Improving Semantic Role Labeling Using High-Level Classification in Complex Networks

Murillo G. Carneiro*, João L. G. Rosa†, Qiusheng Zheng‡, Xiaoming Liu‡ and Liang Zhao§
*Faculty of Computing, Federal University of Uberlândia, Uberlândia, Brazil 38400–902
Email: mgcarneiro@ufu.br
†Institute of Mathematical and Computer Sciences, University of São Paulo, São Carlos, Brazil 13566–590
‡School of Computer Science, Zhongyuan University of Technology, ZhengZhou, China 450007
§Department of Computing and Mathematics, University of São Paulo, Ribeirão Preto, Brazil 14040–901

***Abstract*—While traditional supervised learning methods perform classification based only on the physical features of the data (e.g. distribution, similarity or distance), the high-level classification is characterized by its ability to capture topological features of the input data by using complex network measures. Recent works have shown that a variety of patterns can be detected by combining both features of the data, although the physical features alone are unable to uncover them. In this article we investigate such a hybrid method for the Semantic Role Labeling (SRL) task, which consists of the identification and classification of arguments in a sentence with roles that indicate semantic relations between an event and its participants. Due to its potential to improve many other natural language processing tasks, such as information extraction and plagiarism detection to name a few, we consider the SRL task over a Brazilian Portuguese corpus named PropBank-br, which was built with texts from Brazilian newspapers. Such a corpus represents a challenging classification problem as it suffers with the scarcity of annotated data and very imbalanced distributions, like the majority of non-English corpus. Experiments were performed considering the argument classification task over the whole corpus and, specifically, over the most frequent verbs. Results in the verb-specific scenario revealed that the high-level system is able to obtain a considerable gain in terms of predictive performance, even over a state-of-the-art algorithm for SRL.**

***Index Terms*—Complex networks; High-level classification; Semantic role labeling; PropBank-br; Network-based learning**

I. Introduction

Semantic Role Labeling (SRL) is an important task of the Natural Language Processing (NLP) area. It consists of identifying and classifying the arguments of a predicate (often a verb) with semantic role labels that indicate meaningful relations among the arguments, e.g., who did what to whom, where, when and how [1], [2]. Information extraction [3], question answering [4], plagiarism detection [5] and machine translation [6] are some examples of NLP applications which performance can be improved by considering semantic roles.

Motivated by the SRL potential to improve a wide range of applications, massive lexical resources that include PropBank [7], FrameNet [8] and VerbNet [9] have been built recently to allow the development of efficient SRL systems. Under the PropBank annotation framework, there are two categories of labels, named the core and adjunct roles. The core roles, denoted by Arg0, Arg1, Arg2, etc., are verb-specific, which means their interpretation is specific to each predicate. By contrast, the interpretation of the adjunct roles is common across predicates, e.g., location, manner or time. Sentences 1-3 show a brief example of the SRL task, which includes argument identification in 2 and argument classification in 3. The former consists of the detection of groups of words that are semantic arguments, and the latter aims to provide the specific labels to such groups.

1. *Edison customers have received electric service since April 1985.*

2. [*Edison customers*$_{arg}$] *have received* [*electric service*$_{arg}$] [*since April 1985*$_{arg}$].

3. [*Edison customers*$_{Arg0}$] *have received* [*electric service*$_{Arg1}$] [*since April 1985*$_{ArgM-TMP}$].

From a supervised learning view, the SRL task is divided in two sub-tasks: argument identification and argument classification. In the argument identification, we have a binary classification problem, where each constituent should be predicted as an argument or not. On the other hand, a semantic role, chosen from a pre-defined list of roles, should be assigned for each argument in the classification task. In both sub-tasks, a wide range of features are usually extracted from the sentences, including part-of-speech tags, paths, and so on. Many supervised techniques have been employed in both stages [10]–[13], with logistic regression (LR) method being referred as a state-of-the-art SRL system because of its low computational cost and high predictive performance.

In this article, the SRL task is investigated under a hybrid classification model [14], which employs traditional and complex network-based techniques. The former focuses on the capture of physical features of the input data, being also named low-level technique. The latter considers also topological features of the data, such as the pattern formation, being referred as high-level technique. Examples of low-level classification include very well-known techniques, such as decision tree and neural networks; and examples of high-level classification include new concepts of data classification, e.g., the usage of complex network measures to check the pattern conformity of each test item in relation to the training data.

