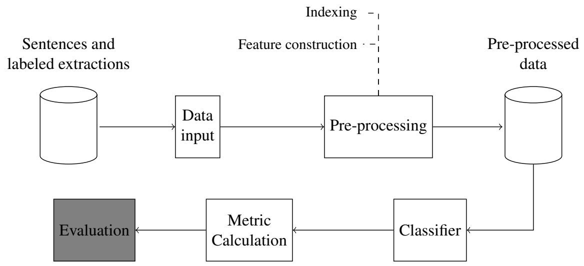


Table 1. Multilingual feature set.

N	**Feature**
1	Relative distance between $arg1$ and rel
2	Relative distance between rel and $arg2$
3	Relative distance between $arg1$ and $arg2$
4	$arg1$ count of each UPOS feature
5	rel count of each UPOS feature
6	$arg2$ count of each UPOS feature
7	$arg1$ count of each UFeat feature
8	rel count of each UFeat feature
9	$arg2$ count of each UFeat feature
10	$arg1$ count of each Dependency Tag with Head pointing to $arg1$
11	$arg1$ count of each Dependency Tag with Head pointing to rel
12	$arg1$ count of each Dependency Tag with Head pointing to $arg2$
13	$arg1$ count of each Dependency Tag with Head pointing to OUT
14	rel count of each Dependency Tag with Head pointing to $arg1$
15	rel count of each Dependency Tag with Head pointing to rel
16	rel count of each Dependency Tag with Head pointing to $arg2$
17	rel count of each Dependency Tag with Head pointing to OUT
18	$arg2$ count of each Dependency Tag with Head pointing to $arg1$
19	$arg2$ count of each Dependency Tag with Head pointing to rel
20	$arg2$ count of each Dependency Tag with Head pointing to $arg2$
21	$arg2$ count of each Dependency Tag with Head pointing to OUT

categories that generalize well across language boundaries. The objective of this feature is to identify valid or invalid relationships between different POS in a sentence. For example, the presence of many verbs in the relation increases the probability that the triple is invalid. Because our classifier algorithm requires a fixed set of features, we create a total of 51 features based on those rules. For each relation we have 17 possible features, one for every single UPOS category.

- **7-9 – UFeat features**
  In the Universal Dependencies (UD), those features distinguish additional lexical and grammatical properties of words, not covered by the POS tags. In UD version 2.3, 50 different features are available, such as *animacy*, *noun type*, *evidentiality* and *type of named entity*. This feature could help to identify for example that a part of a relation has a Named Entity, and this could be an indicator of a valid extraction. A list of features is created composed of all combinations between the existing Ufeat and each relation totaling 150 (50*3) possible features;
- **10-21 – Dependency tree - Tags and Head location**
  This set of features is the count of each 37 universal syntactic relations for each relation and where the head of the relation is located (inside one relation, or OUT if the head is located in a token not located in any relation). For example, the possible categories are *nsubj* (nominal subject) and *advmod* (adverbial modifier). It is created 444 possible features (37 * 12 combinations). This rule is inspired by the work of Oliveira et al [29]. Where they identify a set of hand-crafted rules for Portuguese to identify valid extractions based on the Dependency Tree. For example, they identify a rule that a valid extraction might be composed of a subject (arg1), a verbal phrase (rel) (SV) and one or more arguments (arg2). Where the arg1 have in the dependency tree a *nsubj*.

3.3.3 Classification

In this work, we compared the performance of different interpretable models in the classification task for predicting the quality of Open IE extractions. We compare the performances of the following methods: CatBoost [50], a gradient boosting method for decision trees; SKLearn, the SciKit Learn Learn [47] implementation of Histogram-based Gradient Boosting Classification Tree [41]; Explainable Boosting Machine[44], an Interpretable Gradient Boosting

Classifier; SKOPE-Rules, which uses predictive rule generation over an ensemble of decision trees [23]; and TabNet [1], a tabular-data based explainable Neural Network.

4 Experiments

In this section, we describe the empirical validation of our proposed method to classify Open IE extractions based on language-independent features and interpretable models.

