Universidade FUMEC
Faculdade de Ciências Empresariais
Programa de Pós-Graduação em Sistemas de Informação e Gestão do Conhecimento

An Automatic Semantic Role Labeler for the Portuguese Language

Daniel Henrique Mourão Falci

Belo Horizonte
2018

Daniel Henrique Mourão Falci

An Automatic Semantic Role Labeler for the Portuguese Language

MSc thesis presented to the Programa de Pós-Graduação em Sistemas de Informação e Gestão do Conhecimento of FUMEC University, as partial fulfillment of the requirements for the Master's degree in Information Systems and Knowledge Management.

Advisor: Prof. Dr. Fernando Silva Parreiras

Research Track: Technology and Information Systems

Belo Horizonte

2018

Dados Internacionais de Catalogação na Publicação (CIP)

F178a Falci, Daniel Henrique Mourão, 1982-

An automatic semantic role labeler for the portuguese language / Daniel Henrique Mourão Falci. – Belo Horizonte, 2018.

81 f. : il. ; 29,7 cm

Orientador: Fernando Silva Parreiras

Dissertação (Mestrado em Sistemas de Informação e Gestão do Conhecimento), Universidade FUMEC, Faculdade de Ciências Empresariais, Belo Horizonte, 2018.

1. Linguística - Processamento de dados - Brasil. 2. Redes neurais (Computação) - Análise - Brasil. 3. Algoritmos - Análise. I. Título. II. Parreiras, Fernando Silva. III. Universidade FUMEC, Faculdade de Ciências Empresariais.

CDU: 65.011:681.3.6

Ficha catalográfica elaborada pela Biblioteca Universitária-FUMEC

Acknowledgements

First of all, I would like to thank my parents, particularly my mother, Maria do Amparo Mourão Falci, an incredible human being and a tireless educator who, even in face of difficulties, has always unfolded for the happiness and education of everyone in our house.

I thank my siblings, nephews, nieces, and close friends for the loving support that each of them provided and for understanding my absence in these two years.

My very special gratitude goes out to Prof. Dr. Fernando Silva Parreiras, who gave me the honor of having him as an advisor. Thank you for the countless teachings, instructions, and tips, but mostly for your sincerity, for being an accessible, and simple person.

I am also grateful to Visual Sistemas Eletrônicos LTDA., represented in the figure of its director, Olegário Amorim Pereira. Thank you once again for believing in my work and for providing me another growth opportunity in my life.

Finally, I must express my deep gratitude to my beloved wife and daughter for all the encouragement, support and patience throughout our life together and particularly during the writing process of this thesis. This work would not have been possible without you.

Continuous effort - not strength or intelligence - is the key to unlocking our potential.

Winston Leonard Spencer-Churchill

Resumo

A anotação de papéis semânticas (APS) é uma tarefa do processamento de linguagem natural que fornece os meios para analisar, do ponto de vista semântico, as informações expressas através de texto ou fala. O objetivo é capturar e representar os participantes e as circunstâncias de eventos ou situações descritas no nível sentencial. É tida como um importante passo para a compreensão da linguagem natural.

A maior parte da pesquisa existente sobre a APS é focada na língua inglesa e, portanto, considera suas particularidades sintáticas e semânticas. Este fato impede a transposição direta de seus resultados para outras línguas. Quanto à língua portuguesa, há um pequeno número de estudos dedicados a esta tarefa, e nenhum deles conseguiu um desempenho semelhante ao obtido na língua inglesa. Além disso, ao que sabemos, existe apenas um sistema publicamente disponível capaz de executar a APS automatizada em texto bruto, o que dificulta a pesquisa e detém o potencial inovador para a língua.

O objetivo desta dissertação é avaliar o desempenho de um anotador de papéis semânticos automático para a língua portuguesa construído considerando técnicas abordadas na literatura.

Para atingir este objetivo, o primeiro passo consistiu em uma revisão sistemática da literatura na tarefa de APS que visou identificar as técnicas mais precisas abordadas na literatura. Com base em seus resultados, desenvolvemos e avaliamos um anotador de papéis semânticos para a língua portuguesa. Nossa abordagem é independente de análise sintática e se apóia em uma arquitetura de rede neural recorrente, bidirecional e profunda. As predições da rede são usadas como a entrada de um algoritmo de análise neural recursiva global que foi adaptado para a tarefa de APS.

Nosso método superou, de forma consistente, o sistema mais preciso para a língua portuguesa no Corpus do PropBank-Br por uma margem de 3.05 pontos de F_1-score, reduzindo o erro relativo em 8.74%.

O modelo apresentado nesta pesquisa está disponível publicamente sob licença BSD e pode ajudar estudos futuros focados na língua portuguesa em tarefas que normalmente dependem da análise de conteúdo, que vão desde a tradução automática até os sistemas de perguntas e respostas.

Palavras-chaves: Anotação de Papéis Semânticos, PLN, Aprendizado Profundo, Redes Neurais Recorrentes, LSTM.

Abstract

Semantic Role Labeling (SRL) is Natural Language Processing task that provides the means to analyze, from the semantic point of view, the information expressed through text or speech. Its purpose is to capture and represent the participants and circumstances of events or situations described at the sentential level. It is considered a major step towards natural language understanding.

Most of the existing SRL research is focused on the English language, and thus, considers its syntactic and semantic particularities. This fact prevents a direct transposition of its results to other languages. Regarding the Portuguese language, there is a small number of studies dedicated to the task, and none of them achieved a similar performance to that obtained in the English language. Moreover, to the best of our knowledge, there is only one publicly available system capable of performing automated SRL on raw text what hampers research and detain the innovative potential for the language.

The objective of this thesis is to evaluate the performance of an automatic semantic role labeler for the Portuguese language built considering techniques addressed in the literature.

To achieve this goal, the first step consisted in a systematic literature review on SRL task that intended to identify the most accurate techniques addressed in the literature. Based on its results, we developed and evaluated a semantic role labeler of raw text for the Portuguese language. Our approach is independent of syntactic parsing and relies on a deep bidirectional recurrent neural network architecture. The network predictions are used as the input of a global recursive neural parsing algorithm that was tailored for the SRL task.

Our method consistently outperformed the previous state-of-the-art system for the Portuguese language on PropBank-Br corpus by a margin of 3.05 F_1-score points, reducing the relative error in 8.74%.

The model presented in this research is publicly available under BSD license and may help future studies focused on the Portuguese language in tasks that are typically dependent on content-analysis, ranging from Machine Translation to Question and Answering Systems.

Keywords: Semantic Role Labeling; NLP; Deep Learning; Recurrent Neural Networks; LSTM.

List of Illustrations

List of Tables

List of Abbreviations and Acronyms

NLP	Natural Language Processing
SRL	Semantic Role Labeling
ML	Machine Learning
AI	Artificial Intelligence
BP	Brazilian Portuguese
PTB	Penn TreeBank
RNN	Recurrent Neural Network
LSTM	Long short-term memory
BiLSTM	Bidirection long short-term memory

Summary

1 Introduction

Natural language processing (NLP) is a subfield of artificial intelligence that investigates computational techniques to analyze and represent, at varied linguistic levels, the languages that human uses naturally (1). One of its primary aspirations is to understand, with human-like precision, the message expressed in natural languages. This goal, if achieved, may benefit any task that relies on content-analysis, including question and answering systems, news-gathering, voice activation, and automated text summarization.

The challenge, however, is not small. Comprehending the meaning of the text (the semantics) requires the ability to deal with language-specific issues and, ultimately, depends on cognition - the mental process of acquiring knowledge and understanding through reasoning, experiences, and senses (2). It encompasses phenomena that are not yet fully known by researchers and, therefore, are difficult to be simulated in a computational environment.

Among the NLP tasks dedicated to the semantic analysis, the semantic role labeling (SRL) is of particular interest for this work. Also known as shallow semantic parsing, the purpose of this task is to capture and represent the participants and circumstances of events or situations described at the sentential level (3). Grossly, it allows providing answers to questions such as *who did what to whom where when and how* in text or speech. Events, in this context, are triggered by predicates, which are usually represented by verbs. The participants, in turn, are denoted by words or groups of words (constituents) known as arguments, each of them playing a different abstract role with respect to a given predicate (4).

Consider the sample sentence "*John is building a new house in downtown*". In this case, the predicate *build* evokes a construction event that, implicitly, requires arguments that hold semantic roles such as the constructor (*John*), the thing built (*a new house*), and the construction place (*in downtown*). The idea is that only by the presence of these semantic roles, an event may produce a minimal unit of meaning in its interlocutor.

The annotation of semantic roles has proven to be useful in several NLP applications(5, 6, 7, 8). In a Question and Answering system, for instance, (9), a factoid question may be formulated as *what kills bacteria?*. A semantic approach searches at the available corpus trying to find sentences that contain predicates such as *kill* whose patient is *bacteria*. The sentence elements that play the semantic role of the agent are natural candidate answers.

Since the first automatic SRL approach, presented in the seminal work of (3), the task has been tackled through the use of statistical machine learning techniques (ML). It is usually considered a supervised classification problem where the objective is to choose, from a pre-defined set of possible semantic role labels, the proper ones for each constituent of a sentence. Unsupervised techniques, although rare, can also be employed. In this case, the goal is to discover, from raw text, groups of semantic roles, which are determined by the application of a similarity or distance measure. The resulting clusters are then analyzed by domain specialists to determine their meaning. At last, semi-supervised approaches are considered an option under circumstances where there is not enough annotated data to generalize a supervised ML model. It may employ either self-training algorithms, or unsupervised techniques in order to expand the input data for supervised models (4).

The majority of the SRL techniques described in literature utilizes the syntactic elements of a sentence as input features to their ML models. This choice is supported by the linking theory presented by (10) which states that the predicates that share a similar syntactic behavior tend to exhibit a similar meaning, and consequently, share the same predicate-argument structure. It points to a mutually dependent structure which is based on the observation of the syntactic-semantic occurrence patterns in the English language. Some of the most common syntactic representations are constituency trees, dependency trees, and shallow syntactic trees.

Different studies have corroborated the benefits of using this syntactic information in SRL task (11, 12, 13). A possible drawback is that the use of syntactic information creates a dependency on external tools (syntactic parsers). In this sense, their eventual errors, as expected, are propagated to the SRL model, creating noise and affecting its performance.

As a typically supervised NLP problem, the task requires vast amounts of annotated data. The Proposition Bank, or simply PropBank (14), is the most popular lexical resource for the English language. It adds a hand-tagged semantic layer on top of constituency trees provided by the Penn TreeBank corpus (PTB), thereby taking advantage of the linking theory. This corpus was conceived to support ML approaches and therefore presents a significant size, with approximately one million annotated tokens in its first version.

The success of this corpus inspired similar versions on several languages. Its Portuguese counterpart is known as PropBank-Br (15) and, in large part, follows the original formalism applied by the English version. It was built on top of the Brazilian portion of Bosque, a section of Floresta Sintá(c)tica treebank (the equivalent to the Penn TreeBank for the Portuguese language). Comparatively though, the Portuguese version may be considered a small sized corpus. It accounts for only the seventh part of the annotated sentences contained in the English version.

Hundreds of studies investigated the SRL task and its applications, particularly for the English language where one may find the highest number of systems and results. Regarding the Portuguese language though, the scenario is inverted and SRL task is still in early stages. Although some methods have been proposed, none of them achieved a similar performance to that obtained in the English language. Moreover, to the best of our knowledge, there is only one system capable of performing automated SRL on raw text for the language. In the following lines, we highlight the most relevant studies conducted up until now.

Bick(16) was the first to investigate the SRL task for the European Portuguese language. Their rule-based system (PALAVRAS) uses a set of heuristics to map and disambiguate semantic roles. It employs 500 manually created rules to extract 35 possible semantic roles. Although limited by the manual effort it demands, the author reported excellent results with an F_1-score around 88 points. His method though, was evaluated on a very small dataset comprised of 2500 words taken from the European part of Foresta Sintá(c)tica corpus. This fact, when combined to the high granularity produced by their semantic roles, may have biased the performance analysis. Moreover, the system is distributed under a proprietary license and is not publicly available for NLP community.

Alva-Manchego e Rosa(17) were the first to investigate the SRL task for the Brazilian Portuguese (BP). Their preliminary approach intended to produce a benchmark for future studies. The supervised system utilized classifiers such as Naive Bayes and Decision Trees that were trained on an early version of PropBank-Br. The evaluation though, relied on gold-standard syntactic trees provided by the corpus what produces an inexistent condition under real-usage circumstances.

The first fully functional semantic role labeler for the BP was introduced by Fonseca e Rosa(18). Their system, NLPNET, trained on PropBank-Br corpus, was inspired by the approach of Collobert et al.(19) that obtained a reasonable performance using the English PropBank. It uses word vector representations as input features of a convolutional neural network architecture. Their best single training session yielded 65.13 F_1-score points, a performance that is almost 10 F_1-score points behind the original system. The author points that a possible cause may be related to the scarcity of data on PropBank-Br corpus. To the best of our knowledge, this is the only SRL system for the BP whose source code is publicly available.

1.1 Research Problem

Considering the context exposed above, the following research question emerges: **What is the accuracy of an automatic semantic role labeler for the Portuguese language built considering techniques addressed in literature?**

1.2 Objectives

1.2.1 Main Objective

The main objective of this thesis is to **evaluate the performance of an automatic semantic role labeler for the Portuguese language built considering techniques addressed in the literature**

1.2.2 Specific Objectives

- **OBJ1:** Identify the most accurate semantic role labeling techniques described in the literature.
- **OBJ2:** Analyze the results of an automatic semantic role labeler for the Brazilian Portuguese built considering techniques addressed in the literature.

1.3 Motivation

Due to the existing opportunities in a range of areas, there is a growing interest of foreign investors in emergent economies such as Brazil, an active member of the BRICS (alongside Russia, India, China and South Africa) and the ninth-largest economy in the world (20).

For any external market agent, the local language understanding can represent competitive advantages as it enables informed and consequently, better decision-making process. The speed of information acquisition and processing is also determinant since it provides the chance to take the initiative and act before competitors. Thus, there is a latent necessity to support this interest with new methods and techniques, an objective that can only be achieved through research.

The SRL task is a central technique in this context. It has been successfully employed on several tasks such as question and answering systems (21, 9), open information extraction (6), automatic text summarization (5), and machine translation (7). Portuguese language studies in this field, though, are scarce, particularly when compared to the English language. Considering that the majority of the existing techniques are based on supervised machine learning, the lack of lexical resources combined with the scarcity of freely available tools are limiting factors that hamper the research process and consecutively detain the innovative potential for the Portuguese language. We seek to contribute to changing this scenario by freely releasing all the artifacts produced by this research.

1.4 Adherence to Graduate Program

The FUMEC's graduate program in Information Systems and Knowledge Management is focused on applied research. The program is organized into two main streams: Technology and Information Systems and Information and Knowledge Management. The multidisciplinary approach is a fundamental concept to the program.

This research investigates the performance in semantic role labeling task for the Portuguese language. It is a multi-purpose semantic analysis tool that holds the potential to leverage any task that relies on content-analysis. Our focus is on Decision Support Systems topic under Information Systems area in compliance with FUMEC's graduate program.

The multidisciplinary character arose from the purpose of the research that may enable future applications in several fields of study.

1.5 Document Structure

We structured this thesis in 4 chapters. Chapter 1 presented the introduction. Chapter 2 presents a systematic literature review on semantic role labeling task developed during the thesis project definition. In chapter 3 we describe our semantic role labeler and discuss the experimental results. Lastly, chapter 4 concludes our work.

1.6 Communications of this Thesis

We have communicated the research presented in this thesis through journal papers. In the following, we mention the publications according to the chapters covering the respective contribution.

- Chapter 2: FALCI, D.H.M.; PARREIRAS, F.S. Semantic Role Labeling: A Systematic Literature Review. Computer Speech and Language, 2017. [Under Consideration].

- Chapter 3: FALCI, D.H.M.; PARREIRAS F.S. Applying Recurrent Neural Networks into Semantic Role Labeling for the Portuguese Language. Cognitive Systems Research, 2018. [Under Consideration].

2 Systematic Literature Review

2.1 Introduction

Semantic Role Labeling (SRL) is a Natural Language Processing task (NLP) that intends to reveal the predicate-argument structures described in each sentence of a document. This shallow semantic representation enables computational approaches to capture and represent events, identifying their participants and circumstances (3). It is a major component towards natural language understanding and has been successfully employed in several NLP applications such as question and answering systems (21, 9), automatic text summarization (5), open information extraction (6) and co-reference resolution (8). SRL is particularly useful when applied to Information Extraction activity (22) - extracting limited kinds of semantic content from text and transforming it into structured data - and may benefit any task that relies on semantic knowledge (23).

Since earlier approaches, in an unceasing search for the most accurate methods, hundreds of studies employed complex techniques that combined elements such as several machine learning methods, feature selection procedures, syntactic representations, different lexicons and formalisms, processing stages and others. The overall accuracy though, is still around 83%, illustrating how hard is the task and how much room for improvement there is in this research field.

In this scenario, emerges the following research question: *What are the SRL techniques addressed in the literature and what is their accuracy?* In order to answer this question, we undertake a literature review, revealing research gaps and common practices as well as clarifying the state-of-the-art in this research field. To the best of our knowledge, this is the first study of this kind devoted to this task.

The remainder of this paper is organized as follows: Section 2.2 provides the technical background of SRL task. Section 2.3 describes our methods and materials. Section 2.4 presents our results while section 2.5 discusses them. Section 2.6 examines our limitations and threats to validity. At last, the conclusions and future work are presented in section 2.7.

