Journal of Universal Computer Science, vol. 24, no. 1 (2018), 43-69
submitted: 18/8/17, accepted: 26/1/18, appeared: 28/1/18 © *J.UCS*

A Similarity Grammatical Structures Based Method for Improving Open Information Systems

Erick Nilsen Pereira de Souza
(Federal University of Bahia, Salvador, Brazil
ericknilsen@gmail.com)

Daniela Barreiro Claro
(Federal University of Bahia, Salvador, Brazil
dclaro@ufba.br)

Rafael Glauber
(Federal University of Bahia, Salvador, Brazil
rafaelglauber@gmail.com)

Abstract: Open information extraction (Open IE) discovers facts as triples of relationships in texts. A major challenge to Open IE task is to reduce the proportion of invalid extractions. Current methods based on a set of specific features eliminate many inconsistent and incomplete facts. However, these solutions have the disadvantage of being highly language-dependent. This dependence arises from the difficulty in finding the most representative set of features, considering the peculiarities of each language. These solutions require extensive training sets, usually produced with the aid of a specialized linguistic knowledge. Furthermore, although linguistic knowledge resources are common in English, they are scarce in most other languages. Therefore, we propose a method for classifying extracted facts based on the similarity of grammatical structures, which builds models from morphological structures contained in the extraction through the application of algorithms for the detection of isomorphism in sub-graphs. In particular, Portuguese was chosen for the implementation and validation of the proposed approach as it is one of the languages that lack this type of resource.

Key Words: Information Extraction, Open Information Extraction, Maximum Common Subgraph-isomorphism
Category: M.0, M.7, L.1.0

1 Introduction

The increasing availability of data on the web has enabled an intensive global information exchange. It is estimated that over 3.8 billion people accessed the Internet in 2018[1], and approximately 140,000 new sites appear every day [Valant, 2013]. According to [Fader et al., 2011], over 80% of the data generated in social networks, corporate portals, email exchanges, blogs and news sites are composed of text. Inevitably, most of this vast volume of information is irrelevant to the specific interests of each user, which makes content filtering an

[1] http://www.internetlivestats.com/internet-users/

increasing requirement. However, selecting relevant content presupposes a specific level of semantic knowledge about all or most of the information available. This analysis is impossible for a human being to conduct given the amount and dispersion the information. As a result, the automatic interpretation of this vast volume of data has become a focus of research in critical areas, notably in Information Extraction (IE). Recent studies have been conducted to extract facts with non-limited vocabulary from large-scale corpora characterizing Open Information Extraction (Open IE) [Banko and Etzioni, 2008]. Unlike traditional methods, Open IE may reveal unexpected relationships contained in the unstructured text because the process is not limited to a predefined set of relationships.

One of the great challenges of Open IE is to mitigate the ambiguity inherent in natural language, which is responsible for the excessive number of invalid facts in common methods. A fact is defined as invalid if it is incoherent and/or incomplete [Mausam. et al., 2012]. When the semantics of the relationship between entities is complete but incompatible with the correct interpretation of the sentence, the fact is considered incoherent. An incomplete fact occurs when the interpretation of the association between entities is hampered by the absence of terms that compose the relationship. According to [Zelenko et al., 2003], automatic distinctions between valid and invalid decisions can be modeled as a classification problem. Studies by [Banko and Etzioni, 2008], [Wu and Weld, 2010], [Fader et al., 2011] and [Xu et al., 2013] have applied machine learning algorithms to language features extracted from sentences to increase the precision in Open IE systems. A major drawback of these approaches is the difficulty of selecting appropriate features for the task. Besides, feature-based learning requires relatively large training sets to produce satisfactory results. Resources of this type are common in English but limited in most other languages, including Portuguese. In order to facilitate the application of the Open IE task to texts written in languages that lack these linguistic resources, we propose a method for classifying facts based on the similarity of grammatical structures (SGS) here. Our approach builds models from morphological structures contained in the facts to identify patterns of similarities that can be used to distinguish between valid and invalid facts. For this, we applied algorithms to detect sub-graph isomorphism. The main advantage achieved by the proposed model compared with state-of-the-art classification models is a reduction in the adaptation effort by replacing feature-based training sets with sets of examples with lower construction cost.

The remainder of the present paper is structured as follows. Section 2 summarizes related work. In Section 3 a model is proposed to classify extracted facts based on the SGS. Section 4 addresses methods of validating the proposal through experiments conducted on a representative dataset. Finally, Section 5 concludes the present work and presents suggestions and notes for future studies.

2 Related Work

The precursor Open IE system is called *TextRunner* [Banko et al., 2001] and the extraction is performed using a Bayesian classifier trained on features obtained by a part-of-speech tagger (POS tagger) and noun phrase chunk. In [Banko and Etzioni, 2008], the authors observed improvements in the extraction step when Bayesian classifiers were replaced by conditional random field (CRF) probabilistic models for sequential tagging [Lafferty et al., 2001] and Markov logic networks [Zhu et al., 2009]. However, natural language processing (NLP) tools used in different works do not have maximum precision and this deficiency propagates to the various methods in Open IE. The uncertainty inherent in associating each word of the sentence to a given tag increases the probability of identifying invalid facts, which is a characteristic that has damaged the quality of extractions in larger sentences using this approach.

Subsequently, the results in [Wu and Weld, 2010] that were obtained with the weight of evidence (WOE) tool show that features based on dependency parser (DP) increase the precision and recall of the extraction compared with those obtained with POS tagging. However, the identification of syntactic dependencies in sentences increases the extraction cost of the algorithm and makes the application of this approach not recommended for large-scale corpora. The results presented by [Wu and Weld, 2010] suggest that the use of DP increased the cost of processing by 30x for a slight gain in precision. Opening the second generation of Open IE systems we have the *ReVerb* system [Fader et al., 2011]. This system applies the observations described in [Banko and Etzioni, 2008] in which a small set of POS tags applied to morphological patterns is sufficient to represent many types of facts in English. Thus, rather than identifying all syntactic dependencies of a sentence, only the POS tagging is applied, which improves the efficiency of the extraction algorithm. After the application of the extraction algorithm, a set of relationships between the noun phrases is found in unstructured documents. However, the ambiguity inherent in natural language causes inconsistent and incomplete extractions to be included in the set, which requires a classification step maked by a lexical restriction.

The most recent studies in Open IE make use of dependency analysis techniques and have shown improvements in the amount of extracted facts. Some of the leading exponents of this approach are: *OLLIE* [Schmitz et al., 2012], *ClausIE* [Del Corro and Gemulla, 2013] and *Stanford Open IE* [Angeli et al., 2015]. Unlike previous studies that perform the extraction by firstly identifying the verb-phrase, these new approaches use clause types to identify useful pieces for extraction.

2.1 Portuguese studies

As reported by [de Abreu et al., 2013], the first studies in Open IE for Portuguese was the *DepOE* system [Gamallo et al., 2012]. *DepOE* is a proposal for Open IE multilingual which consists of three steps: DP with DepPattern[2] (multilingual), clause constituents and extraction rules that extract the target relationships (similar to *ClausIE* system). Subsequently, the *LSOE* system [Xavier et al., 2013] was published in 2013. The authors described the system as a set of regular expressions identifying facts in an unsupervised way. The authors in [Collovini et al., 2016] described a method that uses the CRF in a Portuguese Open IE system. However, the extraction method is limited to a set of named entities. To the best of our knowledge, the latest published Portuguese Open IE system is described in [Sena et al., 2017]. This system uses a set of syntactic constraints for identifying relationships and arguments. In the end, the method was enhanced to derive new facts by transitive and symmetry inference. All studies in Portuguese Open IE do not have a filter step for the extracted facts. When we analyze the results of these studies, we can find many invalid facts that are extracted by these systems (at results and error analysis sections). The trade-off between the coverage obtained with Open IE systems and its respective precision remain a great challenge. Our study aims to provide a solution to increase the accuracy of these systems while maintaining high coverage.

3 Proposed Solution

The present study aims to apply and evaluate the incorporation of similarity between grammatical structures in the classification of Open IE in Portuguese texts. Therefore, adjustments were made to the main extraction methods based on morphological patterns, which have good results in this category. In Figure 1 we present the pipeline of our system.

The first two steps of this pipeline apply elementary NLP algorithms. Then, the next step was to use the proposed syntactic constraint in [Fader et al., 2011] in Portuguese. Figure 2 shows the syntactic constraint used to extract facts in Portuguese, where facts can be extracted that contain at least: one verb; one verb and one preposition; or one verb and one noun, adverb, pronoun, or descriptive term followed by a preposition. The final step was to avoid extractions generated by syntactic constraints that are too specific and thus less representative. Previous studies did this by lexical constraints, which posit that a valid fact must be found in various instances of a large *corpus*. Therefore, when a fact is not identified in the *corpus*, it is considered too specific to compose an extraction. The lexical constraint in the present study was performed using a

[2] http://gramatica.usc.es/pln/tools/deppattern.html

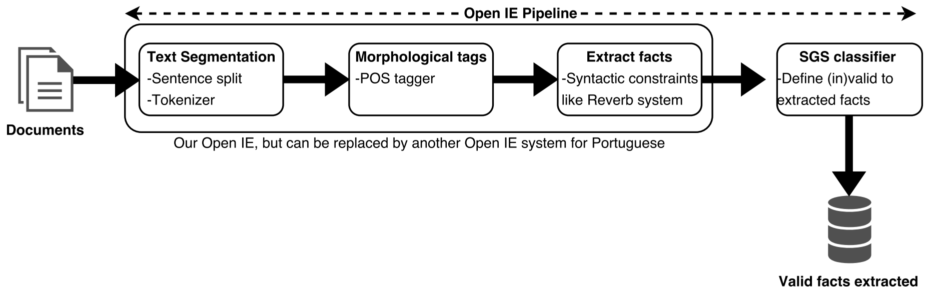

Figure 1: Open IE Pipeline to Portuguese with similarity of grammatical structures classifier.

subset of *Wikipedia* texts written in Portuguese, and according to *Wikimedia*[3] statistics, there are 5.5-fold fewer articles written in Portuguese than in English. This difference means that a greater number of facts were discarded by the lexical constraint in Portuguese because the probability of finding a text sequence in a corpus is proportional to its size. *Wikipedia* was chosen for this task because it was the largest corpus in Portuguese found by the authors. Because of this limitation and considering that less representative facts tend to contain a larger number of words, an additional strategy was used to avoid discarding potentially significant facts. In this strategy, only facts that have more than five POS types are processed by the lexical constraints. From these adaptations, it is estimated that approximately 3/4 of extracted facts based on verbs will be identified in Portuguese (approximately 10% fewer than the percentage obtained by the corresponding approach in English), which can be verified in the recall evaluation described below.

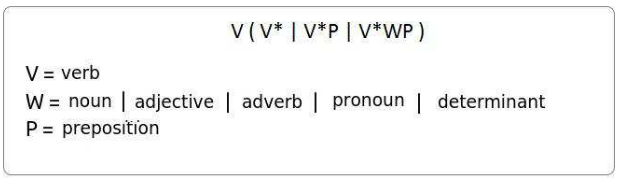

Figure 2: Syntactic restriction adapted for the Portuguese language.

[3] http://stats.wikimedia.org

3.1 Recall Analysis

Because the model for extracting facts in Portuguese cannot identify all types of possible relationships, an analysis was performed to estimate the recall percentage lost after syntactic and lexical constraints were applied, and it considered the adjustments made to enable extractions. In the present analysis, we adopted the method defined in [Fader et al., 2011] in which all valid facts between noun phrases in random text segments are manually identified. From 160 sentences, 177 facts were manually tagged by a specialist, with 76.3% recognized by the automatic extraction method. The remaining facts, i.e. those that were not recognized by the method, correspond to grammatical constructions not covered by the restrictions, which are described below.

– Structures with long unmapped morphological patterns (14.7%):

  - pattern VWWPWP (e.g. X *fez* uma tentativa de acordo *com* Y [X made an attempt according to Y]).

– Sentences with non-contiguous structures (6.2%):

  - coordinated clauses (e.g. X *foi construído* e mantido *por* Y [X was built and maintained by Y]);
  - complements of ditransitive verbs (e.g. X *agradeceu* o presente *a* Y [X thanked Y for the gift]);
  - explanatory clauses (e.g. X, que *comandou* a greve, *foi preso por* Y [X, who led the strike, was arrested by Y]).

– Sentences with relationships that are not among the arguments (2.3%):

  - introductory phrases (e.g. *Sequestrado por* X e Y [Kidnapped by X and Y]);
  - relative clauses (e.g. A empresa X que Y *fundou* [Company X, which Y founded]).

– Relationships excluded by the lexical constraint (0.5%):

  - pattern VVVVWP (e.g. X *alegou estar sendo procurado continuamente por* Y [X claimed to be continuously sought out by Y])

3.2 Grammatical Structures

Similarity approaches have been increasingly applied in language processing tasks, such as in the word sense disambiguation problem [Souza and Claro, 2012]. where the meaning is chosen for an ambiguous word w has a greater similarity to the meanings of neighboring words in contexts where the meaning of w is known. With this assumption, we investigate if the sequences of POS types surrounding the extracted facts contained between noun phrases can be used to differentiate between valid and invalid facts. More precisely, we examine if a new fact can be classified according to structural similarities between sequences of POS types with previously known instances of valid and invalid facts. To verify this hypothesis, a model must be created that represents the sequence of POS types in each extraction and analysis must be performed to estimate the similarity between these sequences. In the present proposal, the model is generated by Algorithm 1, which builds a graph for each sentence using the POS types of tokens between the noun phrases.

Algorithm 1 Builds a graph of morphological structures from the set of classes of a sentence

```
function BuildGraph(listClass)
    ▷ listClass corresponds to a list containing all POS types used in the construction of
    the graph;
    ▷ CreateNode(class) returns a node whose label is a morphological class;
    ▷ SearchNodePerLabel(graph,class) returns a graph node whose label is a certain
    morphological class;
    ▷ AddNode(graph,node) adds a node in the graph;
    ▷ CreateArrow(nodePrevious, nodeAux) creates an edge between two nodes of the graph.
    graph ← ∅;
    nodePrevious ← null;
    for each class in listClass do
        nodeAux ← SearchNodePerLabel(graph, class);
        if nodeAux ≠ null then
            CreateArrow(nodePrevious, nodeAux);
            nodePrevious ← nodeAux;
        else
            no ← CreateNode(class);
            AddNode(graph, node);
            if nodePrevious ≠ null then
                CreateArrow(nodePrevious, node);
            end if
            nodePrevious ← node;
        end if
    end for
    return graph
end function
```

Each morphological class is mapped onto a vertex of the graph, and the links between classes are built based on the order of tokens. Figure 3 shows the structures obtained from examples of sentences with valid facts in Table 1, which are generated by applying the syntactic constraint.

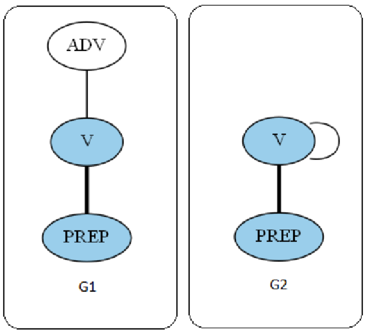

Figure 3: Grammatical structures of valid facts (Table 1) after applying Algorithm 1.

ID	Sentence	Valid fact
1	`A decisão final sobre a UFM <adv>novamente</adv> <v>caberá</v> <prep>ao</prep> prefeito Paulo Maluf.` *(The final decision on the UFM again will be given by the mayor Paulo Maluf.)*	(A decisão final sobre a UFM, caberá ao, prefeito Paulo Maluf) *(The final decision on the UFM, will be given by, the mayor Paulo Maluf)*
2	`As colinas de Golã <v>foram</v> <v>tomadas</v> <prep>por</prep> Israel em 1967.` *(The Golan Heights were taken by Israel in 1967.)*	(As colinas de Golã, foram tomadas por, Israel) *(The Golan Heights, were taken by, Israel)*

Table 1: Valid extracted facts derived from the syntactic constraint.

Graphs G_1 and G_2 (Figure 3) are created by processing sentences in Algorithm 1, and they have structural similarities, i.e. there is a subgraph G' common to G_1 and G_2 formed by vertices interconnected by edges highlighted in Figure 3. More precisely, given $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, subgraphs $G'_1 = (V'_1, E'_1)$ and $G'_2 = (V'_2, E'_2)$, where $E'_1 \subseteq E_1$ and $E'_2 \subseteq E_2$ (with $|E'_1| = |E'_2| > 0$), are isomorphic, i.e. there is a bijection $f : V'_1 \to V'_2$, where u and v are adjacent in G'_1 if, and only if, $f(u)$ and $f(v)$ are adjacent in G'_2, $\forall u, v \in V'_1$. These similarities can also be found in structures generated from the sequence of POS types of sentences that produce invalid facts.

ID	Sentence	Invalid fact
1	`A Alemanha <v>ficou</v> <prep>em</prep> <sub>ruínas</sub> <conj>mas</conj> os EUA viraram uma potência após a Segunda Guerra` *(Germany was in ruins but the US turned a power after World War II)*	(a Alemanha, ficou em, os EUA)*(Germany,was in, the US)*
2	`O Bahia <v>gosta</v> <prep>de</prep> <sub>jogo</sub> <adj>aéreo<adj> <conj>enquanto</conj> <prep>do</prep> Grêmio espera-se mais perigo pelo chão.` *(Bahia like air game while Gremio expected more danger on the ground.)*	(o Bahia, gosta de, o Grêmio) *(Bahia, likes, Gremio)*

Table 2: Invalid extracted facts derived from the syntactic restriction.

In the example contained in Figure 4, G_3 corresponds to a subgraph of G_4, indicating that the graphs obtained from invalid sentences of Table 4 have a high degree of similarity. Also, according to what was described in [Nian et al., 2003], larger isomorph subgraphs between two graphs have a higher degree of similarity.

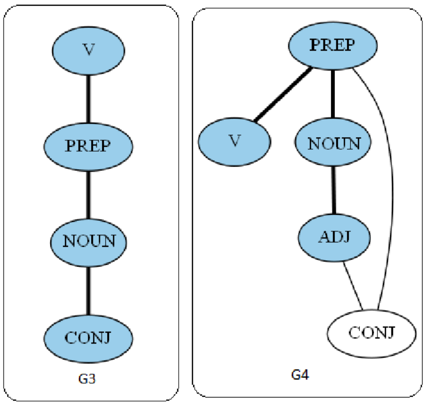

Figure 4: Grammatical structures of invalid facts (Table 2) after the application of Algorithm 1.

In this solution, the graphs of grammatical structures constructed by Algorithm 1 are connected and have a maximum of 10 vertices that correspond to the possible POS types in Portuguese[4]. Given that the proposed model generates

[4] The POS types found in Portuguese are noun, article, adjective, numeral, pronoun, verb, adverb, preposition, conjunction and interjection.

graphs with a small number of vertices, the maximum common subgraph (MCS) can be obtained using exact solutions [Nian et al., 2003, Levi, 1972] within an acceptable processing time. From the identification of MCS between grammatical structure models, the strategies that allow us to infer a numerical value indicating the similarity between them must be defined. Subsequently, the similarity approaches for graphs, which were adapted for the grammatical structure model proposed here, are described.

3.3 Similarity of Grammatical Structures (SGS)

As described in [Nicholson et al., 1987], the similarity between two graphs is equivalent to the distance between the structured objects, concepts or models represented by the graphs. Based on this principle, the similarity calculation between graphs of grammatical structures can identify properties that indicate similarities between the extracted facts that compose these structures. These properties can be useful for extracted fact classification strategies because the similarities can be used to group instances of the same class. Thus, facts belonging to the class of valid facts tend to be closer to other valid facts, and invalid facts tend to have similar structural characteristics. A classic method to estimate the similarity between two graphs consists of identifying the largest pattern between them. The identification of this common pattern, which has been addressed as a subgraph isomorphism problem, must be followed by a strategy that can infer a value of similarity between models represented by the graphs. In grammatical structure graphs, two components of the similarity calculation can be identified: the structure and tags of the common isomorphic subgraph.

3.3.1 Structural Similarity

Intuitively, structural similarity is directly proportional to the size of the maximum clique between two graphs, i.e. a greater number of nodes in the isomorphic graph indicate greater similarity. However, the similarity is not obtained based on the absolute size of the maximum clique but according to its proportion relative to other graphs. This proportional estimation is performed to prevent similarity values in larger graphs from being unduly increased. For example, assuming that the similarity is calculated between two sets of graphs $P_1 = \{G_1(V_1, E_1), G_2(V_2, E_2)\}$ and $P_2 = \{G_3(V_3, E_3), G_4(V_4, E_4)\}$, where $|V_1| = 100$ and $|V_2| = |V_3| = |V_4| = 10$, if the maximum cliques between these pairs of graphs are formed by the same number of vertices, both pairs cannot be considered as having identical similarities. In fact, if the maximum cliques contain ten vertices, then G_3 and G_4 are isomorphic and have maximum structural similarity. Conversely, G_1 and G_2 cannot be isomorphic because they have a different number of vertices, and they should not be considered to have the same similarity as G_3 and G_4. Thus, the similarity is obtained by normalizing the compared graphs. More precisely, if $G_c(V_c, E_c)$ represents the maximum clique between $G_1(V_1, E_1)$ and $G_2(V_2, E_2)$, then the structural similarity in the present study is calculated according to the following equation:

$$SIM_e(G_1, G_2) = \frac{|V_c|}{D}, \tag{1}$$

where D represents a function between the vertices of G_1 and G_2 that can use several normalization approaches, such as decreased similarity proportional to the graph with a number of nodes greater $(max(|V_1|, |V_2|))$ or lower $(min(|V_1|, |V_2|))$ between the compared graphs. In the present study, we used the mean between the vertices, i.e., $D = \frac{|V_1|+|V_2|}{2}$. As $V_c \subseteq V_1$, $V_c \subseteq V_2$ and $|V_c| \geq 1$, then $0 < SIM_e(G_1, G_2) \leq 1$. Thus, the maximum structural similarity occurs when G_1 and G_2 are isomorphic, i.e. $V_1 = V_2 = V_c$. However, the minimum structural similarity occurs when the maximum clique consists of a single vertex.

3.3.2 Similarity of Tags

In addition to the structural component, the differences between the tags of vertices of the MCS obtained from each compared pair of graphs must be considered. Figure 3 shows that isomorphic subgraphs have identical tags, which indicates greater similarity. In Figure 4, the mapping between isomorphic subgraphs has vertices with distinct tags ($CONJ \rightarrow ADJ$). In general, the similarity between the tags is estimated by the Levenshtein distance [Levenshtein, 1966], which is a measure of distance between the character strings on the tags:

$$SIM_r(G_1, G_2) = \frac{\sum_{i=1}^{|V_c|} Levensthein[\psi(V_1, v_i), \psi(V_2, v_i)]^{-1}}{D}, \tag{2}$$

where $\psi(V_1, v_i)$ and $\psi(V_2, v_i)$ are the functions that return the respective tags of the vertices in V_1 and V_2 which correspond to the vertex v_i contained in the maximum clique between G_1 and G_2. The strings of tags consist of POS types in the grammatical structure model; therefore, the Levenshtein distance is a less representative similarity measure in this model because the similarity between characters of POS types does not necessarily imply semantic dependency, and it is replaced by a correlation matrix between POS types:

$$SIM_r(G_1, G_2) = \frac{\sum_{i=1}^{|V_c|} M[\psi(V_1, v_i)][\psi(V_2, v_i)]}{D} \tag{3}$$

The elements from matrix M provide correlations between each pair of possible POS types. Thus, equation 3 uses a subset of elements of M obtained from ψ. The construction of M is based on the proportion of words with the same spelling belonging to different POS types, which characterizes the morphological ambiguity. Intuitively, a higher number of sets with identical words belonging to different classes produces a higher correlation between the classes. An example of morphological ambiguity is found in the word “casa” (*house in English*), which can be classified both as a verb (“casar” - to marry) and noun (moradia - housing). Thus, the word “casa” contributes to the increased correlation between nouns and verbs in Portuguese. Another example is: “Red is my favorite color” and “The red bird is faster” where the word “red” has different uses, a noun and an adjective. Because the process to make M does not limit the accumulation of similarity between POS types, the values of M must be normalized using Equation 4:

$$Normalize(a_{i,j}) = \frac{a_{i,j}}{argmax(M) + k_p}, \tag{4}$$

where $argmax(M)$ represents the maximum frequency obtained among all elements of M and k_p is a weighting factor that limits the elements of M that do not belong to the main diagonal for values lower than 1. Thus, only elements from the main diagonal of M display a maximum correlation, and they correspond to tags that are composed of the same grammatical class. Although there is no limit to k_p, the values that provide the best results were in the range $[0.2, 0.5]$. Thus, the values in M were generated assuming $k_p = 0.2$.

The classification of extracted facts in Open IE systems using the SGS is performed by a classifier algorithm. Similar to the KNN algorithm, our classifier compares a test case with the valid and invalid examples. The threshold ξ_s corresponds to the parameter used by classifier algorithm to decide whether a new instance must be classified, which is based on the difference between the similarity values accumulated between the test case and examples of valid and invalid facts.

4 Experiments

The SGS model is validated by a comparison with current feature-based classification methods using standard IE evaluation measures, such as precision, recall, F-measure and ROC curve. In the Open IE classification problem, precision is defined as the ratio between the number of correctly classified facts and number of facts that were classified. More precisely:

$$\mathtt{P} = \frac{\mathtt{\#(correctly\ classified\ facts)}}{\mathtt{\#(classified\ facts)}} \tag{5}$$

It is hard to establish the complete set of possible extractions in a sentence. In works such as [Del Corro and Gemulla, 2013] all extractions made by the different Open IE systems are considered distinctly. In our experiment, we considered only the set of extractions that our extraction method was able to do. Therefore, recall is calculated according to the fraction of facts that were correctly classified and total valid facts:

$$\mathtt{R} = \frac{\mathtt{\#(correctly\ classified\ facts)}}{\mathtt{\#(valid\ facts)}} \tag{6}$$

The F-measure is calculated by the harmonic mean between precision and recall:

$$\mathtt{F\text{-}measure} = 2 \times \frac{\mathtt{P} \times \mathtt{R}}{\mathtt{P + R}} \tag{7}$$

A receiver operating characteristic curve (ROC curve) measures the ability of a binary classifier when its threshold is varied. The ROC curve is plotted by the true positive rate (TPR) against the false positive rate (FPR). For our experiment, we consider these rates as:

$$\mathtt{TPR} = \mathtt{R\ and\ FPR} = \mathtt{1\ -\ P} \tag{8}$$

Through normalized units the area under the ROC curve (AUC) is the measure equal to the probability of a binary classifier choosing a positive example before a negative example.

4.1 Tools and Resources

The experiments were performed using the *CETENFolha* corpus[5] (Corpus of Electronic Texts Extracts NILC/Folha de S. Paulo - *Corpus de Extratos de Textos Eletrônicos NILC/Folha de S. Paulo*), which contains approximately 24 million words in Portuguese extracted from texts of the *Folha de São Paulo* newspaper. The POS types of words contained in the selected sentences were automatically obtained by the CoGrOO morphosyntactic tagger[6]. We built three datasets with sentences extracted randomly from the CETENFolha. The first one, we used to compute the similarity of grammatical structures. We extracted automatically 116 facts of type (fn_1, rel, fn_2) from 116 sentences, where fn_1 and fn_2 represent noun phrases that are found before and after the relationhip, and rel denotes the relational phrase of the fact. Relational phrases were obtained according to the syntactic constraints, and the noun phrases that contain at least one proper name classified by CoGrOO. The second one was to calculate the values of M for the similarity of tags with 1000 sentences. For the evaluation of the proposed method, we recovered another 500 sentences with 1006 extracted facts automatically. Of these, 582 facts were annotated manually. Each extracted fact was classified as valid or invalid to form the training set (in feature-based

[5] http://www.linguateca.pt/cetenfolha/
[6] http://cogroo.sourceforge.net/

methods) and a set of examples (in the SGS method)[7]. In order to compare the proposed SGS with the current feature-based methods [Fader et al., 2011], 12 training features adapted for Portuguese were selected (Table 3). The grammatical differences between the Portuguese and English languages do not allow direct use of the proposal observed in [Fader et al., 2011]. We therefore adapt only the set of features that were portable for a fair comparison. The feature values were automatically extracted from selected sentences from the corpus. In particular, the feature-based classification was compared to the SGS approach by using the following four machine learning algorithms used recursively in state-of-the-art studies: *J48*, *Lib SVM*, *Multilayer Perceptron* and *Naive Bayes* were applied using the Weka[8] data mining tool.

F_1	size(sentence) - size($fn_1 + rel + fn_2$) $<$ 30 characters?
F_2	The last preposition in rel is “de”?
F_3	The last preposition in rel is “com”?
F_4	The last preposition in rel is “por”?
F_5	The last preposition in rel is “pela”?
F_6	The last preposition in rel is “pelo”?
F_7	The last preposition in rel is “para”?
F_8	The last preposition in rel is “em”?
F_9	The string $fn_1 + rel$ is contained in the sentence?
F_{10}	A string $rel + fn_2$ is contained in the sentence?
F_{11}	A string $fn_1 + rel + fn_2$ is contained in the sentence?
F_{12}	Less than 30 words in the sentence?

Table 3: Features used for the construction of the training dataset in Portuguese.

4.2 Evaluation

We evaluated our proposal using the following four groups of experiments:

1. The experiments from group 1 measure the representativeness (merit) of the set of features adjusted for Portuguese in Table 3. From the results obtained for this set, the adequate subset can be selected from the tested features to classify new instances.

2. The experiments from group 2 evaluate the behavior of the SGS method for different ξ_s threshold values to select the best range in the tested data set.

3. The experiments from group 3 compare the classification models based on the best parameters obtained in groups 1 and 2.

[7] The dataset can be downloaded in `http://formas.ufba.br/uploads/train_sgs_segapp.dump` and `http://formas.ufba.br/uploads/experiments_segapp.dump` in *dump* format for PostgreSQL - `https://www.postgresql.org/`

[8] http://www.cd.waikato.ac.nz/ml/weka

4. The experiments from group 4 evaluate the execution time of each method to verify the feasibility of applying the SGS method to training sets of documents in real applications.

Thus, group 1 and 2 calibrate the parameters used in each approach so that they can be compared in terms of precision, recall and ROC, and these parameters are analyzed in group 3 and evaluated regarding runtime in group 4.

4.2.1 Experiment 1: representativeness of features

In this group, experiments were conducted to determine the most significant attributes (features) of the tested data set. The effectiveness or merit of the attributes was estimated by the Correlation-based Feature Selection (CFS) algorithm [Hall, 1999], which uses a heuristic based on correlations to assess the ability of each attribute to predict the class of a test instance within a training set. The assumption that underlies this algorithm is that useful subsets of attributes must be highly correlated with the prediction class and poorly correlated with each other because attributes that are highly correlated are considered redundant and do not increase the predictive ability of the subset. Therefore, if S is a subset containing k attributes, then the merit of S is calculated by Equation 9:

$$M_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}, \tag{9}$$

where $\overline{r_{cf}}$ is the average correlation between each attribute of S and the class attribute and $\overline{r_{ff}}$ indicates the mean correlation between all combinations of attributes in S. The correlation between the attributes can be estimated by various heuristics, such as the symmetric uncertainty coefficient (based on the concepts of entropy and information gain)[Kononenko and Bratko, 1991] and *Relief* algorithm [Kononenko, 1994] (which uses an approach based on instances to associate weights to iterations between attributes). Figure 5 illustrates the merit of the features described in Table 3 and is based on the entire data set (582 facts obtained from 500 sentences). The features F_9, F_{10} and F_{11} are those with the highest predictive capacity. Moreover, feature F_1 can be eliminated from the set of attributes without affecting the classification quality because it has null merit.

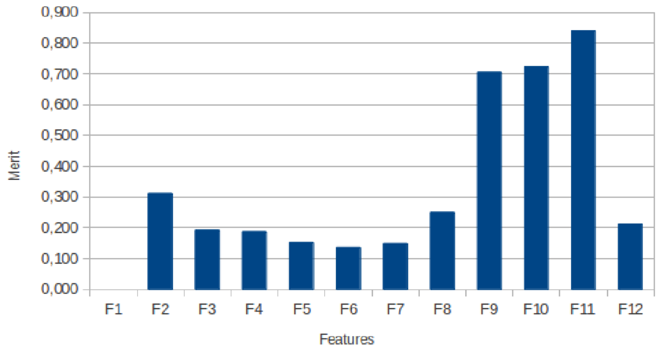

Figure 5: Representation of features in the dataset.

The results in Figure 5 were obtained by executing the CFS algorithm using the *BestFirst* search strategy with parameters $D = 1$ (*forward search*) and $N = 5$ (number of nodes in the stopping criterion), with features selected from the entire training set. Based on these results, four subsets of features (Table 4) were selected for evaluation in the aforementioned machine learning algorithms. The group CF_1 consists of all features that have non-null merit, and groups CF_2, CF_3 and CF_4 correspond to the subsets of best features evaluated by the CFS algorithm.

Subset	Evaluated feature	Elements of best subset
CF_1	-	F_2, F_3, F_4, F_5, F_6, F_7, F_8, F_9, F_{10}, F_{11}, F_{12}
CF_2	F_{11}	$F_9, F_{10}, F_{11}, F_{12}$
CF_3	F_{10}	$F_2, F_6, F_8, F_{10}, F_{11}$
CF_4	F_9	F_4, F_9, F_{11}

Table 4: Features subsets.

Features with the greatest merits do not always form the best subset because they may be highly correlated. Thus, features F_2, F_9, F_{10} and F_{11} do not form a subset with high predictive capacity because of the high correlation between F_2 and F_9.

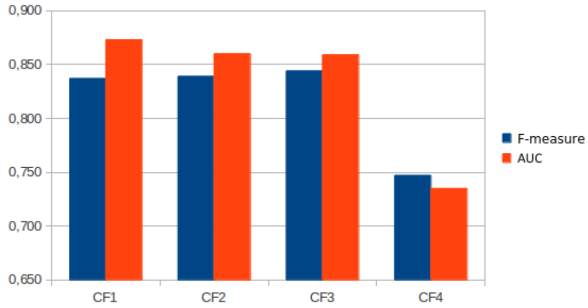

Figure 6: Evaluation of *feature* sets.

Figure 6 shows the mean values of the F-measure and AUC of the four classification algorithms evaluated in each set of features using the 10-fold cross-validation method. The results show approximately equal values for the three groups CF_1, CF_2 and CF_3, and the greatest difference was 0.7% for the F-measure and 1.4% for the AUC between the groups CF_1 and CF_3. These results indicate that the dimensionality of the attributes can be reduced from 11 to 4 (CF_2) or 5 (CF_3) features with minimal losses in the classification quality. Conversely, group CF_4 exhibited mean F-measure values that were 9.5% lower and mean AUC values that were 13.8%, which indicates that it was the least representative set among the evaluated sets.

4.2.2 Experiment 2: threshold ξ_s

As in Experiment 1, this experiment is performed to obtain the best value range for threshold ξ_s in the classification by SGS. In this group of experiments, variations in classification quality were evaluated as a function of the similarity threshold ξ_s in the SGS method. An analysis of the classifier shows that an increase in threshold results in a reduction in the classification recall because the test instances with low similarity differences are not classified. Figure 7 indicates that the behavior of the F-measure and AUC of the SGS method are caused by variations in the threshold, which was determined using the 10-fold cross-validation method.

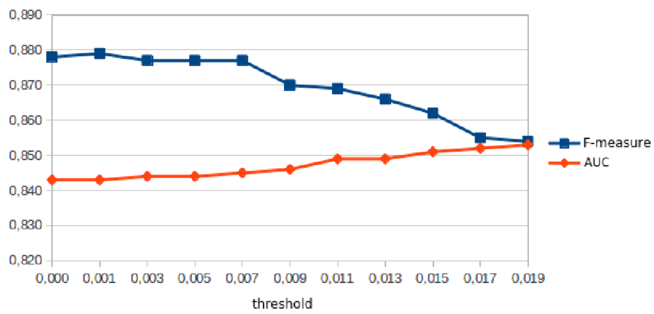

Figure 7: Variation of F-measure and AUC depending on the similarity threshold variation ξ_s.

The results were obtained from the variation of ξ_s in the interval $[0; 0.019]$. At the lower limit of the interval ($\xi_s = 0$) the recall reaches the maximum value, i.e. all instances are classified. However, a threshold value from 0.019 reduces the recall below 80%, which hinders the predictive ability of the model. Consequently, there is a 3.5% reduction in the F-measure between the lower and upper limits of the interval despite a slight increase in precision and AUC (approximately 1%). Figure 8 shows a balance between the F-measure and AUC, which can be used to obtain the best threshold variation interval from the tested data set. We note that the best values are obtained for $\xi_s \in [0; 0.007]$.

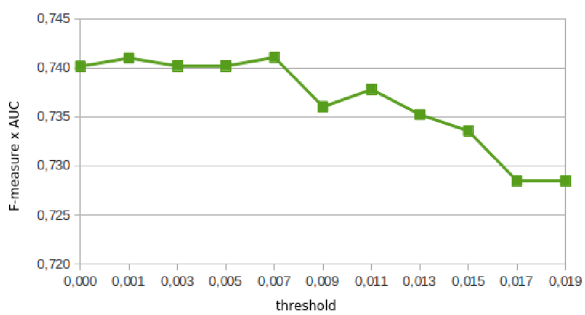

Figure 8: Varying the balance between F-measure and AUC depending on the threshold variation ξ_s.

Identifying the optimal interval of similarity threshold values along with the results of experiments from group 1 allow us to compare the SGS method with feature-based machine learning approaches, the experiments of which are described below.

4.2.3 Experiment 3: classification evaluation

This experiment group compared the proposed classification method based on SGS with the feature-based approach used in current methods and adapted to Portuguese. From the results obtained in the group 1 experiments, the set of features CF_1 was selected for processing in the machine learning algorithms. Similarly, the group 2 experiments provide the adequate similarity threshold interval values for the SGS methods, of which $\xi_s = 0.005$ was used.

Method	**Precision**	**Recall**	**F-measure**
SGS	0.781 ± 0.016	**0.973** ± 0.010	**0.860** ± 0.022
J48	**0.848** ± 0.014	0.841 ± 0.018	0.841 ± 0.018
Lib SVM	**0.848** ± 0.019	0.840 ± 0.018	0.839 ± 0.018
Perceptron	0.823 ± 0.038	0.820 ± 0.041	0.820 ± 0.040
Naive Bayes	0.800 ± 0.037	0.799 ± 0.039	0.799 ± 0.039

Table 5: Average results obtained by cross-validation with 10 folds.

Values in Table 5 show the tested methods sorted by descending performance. Additionally, the curves in the graph of Figure 9 illustrate variations in F-measure with increases in the set of relationships tested in each algorithm. The mean values and corresponding standard deviations are obtained by processing ten sets of sentences of different sizes that range from 57 to 582 extractions. The performance of the classification based on SGS exceeded the J48 algorithm, which is the classifier with the best performance among the feature-based evaluated classifiers, by approximately 2%, and exceeded the Bayesian classifier, which exhibited the lowest value for the F-measure, 6%.

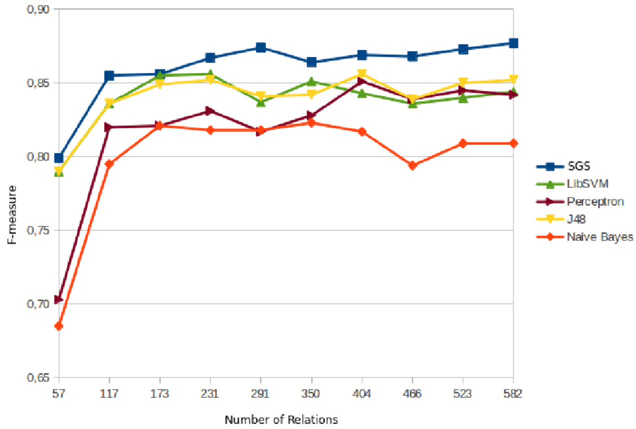

Figure 9: F-measure evaluation on the quantity of classified facts.

The graph in Figure 9 shows that the SGS method is superior to methods that include more than 173 facts, which is equivalent to a set of examples with 156 instances in the cross-validation. Above this value, a significant change in classification performance was not observed, which indicates that a small sample base is sufficient to obtain satisfactory results with the proposed method. Figure 10 shows the behavior of the ROC curves of feature-based classifiers. The curves were obtained from the classification of all 582 extractions of the data set in the cross-validation. Curves that approach the point (0,1) in the ROC space represent good classifiers because they have high true positive rates and low false positive rates; thus, the algorithm that provided the best results was the *Multilayer Perceptron* (AUC = 0.898), followed by the *Bayesian* classifier (AUC = 0.882), *J48* (AUC = 0.867) and *Lib SVM* (AUC = 0.846).

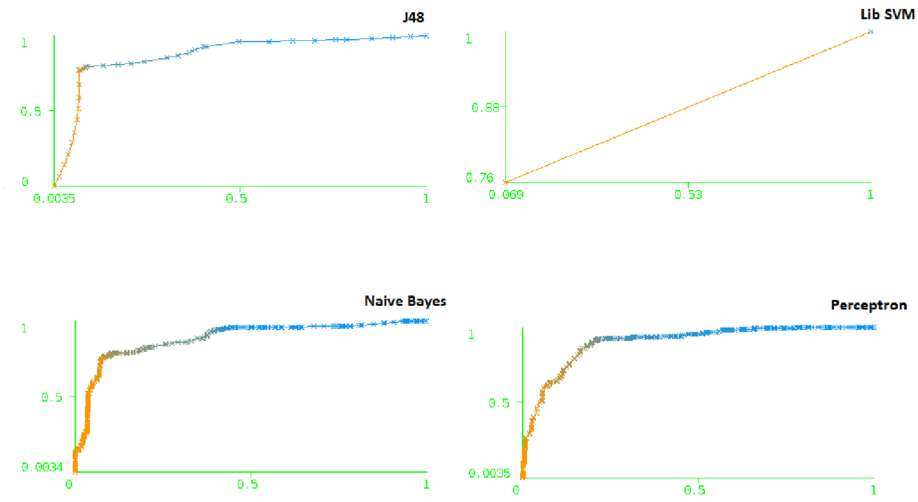

Figure 10: ROC curves of features based classification.

The ROC curve of the SGS method, which can be observed in Figure 11, was obtained using the Jrocfit[9] tool. SGS reduced the AUC by 5.4% compared with the classifier with the best curve (*Multilayer Perceptron*), and the value was equal to the classifier with the worst curve (*Lib SVM*). However, compared with the mean AUC values of the ten evaluated sets of different sizes, the difference dropped to 3%, and the mean AUC of the SGS method was higher compared with the J48 algorithm and lower compared to the other evaluated machine learning algorithms (Figure 12).

[9] www.jrocfit.org

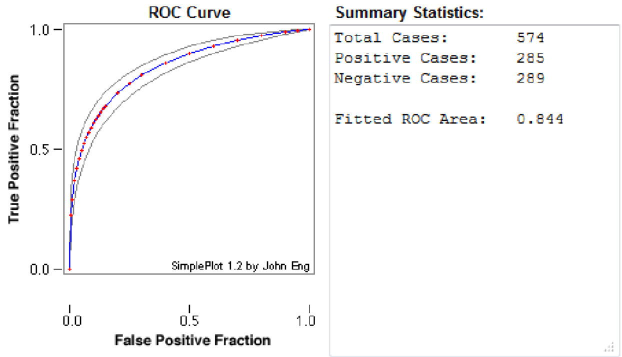

Figure 11: ROC analysis of SGS method.

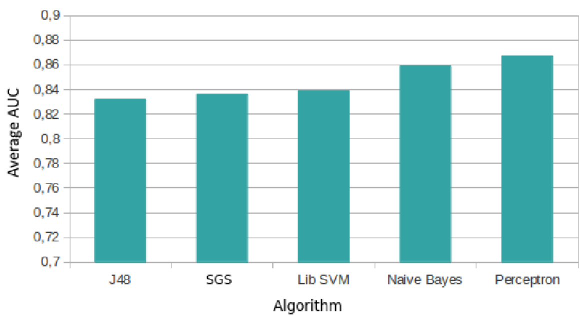

Figure 12: Comparing the average of the areas under ROC curves of classifiers.

An analysis of the graph in Figure 9 shows that the SGS method had the best classification quality among all methods evaluated using the F-measure. Conversely, when the evaluation was performed using the AUC (Figure 12), the proposed method was only better than one of the tested feature-based classifiers. This is because unlike the AUC, the F-measure does not consider the true negative rate (i.e. invalid facts classified as invalid), a characteristic that impacts the evaluation of classifiers, especially binary classifiers (addressed in the present study). The results obtained by the features and SGS methods are close for precision, but a little distant for the sensitivity. Thus, we consider the F-measure suitable for comparison of the different classification models. These results indicate that feature methods are capable of identifying more true negative results than the SGS method. However, similar to the authors in

[Del Corro and Gemulla, 2013] the F-measure is considered adequate to evaluate methods that address this problem.

4.2.4 Experiment 4: time evaluation

In addition to harmonic measurements between precision and recall, execution time is another important aspect that can demonstrate the validity of the proposed method because the structural similarity component of the method uses an algorithm with exponential time complexity. Table 6 shows the time values used to build the model, process each fold in the cross-validation and perform the entire experiment based on ten sets of sentences with different sizes and relationships. The model constructed with the SGS method includes the morphological tagging of sentences and graphs of grammatical structures from the set of examples. The experiments were conducted using a computer with an Intel Core i5 3.2 GHz 64-bit processor and 8 GB of RAM.

Sentences	Facts	Execution time (s)		
		Model	*Per fold*	*Total*
50	57	0.02	1.01	10.13
100	117	0.05	5.71	57.06
150	173	0.07	12.02	120.23
200	231	0.11	25.98	259.79
250	291	0.14	41.11	411.07
300	350	0.17	58.51	585.13
350	404	0.17	67.09	670.93
400	466	0.20	95.30	953.01
450	523	0.21	110.87	1108.71
500	582	0.24	141.78	1417.75

Table 6: Execution time evaluation.

Figure 13 shows a comparison of the construction time of the model based on the number of processed facts in experiments for all of the evaluated methods. The proposed method is the second slowest in the construction of the classification model and only outperforms the *Multilayer Perceptron*, although its time curve is closer to those generated by the most efficient classifiers. Conversely, the most critical asymptotic behavior in the SGS method classification algorithm occurs in the structural similarity calculation, where the maximum isomorphic subgraph is determined among the graphs of grammatical structures of the examples and test sets.

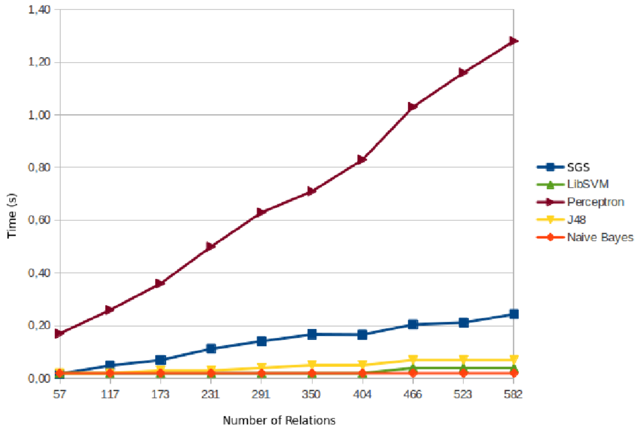

Figure 13: Time to construct the models.

We consider an estimate in the creation of the model of structural and tags similarity as total time. The execution time of the proposed method significantly exceeds the time of feature-based machine learning algorithms because of the exponential growth of the time curve with an increased number of instances (Figure 14). However, because the F-measure of the method approaches the maximum value for a small set of examples (156 instances), the model can be generated from a reduced set and is capable of classifying a significant number of extractions in the time required by feature-based methods.

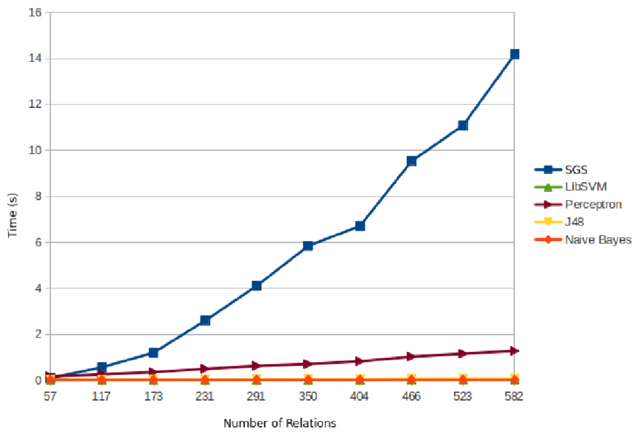

Figure 14: Total time of experiments.

5 Conclusions and Future Studies

The automatic distinction between valid and invalid facts is a recurrent problem of systems that perform Open IE in unstructured text. When the identified relational phrases use a wide vocabulary, the classification task becomes more important to the quality of extracted facts because of the ambiguity inherent in natural language, which can lead to a large number of invalid facts. A greater amount of the Open IE systems extract facts primordially from texts written in English, a language that has the most sophisticated language resources. In this scenario, new approaches have been studied to enable the open relation extraction for languages that have few linguistic resources. In particular, the present study proposes an approach to reduce the language dependence at a particular point in this process wherein extracted facts are classified. The experiments carried out indicate that satisfactory results can be obtained for open relation classifications using SGS and example training sets with a limited number of instances. These results show that the proposed method can replace learning approaches based on language features with training sets that have higher construction costs and are unavailable in most languages. Conversely, aspects related to the time complexity of the SGS approach used in this solution must be further assessed so that it can be applied to large document repositories.

5.1 Future Studies

In future studies, we intend to investigate approaches that can reduce the language dependence at other points of the Open IE by identifying relational phrases and using probabilistic models, such as CRF [Lafferty et al., 2001] and Hidden Markov Model [Levinson et al., 1983]. Such approaches would eliminate the need for morphological tagging and increase the universe of possible facts, which are obtained in the present study through syntactic constraints based on morphological patterns. One of the most important characteristics of our approach is not to use specific words in the Portuguese lexicon. This gives our method the possibility of a lower cost of adaptation to other languages. We would like to advance in new experiments for extractions made in Spanish, French or any other language that has the appropriate requirements. Also, when considering the time constraints imposed by the intractable nature of the exact algorithm used to determine MCS, we raise the cost of our proposal. We must investigate the performance of Open IE classification approaches that are considered more efficient, such as those that use genetic algorithms for the approximate calculation MCS [Gasteiger et al., 2006] and those based on compact representations of graphics, i.e. graphic fingerprints [Teixeira et al., 2012]. Our proposal allows for facts extracted by different Open IE systems for Portuguese to be classified, increasing the precision. In a forthcoming study, we shall map the state of the art and evaluate these systems by applying our proposal.

References

[Angeli et al., 2015] Angeli, G., Premkumar, M. J., and Manning, C. D. (2015). Leveraging linguistic structure for open domain information extraction. *Linguistics*, 1(24):344—-354.

[Banko and Etzioni, 2008] Banko, M. and Etzioni, O. (2008). The tradeoffs between open and traditional relation extraction. *In Proceedings of ACL-08: HLT, pages 28-36, Columbus, Ohio, June. Association for Computational Linguistics.*

[Banko et al., 2001] Banko, M., J. Cafarella, M. J., Soderland, S., Broadhead, M., and Etzioni, O. (2001). Open information extraction from the web. *In the Proceedings of the 20th International Joint Conference on Artificial Intelligence, pages 2670-2676, January.*

[Collovini et al., 2016] Collovini, S., Machado, G., and Vieira, R. (2016). Extracting and structuring open relations from portuguese text. In *International Conference on Computational Processing of the Portuguese Language*, pages 153–164. Springer.

[de Abreu et al., 2013] de Abreu, S. C., Bonamigo, T. L., and Vieira, R. (2013). A review on relation extraction with an eye on portuguese. *Journal of the Brazilian Computer Society*, 19(4):553–571.

[Del Corro and Gemulla, 2013] Del Corro, L. and Gemulla, R. (2013). Clausie: clause-based open information extraction. In *Proceedings of the 22nd international conference on World Wide Web*, pages 355–366. ACM.

[Fader et al., 2011] Fader, A., Soderland, S., and Etzioni, O. (2011). Identifying relations for open information extraction. *In Proceedings of Conference on Empirical Methods in Natural Language Processing.*

[Gamallo et al., 2012] Gamallo, P., Garcia, M., and Fernández-Lanza, S. (2012). Dependency-based open information extraction. In *Proceedings of the joint workshop on unsupervised and semi-supervised learning in NLP*, pages 10–18. Association for Computational Linguistics.

[Gasteiger et al., 2006] Gasteiger, J., Reitz, M., Han, Y., and Sacher, O. (2006). Analyzing biochemical pathways using neural networks and genetic algorithms. *Aust. J. Chem. 2006, 59, 854-858.*

[Hall, 1999] Hall, M. (1999). *Correlation-based Feature Selection for Machine Learning.* PhD thesis, University of Waikato, Hamilton, NewZealand.

[Kononenko, 1994] Kononenko, I. (1994). Estimating attributes: Analysis and extensions of relief. *In Proceedings of the European Conference on Machine Learning.*

[Kononenko and Bratko, 1991] Kononenko, I. and Bratko, I. (1991). Information-based evaluation criterion for classifiers performance. *Machine Learning, 6:67-80.*

[Lafferty et al., 2001] Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the Eighteenth International Conference on Machine Learning (ICML-2001).*

[Levenshtein, 1966] Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady, 10:707-710.*

[Levi, 1972] Levi, G. (1972). A note on the derivation of maximal common subgraphs of two directed or undirected graphs. *Calcolo 9 (1972) 341-352.*

[Levinson et al., 1983] Levinson, S., Rabiner, L., and Sondhi, M. (1983). An introduction to the application of the theory of probabilistic functions of a markov process to automatic speech recognition. *The Bell System Technical Journal, 62(4), pp. 1035-1074.*

[Mausam. et al., 2012] Mausam., Schmitz, M., Bart, R., Soderland, S., and Etzioni, O. (2012). Open language learning for information extraction. *Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CONLL). Jeju, Korea. July 2012.*

[Nian et al., 2003] Nian, Z., Wunsch, D., and Harary, F. (2003). The subcircuit extraction problem. *Potentials, IEEE, 22(3): p. 22-25.*

[Nicholson et al., 1987] Nicholson, V., Tsai, C., Johnson, M., and Naim, M. (1987). A subgraph isomorphism theorem for molecular graphs. *Graph Theory and Topology in Chemistry,(51):226-230, 1987.*

[Schmitz et al., 2012] Schmitz, M., Bart, R., Soderland, S., Etzioni, O., et al. (2012). Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 523–534. Association for Computational Linguistics.

[Sena et al., 2017] Sena, C. F. L., Glauber, R., and Claro, D. B. (2017). Inference approach to enhance a portuguese open information extraction. In *Proceedings of the 19th International Conference on Enterprise Information Systems - Volume 1: ICEIS,*, pages 442–451. INSTICC, ScitePress.

[Souza and Claro, 2012] Souza, E. and Claro, D. (2012). Evaluation of semantic similarity in WSD: An analysis to incorporate it into the association of terms. *WebMedia'12, October 15-28, São Paulo/SP, Brazil.*

[Teixeira et al., 2012] Teixeira, C., Silva, A., and Meira, W. (2012). Min-hash fingerprints for graph kernels: A trade-off among accuracy, efficiency, and compression. *Journal of Information and Data Management, 3(3), 227-242.*

[Valant, 2013] Valant, S. (2013). More than 2 billion people use the internet, here's what they're up to (infographic). *The Culture-Ist: http://www.thecultureist.com, May 9, 2013.*

[Wu and Weld, 2010] Wu, F. and Weld, D. S. (2010). Open information extraction using wikipedia. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL '10, pages 118-127, Morristown.*

[Xavier et al., 2013] Xavier, C. C., de Lima, V. L. S., and Souza, M. (2013). Open information extraction based on lexical-syntactic patterns. In *Intelligent Systems (BRACIS), 2013 Brazilian Conference on*, pages 189–194. IEEE.

[Xu et al., 2013] Xu, Y., Kim, M.-Y., Quinn, K., Goebel, R., and Barbosa, D. (2013). Open information extraction with tree kernels. In *HLT-NAACL*, pages 868–877.

[Zelenko et al., 2003] Zelenko, D., Aone, C., and Richardella, A. (2003). Kernel methods for relation extraction. *Journal of Machine Learning Research 3 1083-1106.*

[Zhu et al., 2009] Zhu, J., Nie, Z., Liu, X., Zhang, B., and Wen, J. (2009). Statsnowball: a statistical approach to extracting entity relationships. *In WWW'09:Proceedings of the 18th international conference on World wide web, pages 101-110, New York, NY, USA. ACM.*