4.1 Dataset

For comparability, in our experiments we employ the same data used by Cabral et al. [11] for their multilingual Open IE classifier. This dataset is composed of relations extracted by five different Open IE systems, namely ClausIE, OLLIE, ReVerb, WOE, and TextRunner, from texts in Portuguese, English, and Spanish languages, and labeled as valid or invalid (z_i) by human judges. A valid extraction ($z_i = 1$) corresponds to a coherent triple with the sentence. These linguistic resources were obtained through the studies of [19] and [25]. The statistics of the dataset are summarized in Table 2.

Table 2. Dataset statistics

	# Sentences	**# Extractions**
Portuguese	200	1856
English	500	7093
Spanish	159	375

4.2 Experimental Setup

Our work uses the AllenNLP [27] library built with the PyTorch [45] framework. The fine-tuned model that extract the UD features is the UDify [35] with the fine-tuned BERT weights available[2].

We implemented our Open IE classifier architecture directly on top of the AllenNLP. We also test with the following classifiers:

- **Scikit-learn (Sklearn)**[46] version 0.23 - A Gradient Boosting Classifier
- **Catboost**[50] version 0.22 - A Gradient Boosting Classifier
- **Skope** [3] - A decision rule Classifier
- **Explainable Boosting Machine (EBM)** - implementation in Interpret[44] version 0.1.22 - A Interpretable Gradient Boosting Classifier
- **TabNet** - Attentive Interpretable Tabular Learning[1] Classifier, version 1.0.6[4]

Among these classifiers, the Skope and Explainable Boosting Machine are considered glass-box classifiers, where they output high interpretable rules. In addition, with the other classifiers, there are blackbox explainers such as SHAP Tree Explainer [38] that are able to explain their outputs.

For all classifiers we utilize the default hyper-parameters, with no additional tuning, only the number of epochs was changed to 300. For each single-language test, we split our corpus into training and testing using a 5-fold cross-validation strategy. However, for the zero-shot test, we train the classifier with the whole corpus, excluding the language to be tested (e.g., the zero-shot test for Portuguese is trained using the whole English and Spanish corpus and evaluated on the whole Portuguese corpus).

Each split on our k-fold strategy is carried on a sentence level. As a consequence, each split has the same number of sentences, but it may differ on the number of extractions. Our results are a weighted average on the number of extracted facts for each test folds using the Precision (P), Recall (R), F1-measure and the Matthews correlation coefficient (MCC) [39]. MCC is employed in machine learning as a quality measure of the classifier. To compute Precision-Recall curves, we select the n extractions with the highest confidence score and compute the classifier's precision. The possible values of confidence considered were: [0.6, 0.7, 0.8, 0.85, 0.9, 0.93, 0.95, 0.98, 0.99, 0.995, 0.999]. The code of our experiments is available at `https://github.com/FORMAS/HybridOIEClassifier`

4.3 Results

We consider three evaluation performances. For monolingual learning, we provide on Table 3 the precision (Prec.), recall (Recall), F1-measure (F1), Accuracy (Acc) and the Matthews metrics [39] (confidence coefficient among the extractions) for each language: Portuguese, Spanish and English. It is important to observe that the Recall measure for an OpenIE task corresponds to the total number of triple extraction performed by all systems. We consider this as a 100% recall. In the scientific community, some researchers are denominating this restriction as a *yield measure* [17].