2.2 Background

2.2.1 Linguistics

In linguistic theory field, semantic roles, also known as *semantic case* or *thematic roles*, are defined as the study of the existing syntactic-semantic relationship between arguments and predicates in a sentence (4). Such predicates are usually represented by the verbs in a sentence and serve as event triggers while its arguments point out to actors, objects, and circumstances related to these events. The sentence (1) illustrate the concept. The verb 'open' triggers an opening event, where the argument 'John' play the role of the opener and the argument 'door', the thing opened.

(1) John opened the door

Semantic roles were one of the first linguistic models, introduced by Panini around the 6th century BC in his *Karaka* theory for the Sanskrit language (24). Since then, few studies were undertaken, and the subject was almost forgotten until its re-introduction in modern linguistics by Fillmore(25) in his *Case Grammar*. In his work, the author claims that sentences, no matter its language, are composed of verbs and its respective arguments which may be classified into six universal semantic roles described in Table 1. In his theory, each semantic role is optional and may occur only once per proposition, while each constituent cannot assume more than one role simultaneously. The author also observes that the structures that modify the main verb such as negation, auxiliary verbs and adverbs are independent of the semantic role structure, not affecting it.

Table 1 – Original set of semantic roles of case grammar

Semantic Role	**Definition**
Agentive	Entity, instigator of the action or state described by the verb.
Instrumental	Inanimate force or object causally involved in the action or state described by the verb.
Dative	Entity, affected by the state or action described by the verb.
Factive	Object or result from the action or state described by the verb.
Locative	Place or spatial orientation of the action or state described by the verb.
Objective	Thing affected by the action or state determined by the verb.

Adapted from Fillmore(25)

The sentences in (2) exemplify the semantic role structure obtained by the application of Fillmore's universal semantic roles. It is noticeable that in the three sentences, despite its verbal time, morphologic and syntactic functions, the event structure remains unchanged. The predicate *open*, in all sentences, accepts three arguments: *Jonh* as *AGENTIVE*, the *door* as *OBJECTIVE* and *key* as *INSTRUMENTAL*. At the same time, we have three different syntactic structures. In the sentence (2a), *John* is the subject while in (2b) and (2c) we have the phrasal components *door* and *key* as the subjects, respectively.

(2) a. [John$_{Agentive}$] *opened* the [door$_{Objective}$] with the [key$_{Instrumental}$]

b. The [door$_{Objective}$] was *opened* with the [key$_{Instrumental}$]

c. The [key$_{Instrumental}$] will not *open* the [door$_{Objective}$]

Fillmore's Case Grammar concepts, though, contained ambiguity. In practice, even when applied in simple sentences, the universal set of semantic roles could mutually overlap, making impossible to determine its semantic function. This fact inspired scientific discussions in the years that followed, and several authors attempted to identify a method to define a reliable set of universal semantic roles (26, 27, 28, 29). The work of Cook(26) for instance, explores Case Grammar theory, eventually expanding it to a set composed of nine thematic roles that are depicted in Table 2.

Table 2 – Current set of universal semantic roles

Semantic Role	**Definition**
Agent	Instigator or main cause of the event
Experiencer	Affected by the action
Instrument	The immediate cause of the event
Object	Entity that changes with the event
Source	Starting point of the action
Goal	Final point of the action
Local	The location where the event occurs
Time	The moment when the event occurs
Benefactive	The beneficiary of an event

Adapted from Cook(26)

In an attempt overcome the limitations of his Case Grammar, Fillmore went in an alternative direction and, instead of working with universal semantic roles or generalization, proposed its specialization, introducing a theory known as *Frame Semantics* (30, 31). The central idea of his theory is that the meaning of the words is strictly dependent on the frames where they have been used. In this context, frames are schematic representations of events or situations that hold a template of specialized thematic roles, evoked in the presence of a particular set of predicates. Consider the sentences in (3). The verbal predicates *buy* and *sell*, although different in perspective, are responsible for triggering the *Commerce* frame. It implicitly requires semantic roles such as the seller, the buyer, the goods, and the amount spent. Our inherent knowledge about its attributes and typical interactions enable us to identify, process, and understand the multiple semantic roles of each participant in the event.

(3) a. [John$_{Buyer}$] *bought* a [car$_{Goods}$] from [Mary$_{Seller}$]

b. [Mary$_{Seller}$] *sold* the [car$_{Goods}$] to [John$_{Buyer}$]

Levin(10) explores the linking theory that is based on the idea that predicates that share a similar syntactic behavior tend to exhibit a similar meaning. Consequently, this synonymy relationship causes similar predicates to trigger the same events or semantic frames, sharing its syntactic structures in a mutually dependent framework. Thus, she suggests that syntactic elements are dominant factors for determining the semantic roles of the arguments of a given verb. In her detailed work, she analyzed the verb occurrence patterns in English language and classified 3100 verbs into 47 classes and 193 subclasses. The classification took into account the possible syntactic alternations of the semantic arguments accepted by each verb so that in the final model, semantic and syntactically similar verbs belong to the same class.

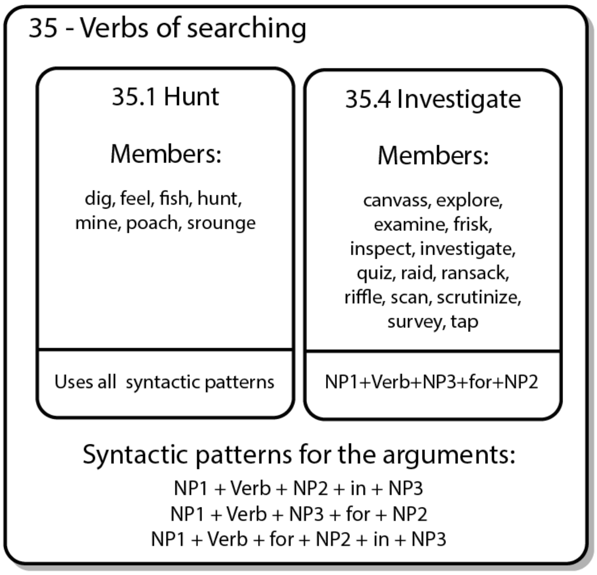
The class 35 for instance (see Figure 1) groups verbs related to 'searching'. The verbs in these class usually have three possible syntactic realizations to express their arguments, but different subclasses use different subsets of these patterns. The subclass 35.1 intends to describe the 'hunt verbs' that may exhibit all the three possible syntactic realizations. Subclass 35.4, on the other hand, is dedicated to 'investigate verbs' and can only use one out of three syntactic alternations.

Her work is known as Levin classes and is particularly relevant in linguistics field because of her discussion about the existing link between syntactic and semantic elements in a sentence, a subject that, although not yet fully understood by scholars, inspired subsequent studies especially in the field of computational linguistics (32, 33).

2.2.2 Lexical resources

Considering SRL research, two corpora for the English language are widely employed: FrameNet and PropBank. Given their importance, these resources are presented individually, each on its subsection. Lexical resources such as NomBank, OntoNotes, VerbNet, and WordNet are either applied as the main semantic resource or to enrich the knowledge base available for computational approaches. The subsection Another Resources offer an overview of them. At last, the subsection Other Languages, as the name suggests, outlines the resource availability for other languages.

Figure 1 – Levin classes - Subclasses for the verbs of searching


35 - Verbs of searching
35.1 Hunt
Members:
dig, feel, fish, hunt, mine, poach, sronge
Uses all syntactic patterns
35.4 Investigate
Members:
canvass, explore, examine, frisk, inspect, investigate, quiz, raid, ransack, riffle, scan, scrutinize, survey, tap
NP1+Verb+NP3+for+NP2
Syntactic patterns for the arguments:
NP1 + Verb + NP2 + in + NP3
NP1 + Verb + NP3 + for + NP2
NP1 + Verb + for + NP2 + in + NP3

NP - indicates a noun phrase. In this sense, NP1 + Verb + NP2 + in + NP3 could be used with the verb hunt in a sentence such as 'John hunt a doe in the forest'. Notice that NP1, NP2, and NP3 are respectively reserved for the roles of agent, patient, and place.

2.2.2.1 FrameNet

The Berkley FrameNet Project, presented in Baker, Fillmore e Lowe(34), is a hand-tagged lexicographical resource following the Frame Semantics theory. It is comprised of a collection of frames and their corresponding set of semantic roles (frame elements) and predicates (lexical units) annotated on top of the British National Corpus[1]. FrameNet contains more than 7,000 annotated predicates, distributed in approximately 1,000 frames. The resource also holds more than 200,000 annotated sentences [2], resulting in an average value of 20 annotated sentences per predicate (35).

Its frames are organized in a hierarchically-related way, with relationships such as inheritance and usage. The sentences (4), (5) and (6) are examples of the frames *Residence*, *Visiting* and *Temporary stay* respectively. The latter one establishes a semantic inheritance relationship with the former ones, in a way that what is true for the parent frame is also true for the child frame. It is also noteworthy that FrameNet also maps nominal predicates, in contrast to other lexical resources that are exclusively concerned with verbal predicates. Sentence (5) provides an example where the predicate *visit* is presented in its nominal form. Another relevant characteristic of this lexical resource is that it does not contain any syntactic annotation. This fact is due to its general purpose, where not only computational domains but also pure linguistic studies must be contemplated. In this sense, FrameNet attempts to avoid the computational bias focusing on its primary goal: semantics.

[1] Avaliable at <http://www.natcorp.ox.ac.uk>

[2] The current annotation status is available at <https://framenet.icsi.berkeley.edu/fndrupal/current_status>

(4) [John$_{Resident}$] still *lives* [with his parents$_{Co-resident}$] [in San Francisco$_{Location}$].

(5) [Harry's$_{Agent}$] landmark *visit* [to Iraq$_{Entity}$] started [yesterday$_{Time}$]

(6) [For a few weeks$_{Duration}$] [Mary$_{Guest}$] *lodged* [at a hotel$_{Location}$]

2.2.2.2 PropBank

The Proposition Bank or PropBank[3] (36, 33), is a lexical resource based on Levin's classes and in the refinements made on them (37). Since its first version, PropBank intended to allow the usage of computational methods and syntactic-semantic structures. The inspiration was the success achieved by the Penn TreeBank II project (PTB) that allowed the development of powerful syntactic parsers for the English language based on machine learning techniques. For this reason, and taking advantage of the linking theory, PropBank added a semantic layer on top of the constituency trees, already existent in PTB. Its first version had about one million sentences annotated on top of the Wall Street Journal section of the PTB corpus.

The formalism implemented in PropBank though diverge from the one adopted by FrameNet. First of all, PropBank is only concerned with predicates in its verbal form. Second, its core arguments are represented by an abstract and optional set of arguments, numbered from *Arg0* to *Arg5* that are strictly dependent on each verb meaning. With the exception of the proto-roles *Arg0*, that is usually reserved for the *Agent*, and *Arg1* that indicates the *Patient* or *Theme* in a proposition, no inferences can be made for the rest of the arguments when comparing predicates. The table 3 exemplifies the PropBank roles for some predicates[4].

Table 3 – PropBank - verb specific roles

Semantic Role	**Open**	**Buy**	**Sell**
Arg0	Opener	The buyer	The seller
Arg1	Thing opened	Thing bought	Thing sold
Arg2	Instrument	The seller	The buyer
Arg3	Benefactive	The price	The price
Arg4	–	Benefactive	Benefactive
Arg5	–	–	–

PropBank also defines a set of functional arguments (*ArgM*) that are common to all predicates. They intend to indicate adjunct arguments such as time, place and manner and can be merged with the verb-specific roles from *Arg2* to *Arg5*. The Table 4 lists the

[3] More information can be found at <http://verbs.colorado.edu/%7Empalmer/projects/ace.html>

[4] Semantic roles were taken from <http://verbs.colorado.edu/propbank/framesets-english-aliases/>

functional arguments and their respective meaning while sentence (7), exemplifies the PropBank annotation formalism.

Table 4 – List of PropBank functional arguments

Semantic Role	**Meaning**	**Examples**
ArgM-LOC	Location	*the hospital, in Dallas*
ArgM-TMP	Temporal	*now, this summer*
ArgM-MNR	Manner	*clearly, very fast*
ArgM-NEG	Negation	*not, n't*
ArgM-CAU	Cause	*In response*
ArgM-DIR	Direction	*to production, to USA*
ArgM-PNC	Purpose	*to pay for, to acquire*
ArgM-EXT	Extension	*30 points, at 200*
ArgM-DIS	Discursive marking	*for instance, additionally*
ArgM-REC	Reciprocity	*each other*
ArgM-PRD	Secondary predication	*to become a doctor*
ArgM-ADV	Adverbials	*none of the above*

(7) [John$_{Arg0}$] *sold* [the car$_{Arg1}$] [to Mary$_{Arg2}$] [last month$_{ArgM-TMP}$]

2.2.2.3 Another Resources

VerbNet (32, 38) is a lexicon that extended the original coverage of Levin classes to more than 470 classes distributed in four hierarchical levels. Such corpus adds a representation of the possible syntactic alternations assumed by the arguments of each predicate (only verbal ones, as the resource name suggests) and their arguments [5]. It also adds selectional restriction information[6] for each verb class what is useful while disambiguating verb senses.

Baker e Ruppenhofer(39) offers a detailed comparison between the VerbNet and the FrameNet. In this study, the authors analyzed aspects such as syntactic realizations, semantic classes and hierarchical structures provided by both formalisms. Their result indicates that despite their comparable verb coverage, the FrameNet formalism produces more consistent categories and richer relationships among them.

As mentioned in the previous subsection, PropBank is only concerned with verbal predicates what may restrict its usage. The NomBank project (40) attempts to fill this gap by annotating the nominal predicates skipped in PropBank sentences. The annotation guidelines are essentially the same ones applied in the PropBank, making of it a complementary corpus for PropBank.

[5] The full semantic roles list is available at <https://verbs.colorado.edu/~mpalmer/projects/verbnet.html>

[6] The kind of concept an argument requires: i.e., Something that is edible is the selectional restriction for the argument *Theme* of the predicate *eat*

WordNet (41) is a lexical resource that groups adjectives, adverbs, nouns and verbs into 117,659[7] sets of synonym classes (synsets) interlinked through their lexical and semantic relations. Nouns and verbs are hierarchically organized observing hypernymy relations which enable the following kind of inference: if armchair is a kind of chair and chair is a piece of furniture then, an armchair is a piece of furniture. For each mapped word, it provides definitions and example sentences. In the NLP field, this corpus is particularly useful in tasks such as word-sense disambiguation, information retrieval, and information extraction. Considering the SRL task, the synonymy relations may improve the overall semantic understanding of a sentence, disambiguating predicates and argument meanings (42).

OntoNotes (43) is a multilingual corpus (English, Chinese, and Arabic) composed by various genres of text, from news to phone talks, annotated with predicate-argument structures, word senses, and co-reference resolution schemes. Regarding the predicate-argument formalism, OntoNotes was inspired by PropBank project, and therefore, follows its guidelines. Another similarity is that it employs the same syntactic representation utilized in PropBank: Constituency trees using the annotation guidelines of PTB). Word senses for verbs and nouns are linked to hypernymy structures that resemble the ones seen on WordNet. The authors also claim that the corpus has a 90% agreement among the annotators.

2.2.2.4 Other Languages

The lexical resources presented so far are targeted at the English language. They enabled successful SRL approaches, creating a favorable environment for its application in several NLP tasks. This fact motivated the development of equivalent corpora in other languages.

The following languages produced their own corpus built considering the FrameNet formalism: Portuguese (44), French (45), German (46), Swedish (47), Spanish (48), Chinese (49), Japanese (50), Korean (51) and Polish (52).

The PropBank formalism was also followed in languages such as Portuguese (53), Arabic (54), Chinese (55), Finnish (56) and Hindi (57).

2.2.3 Machine learning

Since earlier approaches, automatic SRL has been performed through the application of statistical machine learning techniques (4, 22). Machine learning (ML) is a subfield of Artificial Intelligence (AI) concerned with the study of the algorithms and methods that can be used to detect patterns from a dataset. The goal is to perform predictions from

[7] The current WordNet status may be found at <http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html#toc2>

new data (predictive models) or to describe the data behavior (descriptive models), allowing systems to learn from data samples or past experiences even not being explicitly programmed for a task (58, 59, 60). ML algorithms can be classified according to its corresponding learning method into the following categories: Supervised, unsupervised, semi-supervised and reinforcement (see Figure 2 for a taxonomy).