As the high-level classification using complex networks is a recently proposed method, there are also few related works in literature. In [14], data items of each class are mapped as a sub-network in which patterns are represented by a combination of the following network measures: assortativity, coefficient clustering and average degree. In [15], the high-level classification is provided using the same complex network measures as in [14], however, there is no low-level technique; [16] employs the same measures of [14] and introduces a parameter-free graph construction method to the hybrid model; [17] uses tourist walk measures to characterize the patterns of each sub-network; and [18] introduces a bio-inspired framework to optimize the network structure while optimizing a task-oriented quality function.

Despite the majority of the SRL researches has been conducted on the English language for reasons that include its great infrastructure in terms of lexical resources, much work remains to be done in non-English languages. In this article, we focus on the Brazilian Portuguese language, a relatively resource-poor language. To be specific, the SRL task here is investigated over the PropBank-br [19], which is a Brazilian corpus built with texts from newspapers that follows the PropBank annotation framework. The literature contains some investigations about SRL for Brazilian Portuguese language [12], [13], [20]–[22]. Specifically about the PropBank-br, a preliminary study using the corpus is presented in [12], where a general benchmark is provided for the task; in [13], a two-step convolutional neural network is proposed and its predictive performance is compared with a baseline and a logistic regression system; [21] compares the predictive performance of two SRL systems on revised and non-revised syntactic trees; and in [22], the propagation of semantic roles in the PropBank-br is investigated under a semi-supervised framework. In summary, the best results in the PropBank-br has been obtained using the logistic regression, a state-of-the-art algorithm for the task. In addition, the results obtained for the argument identification task in the Brazilian Portuguese corpus are close to those obtained for the English language. However, there is considerable space for improvement in the argument classification task when compared to the English PropBank.

One of the challenging features in the PropBank-br is the scarcity of annotated text. Its size is about one seventh of the PropBank [13]. The corpus represents a difficult scenario for machine learning techniques, which need to deal with very arbitrary and imbalanced distributions. In this way, the hypothesis investigated here suggests that the combination between traditional and complex-network based techniques can improve the predictive performance of the general system. Specifically, we propose in this article a high-level SRL system which is able to consider physical and topological features of the data. In order to evaluate the proposed system, experiments were conducted over the PropBank-br considering the argument classification task over the whole corpus and, specifically, over the most frequent verbs. In both cases, the results reveal a boost in terms of predictive performance, although it is considerably significant just for the verb-specific scenario.

The article is organized as follows. Section II shortly presents some relevant background about the PropBank-br corpus, complex networks and high-level classification; Section III describes the proposed high-level SRL system. Experimental results are presented in Section IV; and Section V concludes the article.

II. Relevant Background

Following we present an overview about the main topics covered in this article: the Brazilian Portuguese corpus, PropBank-br, is presented in Sub-section II-A; and a quick overview about high-level classification using complex networks is given in Sub-section II-B.

A. PropBank-br

The PropBank-br lexical resource was created based on the annotation of the Brazilian Portuguese section (CETENFolha) of the Bosque corpus from the Floresta Sintá(c)tica, which is a corpus annotated by the parser Palavras [23] and manually corrected by linguists. The version of the PropBank-Br used in this article employs the preprocessing steps performed in [12] and it is composed of 3,308 sentences, which results in 5,776 propositions for 1,023 target verbs. Note that a proposition is an instance of a predicate and its arguments, i.e., each predicate in a single sentence is equivalent to one proposition.

As the PropBank-br follows the PropBank annotation style, each verb is associated with core and adjunct roles. As described in sentences 1-3, the core roles (Arg0-Arg5) have specific interpretations to each predicate, while the interpretation of the adjunct roles (ArgM-) are common across predicates. Table I shows the number of arguments per semantic role in the whole PropBank-br. One can see the class distribution in the corpus is very imbalanced.

TABLE I
Distribution of the argument classes in the PropBank-br.

Class	#Arguments
Arg0	3,058
Arg1	5,148
Arg2	1,101
Arg3	113
Arg4	75
Arg5	1
ArgM-ADV (Adverbial)	369
ArgM-CAU (Cause)	156
ArgM-DIR (Directional)	15
ArgM-DIS (Discourse)	294
ArgM-EXT (Extent)	81
ArgM-LOC (Locative)	778
ArgM-MNR (Manner)	410
ArgM-NEG (Negation)	335
ArgM-PNC (Purpose)	171
ArgM-PRD (Predication)	192
ArgM-REC (Reciprocal)	65
ArgM-TMP (Temporal)	1,142
All	13,504

B. High-level Classification in Complex Networks

Networks (or graphs) are effective tools for modeling real systems, such as social and biological networks, Internet and so on [24]. They are called complex when they present non-trivial connection pattern. Gathering concepts from distinct areas, such as complex system, statistics and graph theory [25], complex networks have been applied to a great variety of problems from many areas of science [26]. Specifically about machine learning, graph-based techniques have been largely investigated to unsupervised and semi-supervised learning tasks related to clustering (or community detection), label propagation and dimensionality reduction [27]–[30]. On the other hand, there are also few studies on network-based classification (supervised learning) [14], [31]–[33].

Decision tree, neural networks, support vector machine, instance-based learning and other traditional techniques perform classification considering only the physical features of the data (e.g. distribution, similarity or distance). On the contrary, complex network measures are able to detect the pattern formation of the data by considering also its topological structure. In an attempt to gather low-level and high-level features, [14] proposed a hybrid model $\mathcal{M}$ that combines both techniques and is formally described by:

$$\mathcal{M}_y^{(c)} = (1-\lambda)\mathcal{C}_y^{(c)} + \lambda\mathcal{H}_y^{(c)}, \tag{1}$$

where $\mathcal{C}_y^{(c)} \in [0,1]$ and $\mathcal{H}_y^{(c)} \in [0,1]$ represents respectively the association produced by the low-level and high-level techniques when classifying a test item y to a given class c, and λ is a user-controllable variable which defines the contribution of each technique in the final decision.

About the high-level term $\mathcal{H}$, which compounds the variations calculated by a set of network measures, it is given by:

$$\mathcal{H}_y^{(c)} = \frac{\sum_{u=1}^{Z} \delta(u)[1 - f_y^{(c)}(u)]}{\sum_{g \in L}\sum_{u=1}^{Z} \delta(u)[1 - f_y^{(g)}(u)]}, \tag{2}$$

where $f_y^{(c)}(u)$ provides the variation of a network measure u when checking the pattern conformity of y into a class c, $\delta(u) \in [0,1]$, $\forall u \in \{1 \ldots, Z\}$ is a parameter which controls the contribution of each measure u and it is valid only if $\sum_{u=1}^{Z} \delta(u) = 1$. The denominator of (2) is for normalization.

The variation of each network measure, given by f, is calculated in function of two terms as follows:

$$f_y^{(c)}(u) = \Delta G_y^{(c)}(u)\ p^{(c)}, \tag{3}$$

where $\Delta G_y^{(c)}(u) \in [0,1]$ provides the variation captured by a measure u whenever a test item y is inserted into the sub-network of the class c, and $p^{(c)} \in [0,1]$ is the percentage of the data items that belongs to c.

In summary, $\mathcal{H}$ examines the variations of the complex network measures into distinct sub-networks (classes) before and after the insertion of each test item y. Consequently, the results returned by the high-level technique can be interpreted in function of the degree of variation, i.e., a large value of $\mathcal{H}$ means that y is compliant with the pattern of a given class, and vice versa. The high-level technique employed here is defined in function of three network measures: assortativity, clustering coefficient and average degree, which are described in the next section.

III. Model Description

In this article, we investigate the combination between traditional and complex network-based techniques to perform SRL over the PropBank-br, a Brazilian Portuguese corpus. To the best of our knowledge, complex network-based techniques have not been applied to the SRL task yet. Following, the proposed SRL system is described in details. Sub-section III-A presents a brief overview about the architecture of the system. Sub-section III-B denotes the training phase where each argument is mapped as a node in the underlying network; and Sub-section III-C describes the testing phase where arguments which semantic roles are unknown needs to be classified by the high-level technique.

A. Overview

The general architecture behind the high-level SRL system is exhibited by Fig. 1. Formally, given a set of training arguments $X_{train} = \{(x_1, l_1), \ldots, (x_n, l_n)\}$, each argument x_i is represented by its features extracted directly from the sentence, and $l_i \in \mathcal{L}$ represents the semantic role or label associated to that argument. The test phase aims to predict a set of arguments, denoted by $X_{test} = \{(x_{n+1}, ?), \ldots, (x_{n+m}, ?)\}$, where "?" means the semantic roles are unknown and needs to be predicted.

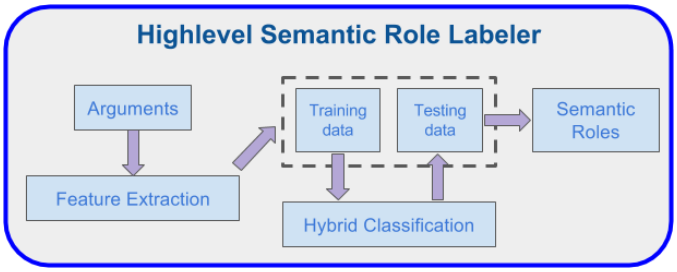

Fig. 1. Architecture of the high-level SRL system.

In order to induce a classifier from $x \rightarrow l$, the hybrid classification model combines low-level and high-level techniques. As the traditional one considers only the physical features of the data input, a myriad of algorithms can be used. By contrast, the high-level technique is characterized by its ability to detect the data pattern conformity using complex network measures. In the next sub-sections we describe in details the high-level SRL system.

B. Graph Construction (Training step)

In the training step, an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is constructed from the training data, where $\mathcal{V}$ denotes the set of nodes and $\mathcal{E}$ the set of edges. Each node $v_i \in \mathcal{V}$ represents an argument $x_i \in \mathbf{X}_{train}$ and $e_{ij} \in \mathcal{E}$ represents a link between node v_i and v_j, which usually indicates some similarity. In a

few words, $\mathcal{G}$ is crucial to the high-level classification because every information is extracted directly from it by using a set of network measures.

In order to deal with the particularities of the PropBank-br, such as sparsity and imbalanced classes, we employ a method based on [14] which constructs the graph from two widely used network formation methods, k-nearest neighbors graph (kNN) and ϵ-radius neighborhood (ϵN) [34]. In kNN graph, a vertex v_i is connected to an other vertex v_j if x_j is one of the k-nearest neighbors of i and if they have the same semantic role, i.e., $l_i = l_j$. In ϵN graph, a vertex v_i is connected to every vertex v_j if both belong to the same class and the distance (or other similarity criterion) between them, named here $S(x_i, x_j)$, is lesser than a pre-defined distance ϵ, i.e., $S(x_i, x_j) < \epsilon$. A graph in the proposed high-level SRL system is constructed using both concepts. Formally, the connections of a vertex x_i is defined as follows:

$$\mathcal{E}(x_i) = \begin{cases} \epsilon\text{N}(x_i), & \text{if } |\epsilon\text{N}(x_i)| > k, \\ k\text{NN}(x_i), & \text{otherwise.} \end{cases} \tag{4}$$

Briefly, this strategy uses kNN method to connect vertices in sparse regions and ϵN method in dense regions.

A general problem with ϵN is to choose an appropriate value for ϵ as distances or similarities can be very dependant in relation to the nature of the data. In view of the PropBank-br particularities, we designed an heuristic to better adjust ϵ. Thus, the neighborhood of a vertex x_i is formally given by $\{x_j,\ v_j \in \mathcal{V} : S(x_i, x_j) < \epsilon \cdot \bar{d}_k/2\}$, where $\bar{d}_k$ is the average distance between all argument and its k-th nearest neighbors.

C. High-level Classification (Testing step)

In (1), $\mathcal{M}$ denotes the final classification produced by our SRL system, which represents the convex linear combination between two classifiers in relation to an unlabeled argument $y \in \mathcal{X}_{test}$. The first classifier, denoted by $\mathcal{C}$, is a traditional (or low-level) technique which captures physical features of each class. The second, represented by $\mathcal{H}$, is a high-level technique which employs network measures to check whether a test data item conforms a data pattern of a class.

In the testing step of the high-level SRL system, the unlabeled arguments $y \in \mathcal{X}_{test}$ are presented to $\mathcal{H}$ one by one. Firstly, the test arguments are temporarily inserted into the network components using the ϵN method, however, without considering class labels as the semantic roles of the test arguments are unknown. Secondly, once y is inserted, its impact is calculated separately for each sub-network (class) using network measures as defined in Eqs. (2) and (3). Finally, the changes that occurred in the pattern formation of each network with the insertion of y are calculated and the test argument gets a high association value for that class in which structure is maintained or is barely modified.

Following, we present the network measures employed here, which are expected to capture complex topological features of the data.

1) Assortativity - $\Delta G_y^{(c)}(1)$*:* This measure quantifies the tendency of connections between nodes with similar degree [35]. It represents the Pearson correlation coefficient between pairs of connected nodes and may assume values between $[-1, 1]$, in which positive values indicate that pairs of vertices are more likely to have similar behavior, while negatives indicate greater probability of connected vertices having different behavior. Let U_c be the set of edges of a sub-network c and i_u, k_u the degree of the vertices i and k which form an edge $u \in U_c$. Formally, the assortativity r of the sub-network c is defined by:

$$r^{(c)} = \frac{L^{-1}\sum_{u\in U_c} i_u k_u - [L^{-1}\sum_{u\in U_c} \frac{1}{2}(i_u + k_u)]^2}{L^{-1}\sum_{u\in U_c} \frac{1}{2}(i_u^2 + k_u^2) - [L^{-1}\sum_{u\in U_c} \frac{1}{2}(i_u + k_u)]^2}. \tag{5}$$

The variation in terms of assortativity is given as follows:

$$\Delta G_y^{(c)}(1) = \frac{\mid r'^{(c)} - r^{(c)} \mid}{\sum\limits_{u\in U} \mid r'^{(u)} - r^{(u)} \mid}, \tag{6}$$

where r' denotes the assortativity calculated after the insertion of a test item y and the denominator is only for normalization.

2) Clustering Coefficient - $\Delta G_y^{(c)}(2)$*:* In many real systems, such as social networks, vertices tend to form cohesive groups. The clustering coefficient measures how close local nodes of a given node are to a complete graph (clique) [36]. Let $\mid e_{us} \mid$ be the number of connections shared by direct neighbors of node i, k_i the degree of node i and V_c the number of vertices of the sub-network c, the average clustering coefficient $CC \in [0, 1]$ is given by:

$$CC_i^{(c)} = \frac{\mid e_{us} \mid}{k_i(k_i - 1)}; \tag{7}$$

$$CC^{(c)} = \frac{1}{V_c}\sum_{i=1}^{V_c} CC_i^{(c)}, \tag{8}$$

The variation in terms of clustering coefficient is given by:

$$\Delta G_y^{(c)}(2) = \frac{\mid CC'^{(c)} - CC^{(c)} \mid}{\sum\limits_{u\in U} \mid CC'^{(u)} - CC^{(u)} \mid}. \tag{9}$$

3) Average Degree - $\Delta G_y^{(c)}(3)$*:* This measure quantifies statistically the relation between edges and vertices. Let $k_i^{(c)}$ be the degree of vertex i and V_c the number of vertices of the sub-network c, the average degree is defined by:

$$\langle k^{(c)} \rangle = \frac{1}{V_c}\sum_{i=1}^{V_c} k_i^{(c)}, \tag{10}$$

The variation in terms of average degree is given by:

$$\Delta G_y^{(c)}(3) = \frac{\mid \langle k'^{(c)} \rangle - \langle k^{(c)} \rangle \mid}{\sum\limits_{u\in \Gamma} \mid \langle k'^{(u)} \rangle - \langle k^{(u)} \rangle \mid}. \tag{11}$$

IV. Experimental Results

Following we present some results of the proposed SRL system on the PropBank-br corpus. This section is divided into three sub-sections. In Sub-section IV-A, a classifier is trained over the whole corpus, i.e., it considers all verbs and their arguments; in Sub-section IV-B, a classifier is trained over arguments of a specific verb; and Sub-section IV-C discusses the main results of both approaches.

In both approaches, classes in which the number of arguments is smaller than ten were removed as a preprocessing step. In addition, Table II presents the set of features extracted from the sentences, which comes from the literature on SRL [10]–[12]. In all simulations, the Euclidean distance is used as the distance measurement.

TABLE II
THE SET OF FEATURES EXTRACTED OF THE ARGUMENTS.

PredLemma+PhraseType	LeftHeadPostag	HeadLemma	FirstPostag
FirstForm+FirstPostag	PostagSequence	FirstLemma	RightHead
LastForm+LastPostag	VoicePosition	PredLemma	LeftHead
PredLemma+Path	TopSequence	RightPhrase	Head

The parameters of the high-level SRL system are defined as follows. The graph construction in the training step is optimized in function of $k \in \{3, 5\}$ for kNN method, $\epsilon \in \{1, 1.25, 1.5, 2\}$ for ϵN method, and $k = 10$ for $\bar{d}_k$ heuristic. For the insertion of each test argument y in the testing phase, the ϵN is optimized in function of $\epsilon \in \{1, 1.25, 1.5, 2\}$. The portion of each measure into the high-level technique, denoted by δ in (2), as well as the convex linear combination between the low-level and high-level classifications, denoted by λ in (1), are optimized using a particle swarm optimization (PSO) method, which swarm size and iteration number are one hundred.

The traditional classification techniques used in the experiments were CART decision tree, k-nearest neighbors (k-NN) and logistic regression (LR) which is a state-of-the-art technique for SRL. In relation to the parameters, CART does not need parameter selection; the k value of k-NN classifier is optimized considering $k \in \{1, 2, 3, \ldots, 30\}$; and the LR parameters are rigorously tuned considering a wide range of parameter configurations, which include $p = \{l1, l2\}$ for the penalization norm and $C = \{2^{-2}, 2^{-1}, \ldots, 2^{12}\}$ for the regularization strength. In the experiments, all parameters were tuned using the grid search algorithm.

A. SRL on the whole PropBank-br

The first experiment is conducted on the whole PropBank-br which means all verbs and arguments in the corpus are considered. Table III presents details about the data, such as number of arguments, features and classes. The high number of features emphasizes the sparsity of the PropBank-br as a consequence of the scarcity of annotated data. Based on previous works [13], the arguments are already divided into training and test data.

TABLE III
METADATA OF THE PROPBANK-BR ARGUMENTS CONSIDERING ALL VERBS IN THE CORPUS.

	#Training / #Testing	#Features	#Classes
PBbr	12967 / 536	39971	17

Table IV presents the predictive results of the traditional techniques isolated and combined with the high-level SRL system. In this table, each cell denotes the F_1-score. One can see the improvement achieved using the proposed method is relatively small.

TABLE IV
PREDICTIVE RESULTS (F_1) OBTAINED OVER PROPBANK-BR BY THE TRADITIONAL TECHNIQUES ($\mathcal{C}$) AND BY THE HIGH-LEVEL CLASSIFICATION ($\mathcal{M}$).

Algs.	CART	k-NN	LR
$\mathcal{C}$	80.0	76.3	83.3
$\mathcal{M}$	80.2	77.1	83.5

B. SRL on the most frequent PropBank-br verbs

The second experiment is conducted over specific verbs in the corpus. We selected the sentences with the most frequent predicates in the PropBank-br: "dar" (to give), "fazer" (to make) and "dizer" (to say). Table V shows the metadata of the three data sets generated: "PBbr-give", "PBbr-do" and "PBbr-say". In the experiments, the results of each technique are averaged over thirty runs using the stratified 10-fold cross-validation process.

TABLE V
METADATA OF THE PROPBANK-BR ARGUMENTS CONSIDERING SPECIFICALLY THE MOST FREQUENT VERBS IN THE CORPUS.

Datasets	#Arguments	#Features	#Classes
PBbr-do	148	1057	3
PBbr-give	397	2118	8
PBbr-say	506	2591	5

The predictive performance of the SRL systems over the PropBank-br verbs is registered in Table VI. In order to statistically analyze the improvements achieved by the high-level SRL system, we adopted the two samples t-test to compare two groups and determine whether their means differ. Using a confidence level of 95%, the null hypothesis is rejected in the most cases, which means the high-level SRL system is able to improve the predictive result of the traditional techniques, even over a state-of-the-art algorithm. In some cases, this improvement is around 3 and 4 points of F_1, which is a considerable boost since the traditional techniques were rigorously tuned.

C. Discussion

Despite the improvement of the SRL performance on the whole PropBank-br is small, the high-level SRL system is able

TABLE VI
PREDICTIVE RESULTS (F_1) OBTAINED BY TRADITIONAL TECHNIQUES AND THE HYBRID MODEL FOR HIGH-LEVEL CLASSIFICATION ($\mathcal{M}$) OVER THE MOST FREQUENT VERBS IN PROPBANK-BR.

Algs.	PBbr-do	PBbr-give	PBbr-say
CART	74.4 ± 1.6	86.7 ± 2.7	91.9 ± 0.8
$\mathcal{M}$	77.6 ± 1.8	87.4 ± 2.7	92.5 ± 0.8
k-NN	72.8 ± 1.7	85.6 ± 1.9	93.3 ± 0.7
$\mathcal{M}$	74.3 ± 1.5	89.4 ± 1.7	93.6 ± 0.8
LR	76.8 ± 1.7	88.4 ± 2.7	93.0 ± 0.8
$\mathcal{M}$	79.8 ± 1.5	89.0 ± 2.6	93.7 ± 0.7

to considerably boost the predictive results of the traditional techniques, including a state-of-the-art algorithm for SRL, when considering the verb-specific approach on the most frequent verbs in PropBank-br.

We believe this contrast between both cases can be explained through two points. The first one relates to the proper core roles (Arg0-Arg5 classes), which are verb-specific, i.e., each core role depends on the verb sense, not only of the extracted features. The second one is the big number of separated components which belong to the same class, but present distinct pattern formation in a local fashion. This may incorporate some noise into the measures variation. However, a properly designed network formation method which takes this into account during the graph construction should be able to alleviate the noise risks.

V. CONCLUSION

This article investigated the application of high-level classification for Brazilian Portuguese semantic role labeling using the PropBank-br. Such corpus suffers from scarcity of annotated data and very imbalanced classes. In the proposed SRL system, the semantic role of an unlabeled argument is predicted by combining the classifications produced by traditional and complex network-based techniques. In the experiments, the high-level classifier is combined with three traditional techniques, namely decision tree, k-nearest neighbors and logistic regression, which is a state-of-the-art algorithm for SRL. The techniques were evaluated over the whole PropBank-br and, specifically, over the most frequent verbs. Despite the improvement was small when considering the argument classification over the whole corpus, in the verb-specific scenario, the proposed system obtained a reasonable improvement in terms of predictive result, even over a state-of-the-art algorithm. Forthcoming works include a detailed study about the effects of graph construction in the SRL task.

ACKNOWLEDGMENT

M.G.C. and L.Z. are grateful to the São Paulo State Research Foundation-FAPESP by the financial support (grants numbers 2012/07926-3, 2011/50151-0 and 2013/07375-0). Authors also thank the support from Brazilian Coordination for the Improvement of Higher Education-CAPES, and Brazilian National Council for Scientific and Technological Development-CNPq.

REFERENCES

[1] C. J. Fillmore, "The case for case," in *Universals in Linguistic Theory*. Holt, Rinehart and Winston, 1968, pp. 1–88.

[2] M. Palmer, D. Gildea, and N. Xue, *Semantic Role Labeling*. Morgan & Claypool Publishers, 2010.

[3] K. Filippova, M. Surdeanu, M. Ciaramita, and H. Zaragoza, "Company-oriented extractive summarization of financial news," in *ACL European Chapter of the Association for Computational Linguistics*, 2009, pp. 246–254.

[4] R. D. Nielsen, J. Masanz, P. Ogren, W. Ward, J. H. Martin, G. Savova, and M. Palmer, "An architecture for complex clinical question answering," in *ACM International Health Informatics Symposium*, 2010, pp. 395–399.

[5] A. H. Osman, N. Salim, M. S. Binwahlan, R. Alteeb, and A. Abuobieda, "An improved plagiarism detection scheme based on semantic role labeling," *Applied Soft Computing*, vol. 12, no. 5, pp. 1493–1502, 2012.

[6] D. Wu and P. Fung, "Semantic roles for SMT: a hybrid two-pass model," in *ACL Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2009, pp. 13–16.

[7] M. Palmer, D. Gildea, and P. Kingsbury, "The proposition bank: An annotated corpus of semantic roles," *Computational Linguistics*, vol. 31, pp. 71–106, 2005.

[8] C. Fillmore, C. Johnson, and M. Petruck, "Background to framenet," *International Journal of Lexicography*, vol. 16, no. 3, pp. 235–250, 2003.

[9] K. Kipper, H. T. Dang, and M. Palmer, "Class-based construction of a verb lexicon," in *AAAI Conference On Artificial Intelligence*. Association for the Advancement of Artificial Intelligence, 2000, pp. 691–696.

[10] D. Gildea and D. Jurafsky, "Automatic labeling of semantic roles," *Computational Linguistics*, vol. 28, no. 3, pp. 245–288, 2002.

[11] S. S. Pradhan, W. Ward, and J. H. Martin, "Towards robust semantic role labeling," *Computational Linguistics*, vol. 34, no. 2, pp. 289–310, 2008.

[12] F. E. Alva-Manchego and J. L. G. Rosa, "Semantic role labeling for brazilian portuguese: A benchmark," in *Advances in Artificial Intelligence – IBERAMIA 2012*. Springer, 2012, pp. 481–490.

[13] E. R. Fonseca and J. L. G. Rosa, "A two-step convolutional neural network approach for semantic role labeling," in *IEEE International Joint Conference on Neural Networks*, 2013, pp. 2955–2961.

[14] T. C. Silva and L. Zhao, "Network-based high level data classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 6, pp. 954–970, 2012.

[15] M. G. Carneiro and L. Zhao, "High level classification totally based on complex networks," in *IEEE BRICS Congress on Computational Intelligence*, 2013, pp. 507–514.

[16] M. G. Carneiro, J. L. G. Rosa, A. A. Lopes, and L. Zhao, "Network-based data classification: combining k-associated optimal graphs and high-level prediction," *Journal of the Brazilian Computer Society*, vol. 20, no. 1, pp. 1–14, 2014.

[17] T. C. Silva and L. Zhao, "High-level pattern-based classification via tourist walks in networks," *Information Sciences*, vol. 294, pp. 109–126, 2015.

[18] M. G. Carneiro, L. Zhao, R. Cheng, and Y. Jin, "Network structural optimization based on swarm intelligence for highlevel classification," in *IEEE International Joint Conference on Neural Networks*, 2016, pp. 3737–3744.

[19] M. S. Duran and S. M. Aluísio, "Propbank-br: a brazilian treebank annotated with semantic role labels," in *International Conference on Language Resources and Evaluation*, 2012, pp. 1862–1867.

[20] E. Bick, "Automatic semantic role annotation for portuguese," in *Workshop on Information and Human Language Technology*, 2007, pp. 1713–1716.

[21] N. S. Hartmann, M. S. Duran, and S. M. Aluísio, "Automatic semantic role labeling on non-revised syntactic trees of journalistic texts," in *Computational Processing of the Portuguese Language*, 2016, pp. 202–212.

[22] M. G. Carneiro, L. Zhao, and J. L. G. Rosa, "Graph-based semi-supervised learning for semantic role diffusion," in *Symposium on Knowledge Discovery, Mining and Learning*, 2016, pp. 1–8.

[23] E. Bick, *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press, 2000.

[24] M. E. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003.
[25] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of Modern Physics*, vol. 74, no. 1, pp. 47–97, 2002.
[26] M. Newman, *Networks: An Introduction*. Oxford University Press, Inc., 2010.
[27] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3–5, pp. 75 – 174, 2010.
[28] O. Chapelle, B. Scholkopf, and A. Zien, *Semi-Supervised Learning*. MIT Press, 2006.
[29] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40 –51, 2007.
[30] T. C. Silva and L. Zhao, *Machine Learning in Complex Networks*. Springer, 2016.
[31] J. R. Bertini, L. Zhao, R. Motta, and A. A. Lopes, "A nonparametric classification method based on k-associated graphs," *Information Sciences*, vol. 181, no. 24, pp. 5435–5456, 2011.
[32] M. G. Carneiro, T. H. Cupertino, and L. Zhao, "K-associated optimal network for graph embedding dimensionality reduction," in *IEEE International Joint Conference on Neural Networks*, 2014, pp. 1660–1666.
[33] T. H. Cupertino, L. Zhao, and M. G. Carneiro, "Network-based supervised data classification by using an heuristic of ease of access," *Neurocomputing*, vol. 149, pp. 86–92, 2015.
[34] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, 2003.
[35] M. E. Newman, "Assortative mixing in networks," *Physical review letters*, vol. 89, no. 20, p. 208701, 2002.
[36] D. J. Watts and S. H. Strogatz, "Collective dynamics of small-world networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.