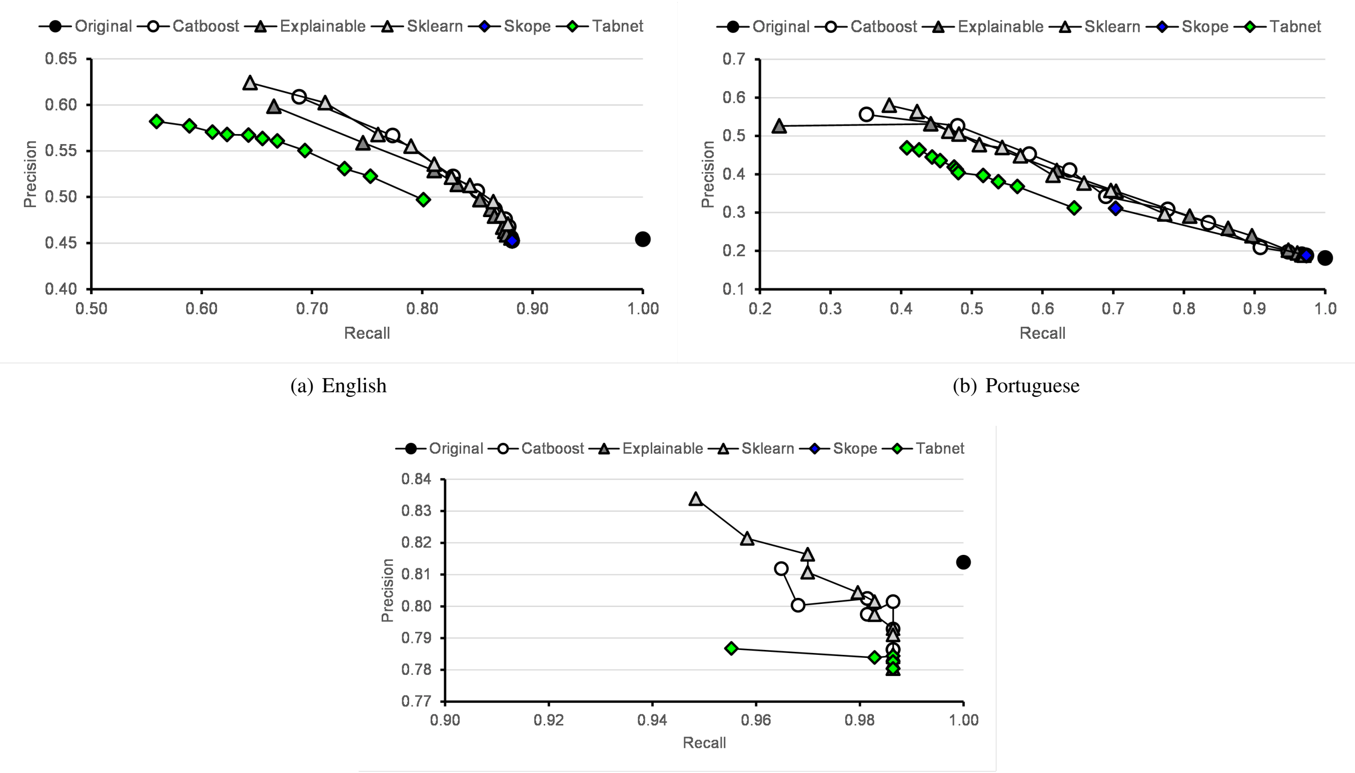
For the Portuguese language, the EBM model achieves a recall of 80.8% in comparison with the 100% from the Original model. However, the best precision performance was achieved by the Sklearn model with over 58%. Taking the Spanish language, we observe that the best results were obtained from Sklearn and no impressive result gathered from EBM model. The F1 measure surpassed all the Portuguese models. For English models, the best F1 results were obtained by Catboost model. All confidence coeficients were over 87% of agreement.

Table 3. Metrics scores for languages classifiers

	Prec.	Recall	F1	Acc.	MCC
Portuguese					
Original	0.181	1.000	0.307	0.181	0.000
Best Precision (C:0.6 - Sklearn)	**0.580**	0.383	0.459	**0.836**	0.976
Best F1 (C:0.8 - Catboost)	0.452	0.581	**0.508**	0.796	0.986
Best Acc. / Recall > 0.8 (C:0.9 - EBM)	0.290	**0.808**	0.427	0.605	**0.990**
Spanish					
Original	0.730	1.000	0.844	0.730	0.000
Best Precision, F1 and Accuracy(C:0.6-Sklearn)	**0.833**	**0.948**	**0.886**	**0.824**	**0.966**
English					
Original	0.454	1.000	0.624	0.454	0.000
Best Precision (C:0.6 - Sklearn)	**0.624**	0.643	0.633	**0.661**	**0.897**
Best F1 (C:0.7 - Catboost)	0.567	0.773	**0.654**	0.628	0.888
Best Acc. / Recall > 0.8 (C:0.9 - Sklearn)	0.536	**0.811**	0.645	0.594	0.875

To evaluate whether our models were able to explore cross-lingual information, i.e. to apply information learned from a set of different languages to a new language, we also performed zero-shot and one-shot classification.

[2] https://github.com/hyperparticle/udify
[3] https://github.com/scikit-learn-contrib/skope-rules
[4] https://github.com/dreamquark-ai/tabnet

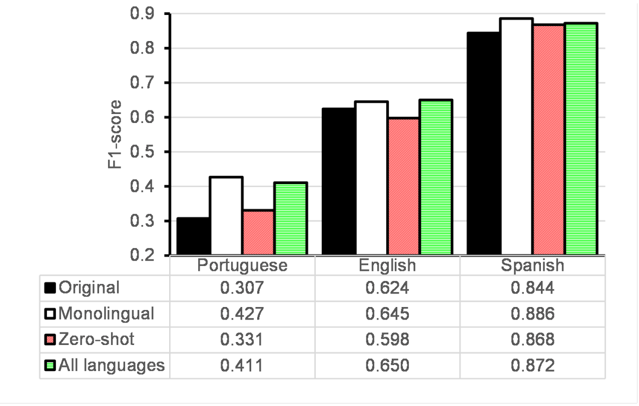


(a) English

(b) Portuguese

(c) Spanish

Figure 3. Language-specific performance



	Portuguese	English	Spanish
■Original	0.307	0.624	0.844
□Monolingual	0.427	0.645	0.886
🟥Zero-shot	0.331	0.598	0.868
🟩All languages	0.411	0.650	0.872

Figure 4. Comparison of the performances for Monolingual, Zero-shot and One-shot

The zero-shot classification is a task where the classifier is evaluated on a language not seen during the training. For Portuguese, we observe an average decrease in F1 performance between 3% (for SKOPE) and 16% (Sklearn and TabNet) on all models, with maximum decrease of 19% on CatBoost at confidence 0.8. Similar behaviour has been observed for zero-shot classification for Spanish - between 2% SKOPE and 20% Sklearn, maximun decrease of 52% with Catboost at 0.6 - and English - betwenn 1% for SKOPE and 15% for SkLearn, with maximum decrease of 24% for SkLearn ate 0.7.

The one-shot classification is a task where the classifier is trained with the data from other languages and part of data on the target language, and tested on the remaining (unseen) data for the target language. For Portuguese, we observe an average decrease in F1 performance between 3% (for SKOPE) and 10% (Sklearn) on all models, with maximum decrease of 15% on SKLearn at confidence 0.99. Similar behaviour has been observed for one-shot classification for Spanish - between 0% SKOPE and 10% TabNet, maximun decrease of 13% with TabNet at 0.7 - and English - betwenn 0% for SKOPE and 1% for EMB, with maximum decrease of 1% for EMB 0.8.

4.4 Discussion

While in the single language experiments, results of the classifier are more robust, in the sense that the decline in Precision is much more nuanced for almost all representations in the three languages, in the zero-shot experiments, however, this decline is much more pronounced. These results indicate that there may be a discrepancy between the datasets for each language regarding the relations extracted. This discrepancy may arise from the fact that the datasets were created using (i) different Open IE systems for each language (ii) annotated by different teams at different times, and (iii) using texts of different linguistic styles - for English, encyclopedic, journalistic and user-generated (Web pages), for Spanish and Portuguese, encyclopedic texts- and domains - multiple domains for English and Spanish and domain-specific for Portuguese. It may also be the case that linguistic parameters of each language, such syntactic structure and stylistic choices of each language community, may play an important role on structuring information through language and, as such, on how this information is extracted.

It is also worth noticing that the English dataset is considerably larger than both datasets for Spanish and Portuguese, thus in the zero-shot learning, it may dominate the training process and can overfit the classifier to the English dataset-specific characteristics. As such, experiments with a higher number of languages to provide the classifier with a more diverse set of examples is recommended.

Considering the multilinguality, we observe that our monolingual model is slightly better than the model trained for three languages, except for the English one. Our results corroborate with the findings of [52] which mention the *curse of multilinguality* from [15] which states that adding mode languages to a model can degrade the performance as the capacity of the model remain the same. For the English language, there is no significant difference from training with monolingual nor multilingual (i.e. three languages) approach.

Observing our results on zero-shot learning, it is important to notice that all three languages achieve a slight learning rate, increasing the original performance indicating a limited but possible exploration of cross-language information. For dissimilar languages such as in the case of training in the extraction from Spanish and Portuguese sentences and testing on extractions from English sentences, the results are less conclusive due probable to their dissimilar linguistic characteristics. Our intuition is that if the models are presented with examples of varied linguistic characteristics, the classifier can be applied to a wide range of low-resource languages - facilitating the development of computational linguistic resources in these languages.

5 Conclusion and Future Work

In this work, we presented the TabOIEC, a language-independent explainable relation extraction binary classifier. The evaluation results demonstrated that a single model could improve the output of multiple state-of-art systems across three languages: Portuguese, English, and Spanish. Our results give evidence that simple and explainable models for extraction quality assessment could be a useful resource for the construction of Open IE datasets systems for different languages.

In the future, we plan to evaluate the use of hand-crafted features by linguist experts. Another point of improvement would test the solution in larger datasets, and utilize some techniques to improve the classifier such as Fine-tuning the classifier on the Open IE tuples.

Once mature, we intend to employ the trained models in an annotation tool, allowing the creation of Open IE and Relation Extraction datasets for different languages. With such a tool, we aim to encourage the development of Relation Extraction techniques and technology for different languages, given the importance of Information Extraction technology for the development of advanced intelligent systems and interfaces.

ACKNOWLEDGEMENTS

We would like to thank CNPQ and CAPES for their financial support.

REFERENCES

[1] Sercan O. Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning, 2019.

[2] Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni, 'Open information extraction for the web', in IJCAI, volume 7, pp. 2670–2676, (2007).

[3] Michele Banko, Oren Etzioni, and Turing Center, 'The tradeoffs between open and traditional relation extraction.', in ACL, volume 8, pp. 28–36, (2008).

[4] George Caique Gouveia Barbosa and Daniela Barreiro Claro, 'Utilizando features linguísticas genéricas para classificação de triplas relacionais em português', in Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology, pp. 132–141, (2017).

[5] David Soares Batista, David Forte, Rui Silva, Bruno Martins, and Mário Silva, 'Extracçao de relaçoes semânticas de textos em português explorando a dbpédia e a wikipédia', Linguamatica, **5**(1), 41–57, (2013).

[6] Emily Bender, 'English isn't generic for language, despite what nlp papers might lead you to believe', in Symposium and Data Science and Statistics, (2019). [Online; accessed 15-may-2020].

[7] Emily M. Bender, 'Linguistically naïve != language independent: Why NLP needs linguistic typology', in Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?, pp. 26–32, Athens, Greece, (March 2009). Association for Computational Linguistics.

[8] Thorsten Brants and Oliver Plaehn, 'Interactive corpus annotation', in Second International Conference on Language Resources and Evaluation LREC-200, (2000).

[9] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen, Classification and regression trees, CRC press, 1984.

[10] Cabral B.S., Glauber R., Souza M., and Claro D.B., 'Crossoie: Cross-lingual classifier for open information extraction', in Computational Processing of the Portuguese Language (PROPOR 2020), ed., Aluísio S. Moniz H. Batista F. Gonçalves T. Quaresma P., Vieira R., volume 12037 of Lecture Notes in Computer Science, 201–213, Springer, Cham, (February 2020).

[11] Bruno Souza Cabral, Rafael Glauber, Marlo Souza, and Daniela Barreiro Claro, 'Crossoie: Cross-lingual classifier for open information extraction', in International Conference on Computational Processing of the Portuguese Language, pp. 368–378. Springer, (2020).

[12] Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie, 'Zero-resource multilingual model transfer: Learning what to share', arXiv preprint arXiv:1810.03552, (2018).

[13] D.B. Claro, M. Souza, C. Castellã Xavier, and L. Oliveira, 'Multilingual open information extraction: Challenges and opportunities', Information, **10**(7), 228, (2019).

[14] Sandra Collovini, Joaquim Santos, Bernardo Consoli, Juliano Terra, Renata Vieira, Paulo Quaresma, Marlo Souza, Daniela Barreiro Claro, and Rafael Glauber, 'Iberlef 2019 portuguese named entity recognition and relation extraction tasks', in Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), volume 2421, pp. 390–410. CEUR-WS.org, (2019).

[15] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale, 2019.

[16] Lei Cui, Furu Wei, and Ming Zhou, 'Neural open information extraction', arXiv preprint arXiv:1805.04270, (2018).

[17] Leandro Souza de Oliveira, Rafael Glauber, and Daniela Barreiro Claro, 'Dependentie: An open information extraction system on portuguese by a dependence analysis', Encontro Nacional de Inteligência Artificial e Computacional, (2017).

[18] Erick Nilsen Pereira de Souza, Daniela Barreiro Claro, and Rafael Glauber, 'A similarity grammatical structures based method for improving open information systems', J. UCS, **24**, 43–69, (2018).

[19] Luciano Del Corro and Rainer Gemulla, 'Clausie: clause-based open information extraction', in Proceedings of the 22nd international conference on World Wide Web, pp. 355–366. ACM, (2013).

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, 'Bert: Pre-training of deep bidirectional transformers for language understanding', arXiv preprint arXiv:1810.04805, (2018).

[21] Anthony Fader, Stephen Soderland, and Oren Etzioni, 'Identifying relations for open information extraction', in Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1535–1545. Association for Computational Linguistics, (2011).

[22] Tobias Falke, Gabriel Stanovsky, Iryna Gurevych, and Ido Dagan, 'Porting an open information extraction system from english to german', in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 892–898, (2016).

[23] Jerome H Friedman, Bogdan E Popescu, et al., 'Predictive learning via rule ensembles', The Annals of Applied Statistics, **2**(3), 916–954, (2008).

[24] Pablo Gamallo, 'An Overview of Open Information Extraction (Invited talk)', in 3rd Symposium on Languages, Applications and Technologies, eds., Maria João Varanda Pereira, José Paulo Leal, and Alberto Simões, volume 38 of OpenAccess Series in Informatics (OASIcs), pp. 13–16, Dagstuhl, Germany, (2014). Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[25] Pablo Gamallo and Marcos Garcia, 'Multilingual open information extraction', in Portuguese Conference on Artificial Intelligence, pp. 711–722. Springer, (2015).

[26] Pablo Gamallo, Marcos Garcia, and Santiago Fernández-Lanza, 'Dependency-based open information extraction', in Proceedings of the joint workshop on unsupervised and semi-supervised learning in NLP, pp. 10–18. Association for Computational Linguistics, (2012).

[27] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer, 'Allennlp: A deep semantic natural language processing platform', (2017).

[28] Rafael Glauber and Daniela Barreiro Claro, 'A systematic mapping study on open information extraction', Expert Systems with Applications, **112**, 372–387, (2018).

[29] Rafael Glauber, Daniela Barreiro Claro, and Leandro Souza de Oliveira, 'Dependency parser on open information extraction for portuguese texts - dptoie and dependentie on iberlef', in Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), volume 2421, pp. 442–448. CEUR-WS.org, (2019).

[30] Rafael Glauber, Leandro Souza de Oliveira, Cleiton Fernando Lima Sena, Daniela Barreiro Claro, and Marlo Souza, '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).

[31] Matthew Honnibal and Ines Montani, 'spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing'. To appear, 2017.

[32] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry, 'Adversarial examples are not bugs, they are features', in Advances in Neural Information Processing Systems, pp. 125–136, (2019).

[33] Robin Jia and Percy Liang, 'Adversarial examples for evaluating reading comprehension systems', in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2021–2031, Copenhagen, Denmark, (September 2017). Association for Computational Linguistics.

[34] Dan Kondratyuk and Milan Straka, '75 languages, 1 model: Parsing universal dependencies universally', in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 2779–2795, Hong Kong, China, (2019). Association for Computational Linguistics.

[35] Daniel Kondratyuk, '75 languages, 1 model: Parsing universal dependencies universally', arXiv preprint arXiv:1904.02099, (2019).

[36] Guillaume Lample and Alexis Conneau, 'Cross-lingual language model pretraining', arXiv preprint arXiv:1901.07291, (2019).

[37] William Léchelle, Fabrizio Gotti, and Philippe Langlais, 'Wire57: A fine-grained benchmark for open information extraction', arXiv preprint arXiv:1809.08962, (2018).

[38] Scott M Lundberg, Gabriel G Erion, and Su-In Lee, 'Consistent individualized feature attribution for tree ensembles', arXiv preprint arXiv:1802.03888, (2018).

[39] Brian W Matthews, 'Comparison of the predicted and observed secondary structure of t4 phage lysozyme', Biochimica et Biophysica Acta (BBA)-Protein Structure, **405**(2), 442–451, (1975).

[40] Tom McCoy, Ellie Pavlick, and Tal Linzen, 'Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference', in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3428–3448, Florence, Italy, (July 2019). Association for Computational Linguistics.

[41] Qi Meng, Guolin Ke, Taifeng Wang, Wei Chen, Qiwei Ye, Zhi-Ming Ma, and Tie-Yan Liu, 'A communication-efficient parallel algorithm for decision tree', in Advances in Neural Information Processing Systems, pp. 1279–1287, (2016).

[42] Timothy Niven and Hung-Yu Kao, 'Probing neural network comprehension of natural language arguments', CoRR, **abs/1907.07355**, (2019).

[43] Joakim Nivre, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, et al. Universal dependencies 2.1, 2017. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

[44] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana, 'Interpretml: A unified framework for machine learning interpretability', arXiv preprint arXiv:1909.09223, (2019).

[45] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala, 'Pytorch: An imperative style, high-performance deep learning library', in Advances in Neural Information Processing Systems 32, eds., H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett, 8024–8035, Curran Associates, Inc., (2019).

[46] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, 'Scikit-learn: Machine learning in Python', Journal of Machine Learning Research, **12**, 2825–2830, (2011).

[47] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al., 'Scikit-learn: Machine learning in python', the Journal of machine Learning research, **12**, 2825–2830, (2011).

[48] Victor Pereira and Vládia Pinheiro, 'Report-um sistema de extração de informações aberta para língua portuguesa', in Proceedings of Symposium in Information and Human Language Technology, pp. 191–200. Sociedade Brasileira de Computação, (2015).

[49] Telmo Pires, Eva Schlinger, and Dan Garrette, 'How multilingual is multilingual bert?', arXiv preprint arXiv:1906.01502, (2019).

[50] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin, 'Catboost: Unbiased boosting with categorical features', in Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS'18, p. 6639–6649, Red Hook, NY, USA, (2018). Curran Associates Inc.

[51] William Radford, Joel Nothman, Matthew Honnibal, James R Curran, and Ben Hachey, 'Document-level entity linking: Cmcrc at tac 2010.', in TAC, (2010).

[52] Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation, 2020.

[53] Cleiton F. L. Sena and D. B. Claro, 'Pragmaticoie: a pragmatic open information extraction for portuguese language', Knowledge and Information Systems, 201–213, (February 2020). https://doi.org/10.1007/s10115-020-01442-7.

[54] Cleiton Fernando Lima Sena and Daniela Barreiro Claro, 'Inferportoie: A portuguese open information extraction system with inferences', Natural Language Engineering, **25**(2), 287–306, (2019).

[55] Cleiton Fernando Lima Sena, Rafael Glauber, and Daniela Barreiro Claro, '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, Porto, Portugal, (2017). INSTICC, ScitePress.

[56] Gabriel Stanovsky and Ido Dagan, 'Creating a large benchmark for open information extraction', in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 2300–2305, (2016).

[57] Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan, ‘Supervised open information extraction’, in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 885–895, (2018).

[58] Mingming Sun, Xu Li, Xin Wang, Miao Fan, Yue Feng, and Ping Li, ‘Logician: a unified end-to-end neural approach for open-domain information extraction’, in Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, pp. 556–564. ACM, (2018).

[59] Clarissa Castellã Xavier, Vera Lúcia Strube de Lima, and Marlo Souza, ‘Open information extraction based on lexical-syntactic patterns’, in Intelligent Systems (BRACIS), 2013 Brazilian Conference on, pp. 189–194. IEEE, (2013).

[60] Clarissa Castellã Xavier, Vera Lúcia Strube de Lima, and Marlo Souza, ‘Open information extraction based on lexical semantics’, Journal of the Brazilian Computer Society, **21**(1), 1–14, (2015).

[61] Sheng Zhang, Kevin Duh, and Benjamin Van Durme, ‘Mt/ie: Cross-lingual open information extraction with neural sequence-to-sequence models’, in Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pp. 64–70, (2017).