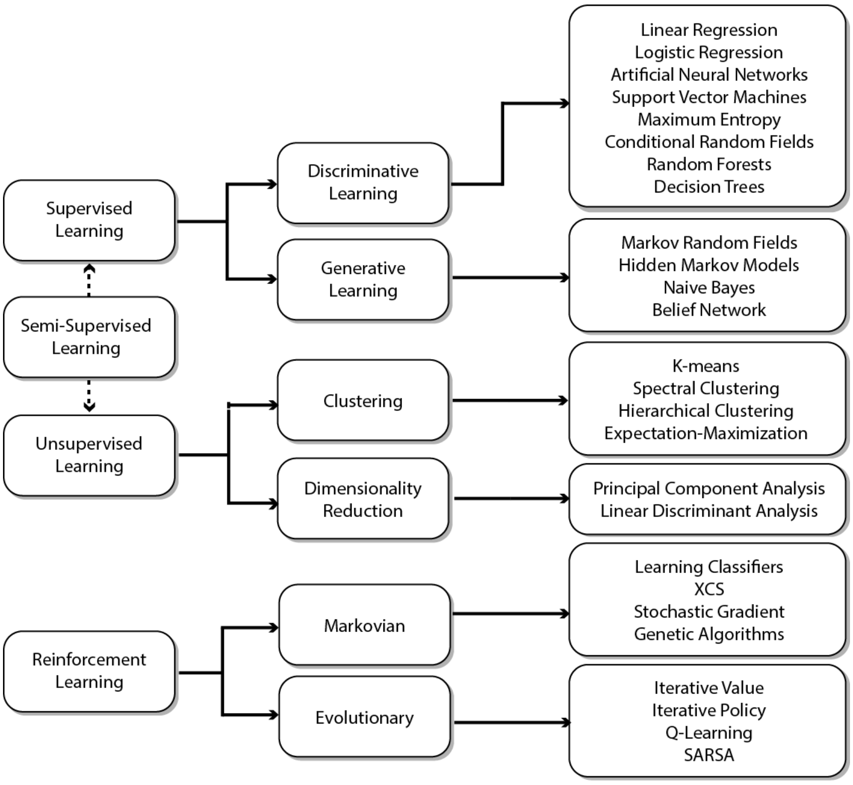
Supervised learning operates through the observation of input-output pairs, learning a model capable of relating them both (58). For this reason, such techniques require labeled data (or the correct output) while inferring (training) the model. Accordingly to their training strategies, one may partition the supervised learning models into two distinct sub-categories: Generative and Discriminative models (61). Generative models attempts to explain how the data was generated describing the joint probability distribution $p(X, Y)$, where X and Y are sets of observable (the input) and hidden (the output) variables, respectively. Discriminative models, on the other hand, calculates the conditional probability $p(Y|X)$ directly, concentrating its efforts on modelling the boundary between the classes in Y known as the decision boundary (59).

The unsupervised learning requires a distinct approach. It consists of discovering patterns from a dataset without providing any other explicit information (58). Therefore, this technique does not require labeled data, in contrast to the supervised learning models. Clustering and dimensionality reduction are popular examples of unsupervised learning techniques. The clustering objective is to partition a dataset into a defined number of data segments in a way that maximizes the number of elements distributed among clusters and minimizes the elements inside of each one of these clusters. Dimensionality reduction in turn, aims at finding the smallest subset of variables that best describes the data behavior. It is usually applied in feature selection or feature extraction procedures. The purpose is to reduce the processing time and storage space required (62).

In semi-supervised learning, which falls between supervised and unsupervised techniques, the models typically combine the usage of big unlabeled and small labeled datasets to achieve proper generalization or to acquire new features that may improve the model's accuracy (63). Techniques such as self-training or bootstraping are some of the most popular examples.

Reinforcement learning is different from the previous approaches presented so far. Instead of requiring labeled examples from the dataset or detecting patterns on unlabeled data, it requires evaluation functions that provide the feedback about a given sequence of actions, so the model can reshape the actions taken in order to maximize a reward notion (60, 58). As with the supervised and unsupervised, reinforcement learning approaches are categorized as Markovian or Evolutionary processes.

Figure 2 – A taxonomy of machine learning algorithms



Source: Adapted from Nicolas(59)

2.2.4 Automatic Semantic Role Labeling

This section describes the procedures adopted while performing the SRL task and is divided into three subsections: syntactic representations and features, processing stages, and evaluation.

2.2.4.1 Syntactic Representations and Features

Since early approaches, the automatic SRL research relied on the hypothesis that the possible syntactic configurations of a constituent in a sentence are determinant factors in evaluating its semantic function (Linking theory). Thus, given a sentence, a crucial step in SRL systems is to extract its syntactic representation whose output is then used as features that model the mentioned relationship.

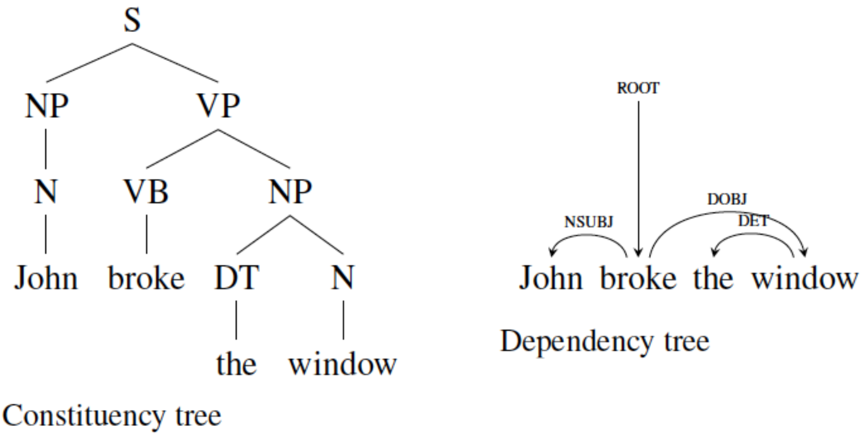
Two common syntactic views are those based on constituency and dependency grammars. Figure 3 exemplifies their structural notation applied in the same sentence. The answer to which syntactic view provides better results is still a matter of debate in scientific inquiry (4).

In the constituency-based representation, a sentence is broken into a tree of sub-phrases where its non-terminal nodes hold the phrase type and the terminal nodes, their morphological categories (N for a noun, VB for verbs, and so forth). The sub-phrases, according to the constituency grammar, may be classified into phrase types[8] such as noun phrases (NP) and verbal phrases (VP). The hierarchical relationship within a tree determines which constituents are modified and those who act as modifiers, disambiguating the sense of sentences. Constituency trees are usually obtained through probabilistic parsers such as that of Collins(64) or Charniak e Johnson(65).

In dependency trees, the words are connected according to their dependency relationship. In this case, each word is a node, and the edges across the nodes map their dependencies. In the example (see Figure 3), the verb "broke" invokes a dependency relation with the words "window" and "John" where the first is qualified as a direct object and the second classified as its subject. The word *window* also evokes a dependency relation to the determiner *the*.

Dependency trees, on average, contain fewer nodes than constituency trees since there are no intermediary levels. It is simpler while it offers deep syntactic information what represents a notable aspect from a computational perspective. Dependency-based representations may be extracted from the output of a constituency-based parser through the usage of rule-based systems (13). However, dependency trees can also be extracted directly from text with parsers such as Nivre, Hall e Nilsson(66) and McDonald et al.(67).

Figure 3 – Different syntactic representations for the same sentence



Shallow syntactic parsers (also known as chunkers) may be understood as a simplified version (and cheaper, from a computational perspective) of a constituency parser (full parser). The basic difference between these representations is that the chunker captures information that responds to just one level of constituency trees (NP, VP, PP, ADVP, ADJP, and so forth.), generally coupled with the IOB tagging scheme (68, 69).

Gildea e Jurafsky(3) were the first to describe a SRL system and their work, based on the FrameNet corpus, explored the syntactic-semantic relationship through con-

[8] A complete set of constituency tree labels used in Penn TreeBank may be found at <http://goo.gl/ghgk17>

stituency trees extracted by the Collins parser (64). Their main contribution stands for the core set of features they described:

- **Phrase type:** The syntactic category of a constituent (noun phrase, prepositional phrase, adverbial phrase, and so on.)
- **Governing category:** Determines, for noun phrases, whether an S or VP, is the closest superior node in the constituency tree.
- **Parse tree path:** Indicates the syntactic relation between the predicate and the argument candidate. It is represented as the path in the tree comprised of the existing phrase types and their respective directions (up or down in the tree).
- **Position:** Registers the target constituent position concerning the predicate in a sentence (before or after).
- **Voice:** Indicates whether the verb is in passive or active voice.
- **Head word:** The head word of a constituent.

Subsequent studies kept these features, adapting them when necessary to different syntactic representations. These studies also introduced new features based on the morphological information. Within a short period, the feature templates could be counted on hundreds (70, 71, 72, 73, 74, 75). Although under complementary perspectives, Surdeanu e Turmo(72), Màrquez et al.(76) and Park e Rim(77) classified the features utilized in SRL into six categories, according to their goal: Argument structure, argument context, predicate structure, predicate context, relationship modeling and sentence structure.

Argument and predicate structures are concerned with the internal syntactic structure of a candidate argument or predicate, respectively. Argument and predicate contexts represent the features intended to capture the structure of the elements surrounding the target candidate argument or predicate. Relationship modeling, on the other hand, groups the features that attempt to capture the relationship between the predicate and the candidate-argument. At last, sentence structure features are concerned with the properties shared by the constituents in a sentence.

The traditional feature engineering process is time-consuming and requires linguistic knowledge and intuition from the researcher (78). This fact inspired Collins e Duffy(79) who introduced a technique known as *Tree Kernels*. It automatically analyzes syntactic patterns in constituency trees, converting them into real-valued vectors. This embedded syntactic representation inherently carries several attributes and relationships that until then, should be explicitly captured by the researcher in a manual process. Several studies explored this technique for relationship modeling features (80, 81), argument and predicate context features (82) and even as the only syntactic representation for the SRL classifier (78). These techniques have reported reasonable results, particularly when exposed to smaller datasets. However, the time required for training is exponential what makes it infeasible when using large lexical resources.

Collobert e Weston(83), Collobert et al.(19) and Zhou e Xu(84) expanded the concept and chose not to use any syntactic structure. Instead, they utilize *distributional semantic models* that transform the words of a given vocabulary into low-dimensional real-valued vectors in a way that similar words produce similar vectors (85, 86, 87, 88, 89, 90). These word vectors are referred to as word embeddings. The theory is based on the idea that the words that co-occur in the same context tend to exhibit a similar meaning. These studies employed the word embeddings as input features of deep neural network architectures (convolutional neural networks and recurrent neural networks). Their results point out to an almost state-of-the-art performance while offering a dramatic reduction in processing time (84).

Jr e Martin(91), FitzGerald et al.(92), and Fonseca e Rosa(18) also investigated the combination of word embeddings and deep neural network models in SRL task. This time, however, they also aggregated traditional syntactic features to their systems (dependency trees). This combination also yielded strong results.

2.2.4.2 Processing Stages

The SRL may be modeled as a sequence labeling task that is arranged in a pipeline architecture. Thus, the system is usually realized as a sequence of smaller and simpler steps. In such setting, the output of the previous stage serves as the input of the next one. In this section, we describe the particular goal of each of these optional processing stages (see Figure 4).

Predicate Identification (PI) is the first and more straightforward stage in the SRL pipeline. As the name suggests, its goal is to identify the predicate tokens in a given sentence. Regarding verbal predicates, this task may be reduced to the application of a Part of Speech tagger (POSTagger) which in turn, reveals the morphologic structure of a sentence. However, nouns, depending on its usage context, may act as a predicate evoking frames or events in the same way as the verbal predicates. In this case, it is treated as a binary classification problem that indicates whether a given token is a predicate or not (93, 94).

Predicates may also evoke different senses. The predicate disambiguation (PD) stage stands for the act of discovering the correct sense of a predicate. The predicate *run*, for instance, may indicate the act of moving rapidly when applied in a sentence like *John runs fast*. It may also assume the *control* sense in a sentence such as *John runs the company*. Each sense of the predicate evokes a different semantic frame. Predicate disambiguation is typically addressed as multi-class classification problem (95, 96, 97). In the early years of SRL though, this task was not popular, and studies used to ignore it, selecting the most common sense observed in the dataset. These two tasks (identification and disambiguation) can also be fused and performed in a single predicate processing stage (PID) (98).

Semantic arguments are comprised of multiple words. For instance, in the sentence 7 the multi-word expression *last month* determines a single argument (ArgM-TMP). The argument identification stage (AI) is concerned with the detection of the predicate-argument boundaries such that a system recognizes the span of the words that constitute each argument in a sentence. Once again, a binary classification mechanism is the regular choice. It indicates whether a token or constituent in a sentence acts as an argument or not (78).

Xue e Palmer(71) analyzed the PropBank corpus and found that the majority of the words in a sentence are not arguments to a given predicate. This fact inspired the authors to create a pruning heuristic (PRU) based on a set of grammar rules that reduce the search space by automatically removing the constituents that are unlikely to be considered as arguments. The idea is to act as a pre-processing stage to the argument identification stage, alleviating the processing requirements for the SRL task. These heuristics were either developed and transposed to other languages (99, 100, 101, 102).

Each argument identified in the previous stage serve as input to the argument classification stage (AC). The goal is to choose, from a pre-defined set of possible semantic role labels, the proper one for that argument. If an approach executes argument identification and argument classification stages individually, then we have a *Two-stage* approach (AI+AC) in argument structure. Otherwise, if one performs both stages jointly, through a single pass on a classifier, we have a *One-Stage* approach (AIC). In this case, the constituents that are not arguments in a sentence are labeled with the special class NULL (19).

The argument classification stage is typically executed in a token by token or constituent by constituent procedure, implying in multiple passes through a classifier for each sentence. As classifier decisions are taken independently of one another, the same semantic role may be attributed to multiple constituents, violating constraints according to the adopted formalism.

In PropBank for instance, although optional, an argument must not be utilized more than once per predicate and if this situation occurs, it invalidates the annotation of the whole sentence. To prevent such problem, some studies adopt a global inference stage (GLOBAL). The objective is to apply a classifier that considers the best global decision for the whole proposition, minimizing eventual inconsistencies produced by the argument classification stage. This procedure what may improve the overall model accuracy. Therefore, a semantic role labeler performs a local inference if the system directly returns the argument classification results. Otherwise, if it uses a post-processing stage with these results, then it performs a global inference stage.

Punyakanok et al.(103) showed competitive results modeling the global stage as an Integer Linear Programming (ILP) problem. In such model, the authors translated the sentence-level constraints to an off-the-shelf ILP solver whose goal is to maximize an objective function considering its constraint functions. Pradhan et al.(101) relaxed the linguistic constraints and, in an attempt to maximize the best global classification, addressed the inference stage using a dynamic programming algorithm (Viterbi) which is intended to find the best sequence of tags for a sentence. Another traditional approach is to employ a re-ranker scheme where a global classifier takes the labels generated in argument classification stage and combines them with a new set of features (104).

The SRL task is tightly related to the quality of the syntactic parser and their respective representation. As expected, the errors on the syntactic parser are propagated to the semantic labeler, degrading its overall accuracy. This phenomenon is known as error propagation and motivated studies to combine the output of different parsers. The idea is that the nature of the mistakes each parser make are different, and therefore, their combination enables the classifier to learn the best representation for a given sentence, reducing the overall error.

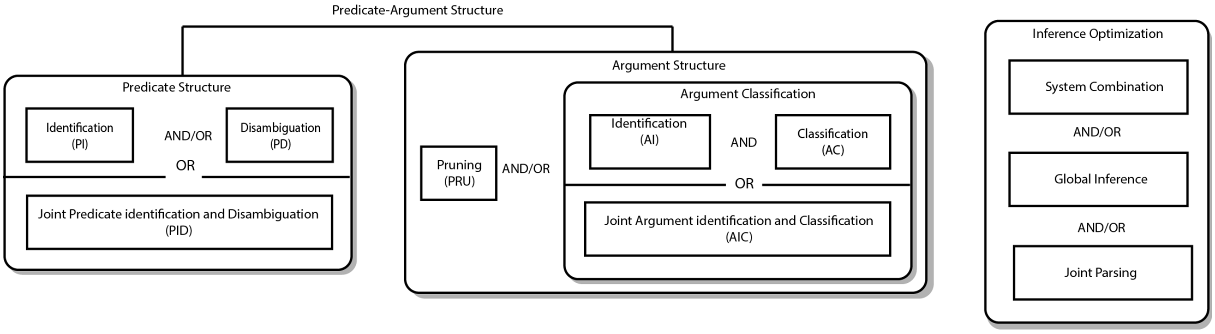
Koomen et al.(105) tested this setting, and their results point out to a significant improvement in the SRL accuracy, confirming such hypothesis. The approach though is computationally expensive and, consequently, impractical in real scenarios. This strategy is named in this work as a System Combination (SYSCOMB) and was employed in several studies (101, 76, 106, 107).

Another global strategy concerns to inferring both, syntactic and semantic labels at the same time, in a joint learning setting. The inspiration is to take advantage of the existing syntactic-semantic relationship, what may mutually benefit each task (108, 109). Several experiments have been conducted based on this hypothesis, reporting a median accuracy (110, 111). Considering the complexity of such approaches and their results, the usage of joint parsing strategy is questionable (112).

2.2.4.3 Evaluation

The SRL task is usually treated as a supervised classification problem and uses the following evaluation measures: Precision (P), Recall (R) and F-Score ($F_{\beta=1}$). They are respectively described in the equations 2.1, 2.2 and 2.3 where T_P means the number o true positives (correctly classified arguments), F_P is the number of false positives (incorrectly classified arguments), and F_N represents the number of false negatives (arguments that

Figure 4 – SRL processing stages



should be classified but were not).

$$P = \frac{T_P}{T_P + F_P} \tag{2.1}$$

$$R = \frac{T_P}{T_P + F_N} \tag{2.2}$$

$$F_{\beta=1} = \frac{(1+\beta^2)PR}{\beta^2 P + R} = \frac{2PR}{P+R} \tag{2.3}$$

The precision is the proportion of correct labels predicted out of the predictions made by the model. Recall, on the other hand, may be understood as the proportion of correct predictions from all the available arguments in the dataset. At last, the F-measure extracts the harmonic mean between precision and recall (113).

To evaluate the performance of a supervised SRL model one must compute these values for each possible semantic role, yielding a local performance measure. To compute the overall performance of the system, one must average local precision and recall. These averaged results are then used as input to the F_1-score equation that will provide the macro-averaged score of the system.

The shared tasks CoNLL-2004 (114), CoNLL-2005 (115), CoNLL-2008 (13) and CoNLL-2009 (116) have an important position toward the evaluation topic. They created a fair comparison environment for SRL task. Each of these tasks offered static groups of resources that included standardized partition sets, lexical resources, and pre-processing tools.

CoNLL-2004 (114) and CoNLL-2005 (115) were designed to employ the PropBank formalism on an early version. Their fundamental exception is that the first offered shallow syntactic parsers while CoNLL-2005 focused on constituency-based parsers.

The CoNLL-2008 (13), also focused on the English language, introduced significant differences from its previous editions. First of all, the corpus utilized a joint version of PropBank and NomBank what implies in the presence of nominal and verbal predicates. The shared task also required results for predicate identification and predicate disambiguation stages in its assessment. The major distinction though concerns the usage of dependency-based grammars as the syntactic representation offered to researchers. Although not obligatory, the authors also encourage the joint parsing of syntactic and semantic structures.

CoNLL-2009 (116) followed most of the settings from the previous version. This time, however, instead of working only with the English language, the task introduces a multilingual setting covering the following languages: Catalan, Chinese, Czech, English, German, Japanese and Spanish. The F_1-score achieved in each language is averaged, designating the overall accuracy achieved by a participant paper. The predicate structure investigation was also modified, and this time it only requires investigation on predicate disambiguation stage.

2.3 Methods

This systematic literature review observes the general guidelines proposed by Levy e Ellis; Moher et al.(117, 118) and is divided into the following steps: The definition of the research questions, the search strategy, the paper screening procedure, and classification.

2.3.1 Research questions

The research questions are crucial elements while conducting a systematic literature review. They are responsible for the scope, guiding the research choices in multiple levels (117). The Table 5 lists the research questions of the present study. The primary aspiration of these questions is to provide the understanding of how scientific literature addresses the SRL task.

Table 5 – Research questions

ID	Research Question
RQ1	What aspects of SRL techniques impact its accuracy?
RQ2	What are the most accurate studies reported in literature?
RQ3	How often has each processing stage been employed in SRL?
RQ4	Which are the machine learning techniques utilized in SRL?
RQ5	Which are the dominant syntactic representations in SRL experiments?

The goal of **RQ1** is to determine which factors (syntactic representation, learning approach, syntactic views, target languages) may impact the SRL accuracy. The **RQ2**, on the other hand, intends to identify the state-of-the-art accuracy of SRL studies reported in the literature. Thereby, in both questions, our analysis took into account the measures and procedures described in section 2.2.4.3.

The **RQ3** aims to identify how often each processing stage is utilized. In our analysis, we considered the processing stages listed in section 2.2.4.2.

The **RQ4** seeks to verify tendencies and research gaps regarding ML approaches in SRL. This analysis considered the ML taxonomy presented in section 2.2.3.

Finally, **RQ5** serves the purpose of identifying eventual trends in the adoption of syntactic representations in SRL research. The analysis took into account the syntactic representations presented in section 2.2.4.1.

2.3.2 Search Strategy

As a strategy in the search for papers, we first created the search strings. We later defined the online databases to be used and their respective retrieval settings. At last, we determined the inclusion and exclusion criteria on the search results. These steps are detailed in the following subsections.

2.3.2.1 Search Terms

The search terms adopted in this work are presented in the Table 6. With them, we attempted to cover the most frequent terms and its corresponding variations employed in the literature: *Semantic role labeling*, *Shallow semantic parsing*, *Semantic parser*, *Semantic labeler* (3, 119). The search terms also comprise suffix variations such as parser and parsing, labeler and labeling, and so on.

Table 6 – Conceptual search string used in data retrieval procedure

Conceptual search string
"semantic role*" OR "semantic pars*" OR "semantic label*" OR "shallow pars*"

2.3.2.2 Data retrieval

In this study we chose to work with the following digital libraries: *ACM Digital Library*[9], *IEEEXplore*[10], *Science Direct*[11] and *Springer Link*[12].

[9] <http://dl.acm.org>

[10] <http://ieeexplore.ieee.org/Xplore/home.jsp>

[11] <http://www.sciencedirect.com>

[12] <http://link.springer.com>

The data retrieval process consisted in querying each selected digital library for the terms mentioned in the conceptual search string (Table 6). The results were then stored in a reference management application named EndNote[13].

2.3.2.3 Selection Criteria

This literature review covers experiments published in journal articles and conference papers from the previously mentioned digital libraries, written in the English language, which had reported results for the SRL task. Furthermore, we seek papers published in the period comprised from 2002, the year of the first successful approach for automatic SRL (3), to August 2016, when we conducted the search.

Books, thesis, dissertations, posters, letters, interviews, newspaper articles and other publication means were not considered in this review. We also excluded duplicated articles or those that do not mention the usage of SRL technique.

2.3.3 Screening of Papers

The Figure 5 illustrates the procedures that have been followed to determine the articles contained in this systematic literature review. The search in online databases resulted in a total of 2584 papers that ranked according to their origin as follows: *ACM Digital Library* (1011 or 39.12%), *SpringerLink* (859 or 33.23%), *IEEEXplore* (584 or 22.60%) and *Science Direct* (130 or 5.03%). It is worth mentioning that we directly applied, in each of the target search engines, selection criteria such as the publication period and publication mean.

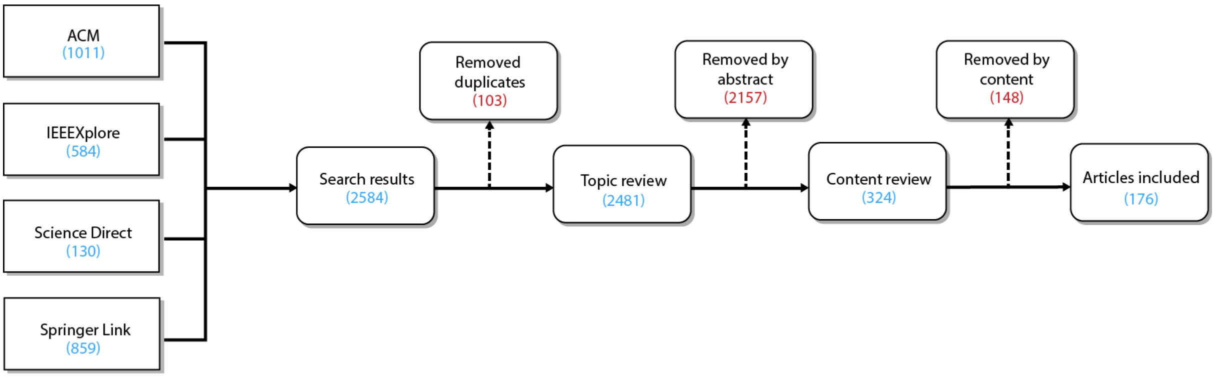
With this initial set of papers, we proceeded to the removal of duplicate references found when joining the results of the aforementioned digital libraries. To do so, we applied the function *Find Duplicates* of EndNote software, accepting its suggestions on which version of each duplicate document should be discarded. In this stage, 103 references (or 3.98% of the total) were excluded resulting in 2481 papers for the sequence.

The following stage consisted of a topic filtering procedure when we read the title and abstract of each one of the 2481 remaining papers. The goal was to identify papers noncompliant with the topic defined in the selection criteria. In this stage, we excluded 2157 articles (86.94%). This amount of exclusions is due to the broad coverage of the terms that we utilized in the search string. Semantics is an important topic discussed in a range of research fields such as linguistics, computer vision and even in neurology. For this reason, the results contained a considerable amount of papers that did not fit the selection criteria.

[13] See <http://endnote.com>

The 324 articles left were the subject of further analysis, through a full read of its content. In this last stage, we seek to verify the compliance of its content with the selection criteria. In this step, we removed another 148 articles (5.96%) resulting in a total of 176 papers included in this review (7.09%). The amount of excluded papers in this last stage is due to the selection criteria that requires that the experiment report results for the SRL task.

Figure 5 – The procedures followed in the screening of papers



2.3.4 Data Extraction

The data extraction procedures adopted in this study were based in (120) and have been adjusted to our research questions and methods. The 176 articles included in the review were read in full twice, and both text and meta-data used to classify the papers in line with the defined classification.

2.3.5 Classification

The categories analyzed in this study were: *Experiment accuracy*, *evaluation set*, *processing stages*, *machine learning approach*, *machine learning algorithms*, *syntactic structure*, *formalism*, and *target languages*. During the data extraction, papers assumed a single value in the following categories: *evaluation set*, *machine learning approach* and *syntactic structures*. The remaining categories allowed the assignment of multiple values for the same paper.

The categories *experiment accuracy* and *evaluation set* were collected from a subset of selected papers considered comparable. By comparable, we admit the papers that participated or adhered to a test set proposed by shared tasks such as CoNLL or SemEval. This choice is due to the static nature of the lexical resources offered by these events, which facilitates the comparison process (as explained in section 2.2.4.3). Lexicons such as PropBank and FrameNet are dynamic and have been augmented in time with new releases and versions that updated its coverage and corrected eventual mistakes. Dynamic evaluation sets hamper the comparison since they may create an unfair ground among experiments carried in different epochs. We also considered as comparable only the experiments that utilized evaluation sets from shared tasks that were used by more than one participant. For this reason, the four SemEval shared tasks identified in our study were excluded from this subset since each of them accounted to one experiment.

The *experiment accuracy* is pointed by its F_1*-score* (more detail in section 2.2.4.3). As we are concerned with automatic SRL, results derived from approaches based on gold-standard syntactic resources were discarded. The argument is that this is an inexistent condition in real-world usage. If an article reported accuracy to multiple comparable evaluation sets, the one with the highest F_1-score was collected. If the study presented results for various languages, the average value was taken as the overall accuracy of the study (following the procedures adopted by (116) in CoNLL-2009). We also registered the score reached for each language, individually. If an experiment presented corrected results after the submission to a shared task, we considered the corrected ones. The *evaluation set* category indicates the shared task utilized by the experiment. For the remaining categories in this classification, the *comparable* constraint is not applicable.

The *processing stages* category is comprised of the stages presented in the section 2.2.4.2. We collected this information in all the experiments that have reported its usage. If an article reported results for multiple framework configurations, we collected those employed in the most accurate one.

The *machine learning approach* category concerns to the learning type described by each paper (supervised, unsupervised, and so forth). The *machine learning algorithms* category was collected whenever the article mentioned its usage in one of the processing stages. If a given article tests different ML algorithms in a processing stage, the algorithms that participated in the most accurate approach are collected. The taxonomy applied in such categories was exhibited in section 2.2.3.

The *syntactic structure* category encompass the syntactic representations introduced in section 2.2.4.1. We classified papers in this category whenever a paper explicitly reports the usage of a syntactic view. If a paper reported results to multiple syntactic views, we collected the ones utilized by the most accurate approach.

The employed *formalism* and the *target languages* were collected for all papers that have reported its usage.

2.4 Results

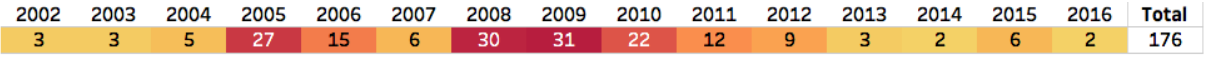
This section summarizes the descriptive results of this study.

The Table 18 (in Appendix A) lists the 176 papers included in this systematic literature review. Conference papers responded for a share of 85.22% (150) while journal articles kept the remaining 14.78% (26) from the selected studies.

During the data extraction procedure, the publication year was considered useful to understand how active is the research on SRL task. Figure 6 shows the number of collected experiments per year. One may observe that the period comprised between the years 2005 and 2012 exhibited a higher research activity. This fact may relate to the occurrence of shared tasks such as CoNLL and SemEval, dedicated to SRL. Considering the whole period, the average number of experiments per year is 11.73 while its median amounts to 6. Notice that we collected data on August-2016 so, the last year is likely to be incomplete.

In regards to the comparable subset of papers, following our method, this study identified 74 experiments (or 42.04% from the total), distributed according to the test set as follows: CoNLL 2004 (2), CoNLL 2005 (37), CoNLL 2008 (20) and CoNLL 2009 (15).

Figure 6 – Number of collected experiments per year



2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	Total
3	3	5	27	15	6	30	31	22	12	9	3	2	6	2	176

2.5 Discussion

This section describes and discusses our findings from the data extraction and classification procedures in the context of the research questions.

2.5.1 RQ1 - What aspects of SRL techniques impact its accuracy?

As the question is concerned with the factors of accuracy, in this section we exclusively utilized the comparable subset of the selected papers (74 papers).

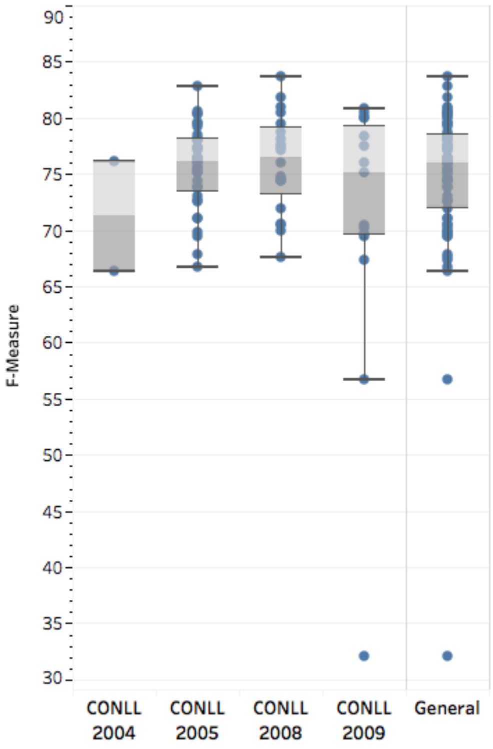
Figure 7 aims at examining the impact of the existing differences among the evaluation sets on the overall accuracy of SRL experiments (CoNLL shared tasks described in section 2.2.4.3). The visual analysis suggests that the test set choice does not cause a significant impact on F_1-score. To test such hypothesis, without assuming that our sample is derived from a normally distributed population[14], we conducted a Kruskal-Wallis H test[15] which failed to reject the null hypothesis ($\textit{chi-squared} = 2.3697$, $d.f. = 3$, $\textit{p-value} = 0.4993$) which states that the samples came from populations with identical locations. In other words, one may say that, considering our sample, the test set choice does not affect the SRL accuracy what is contrary to common sense. The differences among the settings of each shared task (section 2.2.4.3) led us to believe that they would produce different results. For this reason, from now on, we analyze the comparable experiments

[14] See the F_1-score Q-Q plot in Appendix A

[15] A nonparametric test also known as *One-way ANOVA on ranks*

disregarding the evaluation set utilized. When analyzing the general F_1-score distribution on SRL we found that the median is ≈ 76, while the standard deviance is ≈ 6.8 and the mean is ≈ 75.71.

Figure 7 – The F_1-score collected in comparable experiments grouped by test set

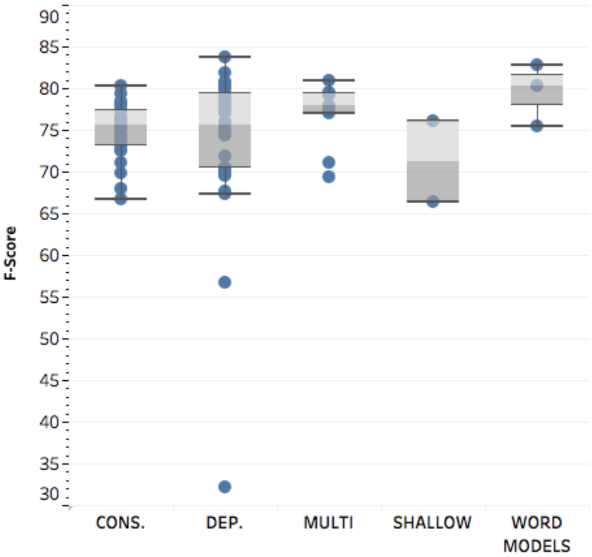


Regarding learning type category, all comparable studies were based on supervised approaches with the exception of Deschacht e Moens(121) that employed a semi-supervised approach based on a distributional semantic model. A comprehensible result particularly considering the nature of the shared tasks. Another point of view is explored in Figure 8 which analyzes the F_1-score distribution regarding the syntactic representations adopted by each comparable experiment. One may observe that dependency trees were the most popular choice among the comparable experiments with 34 occurrences, while 26 adopted the constituency-based representation. The syntactic combination strategy[16] was utilized on nine occasions while word embedding models accounted for three experiments. Finally, two studies chose shallow syntactic parsers. We verified the impact on accuracy caused by the adoption of different syntactic representations. In this sense, we performed a Kruskall-Wallis H test, which failed to reject the null hypothesis (*chi-squared* $= 5.8655$, *df* $= 4$, *p-value*$= 0.2094$), and therefore, regarding our sample, no syntactic representation causes a statistically significant difference on location shifts. The result indicates that the simple choice of a syntactic representation is not enough to produce a significant accuracy gain in an SRL model. This finding, however, contrasts with the

[16] See Table 18 in Appendix A for the stratified view on syntactic combinations

findings of (19, 122, 11, 101) that achieved an improved accuracy by combining multiple syntactic representations. The main reason, for this divergence, may be related to the fact that these studies relied only on their local models while our method analyzed several studies at once.

Figure 8 – Syntactic View



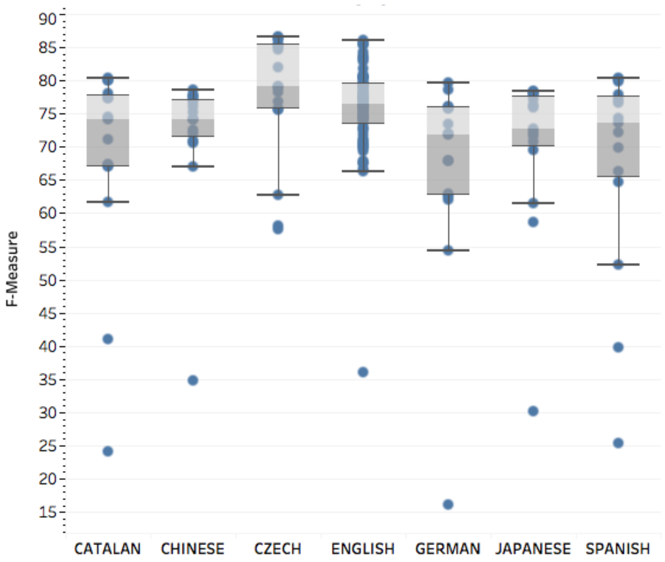
The target language of each experiment is another point addressed in our work (see Figure 9). All the 74 comparable experiments reported results to the English language. From this number, we identified 15 multilingual studies, which reported results for Catalan, Chinese, Czech, English, Spanish, German and Japanese following the CoNLL-2009 shared task settings. We also notice that all multilingual studies were based on purely supervised approaches. The Kruskall-Wallis H test (*chi-squared* = 21.874, *d.f.* = 6, *p-value*= 0.001276) followed by a multiple comparison post-hoc test [17] revealed that the difference in location shifts are statistically significant when comparing the German experiments with English or Czech based studies. As expected, our results corroborate the notion that the accuracy in SRL, particularly for the techniques based on supervised learning, is dependent on the researcher ability to capture features that translate linguistic aspects into features for the model (78). Given its characteristics, some languages are harder to parse than others. Previous studies targeted at the Chinese language, for instance, attempted to employ the same feature templates utilized for the English language and obtained a degraded performance (124, 125, 102).

2.5.2 RQ2 - What are the most accurate studies reported in literature?

As this question concerns the accuracy in SRL task, we exclusively analyzed the comparable subset of the selected papers (74 papers).

[17] Non parametric test described by (123) and implemented in the *pgirmess* package for R language

Figure 9 – The F_1-Score distribution grouped by target language



The Table 7 ranked the ten most accurate experiments, elucidating some of its fundamental characteristics. The most accurate approach achieved an F_1-score of 83.75%, allowing a good margin for improvements in future studies. We highlight two main points on the rank: First, dependency-based experiments dominates the rank with eight occurrences, followed by word embeddings with three appearances. Constituency-based studies, on the other hand, exhibits only one entry in the rank. Second, seven out of ten studies in the rank utilized a global inference stage what may indicate that its presence is essential for a state-of-the-art accuracy.

Table 7 – The 10 most accurate experiments reported in literature

Experiment	F_1-Score	Learning strategy	Target language	Syntactic view	Framework
Johansson e Nugues(94)	83.75	Supervised	English	Dep.	PI+PD+AI+AC+GLOBAL
Zhou e Xu(84)	82.84	Supervised	English	No Synt.(WE)	AIC+GLOBAL
Lim, Lee e Ra(126)	81.87	Supervised	English	Dep.	PI+PD+AIC+GLOBAL
Deschacht e Moens(121)	80.98	Semi-supervised	English	WE+Dep.	AIC
Björkelund, Hafdell e Nugues(127)	80.80	Supervised	Multilingual	Dep.	PD+AI+AC+GLOBAL
Johansson e Nugues(94)	80.61	Supervised	English	Dep.	JP(PD+PRU+AI+AC+GLOBAL)
Zhao, Chen e Kit(128)	80.53	Supervised	English	Dep.	PD+PRU+AIC
Zhao et al.(129)	80.47	Supervised	Multilingual	Dep.	PD+PRU+AIC
Toutanova, Haghighi e Manning(130)	80.32	Supervised	English	Cons.	AI+AC+GLOBAL
FitzGerald et al.(92)	80.30	Supervised	English	No Synt.(WE)	PI+AIC+GLOBAL

Syntactic View - *Dep.*:Dependency trees. *Cons.*: Constituency Trees. *WE*: Word Embeddings. *No Synt.*: No Syntactic View.
Framework - *PI*: Predicate Identification. *PD*:Predicate Disambiguation. *AIC*: Joint argument identification and classification. *PRU*: Pruning. *AI*: Argument Identification. *AC*: Argument Classification. *GLOBAL*: Global Inference *JP*: Joint Parsing

2.5.3 RQ3 - How often has each processing stage been employed in SRL?

We identified the processing stages in 159 out of the 176 selected experiments. Concerning the predicate structure (see Table 8), we observe that while only 36.4% of the experiments performed at least one of its stages, the disambiguation stage is the most frequently used. The argument structure (presented in Table 9) on the other hand, was explored in 94.5% of the experiments. In this sense, we highlight the low representativeness of the pruning stage when compared to the other stages (which are mutually exclusive). Concerning the inference optimization strategies (see Table 10), we observe that the Global inference stage championed the rank, while almost a half of the experiments attempted to improve their results by employing at least one of such strategies (46.5%).

Table 8 – Predicate Strcture

Stage	# Experiments (%)
PI	33 (20.7%)
PD	40 (25.1%)
PID	2 (1.2%)
Total	58 (36.4%)

Table 9 – Argument Structure

Stage	# Experiments(%)
AI+AC	89 (55.9%)
AIC	63 (40.6%)
PRU	45 (28.3%)
Total	151 (94.9%)

Table 10 – Inference Optimization

Stage	# Experiments(%)
Global inference	52 (32.7%)
Joint Parsing	15 (9.6%)
System Combination	13 (8.1%)
Total	74 (46.5%)

With the intention of elucidating how the processing stages are arranged together, our study also collected the configuration presented in each of the experiments (identified in 155 out of the 159 studies in this section). Table 11 ranks the ten most popular pipelines configurations. When combined, the rank responds to more than 60% of the total number of pipeline configurations reported in the literature. One may notice that simple frameworks, which investigated only the argument structure, are the most popular choice in literature. The global inference stage is also widespread.

2.5.4 RQ4 - Which are the machine learning techniques utilized in SRL?

Our study identified the learning strategy for the 176 selected papers. The machine learning algorithms have been identified in 162 occasions.

Table 11 – The ten most frequently addressed frameworks in literature

Framework	# Exp.(%)
AIC	17
AI+AC	16
AI+AC+GLOBAL	14
AIC+GLOBAL	10
PRU+AI+AC	8
PRU+AI+AC+GLOBAL	8
PD+AI+AC	5
PD+AIC	5
PI+AI+AC	5
PRU+AIC	4
Total	94 (60.6%)

Regarding the learning method, the selected papers were classified as follows: Supervised (147), semi-supervised (17) and unsupervised (12). One may observe that, while there is an expressive preference for purely supervised approaches (83% of the studies), no studies utilizing reinforcement learning have been identified what may represent a research gap. Maximum Entropy (55), Support Vector Machines (46), Conditional Random Fields (8) and Neural Networks (8) are the most popular machine learning algorithms employed in the SRL task, being utilized in 72.22% of the approaches. All these algorithms are considered discriminative classifiers. Table 12 stratifies the machine learning algorithms when applied on the argument structure, demonstrating that this pattern is also repeated.

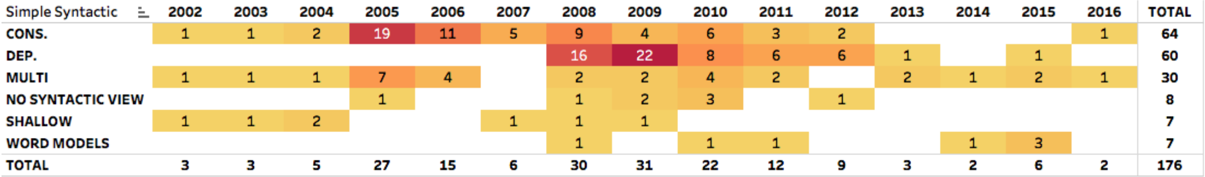
Table 12 – The frequently utilized algorithms in argument structure

AIC	AI	AC
Maximum Entropy (20)	SVM (38)	SVM (36)
SVM (12)	Maximum Entropy (24)	Maximum Entropy (25)
Neural Networks (6)	Logistic Regression (5)	Logistic Regression (5)
ADABOOST (3)	CRF (3)	CRF (3)
IB1 (3)	Snow Learning (3)	Snow Learning (3)

2.5.5 RQ5 - Which are the dominant syntactic representations in SRL experiments?

In this study, we were able to identify the syntactic representation employed in all 176 selected experiments. Figure 10 presents the adoption of each syntactic view grouped by year. One may observe that constituency-based representation was slightly more frequent than the dependency-based representation. We also notice that when combined, constituency and dependency representations responds for 70.45% of the total. We also observe that constituency-based views were most common in the first years, reaching its peak in 2005 and gradually losing relevance. Dependency-based representations, in turn, reached its peak in 2008 and 2009. In recent years, though, distributional word models and the combination of multiple syntactic views have become the dominant representations.

Figure 10 – The adoption of each syntactic representation per year



Simple Syntactic	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	TOTAL
CONS.	1	1	2	19	11	5	9	4	6	3	2				1	64
DEP.							16	22	8	6	6	1		1		60
MULTI	1	1	1	7	4		2	2	4	2		2	1	2	1	30
NO SYNTACTIC VIEW				1			1	2	3		1					8
SHALLOW	1	1	2			1	1	1								7
WORD MODELS							1		1	1			1	3		7
TOTAL	3	3	5	27	15	6	30	31	22	12	9	3	2	6	2	176

2.6 Limitations and Threats to Validity

We focused our automated search on four digital libraries in the computer science domain. However, it is possible that such databases do not cover all the relevant papers in this field of study. Besides that, our research only covers material written in the English language published before August-2016, when we performed the search on databases. It is probable that there are relevant studies published in other languages or even new ones, published after our search for papers.

2.7 Conclusions and Future Work

This article summarized the state-of-the-art of SRL task through a systematic review process. We selected 176 experiments from an initial set of 2584 articles. Our method revealed that the best overall accuracy reported for SRL is around 83%. The following possible factors have been analyzed regarding the accuracy:

- The adoption of different evaluation sets do not cause a significant impact on SRL accuracy.
- The choice of syntactic representations does not produce relevant impact on SRL accuracy.
- SRL performance is language-dependent and some languages are significantly harder for SRL systems.

Our study was not able to identify papers reporting the use of reinforcement learning algorithms, what constitutes a research gap to be addressed in future investigations.

Concerning the adoption of syntactic representations, we found that constituency based views have lost its significance with time. In this sense, distributional semantic models and the combination of multiple syntactic representations are the dominant syntactic choice in recent years.

3 Results

3.1 Introduction

Semantic role labeling (SRL) is a natural language processing task (NLP) whose goal is to capture and represent the core structures - participants and circumstances - of events or situations typically expressed in human languages. These event structures are revealed by providing answers to a question such as *who did what to whom, where, when and how*. Formally, the task is to determine the semantic role played by each argument of the predicates in a sentence. Sentence 8 illustrates the concept:

(8) [Throughout his life$_{When}$], [Einstein$_{Who}$] [published$_{did}$] hundreds of [articles$_{What}$]

SRL is usually employed as an intermediary technique and is considered a significant step towards the natural language understanding (11). Many previous studies had proven its utility in a wide range of NLP tasks such as question and answering systems (21, 9), text summarization (5), open information extraction (6), machine translation (7), and co-reference resolution (8), to cite a few.

However, this is a challenging task. The model's performance depends on its ability to deal with language aspects such as syntactic alternations, selectional restrictions, and ambiguity. Even when considering the English language - the most addressed language in the literature - the performance is still around 83 points (94, 131). This fact illustrates how hard is the task and how much room for improvement there is in this field. Regarding the Portuguese language, the research is still incipient. There is a scarcity of resources and publicly available tools what hampers research and, consecutively, detain the innovative process for the language.

Earlier approaches employed statistical machine learning methods that relied on the extraction of complex morphosyntactic features and a series of declarative constraints (11, 132, 133). In contrast, for the past five years, there has been a rapid rise in the usage of neural networks, exploring its feature induction capabilities. It reduces the overall complexity while achieving competitive results(19, 84, 131). In this context, Recurrent Neural Networks have been receiving much attention, and recent inquiry demonstrated that its natural ability to articulate long-range dependencies is particularly beneficial for several NLP tasks (134, 135, 136). This network architecture though has not yet been tested in the SRL task for the Portuguese language.

In this paper, we present an end-to-end semantic role labeler for the Portuguese language that outlines the problem as a supervised sequence labeling task. The one-step system uses the IOB tagging schema and applies a word embedding model in a deep bidirectional long short-term memory neural network (deep BiLSTM). The network predictions serve as input to an inference mechanism that uses a global recursive neural parsing algorithm, specifically tailored for the task. We also provide a detailed investigation of the effects of word embedding dimensionality and network depth on the overall performance of the system. Our method requires a minimal feature engineering process and does not depend on syntactic parsing. The resulting system improved the previous state-of-the-art for the Portuguese language in around 3 points using the PropBank-Br corpus (15), reducing the relative error in 8.74%. Moreover, our investigation revealed that choosing proper word embedding dimensionality and network depth are critical aspects for optimizing the system's performance.

- Our system pushes the previous state-of-the-art for the Portuguese language in around 3 F_1-score points, reducing the relative error in 8.74%.
- To the best of our knowledge this is the first study to apply a recurrent neural network architecture on the SRL task for the Portuguese language.
- We provide the tool support for the SRL task in the Portuguese language. The system was released under BSD license and its source code is available at <https://github.com/dfalci/deep_pt_srl>.

This paper is organized as follows: section 3.2 provides the background for the task, exposing some fundamentals and the relevant prior work in this field. Section 3.3 presents the PropBank-Br corpus, used in our experiment. Section 3.4 depicts our approach, detailing the relevant parts of our system. Section 3.5 shows the experimental setup while section 3.6 reports our results, followed by discussion. At last, section 3.7 concludes our work.

3.2 Background

In this section, we briefly describe some fundamentals of the semantic role labeling task and discuss the relevant prior work.

3.2.1 Semantic Role Labeling

From a linguistic perspective, the goal of the task is to discover the existing predicate-argument structures in a sentence that, in this work, are referred to as *propositions*. A predicate, in this case, is represented by the words, usually verbs, responsible for evoking semantic frames that in turn, triggers the description of an event or situation. The verbal predicate *bought* in the sentence *John bought a new house* evokes the purchase frame that requires a particular set of optional semantic arguments such as the buyer, the seller, the goods, and the price paid. In humans, these relationships are implicitly inferred considering the individual characteristics such as linguistic knowledge and past experiences. Only in the presence of these semantic arguments, an event may produce a minimal unit of meaning in the interlocutor.

From a machine learning point of view, SRL task is usually treated as a supervised classification problem. The task is to choose, from a pre-defined set of possible semantic role labels, the proper ones for each token or constituent with respect to a given predicate.

Models may be classified according to its approach to the problem. Some authors treat the task as an atomic unit, where final predictions are obtained after a single pass of each token or sentence segment into a classifier. In this one-step framework, an exceptional *null* semantic role indicates the absence of semantic attachment of a candidate argument with the predicate. On the other hand, there are authors who argue that considering its high complexity, the task must be decomposed into a sequence of smaller and specialized steps arranged in a pipeline where the output of the previous step is an input component of the following one. Two of the most frequent steps used in these pipelines are the argument *identification* and *classification*. The former performs a binary classification that indicates whether a sentence segment act as an argument or not. Then, the latter predicts the semantic role for each item positively labeled by the identification step. This pipelined approach is known as a *two-step framework* and which approach is better is still a matter of debate in scientific inquiry.

As usual in any supervised problem, the task requires a representative annotated lexical resource, built considering a specific formalism. For the English language, PropBank (14) is the most used corpus and, it adds a semantic layer on top of syntactic trees annotated by *Penn TreeBank*, following the theory presented by Levin(10) and Dang et al.(37). Sentences 9 and 10 illustrates its annotation formalism.

(9) [The stock's $_{A1}$] [accelerated $_{V}$] [from a price of \$8 a share $_{A3}$], reaching [its peak at \$10 $_{A4}$].

(10) [Jobs $_{A0}$] [built $_{V}$] [the Apple I $_{A1}$] [in a garage $_{AM-LOC}$].

In sentence 9, the predicate *accelerated* evokes the *accelerate* frame, with the *acceleration* sense (mapped in PropBank frame files[1]), that in spite of its many syntactic configurations, expects the following set of core semantic arguments: the agent (*A0*), the thing accelerating (*A1*), the extension (*A2*), the start point (*A3*), and the end point (*A4*). At the same time, the predicate *built* at sentence 10 expects arguments such as the

[1] PropBank frame files are described in <https://github.com/propbank/propbank-frames/tree/master/frames>

builder (*A0*), the construction (*A1*), the materials (*A2*), and the end state (*A3*). Note that each predicate assigns a different meaning to the core labels (*A0-A5*), and, with the exception of the arguments *A0* and *A1* that usually designate the agent and the patient of an action, no inferences can be made about the meaning of the other roles. There is also a set of optional adjunct arguments (AM-X) that are shared by all predicates. They modify the proposition adding information such as location, time, and manner[2].

To evaluate the performance of a supervised SRL model one may apply standard measures such as precision, recall, and f-measure, described in equations 3.1, 3.2, and 3.3, respectively. The measurement is made for each semantic role, yielding a local precision and recall. To compute the overall performance of the system, one must average local precision and recall. These averaged results serve as input to compute the F_1-score of the system (macro-averaged F-score), which indicates the harmonic mean between the averaged precision and recall.

$$P = \frac{correct\ positive}{all\ positive\ tagged} \tag{3.1}$$

$$R = \frac{correct\ positive}{all\ existing\ positive} \tag{3.2}$$

$$F_{\beta=1} = \frac{(1+\beta^2)PR}{\beta^2 P + R} = \frac{2PR}{P+R} \tag{3.3}$$

3.2.2 Related work

Collobert et al.(19), pioneered the usage of neural networks for the SRL task. Their one-step system (*SENNA*) uses a convolutional neural network over windows of words in a proposition to infer its semantic labels. An inference stage based on a Viterbi decoder determines the final tags from the network output. Differing from the previous approaches, this system uses a word embedding model in replacement of traditional syntactic features, hitherto widely employed. Their results point to a reasonable performance while dramatically reducing the processing time. The SENNA approach also inspired subsequent studies that expanded the investigation by adding syntactic features based on dependency trees and morphological information what significantly boost the model's accuracy (91). While convolutional layers allow one to work with arbitrary large vectors capturing their most relevant features, it sacrifices most of the structural information on an input sequence and therefore, are not the best way to capture long-term relationships. In other words, this architecture does not preserve the input order in sequential data such as text and only encompasses words inside of its sliding window context (138).

[2] The full list of semantic arguments in PropBank is defined in (137)

Considering these aspects, Zhou e Xu(84) proposed a deep BiLSTM model that does not resort to syntactic features. Their approach, also based on word embeddings, outperformed previous studies based on syntactic features in the English language using CoNLL-2005 and CoNLL-2012 data sets. Unlike this work though, they employed a Conditional Random Field (CRF) layer at the inference stage and did not investigate the effects of word embedding dimensionality on the model's performance. In a similar architecture, Wang et al.(139) added feature templates based on part-of-speech information and produced a state-of-the-art labeler for the Chinese language. He et al.(131) also proposed a BiLSTM architecture for the SRL task on the English language. This time though, the focus was at network initialization, hyper-parameter optimization, and in the incorporation of recent training techniques such as highway connections and recurrent dropout. Our system may be seen as a hybrid of these approaches, considering their useful observations and experiences.

Regarding the Portuguese language, Alva-Manchego e Rosa(17) proposed a preliminary architecture for the SRL task. Their supervised approach, dependent on morphosyntactic features, consists of a pipeline that uses Naive Bayes and Decision Trees as classifiers. The system though uses an early version of PropBank-Br (v1.0) and, most importantly, relies on golden syntactic trees provided by the corpus, an inexistent condition under real-world circumstances. These factors prevent a direct comparison with our results.

Fonseca e Rosa(18) were the first to provide a fully automated semantic role labeler for the Portuguese language. The system (NLPNET[3]) was trained on PropBank-Br v.1.1 and is heavily based on SENNA's approach. The most fundamental difference between these systems regards the number of stages they employed: As mentioned before, SENNA utilizes a one-step approach while NLPNET, after evaluating the on-step strategy, adapted its architecture transforming it into a two-step pipeline. Experiments were also conducted to verify the impact of the addition of syntactic chunks to the feature templates used in the original system. Their best single training session, due to data scarcity, yielded 65.13 F_1-score points, an overall performance far inferior compared to that of SENNA (a margin of 10 F_1-score points). To the best of our knowledge, this is the only functional SRL system for the Portuguese language whose source code is publicly available and, therefore, is referred throughout this work as our baseline system.

Hartmann, Duran e Aluísio(140) compared both approaches (the preliminary approach of Alva-Manchego e Rosa(17) and Fonseca e Rosa(18)) in a hybrid lexicon, specifically created for this task. This new corpus incorporated two subsequent versions of PropBank-Br. Their main goal was to evaluate the accuracy of these systems under revised and non-revised syntactic trees using a larger and balanced corpus for the Brazilian

[3] Available at <http://nilc.icmc.usp.br/nlpnet/>

Portuguese. Their results indicate that NLPNET systematically yielded an inferior performance when compared to the system of Alva-Manchego e Rosa(17). Our results cannot be compared to this system since the new corpus is not publicly available.

Semi-supervised learning has also been investigated in the Portuguese language. Alva-Manchego e Rosa(141) proposed an architecture based on a self-training strategy in a three-stage pipeline which uses maximum entropy classifiers. In this paper, however, the authors focused on the discussion of topics such as data preparation, feature extraction, and methodology, without providing practical results. Carneiro et al.(142) materialized the self-training strategy. This article though, considered just three commonly used verbs in Portuguese language (*give*, *say*, and *do*). Their results point that a supervised method must be exposed to over at least 40% more labeled arguments to achieve a comparable performance level, a promising observation considering the limited size of the PropBank-Br corpus.

3.3 The PropBank-Br Corpus

In this section, we present the corpus used in our experiment and describe the procedures that have been executed on it.

Inspired by its English counterpart (14), PropBank-Br (15) is a Brazilian Portuguese training corpus specifically designed for the SRL task. It was built on top of the Brazilian portion of *Bosque*, a section of *Floresta Sintá(c)tica* treebank. The corpus, in its version 1.1[4], contains 6,142 propositions distributed in 4,213 sentences extracted from Brazilian newspapers. The corpus follows in large part, the annotation guidelines employed by its original version.

When compared to other languages though, PropBank-Br may be considered a small-sized corpus, providing 13,138 annotated roles against the 95,438 registered by the English version - a number 7.2 times smaller. Table 13 lists the semantic roles mapped by the Portuguese version sorted by the number of occurrences in the corpus. Note the bias in favor of the core arguments (*A0*, *A1*, and *A2*) that, when combined, cover more than 70% of the total number of labels. On the other hand, the last three labels are practically insignificant, appearing less than ten times. Such bias is undesired and hurts generalization capacity of models based on machine learning.

The following pre-processing operations where performed on the corpus:

- The original corpus unpacked all words formed by prepositional contraction (i.e.: dele = de + ele), as opposed to what is routinely practiced in the Portuguese

[4] Downloadable in the CONLL format at <http://www.nilc.icmc.usp.br/portlex/index.php/en/projects/propbankbringl>

Table 13 – Semantic role distribution in PropBank-Br

Label	# of roles	% of total
A1	5061	38.52%
A0	2891	22.00%
A2	1290	9.82%
AM-TMP	1082	8.24%
AM-LOC	672	5.11%
AM-MNR	384	2.92%
AM-ADV	346	2.63%
AM-NEG	322	2.45%
AM-DIS	288	2.19%
AM-PRD	169	1.29%
AM-PNC	143	1.09%
AM-CAU	141	1.07%
A3	139	1.06%
A4	111	0.84%
AM-EXT	74	0.56%
AM-DIR	13	0.10%
AM-REC	8	0.06%
AM-MED	3	0.02%
A5	1	0.01%

language, even in formal writing. For this reason, we have re-constructed these connections.

- Using character '_', PropBank-Br artificially concatenates tokens used in multiword nouns such as organization names (i.e.: Secretaria Municipal = Secretaria_Municipal). We have identified and broke these nouns.

- To preserve argument contiguity, we excluded all the sentences that contained at least one predicate whose arguments were mapped as continuation roles (*C-ARG* roles)[5]. Thus, 536 propositions were removed from the training corpus.

- Six propositions presented overlapping labels what contradicts the constraints imposed by the PropBank formalism. These propositions were considered labeling mistakes and were also removed from the training corpus.

In order to produce a fair comparison ground, we divided the corpus using the same proportion pointed out by baseline Fonseca e Rosa(18). Thus, after shuffling propositions, 95% of them (5320) were used for training our model while the remaining 5% (280) were used for evaluation purposes.

[5] Continuation arguments indicate that a given sentence chunk acts as a continuation element of the sense of another, as long as both of them are separated by other arguments

3.3.1 IOB Conversion

As mentioned earlier, our approach uses the IOB format to express semantic roles. Formally, this schema represents non-overlapping sentence chunks where each of them delimits an argument boundary for a given predicate. Practically, the tokens that are not part of an argument are tagged with the outside label (*O*). Otherwise, words within the boundaries of an argument of type *X* are mapped as follows: The first word receives the begin-of-X tag (*B-X*), and the remaining words inside the same argument structure are labeled with inside-of-X tag (*I-X*). Figure 11 illustrates the differences between IOB annotation schema and the constituent based annotation for SRL task.

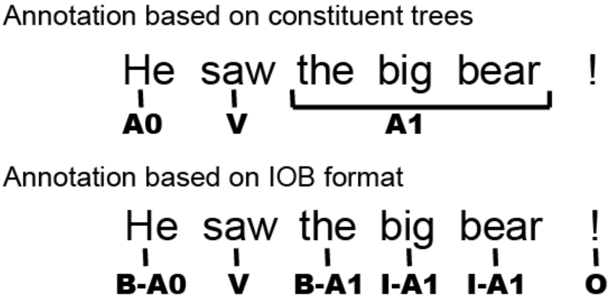

Annotation based on constituent trees
He saw the big bear !
A0
V
A1
Annotation based on IOB format
He saw the big bear !
B-A0
V
B-A1
I-A1
I-A1
O

Figure 11 – Differences between annotation produced using IOB schema and constituent trees

The IOB schema is particularly useful when token-by-token processing is desired instead of the traditional constituent-by-constituent. Notice that this transformation maintains an interchangeable label alignment structure between these formats. In our case, the most significant motivation for its usage is that this format allows us to eliminate dependencies on an eventual syntactic parser that, if used, would be responsible for extracting constituency trees from propositions.

Applied to PropBank-Br, the IOB schema produced 39 different roles (B and I tags for the items in the previous table, which accounts for 37 items, plus the outside label *O* and the verb *V*)

3.4 Our Model

This section outlines our approach, providing details about the relevant parts of our semantic role labeler.

3.4.1 Word Representations

Distributional semantics is based on the hypothesis that the words that co-occur in the same context tend to exhibit a similar meaning. Therefore, the meaning of a word is dependent on its usage context (143). Computationally, this theory provides the foundation for the unsupervised creation of word representations from the co-occurrence analysis in vast amounts of raw text. In this case, words are represented as low dimensional real-valued vectors (word embeddings), so that the similarity of vectors indicate the semantic similarity of the terms. Thus, vector operations in Euclidean space (typically the cosine) may be used to compute the relatedness between word pairs.

There are several models to learn word representations from large-scale unlabeled corpora. These methods usually belong to two categories: The ones based on the decomposition of co-occurrence matrices (as in latent semantic analysis (144)) and the ones that explore neural networks (87, 86, 145, 89), a method pioneered by Bengio et al.(146). Previous research though has demonstrated that these methods, although very different, tend to produce equivalent word representations (88). A systematic comparison of different word embedding methods is beyond the scope of this study.

The word representations utilized in this paper[6] were obtained by the application of the *skip-gram* model (86) on the full dump of the Brazilian Portuguese version of *Wikipedia* corpus[7]. In this model, given a sliding window of words, one attempts to predict the adjacent words (the context) based on the central word (the target token). It offers good representation for rarely seen tokens and outperformed other models in NLP tasks such as sentiment analysis and syntactic parsing (88). We relied on the implementation provided by the *Gensim* library (147) for the Python language.

Text preparation for training took place as follows: After extracting the raw text from the *Wikipedia* corpus[8], we used the *NLTK Punkt tokenizer* (148) for sentence splitting on each of its articles. Sentences obtained were then lowercased, followed by a series of transformations that included accents removal, punctuation separation, and substitutions[9]. At last, each resulting sentence was tokenized, feeding the *skip-gram* training algorithm.

Recent research points that the dimensionality of word embeddings is a determinant factor for the overall performance on NLP tasks such as named entity recognition, dependency parsing, sentiment analysis and co-reference resolution (149). Larger vector dimensionalities, while beneficial in semantic relation tasks (intrinsic tasks), ended up hurting the performance on NLP tasks (extrinsic). The results suggest that a dimensionality between 50 and 150 yields the best accuracy values for extrinsic tasks and it is worthwhile to carefully choose word embedding dimensionality for extrinsic tasks.

[6] The source code is available at <https://github.com/dfalci/pt_embeddings>

[7] Freely available at <https://dumps.wikimedia.org/>

[8] We used *wikiextractor*, a tool for extracting plain text from Wikipedia dumps: available at <https://github.com/attardi/wikiextractor>

[9] Sequences of numbers were transformed into the '#' token while email addresses and URLs were shortened to '*email*' and '*url*' tokens, respectively

Following these observations, we trained three distinct word representations with 50, 100 and 150 dimensions, respectively. In all three models, we employed a context-window of size 5, discarding the tokens with a total frequency lower than 5. Models were trained for ten iterations with an initial learning rate of 0.025 that linearly decays until it reaches the minimum learning rate of 0.0001. After traversing 10,690,000 sentences distributed in 957,206 *Wikipedia* articles for approximately 2 hours of training per iteration[10], we obtained a vocabulary containing 436,190 unique tokens. This number covers more than 99% of the tokens used in PropBank-Br (we missed 138 tokens). Further analysis on these missing tokens revealed that they are primarily composed of rarely seen nouns and first-person verbs - an infrequent narrative style in *Wikipedia*, but common in journalistic and opinion texts such as those in PropBank-Br. We chose to represent these words by randomly generated vectors.

Table 14 provides a sample of the word similarities achieved following our method.

Table 14 – Word similarities of our word representations

comprar ***(buy)***	**paris** ***(paris)***	**característica** ***(characteristic)***	**python**	**mestrado** ***(master's)***
vender (sell)	bruxelas *(brussels)*	peculiaridade *(peculiarity)*	c++	doutorado *(doctorate)*
adquirir *(purchase)*	grenoble *(grenoble)*	distintiva *(distinctive)*	javascript	bacharelado *(baccalaureate)*
alugar *(rent)*	marselha *(marseille)*	particularidade *(particularity)*	smalltalk	pos-graduacao *(post-graduation)*
gastar *(spend)*	lyon *(lyon)*	diferenciado *(differentiating)*	lisp	licenciatura *(graduation)*

3.4.2 Deep BiLSTM Model

A recurrent neural network (RNN) is a neural network architecture designed to learn tasks whose output is not only dependent on the current input, but also from previous input events. These networks usually have a form of a chain of cell instances (also known as memory blocks) where feedback connections are responsible for transmitting the weight of previous events throughout its structure. Standard RNN implementation though suffers from exploding and vanishing gradient problems that, during the training stage, prevents the network from learning long-term dependencies (150). To overcome the vanishing gradient problem (151) proposed a kind of RNN architecture based on long short-term memory (LSTM) cells in its hidden units. Each LSTM cell has a gating mechanism responsible for controlling the portion of information that will be propagated to its internal structures and the rest of the chain. One of its most distinctive abilities concerns preserving sequential information over very long time periods.

[10] The training time varies according to dimensionality

The following equations explain the internal mechanism of each LSTM cell:

$$
\begin{aligned}
i_t &= \sigma(x_t U^i + h_{t-1} W^i) \\
f_t &= \sigma(x_t U^f + h_{t-1} W^f) \\
o_t &= \sigma(x_t U^o + h_{t-1} W^o) \\
g_t &= \tanh(x_t U^g + h_{t-1} W^g) \\
c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\
h_t &= o_t \odot \tanh(c_t)
\end{aligned}
$$

Formally, let S represent a sequence of input vectors x with an arbitrary length n, such that $S = \{x_1, x_2, ..., x_n\}$. In this case, t designates a given time step in S. Weight matrices U and W are adjusted during the training phase. The σ symbol indicates a logistic sigmoid function and $\odot$ represents an element-wise multiplication. The input gate i_t determines whether or not the current input worth preserving while the forget gate f_t computes the proportion of the previous hidden state that must be forgotten. The cell state c_t is obtained through an operation that requires the multiplication of the new memory state g_t and input gate i_t added with the previous cell state c_{t-1} multiplied by the forget gate f_t. The hidden state h_t uses the output gate o_t to discover the part of the cell state c_t that will be exposed to the rest of the chain.

Traditional LSTM architecture (unidirectional, left-to-right or right-to-left) only considers information from the previous time steps to produce each output. Bidirectional LSTM (BiLSTM) architecture (152, 135) in contrast, considers both historical and future steps in order to learn information from preceding as well as future input events. It contains forward (left-to-right) and backward (right-to-left) LSTM layers whose outputs are merged by concatenation in a new layer that, intuitively, encodes past and future information. BiLSTM layers are typically stacked in k bidirectional layers. This arrangement, as occur in other types of multi-layer networks, enables capturing higher levels of abstraction yielding superior performance in sequence labeling tasks such as part of speech tagging, chunking, and named entity recognition (153, 154).

Our approach is depicted in Figure 12. Given a proposition expressed in natural language and its respective predicate, we start by the feature extraction stage. Our features, listed below, were inspired by (84, 139, 131) and are performed for each token in the proposition.

- **Word embeddings:** we capture the word representations for each token in a sentence, including the predicate. A look-up table operation is performed in an embedding matrix initialized with all the word vectors computed as mentioned in subsection 3.4.1.

- **Predicate embeddings:** through a look-up table operation, the word vector for the given predicate is extracted and repeated for each token in a proposition. This time, however, to save memory, the embedding matrix contains only the predicates used by PropBank-Br.

- **Capitalization:** as our word representations are all lowercased, capitalization is not naturally encoded by our model. To overcome this issue we created a set of binary features that indicate whether all characters in a given token are capitalized, contain any capital letter, or are lowercased.

- **Path to predicate:** the path to the predicate is given by the relative position of a token in a sentence with respect to the predicate position. Thereby, the token whose position coincides the predicate position is valued as 0 while the tokens that occur right before and after the predicate are represented with negative and positive values, respectively. Practically, a sentence containing five tokens whose predicate occur in the fourth position would have its tokens labeled as $\{-3, -2, -1, 0, 1\}$.

- **Predicate context:** this binary feature indicates whether a given token is inside the predicate context. To compute it, we apply a fixed window of size five where the predicate occupies its center. If a token is inside this window, then the token is said to be a member of predicate context (the value one is assigned).

These features are concatenated and feed the deep BiLSTM network that will compute abstract representations from propositions. The output of the last BiLSTM layer is attached to a softmax layer that, for each input token contained in the original proposition, yields the probability distribution over all the possible semantic roles (39 roles), creating a probability matrix. At last, in order to obtain the final prediction for the whole proposition, this probability matrix is sent to the global recursive neural parsing algorithm, explained in more details in the following subsection.

3.4.3 Global Recursive Neural Parsing

As mentioned before, BiLSTM networks can make decisions based on contextual information from previous and future input events. However, its output does not explicitly encode the functional dependencies and constraints that exist at the sentence level (global level). For instance, PropBank formalism states that core roles can occur at most once per proposition, but a network, due to its localized nature, may assign the same role for multiple tokens in the same proposition. Under these circumstances, if our final predictions are made by using only network predictions, we are exclusively relying on the model's ability to indirectly learn global dependencies. In this context, a minimal mistake may invalidate the whole sentence tagging.

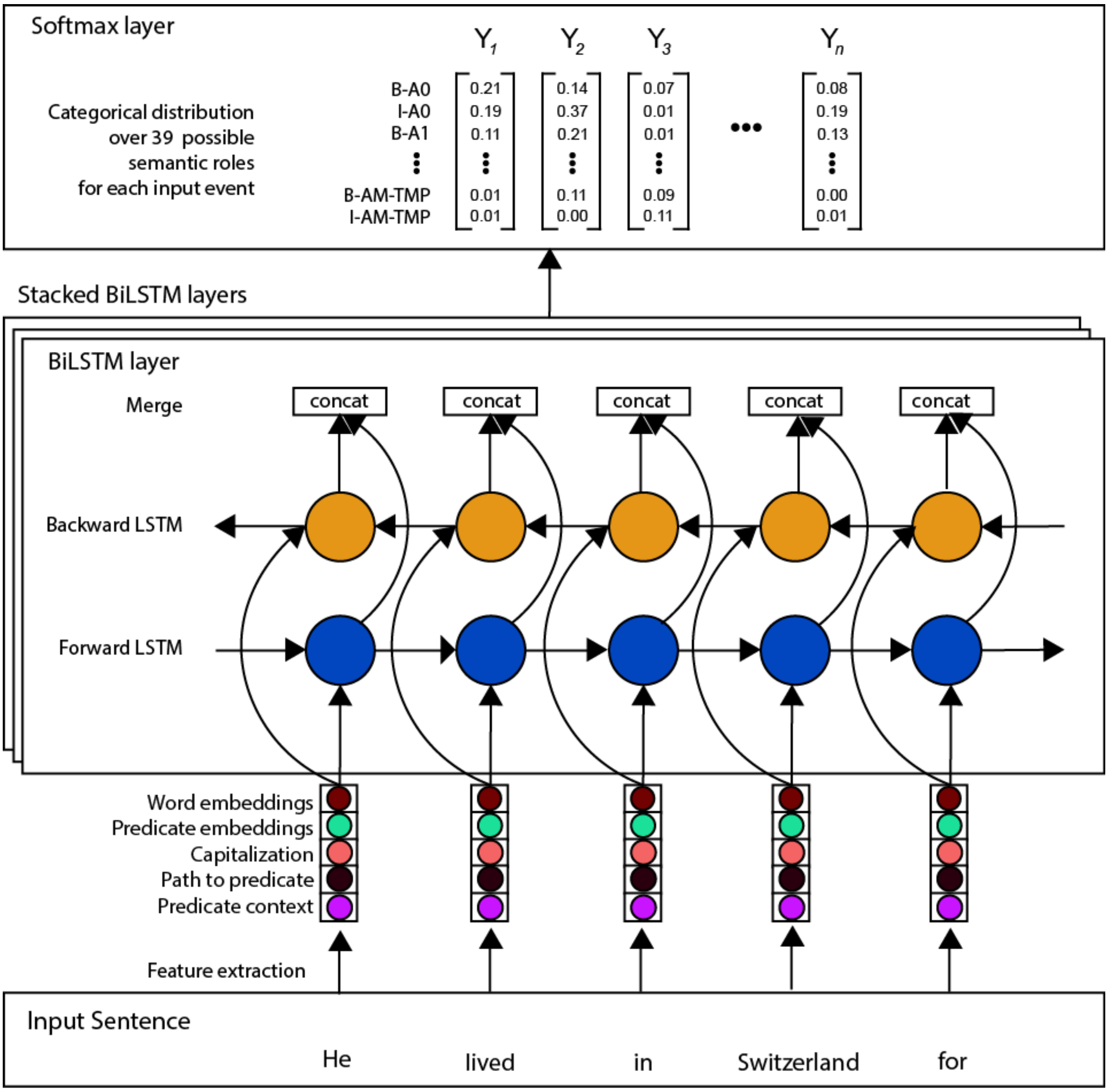


Figure 12 – Our deep BiLSTM architecture

Existing literature addresses such problem by applying a global inference mechanism whose objective is to find the best overall labeling for a given sentence. Dynamic programming algorithms such as Viterbi are candidates for solving this type of problem. The argument is that the usage of a transition state matrix naturally excludes violating sequences (132, 133). The inference stage is also modeled as an Integer Linear Programming (ILP) problem where one attempts to maximize the sentence labeling probability observing the formalism constraints that are translated to an off-the-shelf ILP solver (11). At last, Some authors rely on a reranking strategy that uses a second classifier which aggregates features from sentence and frame level features (12).

Our global inference stage is based on the recent work of Lee, Lewis e Zettlemoyer(155) that proposed the global recursive neural parsing algorithm. It directly searches the space of all possible labels derived from the network predictions with no dynamic programming techniques. The approach may be seen as a special case of A* algorithm and was tested on CCG parsing. The results point to an accurate and efficient model, finding optimal parse in 99.9% of sentences while exploring only 190 subtrees on average.

As in a standard A* search algorithm, the score function s of a partial sequence of nodes until the time step t is given by the equation $s(t) = g(t) + h(t)$ where g function is the cost of the path from starting node to node t and h function indicates an admissible heuristic for best path. Regular opening cost function g was modified by the introduction a constraint function c that yields a non-negative score whenever the candidate sequence violates any global constraint and 0 otherwise (Eq. 3.4). Hence, the opening cost is given by summing over the network probabilities output (represented as $\log p$) subtracted from the violation cost from the starting node until time step t.

$$g(w, y_t) = \sum_{i=1}^{t} \log p(y_i|w) - c(w, y_i) \tag{3.4}$$

The role of constraint function c is to discourage node exploring whose partial path leads to an invalid sequence of tags. The following global rules have been encoded into this function :

- **PropBank constraint:** As described in (15), core semantic roles (A0-A5) and adjunct arguments (AM) must be utilized at most once in a given proposition. Therefore, starting from second appearance, repeating semantic roles yields the violation score of 10.

- **IOB schema:** The constraints implemented here penalizes any partial sequence that does not produce a valid IOB sequence, such as the case where an inside tag (I) is not preceded by the begin tag (B). Here, the violation score is also 10.

The heuristic function h utilized in our work (see equation 3.5) is the same used by (131) and is given by the summation over the most probable labels for all timesteps after t.

$$h(w, y_t) = \sum_{i=t+1}^{T} \max_{y_i \in T} \log p(y_i|w) \tag{3.5}$$

3.5 Experimental Setup

This section exposes the details and settings shared across the experiments reported in this paper.

Our system was fully implemented using the *Python* language. The neural network uses libraries such *Keras* (156) and *TensorFlow* (157) whereas global recursive neural parsing algorithm was implemented from scratch in pure *Python*. Our experiments were performed on a machine equipped with an Intel Xeon E5-2686 v4 CPU, 64*GB* of *RAM*, and an *NVIDIA K80 GPU*.

Our models were trained using *Adam*, an efficient adaptative algorithm for gradient-based optimization of stochastic objective functions that is typically suited for high-dimensional parameter models (158), as it is the case. The algorithm was initialized with the default settings suggested by the original paper ($\alpha = 0.001, \beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$). To prevent overfitting and to improve the overall performance of our model, we used the *Dropout* technique introduced by (159). After some experimentation, we chose to drop 35% of the units at the input of each BiLSTM layer and another 20% between recurrent connections of the layers (160).

Each training session lasted up to a hundred epochs with an early stopping policy that ends the session after ten epochs without any improvement in the model's overall performance. We saved the network state whenever the current epoch result beats the one obtained by the previous best model.

Performance evaluation was executed after the end of each training epoch. The process was carried through the usage of the official evaluation script for the CoNLL-2005 Shared Task (*srl-eval.pl*[11]), that was dedicated to the SRL task (161). As stated in related work (subsection 3.2.2), our results were compared to those reported by Fonseca e Rosa(18) referred to in this paper as our baseline.

3.6 Results

Our first experiment investigates the optimal word embedding dimensionality applicable to our model. We prepared three distinct models with an almost identical setup where the only exception concerns the choice of word vector dimensionality. These model's used distinct pre-computed word vector representations with 50, 100, and 150 dimensions, respectively. All of them used four stacked BiLSTM layers (following the setup of Zhou e Xu(84)), each of them containing 300 hundred LSTM cells equally distributed between internal forward and backward layers. The remaining hyperparameters strictly followed the experimental setup described in the previous section.

In order to provide a robust evaluation, we chose to employ the cross-validation technique in a 20-fold configuration (162), what maintains the same partition size used by baseline. Hence, we randomly divided the original corpus into 20 equal sized folds and performed 20 separate training sessions, each using 19 folds for training (95% of data) and 1 fold for testing (the remaining 5%). We rotate the fold selection in a way that all

[11] Available at <http://www.lsi.upc.edu/~srlconll/soft.html>

folds are used as the test set exactly once. Therefore, considering our experiment, in this stage we conducted 60 training sessions that took three and half days to run.

From Figure 13 one may observe that the model's performance is sensitive to changes in word embedding dimensionality. Averaged results indicate a difference in performance that surpassed 4 F_1-score points. A Kruskal-Wallis H-test confirmed this observation as it rejects the null hypothesis that the population median of all the groups is equals ($p - value = 4.71 * 10^{-7}$). A post-hoc comparison[12] points that, considering our model, the usage of word vectors with 50 dimensions systematically produces inferior results when compared to the other models, based on a 100 ($p - value = 2.13 * 10^{-6}$) and a 150 dimensions ($p - value = 2.27 * 10^{-6}$). On the other hand, when directly comparing the performance of models based on word embeddings with 100 and 150 dimensions, we fail to reject the null hypothesis (*p-value* = 0.11). Thereby, despite a slightly better averaged F_1-score obtained by the model with 150 dimensions, there is no significant difference when compared to the result of the model based on 100 dimensions.

These results corroborate the findings of Melamud et al.(149) that suggest that picking the optimal dimensionality is critical for obtaining the best performance on extrinsic tasks such as SRL. In our case, the optimal level of semantic expressiveness was reached using vectors with 150 dimensions.

In the next experiment, we analyze the effect caused by the depth of stacked BiSLTM layers in the overall performance of our system. This time, we trained four identical models whose only exception regards its number of layers (1, 2, 3, and 4 layers). These models used pre-computed word embedding models with 150 dimensions (the best performance on the previous experiment) and BiLSTM layers with 300 LSTM cells each. Once again, we used a 20-fold cross-validation technique, and the remaining hyper-parameters followed the experimental setup described in the previous section. This experiment took four days to run.

Table 15 presents the results. Notice that the model based on just one BiLSTM layer yields an inferior performance when compared to the remaining models, based on more layers ($p - value = 0.01$). The accuracy reaches its peak in the model based on two layers (65.63) and, as we stack more layers, deepening the network architecture, one may observe a slight performance degradation. However, after comparing results from groups based on 2, 3, and 4 stacked layers we observe that notwithstanding the model based on two layers have achieved a slightly better averaged F_1-score, there is no significant difference on the accuracy of these groups ($p - value = 0.65$).

These observations converged into our final model that uses word vector representations of 150 dimensions and a neural network architecture composed by two stacked

[12] We used Wilcox Mann-Whitney test

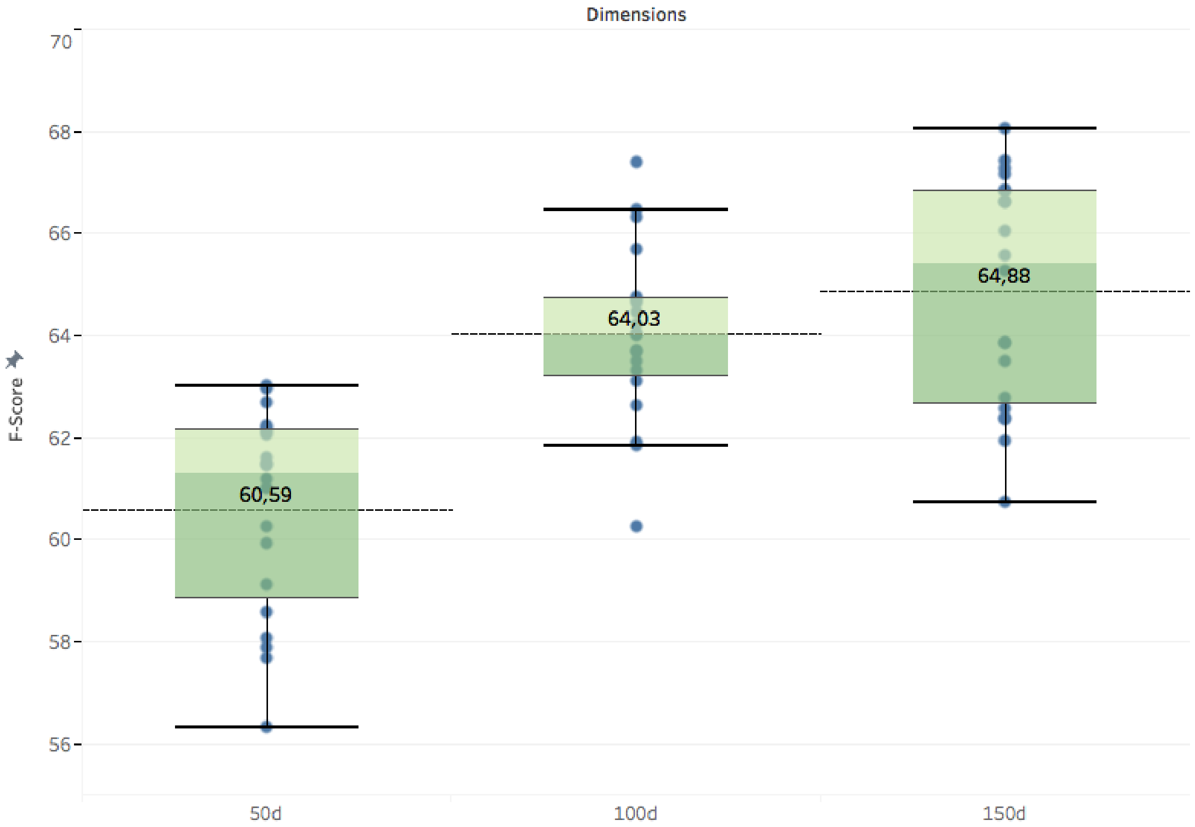


Figure 13 – Results considering different dimensionality

Table 15 – Model's performance variation according to the number of stacked layers

Layers	Averaged F_1-Score
1	61.76
2	**65.63**
3	65.22
4	64.76

BiLSTM layers.

Table 16 provides a comparison of our final model with the baseline system based on one-step and two-step frameworks (18). When we confront the one-step framework from the baseline with our system, one may observe that their best result (the best performance on a single training session) was consistently outperformed by our model's best result (62.31 vs 68.18), creating a margin of almost 6 F_1-score points. Even if we compare the baseline best model with our averaged score, the margin is still consistent, yielding a difference of 3.32 F_1-score points.

Our model also produced superior results when compared to the two-stage model of the baseline, albeit with a smaller margin. This time, the difference between the best models was 3.05 points, what points to a reduction on relative error of 8.74%. Again, our averaged result surpassed their best single model by a margin of 0.5 points.

Notice that the baseline paper (18) reported results only for their best single training sessions. For this reason, we could not produce a direct comparison based on k-fold cross-validation, that would produce a more robust evaluation.

Table 16 – SRL Performance comparison

Model	Precision	Recall	F_1-Score
Our best	**67.62**	**68.75**	**68.18**
Ours (averaged after 20-fold)			**65.63**
Baseline best (One-step)	64.41	60.34	62.31
Baseline best (Two-step)	67.06	63.31	65.13

Finally, the results of our best single model are detailed in Table 17. As expected, the system was more performant in well-defined and numerous semantic roles such as A0, A1, AM-TMP, and AM-NEG. On the other hand, roles such as A2 and A3 yielded inferior results. This reduction may be due to the lack of standardization in the semantic role definitions. As mentioned in the background (see section 3.2), only A0 and A1 exhibit a shared the meaning across different predicates. The meaning of the remaining core roles varies according to the predicate and its senses and can be even fused with adjunctive roles. This ambiguity may act as a noise factor for the neural network consequently causing a performance drop in the remaining roles.

Table 17 – Overall results for our best model

	Precision	Recall	$F_{\beta=1}$
Overall	67.62%	68.75%	68.18
A0	81.82%	86.90%	84.28
A1	71.15%	72.29%	71.71
A2	52.73%	42.03%	46.77
A3	28.57%	40.00%	33.33
A4	100.00%	50.00%	66.67
AM-ADV	42.86%	50.00%	46.15
AM-CAU	50.00%	33.33%	40.00
AM-DIS	44.44%	28.57%	34.78
AM-EXT	0.00%	0.00%	0.00
AM-LOC	54.17%	72.22%	61.90
AM-MED	0.00%	0.00%	0.00
AM-MNR	34.78%	47.06%	40.00
AM-NEG	90.00%	94.74%	92.31
AM-PNC	42.86%	66.67%	52.17
AM-PRD	100.00%	33.33%	50.00
AM-TMP	66.67%	73.47%	69.90
V	100.00%	100.00%	100.00

Our final result is still far from the best results reported for the English language.

However, we believe that our system, if exposed to a more balanced corpus, with a comparable number of sentences, would be able to achieve competitive performance.

3.7 Conclusion

In this paper, we described an end-to-end semantic role labeler for the Portuguese language. The one-step system was built on top of a BiLSTM neural network architecture tied to an inference stage based on a global recursive neural parsing algorithm that was specifically tailored for the SRL task. Seeking an optimal structure, we also conducted an extensive investigation about the effects of two crucial factors on our structure: The depth of network architecture and the proper word embedding dimensionality.

Our model consistently outperformed the previous state-of-the-art by 3.05 F_1-score points, reducing the relative error in 8.74%. We also confirmed the hypothesis that picking the optimal embedding dimensionality is critical for obtaining the best accuracy on SRL task. Our final model was based on word vectors with 150 dimensions passing through a deep network with two BiLSTM layers.

Future research may invest in the expansion of PropBank-Br corpus what, in our point of view, is essential for reaching a competitive performance. Moreover, we believe that a promising direction point to an architecture designed to attenuate the impact of ambiguity in semantic role definitions of PropBank formalism.

4 Conclusion

The objective of this thesis was to evaluate the performance of a semantic role labeler for the Portuguese language built considering techniques addressed in the literature. We evaluated an end-to-end semantic role labeler based on a deep bidirectional long short-term neural network whose predictions serve as input to a recursive neural parsing algorithm, specifically tailored for the task. The first specific objective of this research was to "Identify the most accurate semantic role labeling techniques described in the literature." and have been achieved on the systematic literature review (chapter 2). The second specific objective was to "Analyze the results of an automatic semantic role labeler for the Portuguese language built considering techniques addressed in the literature." and has been achieved with the method presented in chapter 3.

The results demonstrated that our semantic role labeler consistently outperformed the previous state-of-the-art on PropBank-Br corpus by 3.05 F_1-score points, reducing the relative error in 8.74%. The performance though is only modest, still far from the one reached by techniques targeted at the English language. We believe that our system would be able to yield a superior performance if exposed to larger and more balanced data.

The source code of the proposed model, as well as the trained word representations, were made publicly available on the internet, under BSD license, and may be used by future investigations focused on content-analysis for the Portuguese language.

Future research may also invest in the expansion of PropBank-Br corpus what, in our point of view, is an essential element for reaching a competitive performance.

References

1 LIDDY, E. D. *Natural language processing*. 2. ed. [S.l.]: Marcel Drecker, Inc, 2001. v. 1. Encyclopedia of library and information science. 11

2 OXFORD, E. D. *Cognition*. Oxford: Oxford University Press, 1989. Disponível em: <https://en.oxforddictionaries.com/definition/cognition>. 11

3 GILDEA, D.; JURAFSKY, D. Automatic labeling of semantic roles. *Computational linguistics*, MIT Press, v. 28, n. 3, p. 245–288, 2002. 11, 16, 26, 33, 34

4 PALMER, M.; GILDEA, D.; NIANWEN, X. *Semantic Role Labeling (Synthesis Lectures on Human Language Technologies)*. 1. ed. Morgan & Claypool Publishers, 2010. ISBN 1598298313,9781598298314. Disponível em: <http://gen.lib.rus.ec/book/index.php?md5=A63331E1CD522417B8774FDD5817D0EA>. 11, 12, 17, 23, 25

5 KHAN, A.; SALIM, N.; KUMAR, Y. J. A framework for multi-document abstractive summarization based on semantic role labelling. *Applied Soft Computing*, Elsevier, v. 30, p. 737–747, 2015. 11, 14, 16, 45

6 CHRISTENSEN, J. et al. Semantic role labeling for open information extraction. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*. [S.l.], 2010. p. 52–60. 11, 14, 16, 45

7 WU, D.; FUNG, P. Can semantic role labeling improve smt. In: *Proceedings of the 13th Annual Conference of the European Association for Machine Translation*. [S.l.: s.n.], 2009. p. 218–225. 11, 14, 45

8 PONZETTO, S. P.; STRUBE, M. Exploiting semantic role labeling, wordnet and wikipedia for coreference resolution. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. [S.l.], 2006. p. 192–199. 11, 16, 45

9 SHEN, D.; LAPATA, M. Using semantic roles to improve question answering. In: *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*. [S.l.: s.n.], 2007. 11, 14, 16, 45

10 LEVIN, B. *English verb classes and alternations: A preliminary investigation*. [S.l.]: University of Chicago press, 1993. 12, 19, 47

11 PUNYAKANOK, V.; ROTH, D.; YIH, W.-t. The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics*, MIT Press, v. 34, n. 2, p. 257–287, 2008. 12, 39, 45, 57

12 TOUTANOVA, K.; HAGHIGHI, A.; MANNING, C. D. Joint learning improves semantic role labeling. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. [S.l.], 2005. p. 589–596. 12, 57

13 SURDEANU, M. et al. The conll-2008 shared task on joint parsing of syntactic and semantic dependencies. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the Twelfth Conference on Computational Natural Language Learning.* [S.l.], 2008. p. 159–177. 12, 26, 31

14 PALMER, M.; GILDEA, D.; KINGSBURY, P. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, MIT Press, v. 31, n. 1, p. 71–106, 2005. 12, 47, 50

15 DURAN, M. S.; ALUÍSIO, S. M. Propbank-br: a brazilian treebank annotated with semantic role labels. In: *LREC*. [S.l.: s.n.], 2012. p. 1862–1867. 12, 46, 50, 58

16 BICK, E. Automatic semantic role annotation for portuguese. In: *Proceedings of TIL 2007-5th Workshop on Information and Human Language Technology.* [S.l.: s.n.], 2007. p. 1713–1716. 13

17 ALVA-MANCHEGO, F. E.; ROSA, J. L. G. Semantic role labeling for brazilian portuguese: A benchmark. In: SPRINGER. *IBERAMIA*. [S.l.], 2012. p. 481–490. 13, 49, 50

18 FONSECA, E. R.; ROSA, J. L. G. A two-step convolutional neural network approach for semantic role labeling. In: IEEE. *Neural Networks (IJCNN), The 2013 International Joint Conference on.* [S.l.], 2013. p. 1–7. 13, 28, 49, 51, 59, 61, 62

19 COLLOBERT, R. et al. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, v. 12, n. Aug, p. 2493–2537, 2011. 13, 28, 29, 39, 45, 48

20 The World Bank. World Development Indicators. 2016. Website. <http://databank.worldbank.org/data/reports.aspx?Code=NY.GDP.MKTP.CD&id=af3ce82b&report_name=Popular_indicators&populartype=series&ispopular=y> [Accessed: 2017-11-21]. 14

21 BILOTTI, M. W. et al. Rank learning for factoid question answering with linguistic and semantic constraints. In: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management.* New York, NY, USA: ACM, 2010. (CIKM '10), p. 459–468. ISBN 978-1-4503-0099-5. Disponível em: <http://doi.acm.org/10.1145/1871437.1871498>. 14, 16, 45

22 JURAFSKY, D.; MARTIN, J. H. *Speech and language processing.* 2. ed. [S.l.]: Pearson, 2014. 16, 23

23 MÀRQUEZ, L. et al. Semantic role labeling: an introduction to the special issue. *Computational linguistics*, MIT Press, v. 34, n. 2, p. 145–159, 2008. 16

24 BUTT, M. *Theories of case.* [S.l.]: Cambridge University Press, 2006. 17

25 FILLMORE, C. The case for case. *Universals in Linguistic Theory*, Rinehart and Winston, p. 1–88, 1968. 17

26 COOK, W. A. *Case grammar theory.* [S.l.]: Georgetown University Press, 1989. 18

27 DOWTY, D. Thematic proto-roles and argument selection. *Language*, JSTOR, p. 547–619, 1991. 18

28 JACKENDOFF, R. *Semantic structures*. [S.l.]: MIT press, 1992. v. 18. 18

29 SCHULER, K. K. Verbnet: A broad-coverage, comprehensive verb lexicon. 2005. 18

30 FILLMORE, C. J. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, Wiley Online Library, v. 280, n. 1, p. 20–32, 1976. 18

31 FILLMORE, C. Frame semantics. *Linguistics in the morning calm*, Hanshin Publishing Co., p. 111–137, 1982. 18

32 KIPPER, K. et al. Class-based construction of a verb lexicon. In: *AAAI/IAAI*. [S.l.: s.n.], 2000. p. 691–696. 19, 22

33 PALMER, M.; GILDEA, D.; KINGSBURY, P. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, MIT press, v. 31, n. 1, p. 71–106, 2005. 19, 21

34 BAKER, C. F.; FILLMORE, C. J.; LOWE, J. B. The berkeley framenet project. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*. [S.l.], 1998. p. 86–90. 20

35 RUPPENHOFER, J. et al. *FrameNet II: Extended theory and practice*. 2016. <https://framenet.icsi.berkeley.edu/fndrupal/sites/default/files/book2016.11.01.pdf> [Accessed: 2016-12-18]. 20

36 KINGSBURY, P.; PALMER, M. From treebank to propbank. In: CITESEER. *LREC*. [S.l.], 2002. 21

37 DANG, H. T. et al. Investigating regular sense extensions based on intersective levin classes. In: *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*. Stroudsburg, PA, USA: Association for Computational Linguistics, 1998. (ACL '98), p. 293–299. Disponível em: <http://dx.doi.org/10.3115/980845.980893>. 21, 47

38 KIPPER, K. et al. A large-scale classification of english verbs. *Language Resources and Evaluation*, Springer, v. 42, n. 1, p. 21–40, 2008. 22

39 BAKER, C. F.; RUPPENHOFER, J. Framenet's frames vs. levin's verb classes. In: *Proceedings of the 28th annual meeting of the Berkeley Linguistics Society*. [S.l.: s.n.], 2002. p. 27–38. 22

40 MEYERS, A. et al. The nombank project: An interim report. In: MEYERS, A. (Ed.). *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*. Boston, Massachusetts, USA: Association for Computational Linguistics, 2004. p. 24–31. 22

41 MILLER, G. A. Wordnet: a lexical database for english. *Communications of the ACM*, ACM, v. 38, n. 11, p. 39–41, 1995. 23

42 SHI, L.; MIHALCEA, R. Open text semantic parsing using framenet and wordnet. In: *Demonstration Papers at HLT-NAACL 2004*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2004. (HLT-NAACL–Demonstrations '04), p. 19–22. Disponível em: <http://dl.acm.org/citation.cfm?id=1614025.1614031>. 23

43 WEISCHEDEL, R. et al. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 2013. 23

44 SALOMÃO, M. M. M.; TORRENT, T. T.; SAMPAIO, T. F. A linguística cognitiva encontra a linguística computacional: notícias do projeto framenet brasil. *Cadernos de Estudos Lingüísticos*, Universidade Estadual de Campinas-UNICAMP, IEL-Cadernos de Estudos Linguísticos, v. 55, n. 1, 2013. 23

45 CANDITO, M. et al. Developing a french framenet: Methodology and first results. In: *LREC-The 9th edition of the Language Resources and Evaluation Conference*. [S.l.: s.n.], 2014. 23

46 BURCHARDT, A. et al. The salsa corpus: a german corpus resource for lexical semantics. In: *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*. [S.l.: s.n.], 2006. p. 969–974. 23

47 HEPPIN, K. F.; GRONOSTAJ, M. T. The rocky road towards a swedish framenet-creating swefn. In: *LREC*. [S.l.: s.n.], 2012. p. 256–261. 23

48 SUBIRATS, C. Spanish framenet: A frame-semantic analysis of the spanish lexicon. *Multilingual FrameNets in Computational Lexicography. Methods and Applications, Mouton de Gruyter, Berlin/New York*, p. 135–162, 2009. 23

49 YOU, L.; LIU, K. Building chinese framenet database. In: IEEE. *Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE'05. Proceedings of 2005 IEEE International Conference on*. [S.l.], 2005. p. 301–306. 23

50 OHARA, K. H. et al. The japanese framenet project: An introduction. In: *Proceedings of LREC-04 Satellite Workshop "Building Lexical Resources from Semantically Annotated Corpora"(LREC 2004)*. [S.l.: s.n.], 2004. p. 9–11. 23

51 NAM, S. et al. Korean framenet for semantic analysis. In: *Proceedings of the 13th International Semantic Web Conference*. [S.l.: s.n.], 2014. 23

52 ZAWISŁAWSKA, M.; DERWOJEDOWA, M.; LINDE-USIEKNIEWICZ, J. A framenet for polish. In: *Converging Evidence: Proceedings to the Third International Conference of the German Cognitive Linguistics Association (GCLA'08)*. [S.l.: s.n.], 2008. p. 116–117. 23

53 DURAN, M. S.; ALUÍSIO, S. M. et al. Propbank-br: a brazilian portuguese corpus annotated with semantic role labels. In: *Proceedings of the 8th Symposium in Information and Human Language Technology, Cuiabá/MT, Brazil*. [S.l.: s.n.], 2011. 23

54 ZAGHOUANI, W. et al. The revised arabic propbank. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the Fourth Linguistic Annotation Workshop*. [S.l.], 2010. p. 222–226. 23

55 XUE, N.; PALMER, M. Adding semantic roles to the chinese treebank. *Natural Language Engineering*, Cambridge Univ Press, v. 15, n. 01, p. 143–172, 2009. 23

56 HAVERINEN, K. et al. The finnish proposition bank. *Language Resources and Evaluation*, Springer, v. 49, n. 4, p. 907–926, 2015. 23

57 PALMER, M. et al. Hindi syntax: Annotating dependency, lexical predicate-argument structure, and phrase structure. In: *The 7th International Conference on Natural Language Processing*. [S.l.: s.n.], 2009. p. 14–17. 23

58 RUSSELL, S. J. et al. *Artificial intelligence: a modern approach*. [S.l.]: Prentice hall Upper Saddle River, 2003. v. 2. 24

59 NICOLAS, P. R. *Scala for Machine Learning*. [S.l.]: Packt Publishing Ltd, 2015. 24, 25

60 ALPAYDIN, E. *Introduction to machine learning*. [S.l.]: MIT press, 2009. 24

61 SUTTON, C.; MCCALLUM, A. et al. An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, Now Publishers, Inc., v. 4, n. 4, p. 267–373, 2012. 24

62 MURPHY, K. P. *Machine learning: a probabilistic perspective*. [S.l.]: MIT press, 2012. 24

63 ZHU, X. *Semi-supervised learning literature survey*. [S.l.], 2005. 24

64 COLLINS, M. *Head-driven statistical models for natural language parsing*. Tese (Doutorado) — University of Pennsylvania, 1999. 26, 27

65 CHARNIAK, E.; JOHNSON, M. Coarse-to-fine n-best parsing and maxent discriminative reranking. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. [S.l.], 2005. p. 173–180. 26

66 NIVRE, J.; HALL, J.; NILSSON, J. Maltparser: A data-driven parser-generator for dependency parsing. In: *Proceedings of LREC*. [S.l.: s.n.], 2006. v. 6, p. 2216–2219. 26

67 MCDONALD, R. et al. Non-projective dependency parsing using spanning tree algorithms. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. [S.l.], 2005. p. 523–530. 26

68 HACIOGLU, K. Semantic role labeling using dependency trees. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 20th international conference on Computational Linguistics*. [S.l.], 2004. p. 1273. 26

69 MITSUMORI, T. et al. Semantic role labeling using support vector machines. In: *Proceedings of the Ninth Conference on Computational Natural Language Learning*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005. (CONLL '05), p. 197–200. Disponível em: <http://dl.acm.org/citation.cfm?id=1706543.1706580>. 26

70 THOMPSON, C. A.; LEVY, R.; MANNING, C. D. A generative model for semantic role labeling. In: SPRINGER. *European Conference on Machine Learning*. [S.l.], 2003. p. 397–408. 27

71 XUE, N.; PALMER, M. Calibrating features for semantic role labeling. In: *EMNLP*. [S.l.: s.n.], 2004. p. 88–94. 27, 29

72 SURDEANU, M.; TURMO, J. Semantic role labeling using complete syntactic analysis. In: *Proceedings of the Ninth Conference on Computational Natural Language Learning*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005. (CONLL '05), p. 221–224. Disponível em: <http://dl.acm.org/citation.cfm?id=1706543.1706586>. 27

73 JOHANSSON, R.; NUGUES, P. Automatic annotation for all semantic layers in framenet. In: *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters &#38; Demonstrations*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006. (EACL '06), p. 135–138. Disponível em: <http://dl.acm.org/citation.cfm?id=1608974.1608991>. 27

74 BETHARD, S. et al. Semantic role labeling for protein transport predicates. *BMC bioinformatics*, BioMed Central, v. 9, n. 1, p. 277, 2008. 27

75 WANG, H. et al. