Semi-supervised Semantic Role Labeling for Brazilian Portuguese

Murillo G. Carneiro[1], Thiago H. Cupertino[2], Liang Zhao[3], João L. G. Rosa[3]

[1] Universidade Federal de Uberlândia, Brazil
mgcarneiro@ufu.br
[2] Secretaria da Fazenda do Estado de São Paulo, Brazil
thcupertino@fazenda.sp.gov.br
[3] Universidade de São Paulo, Brazil
zhao@usp.br, joaoluis@icmc.usp.br

Abstract. Semantic Role Labeling (SRL) is a natural language processing task that detects the arguments of predicates (usually verbs) and their semantic roles. Such roles characterize semantic relationships between an event and its participants, as who did what to whom, where, when and how, which is very useful to improve a wide range of tasks, such as information extraction and plagiarism detection to name a few. Commonly, a supervised classifier is trained over large English annotated resources in order to perform the prediction of unlabeled sentences. However, most part of non-English languages suffers from scarcity of annotated data, as the labeling process is expensive, time consuming and requires the efforts of human annotators. Although such limitation makes harder the training of supervised methods for those languages, it indicates an appropriate scenario to apply semi-supervised learning (SSL) methods, which are able to learn not only from labeled data, but also from the unlabeled ones. In this article, we investigate SSL methods in the classification of semantic roles for the Brazilian Portuguese, a relatively resource-poor language. Specifically, a representative set of SSL methods based on low density separation, label propagation and self-training are considered. Experiments on the PropBank-br, a Brazilian Portuguese corpus built with text from Brazilian newspapers, were performed varying the number of labeled arguments. Additionally, the SSL methods were compared against state-of-the-art SRL methods. The results demonstrated that self-training heuristic outperforms other SSL and supervised methods, even when the latter ones are trained on a large number of labeled arguments.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications; I.2.6 [**Artificial Intelligence**]: Learning; I.2.7 [**Artificial Intelligence**]: Natural Language Processing

Keywords: Brazilian Portuguese SRL, PropBank-br, Semantic Role Labeling, Semi-Supervised Learning

1. INTRODUCTION

Semantic roles is a natural language processing term which denotes the logical relations between a group of predicates (often a verb) and arguments [Fillmore 1968; Dowty 1991]. Semantic Role Labeling (SRL) is the task of automatically identifying and classifying the participants and properties associated with a given predicate using semantic roles [Màrquez et al. 2008; Palmer et al. 2010]. Machine translation [Wu and Fung 2009], plagiarism detection [Osman et al. 2012] and sentiment analysis [Mohammad et al. 2014] are examples of some applications which performance can be improved by including analysis of semantic roles.

Motivated by the SRL potential to improve a wide range of applications, massive lexical resources, such as PropBank [Palmer et al. 2005] and FrameNet [Fillmore et al. 2003] have been built to allow the development of efficient semantic role labelers for the English language. The common approach

Authors thank the financial support given by the São Paulo State Research Foundation - FAPESP (grants numbers 2012/07926-3, 2011/50151-0 and 2013/07375-0). Authors also acknowledge support from CAPES and CNPq.

is the training of supervised learning methods over large annotated resources in order to construct classifiers that perform well in the prediction of new sentences. The abundance of labeled data plays a key role in the success of such an approach. Thus, most SRL works have been conducted over the English language, while much remains to be done in other languages, which suffers from scarcity of annotated data. Although such limitation makes harder the training of supervised methods for those languages, it indicates an appropriate scenario to apply semi-supervised learning (SSL) methods.

SSL has recently attracted a considerable amount of research in machine learning [Chapelle et al. 2006; Zhu and Goldberg 2009]. While unlabeled data is far easier to obtain, the labeling process is often expensive, time consuming and requires the efforts of human annotators, who must often be quite skilled. Different from supervised methods, semi-supervised ones are able to learn not only from labeled data, but also from the unlabeled ones. In this article, we evaluate the usage of SSL methods in the classification of semantic roles for the Brazilian Portuguese, a relatively resource-poor language [Fonseca and Rosa 2013]. The SRL task here is investigated over the PropBank-br [Duran and Aluísio 2012], which is a corpus built with text from Brazilian newspapers that follows the PropBank style. The objective is to analyze the predictive performance of distinct categories of SSL methods in dealing with particular characteristics of the corpus, such as scarcity of labeled data and imbalanced classes.

As the semantic roles are verb-specific and in order to have a bigger number of unlabeled arguments, we conduct experiments using three of the most frequent verbs in the PropBank-br, “dar” (to give), “fazer” (to do) and “dizer” (to say). They are tested by SSL and supervised methods over distinct numbers of labeled arguments. To be specific, a representative set of SSL methods based on low density separation [Chapelle and Zien 2005], graph-based learning [Zhou et al. 2004] and self-training [Yarowsky 1995] is evaluated as well as compared against state-of-the-art SRL supervised methods.

The remainder of the article is organized as follows: Sect. 2 briefly discusses SRL lexical resources and the main related works; Sect. 3 describes the main SSL approaches considered in this work; Empirical results are presented in Sect. 4; and Sect. 5 concludes the article.

2. RELATED WORK

There has been a lot of research in semantic role labeling to detect events and activities happening in sentences and texts. Following we review related lexical resources and learning techniques.

2.1 Lexical Resources and the PropBank-br

VerbNet [Kipper et al. 2000], FrameNet [Fillmore et al. 2003] and PropBank [Palmer et al. 2005] are important lexical resources of the English language which have decisive contribution in the progress achieved by the semantic role labeling task in the last decades. The VerbNet groups verbs into syntactically and semantically similar classes by extending Levin classes. Based on an attempt to represent the frame semantics theory proposed by Fillmore [1968], the FrameNet assumes that a word evokes a frame of semantic knowledge related to the specific meaning it refers to. By contrast, the PropBank captures predicate-argument structure by annotating predicates and the semantic roles of their arguments. As it is a verb-specific approach, a fixed set of roles are specified for each verb and a different label is assigned to each role.

Under the PropBank annotation framework, which PropBank-br is based on, each predicate is associated with a set of core roles (named Arg0, Arg1, Arg2, and so on) which interpretation is specific to that predicate. Despite the same core roles are adopted to label the semantic roles of all verbs, such roles are verb-specific, which means arguments of different verbs that present the same core role can be very distinct in terms of semantic similarity. In addition, there is also a set of adjunct roles (named ArgM) which interpretation is common across predicates, e.g., location, manner or time. Table I presents some examples of core and adjunct roles in PropBank. In the case of core roles, one can see different semantic labels are assigned to each predicate.

Table I: Examples of core (Arg) and adjunct roles (ArgM) in PropBank.

Core Roles				Adjunct Roles	
Arg	Meaning			ArgM	Meaning
	to give	to do	to say		
Arg0	Giver	Agent	Sayer	ADV	Adverbial
Arg1	Thing given	Thing done	Utterance	CAU	Cause
Arg2	Entity given to	Benefactive	Hearer	DIR	Directional
Arg3		Instrumental	Attributive	DIS	Discourse
Arg4		Comitative		EXT	Extent
Arg5				LOC	Locative
				MNR	Manner
				NEG	Negation
				PNC	Purpose
				PRD	Predication
				REC	Reciprocal
				TMP	Temporal

The PropBank-br corpus 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 [Bick 2000] and manually corrected by linguists. The PropBank.Br version used in this article employs the preprocessing steps performed in Alva-Manchego and Rosa [2012b] 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. Following we present an example about the SRL task through the sentences 1, 2 and 3. In 2, the arguments are identified, and the argument classification is shown in 3. The former aims to identify groups of words in a sentence that represent semantic arguments, and the latter aims to assign specific labels to the identified groups. In both sub-tasks, a wide range of features are usually extracted from the corpus, including part-of-speech tags, paths, and so on. At the end, in order to obtain the semantic meaning of each core role, the frameset information can be obtained through the verb sense[1]. In sentence 3, Arg0 e Arg1 denote the Receiver and the Thing gotten roles, respectively.

1. *Ele receberá o valor à vista após 30 dias. /*He will receive the value in cash after 30 days.*/*

2. [Ele$_{arg}$] receberá [o valor à vista$_{arg}$] [após 30 dias$_{arg}$] .

3. [Ele$_{Arg0}$] receberá [o valor à vista$_{Arg1}$] [após 30 dias$_{TMP}$] .

Table II shows sentence 1 in PropBank-br. One can observe each word in column "Form" has an identification number "ID" as well as its lemma, part-of-speech tag (GPOS) and morphological features. The table also exhibits the full syntactic tree and the predicate of the sentence. The last column denotes the semantic roles assigned to each argument. Also, regarding sentence 1, Fig. 1 illustrates its syntactic tree. In the figure, "NP", "VP" and "PP" refer to the noun phrase, verb phrase and prepositional phrase, respectively.

In contrast with PropBank (English), PropBank-br (Brazilian Portuguese) suffers from scarcity of annotated data and imbalanced classes. For example, Arg5 class has only one argument in the whole corpus. If such characteristics make the Brazilian Portuguese SRL a yet harder task, they also represent a motivating scenario for the development of efficient learning solutions.

[1] https://verbs.colorado.edu/propbank/framesets-english/

Table II: Example of annotated sentence in PropBank-br.

ID	Form	Lemma	GPOS	Feat	Clause	FClause	Synt	Pred	Arg
1	Ele	ele	PRON-PERS	M|3S|NOM	(S*	(FCL*	(FCL(NP*)	-	(A0*)
2	receberá	receber	V-FIN	FUT|3S|IND	*	*	(VP*)	receber	(V*)
3	o	o	ART	M|S	*	*	(NP*	-	(A1*
4	valor	valor	N	M|S	*	*	*	-	*
5	a	a	PRP	-	*	*	(PP*	-	*
6	a	o	ART	F|S	*	*	(NP*	-	*
7	vista	vista	N	F|S	*	*	*)))	-	*)
8	após	após	PRP	-	*	*	(PP*	-	(TMP*
9	30	30	NUM	M|P	*	*	(NP*	-	*
10	dias	dia	N	M|P	*	*	*))	-	*)
11	.	-	PU	-	*)	*)	*)	-	*

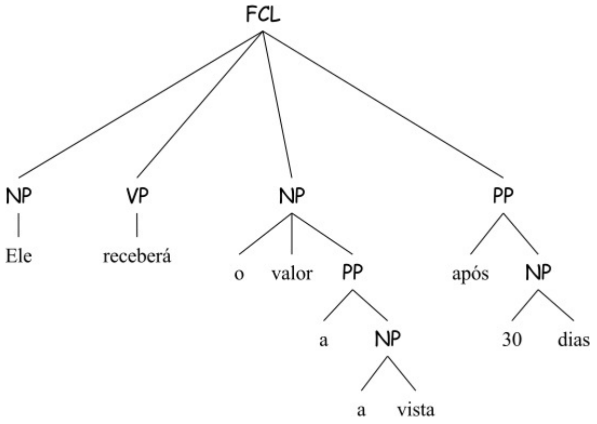


Fig. 1: Syntactic tree of the annotated sentence shown in Table II.

2.2 SRL Learning Systems

Most part of the techniques developed for SRL lies in the supervised category, where large (and expensive) amount of human annotated data are used to train classifiers. The seminal work proposed by Gildea and Jurafsky [2002] modeled the SRL task into a classification framework. Firstly, a statistical technique is trained on the annotated data, then the learned function is employed to identify and classify the arguments of predicates in new sentences. Most SRL systems are based on such a methodology: Pradhan et al. [2005] investigated the SRL task over different domains by designing new features and using a support vector machine technique; Punyakanok et al. [2008] enforced global consistency by combining statistical learning and integer linear programming; Collobert et al. [2011] presented the first neural network system for SRL, which employs a feed-forward network and uses a convolution function to model the context window. Regarding more recent SRL systems, neural networks have been largely investigated [FitzGerald et al. 2015; Roth and Lapata 2016; He et al. 2017] as well as global inference methods [Täckström et al. 2015].

In the semi-supervised learning category, SRL works include the investigation of semi-supervised and "semi-unsupervised" approaches. For instance, bootstrapping approaches such as self-training and co-training methods were proposed in He and Gildea [2006] and Zadeh Kaljahi [2010] while a semi-unsupervised approach which employs a small number of labeled data to build an informed prior distribution over an unsupervised method was presented in Titov and Klementiev [2012b]. In addition, some approaches increase the manually annotated instances with unlabeled instances which roles are inferred through projection [Fürstenau and Lapata 2012] or by integrating several lexical resources [Hartmann et al. 2016]. By exploiting word representations, some researches proposed the reduction of lexical features sparsity [Deschacht and Moens 2009], while others dealt with large corpus by combining semi-supervised methods and deep learning [Weston et al. 2012].

SRL literature contains also a few works regarding the unsupervised learning category. Grenager and Manning [2006] proposed a generative model which considers linguistic priors on syntactic-semantic interface. Lang and Lapata [2011a] and Lang and Lapata [2011b], which refer to the task as semantic role induction, presented a similarity-driven approach in order to cluster related argument instances of each verb; Titov and Klementiev [2012a] followed a similar approach, but using multilingual data to improve the role induction.

Although Brazilian Portuguese is a resource-poor language, supervised learning systems have been explored using PropBank-br. A preliminary study using the corpus was presented in Alva-Manchego and Rosa [2012b], where a general benchmark was provided for the task based on the state-of-the-art of English SRL systems. Fonseca and Rosa [2013] developed a two-step convolutional neural network based in Collobert et al. [2011]; despite outperformed by the logistic regression method, the technique has important features, such as independence of syntactic parser; authors also believed that the predictive performance could be improved if more annotated data was provided in PropBank-br. Hartmann et al. [2016] evaluated the performance of both SRL techniques (logistic regression and neural networks) on revised and non-revised syntactic trees; in common both techniques achieved better performance after training over non-revised syntactic trees. The high-level classification has also been designed for the Brazilian Portuguese SRL task. In such a technique, the classification produced by a traditional technique (named low-level) is combined with the classification provided by complex network measures (named high-level). By taking the logistic regression as the low-level technique and a set of network measures as the high-level, Carneiro et al. [2017] evaluated the designed SRL system in two scenarios: the whole PropBank-br and the verb-specific case. In addition, Carneiro et al. [2016] presented a structural framework designed to optimize the graph connections in order to achieve better predictive performance; as the results, both techniques were able to boost the predictive performance of the logistic regression technique in the verb-specific scenario.

By contrast, to the best of our knowledge, semi-supervised learning is also a barely explored topic for Brazilian Portuguese SRL. Alva-Manchego and Rosa [2012a] represents the first attempt to design semi-supervised learning in the PropBank-br. In that article, authors discuss about data preparation, feature extraction, methodology and a self-training system. Other related article is presented in Carneiro et al. [2016], which investigated the propagation of semantic roles under a graph-based semi-supervised framework; the framework achieves reasonable performance even in the simplest case when there is only one labeled argument per class.

Given the relevance of the SRL task in the semi-supervised context, the few works dealing with and the scarcity of annotated data for Brazilian Portuguese, this article extends SSL related works by investigating not only graph-based SSL methods, but also other categories of SSL, such as low-density separation and self-training. By exploring such methods, it is expected the designing of alternative systems able to learn also from the unlabeled data as data annotation is expensive and time-consuming.

3. MODELS DESCRIPTION

In this section we present an overview about the SSL methods investigated. Sub-sect. 3.1 defines the problem addressed here; and Sub-sects 3.2, 3.3 and 3.4 describe respectively the graph-based approach, the low density separation technique and the self-training heuristic.

3.1 Problem Definition

Given a set of arguments $\mathbf{X} = \{a_1, \ldots, a_l, a_{l+1}, \ldots, a_n\}$ and a set of semantic roles $\mathcal{L} = 1, \ldots, c$, the first l arguments are labeled $\{y_1, \ldots, y_l\} \in \mathcal{L}$ and the remaining arguments $(u = n - l)$ are unlabeled.

Tipically, $l \ll u$, i.e., the great majority of arguments does not possess labels. The goal is to predict the semantic roles of the unlabeled arguments.

3.2 Local and Global Consistency (LGC)

Based on the results presented in Carneiro et al. [2016], local and global consistency (LGC) has been selected to represent the SSL graph-based approaches. It is a simple label propagation algorithm in which every argument iteratively spread its label information to its neighbors until a global stable state is achieved [Zhou et al. 2004]. Following we summarize the technique used in our investigation:

(1) Build an undirected graph $\mathcal{G}$ from $\mathbf{X}$ using symmetric kNN;
(2) Generate a weighted matrix $\mathbf{W}$ from $\mathcal{G}$ and a Gaussian kernel $\mathcal{K}$;
(3) Compute the graph Laplacian matrix $\mathbf{L}$ from $\mathbf{W}$;
(4) Label the unlabeled points from the output matrix $\mathbf{F}$ obtained by using LGC.

In order to spread label information, the arguments must be linked. Intuitively, a fully connected weighted matrix (a graph) can be obtained from the affinity matrix $\mathbf{W}$, however, the usage of a graph construction heuristic, such as k-nearest neighbors, promotes sparsification in $\mathbf{W}$, which considerably reduces the computational complexity and usually improves the performance. Thus, consider an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each node $v_i \in V$ represents an argument $a_i \in \mathbf{X}$. Let $\mathbf{S}$ be a distance matrix in which $\mathbf{S}_{ij} = \delta(a_i, a_j)$ and $k\mathrm{NN}_i$ be the set of k nearest neighbors of a_i, the adjacency matrix $\mathbf{A}$ of a kNN-graph is obtained as follows:

$$\mathbf{A}_{ij} = \begin{cases} 1, & \text{if } a_j \in k\mathrm{NN}_i \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Consequently, the symmetric kNN is obtained by: $\mathbf{A} = max(\mathbf{A}, \mathbf{A}^T)$.

Now it is possible to obtain a sparse weighted matrix $W \subset \mathbb{R}^{n \times n}$:

$$\mathbf{W}_{ij} = \mathbf{A}_{ij}\mathcal{K}(a_i, a_j), \tag{2}$$

where $\mathcal{K}(a_i, a_j)$ denotes the Gaussian kernel.

In order to ensure the convergence of the LGC method, W is normalized symmetrically:

$$\mathbf{L} = \mathbf{I}_n - \mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}, \tag{3}$$

where $\mathbf{I}_n$ is the identity matrix and $\mathbf{D}$ is the diagonal matrix which contains the vertices degree.

Given a graph Laplacian $\mathbf{L}$ and let $\mathbf{Y} \in \mathbb{B}^{n \times c}$ be a label matrix in which $\mathbf{Y}_{ij} = 1$ if and only if a_i has label $y_i = j$, LGC generate the output matrix $\mathbf{F}$ by iterating the following system:

$$\mathbf{F}(t+1) = \alpha\mathbf{L}\mathbf{F}(t) + (1-\alpha)\mathbf{Y} \ , \tag{4}$$

where α defines the relative amount of the information from its neighbors and its initial label.

3.3 Low Density Separation (LDS)

Low density separation approaches assume the separation among classes lies in low density regions. Consequently the decision boundary for classification go through such regions. In this SSL category, we consider the low density separation (LDS) method. LDS combines graph-based distances and transductive SVM (TSVM) in order to enforce the cluster assumption [Chapelle and Zien 2005; Chapelle et al. 2006]. While many graph-based SSL algorithms employ the graph to enforce the smoothness of the solution, LDS employs the graph to derive pairwise similarities, in which points from the same cluster are shrunk, otherwise dominated by the inter-cluster distance. The main steps of LDS are summarized as follows:

(1) Compute ρ-distances.
   (a) Build a fully connected graph $\mathcal{G}$ with edge lengths $w_{ij} = \exp(\rho d(i,j)) - 1$.
   (b) Compute the shortest path lengths d_{SP} for all edges in $\mathcal{G}$.
   (c) Obtain matrix D of squared ρ-distances.

$$D_{ij} = \left(\frac{1}{\rho} \log(1 + d_{SP}(i,j)) \right)^2 . \tag{5}$$

(2) Apply multidimensional scaling to find a Euclidean embedding of D.

$$U \Lambda U^\top = -HDH; \text{ where } H_{ij} = \delta_{ij} - \frac{1}{n+m}. \tag{6}$$

(3) Obtain the new representation of a_i, given by: $\bar{a}_{ik} = U_{ik}\sqrt{\lambda_k}$, $1 \leq k \leq p$; where p means the number of eigenvectors taken.
(4) Train TSVM.

3.4 Self-Training (S.T.)

Self-training is an SSL wrapper-algorithm, which uses a supervised learning method as a base classifier [Chapelle et al. 2006]. It is divided into two stages: learning and prediction. At the learning stage, it trains over the labeled data. At the prediction, it classifies the unlabeled data, although only those predictions that reach a confidence threshold gets a label. In summary, the base classifier is iteratively retrained using its own predictions as additional labeled points. Let L be the set of labeled data and U the set of unlabeled, the main steps of S.T. are summarized as follows:

(1) Repeat until $U = \emptyset$:
(2) Train a base classifier $\mathcal{C}$ over the labeled data L.
(3) Classify the unlabeled data U using $\mathcal{C}$.
(4) Labels the unlabeled arguments U' that satisfy a confidence threshold p.
(5) If $U' = \emptyset$, decrease the confidence threshold.
(6) Else, $L \leftarrow L \cup U'$; $U \leftarrow U - U'$.

In this article, the logistic regression (LR) technique has been selected as the base classifier of S.T.. LR is an state-of-the-art technique for supervised semantic role labeling, which has also been included in our experiments to better evaluate the predictive performance of the SSL methods.

4. RESULTS

This section provides experimental results for Brazilian Portuguese SRL using the techniques described before. Each simulation was performed in a transductive setting using configurations with different constraints about the numbers of labeled arguments. For each data set configuration, we perform 30 trials randomly varying the labeled data. The error rate averaged over the trials is used to evaluate the quality of the semantic role diffusion process.

In order to obtain the initial labeled arguments, two approaches are considered in the study. The *argument-based labeling* comprehends the same approach used in Carneiro et al. [2016], which is directly based on the arguments. In such approach, the number of labeled arguments l^{arg} is calculated from the number of argument classes c, i.e., $l^{arg} = \eta c$ with $\eta \in \{1, 2, 3\}$ in this article. Thus, in each trial, we randomly sample labeled data from the entire data set ensuring there is at least one labeled argument per class. By contrast, the *sentence-based approach* labels all arguments of a randomly selected sentence in the corpus. Here, the number of labeled arguments l^{sent} is also dependant of the number of argument classes c, but η works as a constraint to establish the minimum number of

Table III: Summary of the features used to extract the attributes of the arguments in PropBank-br.

Features	Description
FirstPostag	the part-of-speech tag of the first word of the phrase
FirstPostag+FirstForm	the first word of the phrase + FirstPostag
FirstLemma	the lemma of the first word of the phrase
HeadLemma	the head lemma of the phrase
HeadWord	the syntactic head of the phrase
LastPostag+LastForm	the last word of the phrase + its part-of-speech tag
LeftHead	the head word of the left sibling
LeftHeadPostag	the part-of-speech tag of the LeftHead
PostagSequence	the sequence containing the part-of-speech tag of the words that compose the phrase
PredLemma	the predicate lemma
PredLemma+Path	PredLemma + the syntactic path from the parse constituent to the predicate
PredLemma+PhraseType	PredLemma + the syntactic category of the phrase
RightHead	the head word of the right sibling
RightPhrase	the phrase label of the right sibling
TopSequence	the right-hand side of the rule expanding the constituent node
Voice+Position	the voice is active or passive + the phrase position is before or after the predicate

labeled arguments per class, i.e., $l^{sent} \geq l^{arg}$. Therefore, in each trial we randomly sample labeled arguments from the sentences of the corpus ensuring there is at least η labeled arguments per class. The main difference between both approaches lies on the constraints adopted in order to ensure that there is at least one or η labeled argument per class. While the argument-based approach seems more suitable for machine learning purposes such as active learning, the sentence-based one follows most part of works from the natural language processing area.

Next sub-sections are divided as follows: Sub-sect. 4.1 describes the features used to extract attributes from sentences as well as data preprocessing steps; Sub-sect. 4.2 shows the parameters of the techniques under study, including the supervised ones; Sub-sect. 4.3 present discussions about results obtained from initial labeled data provided by the argument-based and sentence-based approaches; Sub-sect. 4.4 compares the SSL method with better performance and the state-of-the-art supervised method through a wide variation in the numbers of labeled arguments; and Sub-sect. 4.5 presents the final remarks about the experiments.

4.1 Features and Data Preprocessing

In this article we present results for SRL task using the PropBank.br. For the experiment, we select all sentences related to the predicate "to give", "to do" and "to say", which are among the most frequent verbs in the corpus, and extract the attributes of each argument by using a set of features from the literature [Gildea and Jurafsky 2002; Pradhan et al. 2008; Alva-Manchego and Rosa 2012b]. Table III summarizes the employed features.

In the preprocessing step, argument classes smaller than ten instances were excluded. As we have a very large imbalance among the semantic role labels, related to the scarcity of annotated data in PropBank-br, this treatment is performed in order to establish some smooth reduction of class imbalance. Table IV presents a brief description of the data sets obtained, named here "PBbr-give", "PBbr-do" and "PBbr-say", in terms of the number of sentences, arguments and classes. As a data preparation step, each instance attribute vector was normalized to have a magnitude of one and the Euclidean distance was used in all simulations as the distance measurement. In order to avoid the dimensionality curse problem, we run principal component analysis (PCA) to reduce the dimensionality of the data to a hundred features.

Table IV: Brief description of the PropBank-br data sets in terms of number of sentences, arguments, attributes and classes.

Name	#Sentences	#Arguments	#Features	#Classes
PBbr-give	80	148	1057	3
PBbr-do	181	397	2118	8
PBbr-say	254	506	2591	5

4.2 Parameters and Baselines

The following parameters are employed in the computer simulations presented here.

1NN As in related works in literature, 1NN classifier is used for comparison purposes as a baseline.

LGC: the SkNN network formation method is optimized over the set $k \in \{1, 2, \ldots, 60\}$; the kernel bandwidth σ is defined as suggested in Jebara et al. [2009] by $\sigma = \bar{d}_k/3$ where $\bar{d}_k$ is the average distance between each sample and its k-th nearest neighbor; and the parameter α is chosen at range $\{0.001, 0.01, 0.05, 0.1, 0.2, 0.5, 0.8, 1.0\}$.

LDS: the parameter to compute the ρ-distances is optimized over the set $\rho \in \{0, 2^0, 2^1, \ldots, 2^4\}$, and the soft margin parameter C is chosen at range $\{10^{-1}, 10^1, \ldots, 10^3\}$.

S.T.: the confidence threshold is defined as $p = 0.99$, decaying by $e = 0.01$. The base classifier parameters are the same as LR.

LR: the state-of-the-art method for supervised SRL is optimized over the set $C = \{2^0, 2^1, \ldots, 2^8\}$, with penalization defined by $l1$-norm.

4.3 Results on Argument and Sentence-Based Labeling Approaches

Following we present results using the argument-based approach to provide the initial labeled arguments. There are three groups of simulations which corresponds to each value taken by $\eta \in \{1, 2, 3\}$. Table V presents the results obtained by each group of simulation. At the table, each cell corresponds to the average error and standard deviation obtained over thirty runs. After the name of each data set is also shown the total number of initial labeled arguments for $\eta = 1$. Indeed, most of our attention is on $\eta = 1$ scenario, which represents a very challenging problem given the lower number of labeled arguments per verb. In terms of results, LGC performs well on such scenario as also pointed in Carneiro et al. [2016], however, the S.T. performance is outstanding. It largely outperforms all other techniques, including both SSL and supervised ones. Despite the poor performance in the first scenario, the LR is able to obtain competitive and also better results than LGC and LDS as the number of labeled arguments increases, e.g. $\eta = 3$.

In order to better analyze the results, a statistical test that compares each two methods over multiple data sets has been adopted, the Wilcoxon Signed Ranks. The test is a non-parametric alternative to the paired t-test when its assumptions, such as normal distribution, can not be assured. Basically, the Wilcoxon test ranks the differences in performances of two techniques for each data set, ignoring the signs, and compares the ranks for the positive and the negative differences [Demšar 2006]. A confidence level of 95% has been considered in the statistical tests presented here. The tests show that S.T. has better predictive performance than other techniques under comparison. They also revealed that LGC and LR perform statistically better than LDS and that LGC also outperforms the baseline.

Now we move on to analyze the performance of the techniques over other labeling approach in which all arguments of randomly selected sentences are labeled. Such approach is more realistic as it considers the experts annotate the argument labels after receiving a sentence. In addition, the number of labeled data provided by such approach is bigger than the argument-based one as $\eta \in \{1, 2, 3\}$ denotes the minimum number of labeled arguments per class here. Table VI shows the results obtained by the

Table V: Comparative results in terms of average error rates and standard deviations (over thirty runs) using $\eta = \{1, 2, 3\}$. The argument-based approach is employed to generate the initial labeled data. Best result for each data set is in bold face.

η	Algs.	Argument-based approach		
		PBbr-give (3)	PBbr-do (8)	PBbr-say (5)
	1NN	45.31 (11.37)	63.52 (9.80)	35.41 (14.97)
	LGC	37.65 (11.20)	47.02 (2.90)	30.31 (19.36)
1	LDS	41.68 (13.05)	65.45 (8.76)	42.50 (11.73)
	S.T.	**19.36 (6.17)**	**30.85 (2.01)**	**16.24 (6.62)**
	LR	42.71 (12.69)	55.78 (0.00)	41.35 (16.33)
	1NN	32.98 (8.96)	47.23 (3.77)	22.65 (13.15)
	LGC	30.96 (9.43)	43.57 (4.18)	22.91 (10.83)
2	LDS	30.47 (15.12)	51.57 (9.89)	28.84 (10.41)
	S.T.	**13.92 (2.07)**	**28.77 (1.79)**	**11.54 (4.77)**
	LR	30.40 (13.21)	42.03 (5.74)	20.57 (9.29)
	1NN	33.76 (9.51)	43.22 (6.26)	15.48 (6.57)
	LGC	28.94 (8.36)	41.38 (5.07)	16.71 (8.69)
3	LDS	26.26 (10.69)	42.02 (10.57)	19.63 (8.50)
	S.T.	**13.67 (2.36)**	**27.78 (1.67)**	**9.15 (3.41)**
	LR	26.88 (9.92)	38.80 (4.02)	15.09 (6.49)

techniques. Again, S.T. presented very good predictive performance, also better than using the labeled data provided by the argument-based approach. The same comment is also true for 1NN and LR, and also for most cases of LGC and LDS, although their improvement in terms of performance is smaller than the supervised ones. Furthermore, similar characteristics presented in Tab. V can be seen in Tab. VI, such as the better performance of LGC over LDS, 1NN and LR when $\eta = 1$, and the considerable improvement of LR as the amount of labeled arguments increase. However, S.T. largely outperforms those techniques again. Statistical tests confirm such method is superior to others, and they also show that LDS is outperformed by LGC and LR while LGC also exceeded 1NN.

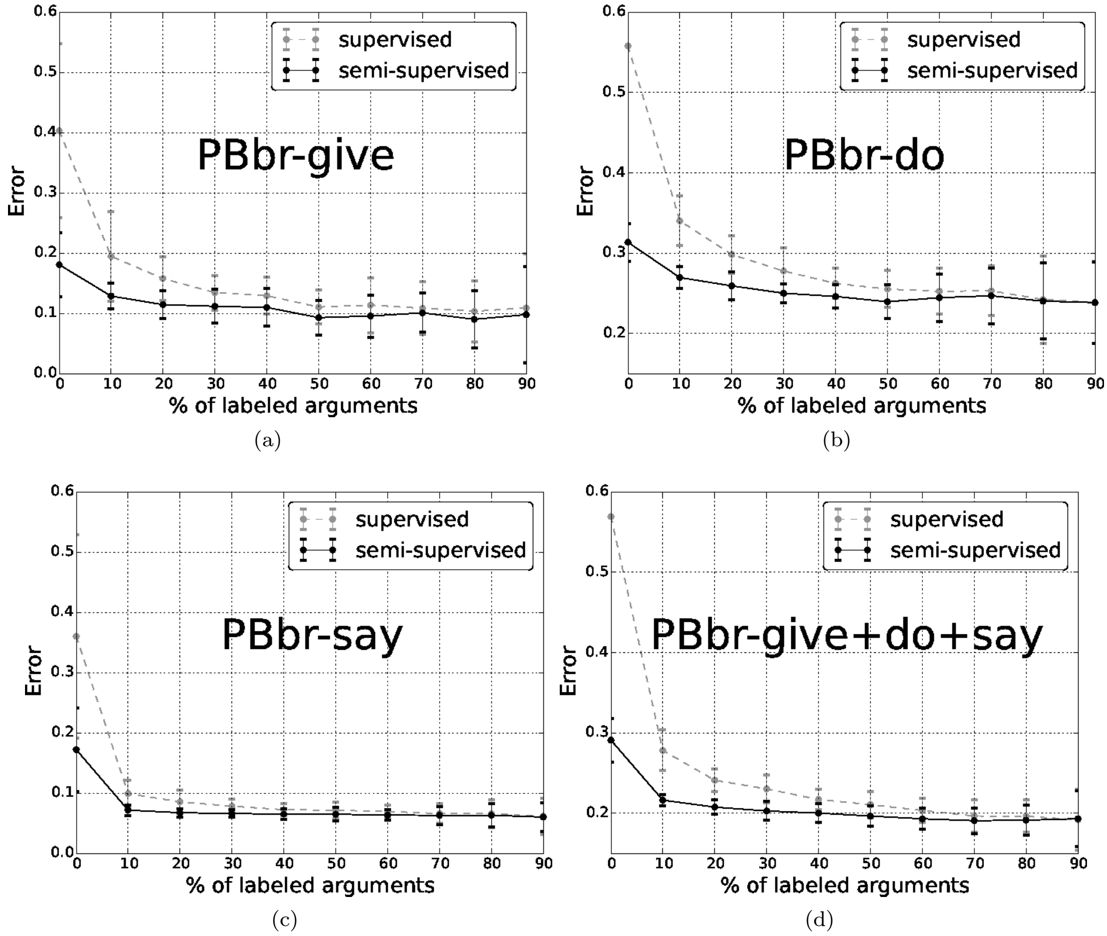
Table VI: Comparative results in terms of average error rates and standard deviations (over thirty runs) using $\eta = \{1, 2, 3\}$. The sentence-based approach is employed to generate the initial labeled data. Best result for each data set is in bold face.

η	Algs.	Sentence-based approach		
		PBbr-give ($\sim$ 4)	PBbr-do($\sim$ 13)	PBbr-say($\sim$ 7)
	1NN	37.66 (8.99)	51.58 (10.36)	29.59 (12.87)
	LGC	36.83 (11.72)	48.18 (8.78)	27.03 (13.62)
1	LDS	39.13 (12.90)	60.22 (11.98)	38.27 (15.22)
	S.T.	**15.82 (3.20)**	**29.25 (1.48)**	**15.14 (7.17)**
	LR	40.79 (10.71)	48.37 (10.84)	31.67 (14.13)
	1NN	30.91 (10.81)	45.76 (5.84)	19.83 (8.72)
2	LGC	28.13 (8.17)	43.29 (5.60)	19.58 (8.94)
	LDS	26.35 (9.91)	51.22 (10.58)	30.55 (8.66)
	S.T.	**12.24 (1.40)**	**27.70 (1.65)**	**9.00 (3.02)**
	LR	25.88 (11.10)	41.49 (6.82)	19.20 (10.15)
	1NN	28.91 (8.66)	42.08 (4.28)	13.87 (6.99)
	LGC	24.76 (7.10)	40.20 (5.16)	14.73 (7.66)
3	LDS	23.59 (8.99)	44.98 (9.08)	23.91 (11.50)
	S.T.	**11.83 (2.11)**	**26.64 (1.61)**	**7.86 (2.21)**
	LR	22.06 (8.01)	36.44 (4.40)	12.78 (5.44)

4.4 Semi-supervised versus supervised learning

The results presented before demonstrate that self-training (equipped with logistic regression as a base classifier) has good performance for the Brazilian Portuguese SRL task. However, an interesting question is how such semi-supervised method can evolve as more examples are labeled? Is its predictive performance also better than a logistic regression method trained over a considerable amount of labeled arguments? In this subsection, we investigate these questions by varying the number of labeled data through a set of ten experiments containing $l = \{0\%, 10\%, 20\%, \ldots, 90\%\}$ of labeled arguments, respectively. For the sake of clarity, $l = 0\%$ means an approximation for the first experiment, as it considers only one labeled argument per class. The initial labeled data are obtained using the argument-based approach.

Figure 2 exhibits the average error rates and standard deviations over thirty runs for each experiment. In the figure, PBbr-give+do+say refers to the inclusion of the arguments from the three verbs in a unique data set, semi-supervised to the self-training heuristic, and supervised to the LR method. One can see in the figure that the semi-supervised method outperforms the supervised one in all data sets under study. The lower the percentage of labeled arguments, the higher is the predictive difference between both methods. By taking at most 10% of labeled arguments, the semi-supervised approach is able to provide competitive results against the supervised one trained over at least 50% of labeled



(a)

(b)

(c)

(d)

Fig. 2: Comparison of predictive performance between semi-supervised (self-training equipped with logistic regression as base classifier) and supervised (logistic regression) methods in function of the number of labeled arguments.

Table VII: Number of arguments and classes before and after the preprocessing.

Name	#Orig.Args	#Preproc.Args	#Orig.Class	#Prepoc.Class
PBbr-give	187	148	12	3
PBbr-do	428	397	14	8
PBbr-say	547	506	13	5
PBbr-give+do+say	1162	1154	14	12

arguments in general, which is a really considerable amount. Considering also that the PropBank-br has more than one thousand verbs, such difference may represent significant time and money.

4.5 Final Remarks

In relation to our data preprocessing step, Table VII shows the number of arguments and classes before and after the preprocessing step. One can see in the table that despite the high number of semantic role labels in PBbr-give (12), most of its classes are removed because they do not have a representative number of arguments. By contrast, PBbr-give+do+say presents a high number of classes and a considerable number of arguments for most classes. Intuitively, one can assume that PBbr-give, PBbr-do and PBbr-say are easier to classify than PBbr-give+do+say as they have some labels removed. However, that is not the case. One can observe in Fig. 2 that our preprocessing step is smooth, i.e., it does not cause large variations in terms of predictive performance. This can be explained by taking into account that most of the removed classes are not core roles, but adjunct roles, which interpretation is common across verbs.

About the predictive performance, although self-training is a simple semi-supervised heuristics, the experimental results presented in this article have shown that such a method is very promising for Brazilian Portuguese semantic role labeling. The usage of logistic regression as a base classifier in order to learn the unlabeled data iteratively from a very few labeled data allows good predictive performance. In addition, common problems related to self-training, such as the inclusion of errors by labeling unlabeled data wrongly, do not affect too much the learning process as we can see by examining the execution step by step. Likely reasons are the nature of the data which is too sparse and the simplicity of the logistic regression which avoids overfitting.

5. CONCLUSION

In this article we investigated the design of semi-supervised learning systems for semantic role labeling task. To be specific, we are interested in the Brazilian Portuguese, a resource-poor language which suffers from scarcity of annotated data. Despite the main SRL approaches proposed to Brazilian Portuguese employ supervised learning methods, we have considered a representative set of SSL methods, which is based on low density separation, label propagation and self-training. Those methods have relevant advantages over the supervised ones as they learn not only from labeled data, but also from the unlabeled ones, which means less annotated data is required.

Experiments were conducted considering SSL and supervised methods over the most frequent verbs in the PropBank-br, a Brazilian Portuguese corpus built with text from Brazilian newspapers. The results revealed self-training has an outstanding performance in comparison with other methods, including the supervised ones. Additional experiments attested that the technique also evolves very well as more labeled examples are considered. Interestingly, the results demonstrate that a state-of-the-art supervised SRL method needs to be trained over at least 40% more labeled arguments to obtain results comparable to the self-training method. Taking into account the limitations of the PropBank-br, this can be considered a promising result.

Besides the progress obtained in the argument classification task, much remains to be done in terms of SSL for Brazilian Portuguese SRL: the combination of argument identification and classification, the design of more efficient SRL systems and the evaluation of more verbs, just to name a few. Another important investigation includes co-training [Blum and Mitchell 1998], which assumes that each argument can be described using two distinct sets of features that provide different (complementary) information about it; in a few words, we believe the predictive performance of self-training can also be enhanced by considering not only the analysis of constituent, but also the analysis of dependency.

REFERENCES

Alva-Manchego, F. E. and Rosa, J. L. G. Towards Semi-Supervised Brazilian Portuguese Semantic Role Labeling: building a benchmark. In H. Caseli, A. Villavicencio, A. Teixeira, and F. Perdigão (Eds.), *Computational Processing of the Portuguese Language.* Lecture Notes in Computer Science, vol. 7243. Springer Berlin Heidelberg, pp. 210–217, 2012a.

Alva-Manchego, F. E. and Rosa, J. L. G. Semantic Role Labeling for Brazilian Portuguese: a benchmark. In J. Pavón, N. D. Duque-Méndez, and R. Fuentes-Fernández (Eds.), *Ibero-American Conference on Artificial Intelligence.* Lecture Notes in Computer Science, vol. 7637. Springer Berlin Heidelberg, pp. 481–490, 2012b.

Bick, E. *The Parsing System "Palavras": automatic grammatical analysis of portuguese in a constraint grammar framework.* Aarhus University Press, 2000.

Blum, A. and Mitchell, T. Combining Labeled and Unlabeled Data with Co-Training. In *ACM Annual Conference on Computational Learning Theory.* Madison, USA, pp. 92–100, 1998.

Carneiro, M. G., Rosa, J. L. G., Zheng, Q., Liu, X., and Zhao, L. Improving Semantic Role Labeling Using High-Level Classification in Complex Networks. In *International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery.* Guilin, China, pp. 2185–2191, 2017.

Carneiro, M. G., Zhao, L., Cheng, R., and Jin, Y. Network Structural Optimization Based on Swarm Intelligence for Highlevel Classification. In *IEEE International Joint Conference on Neural Networks.* Vancouver, Canada, pp. 3737–3744, 2016.

Carneiro, M. G., Zhao, L., and Rosa, J. L. G. Graph-Based Semi-Supervised Learning for Semantic Role Diffusion. In *Symposium on Knowledge Discovery, Mining and Learning.* Recife, Brazil, pp. 108–115, 2016.

Chapelle, O., Schölkopf, B., and Zien, A. *Semi-Supervised Learning.* MIT Press, 2006.

Chapelle, O. and Zien, A. Semi-Supervised Classification by Low Density Separation. In *International Workshop on Artificial Intelligence and Statistics.* Barbados, pp. 57–64, 2005.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. P. Natural Language Processing (almost) from Scratch. *Journal of Machine Learning Research* vol. 12, pp. 2461–2505, 2011.

Demšar, J. Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research* vol. 7, pp. 1–30, 2006.

Deschacht, K. and Moens, M.-F. Semi-Supervised Semantic Role Labeling Using the Latent Words Language Model. In *ACL Conference on Empirical Methods in Natural Language Processing.* Singapore, pp. 21–29, 2009.

Dowty, D. Thematic Proto-Roles and Argument Selection. *Language* 67 (3): 547–619, 1991.

Duran, M. S. and Aluísio, S. M. Propbank-Br: a Brazilian treebank annotated with semantic role labels. In *International Conference on Language Resources and Evaluation.* Istanbul, Turkey, pp. 1862–1867, 2012.

Fillmore, C., Johnson, C., and Petruck, M. Background to FrameNet. *International Journal of Lexicography* 16 (3): 235–250, 2003.

Fillmore, C. J. The Case for Case. In E. Bach and R. Harms (Eds.), *Universals in Linguistic Theory.* Holt, Rinehart and Winston, pp. 1–88, 1968.

FitzGerald, N., Täckström, O., Ganchev, K., and Das, D. Semantic Role Labeling with Neural Network Factors. In *ACL Conference on Empirical Methods in Natural Language Processing.* Lisbon, Portugal, pp. 960–970, 2015.

Fonseca, E. R. and Rosa, J. L. G. A Two-Step Convolutional Neural Network Approach for Semantic Role Labeling. In *IEEE International Joint Conference on Neural Networks.* Dallas, USA, pp. 2955–2961, 2013.

Fürstenau, H. and Lapata, M. Semi-Supervised Semantic Role Labeling via Structural Alignment. *Computational Linguistics* 38 (1): 135–171, 2012.

Gildea, D. and Jurafsky, D. Automatic Labeling of Semantic Roles. *Computational Linguistics* 28 (3): 245–288, 2002.

Grenager, T. and Manning, C. D. Unsupervised Discovery of a Statistical Verb Lexicon. In *ACL Conference on Empirical Methods in Natural Language Processing.* Sydney, pp. 1–8, 2006.

Hartmann, N. S., Duran, M. S., and Aluísio, S. M. Automatic Semantic Role Labeling on Non-revised Syntactic Trees of Journalistic Texts. In J. Silva, R. Ribeiro, P. Quaresma, A. Adami, and A. Branco (Eds.), *Computatio-*

nal Processing of the Portuguese Language. Lecture Notes in Computer Science, vol. 9727. Springer International Publishing, pp. 202–212, 2016.

Hartmann, S., Eckle-Kohler, J., and Gurevych, I. Generating Training Data for Semantic Role Labeling Based on Label Transfer from Linked Lexical Resources. *Transactions of the Association for Computational Linguistics* vol. 4, pp. 197–213, 2016.

He, L., Lee, K., Lewis, M., and Zettlemoyer, L. Deep semantic role labeling: What works and what's next. In *Annual Meeting of the Association for Computational Linguistics.* Vancouver, Canada, pp. 1–11, 2017.

He, S. and Gildea, D. Self-Training and Co-Training for Semantic Role Labeling: primary report. Tech. rep., 2006.

Jebara, T., Wang, J., and Chang, S.-F. Graph Construction and b-Matching for Semi-Supervised Learning. In *International Conference on Machine Learning.* Montreal, Canada, pp. 441–448, 2009.

Kipper, K., Dang, H. T., and Palmer, M. Class-Based Construction of a Verb Lexicon. In *AAAI Conference On Artificial Intelligence.* Austin, USA, pp. 691–696, 2000.

Lang, J. and Lapata, M. Unsupervised Semantic Role Induction via Split-Merge Clustering. In *Annual Meeting of the Association for Computational Linguistics.* Portland, USA, pp. 1117–1126, 2011a.

Lang, J. and Lapata, M. Unsupervised Semantic Role Induction with Graph Partitioning. In *ACL Conference on Empirical Methods in Natural Language Processing.* Edinburgh, UK, pp. 1320–1331, 2011b.

Màrquez, L., Carreras, X., Litkowski, K. C., and Stevenson, S. Semantic Role Labeling: an introduction to the special issue. *Computational Linguistics* 34 (2): 145–159, 2008.

Mohammad, S., Zhu, X., and Martin, J. Semantic Role Labeling of Emotions in Tweets. In *Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis.* Baltimore, USA, pp. 32–41, 2014.

Osman, A. H., Salim, N., Binwahlan, M. S., Alteeb, R., and Abuobieda, A. An improved plagiarism detection scheme based on semantic role labeling. *Applied Soft Computing* 12 (5): 1493–1502, 2012.

Palmer, M., Gildea, D., and Kingsbury, P. The Proposition Bank: an annotated corpus of semantic roles. *Computational Linguistics* vol. 31, pp. 71–106, 2005.

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

Pradhan, S., Hacioglu, K., Krugler, V., Ward, W., Martin, J. H., and Jurafsky, D. Support Vector Learning for Semantic Argument Classification. *Machine Learning* 60 (1-3): 11–39, 2005.

Pradhan, S. S., Ward, W., and Martin, J. H. Towards Robust Semantic Role Labeling. *Computational Linguistics* 34 (2): 289–310, 2008.

Punyakanok, V., Roth, D., and Yih, W.-t. The Importance of Syntactic Parsing and Inference in Semantic Role Labeling. *Computational Linguistics* 34 (2): 257–287, 2008.

Roth, M. and Lapata, M. Neural Semantic Role Labeling with Dependency Path Embeddings. In *Annual Meeting of the Association for Computational Linguistics.* Berlin, Germany, pp. 1192–1202, 2016.

Täckström, O., Ganchev, K., and Das, D. Efficient Inference and Structured Learning for Semantic Role Labeling. *Transactions of the Association for Computational Linguistics* vol. 3, pp. 29–41, 2015.

Titov, I. and Klementiev, A. Crosslingual Induction of Semantic Roles. In *Annual Meeting of the Association for Computational Linguistics.* Jeju, Korea, pp. 647–656, 2012a.

Titov, I. and Klementiev, A. Semi-Supervised Semantic Role Labeling: approaching from an unsupervised perspective. In *International Conference on Computational Linguistics.* Mumbai, India, pp. 2635–2652, 2012b.

Weston, J., Ratle, F., Mobahi, H., and Collobert, R. Deep Learning via Semi-Supervised Embedding. In G. Montavon, G. B. Orr, and K.-R. Müller (Eds.), *Neural Networks: Tricks of the Trade.* Springer Berlin Heidelberg, pp. 639–655, 2012.

Wu, D. and Fung, P. Semantic Roles for SMT: a hybrid two-pass model. In *ACL Annual Conference of the North American Chapter of the Association for Computational Linguistics.* Los Angeles, USA, pp. 13–16, 2009.

Yarowsky, D. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *Annual Meeting of the Association for Computational Linguistics.* Cambridge, USA, pp. 189–196, 1995.

Zadeh Kaljahi, R. S. Adapting Self-Training for Semantic Role Labeling. In *ACL Student Research Workshop.* Uppsala, Sweden, pp. 91–96, 2010.

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., and Schölkopf, P. B. Learning with Local and Global Consistency. In S. Thrun, L. K. Saul, and P. B. Schölkopf (Eds.), *Advances in Neural Information Processing Systems.* MIT Press, pp. 321–328, 2004.

Zhu, X. and Goldberg, A. B. Introduction to Semi-Supervised Learning. *Synthesis lectures on artificial intelligence and machine learning* 3 (1): 1–130, 2009.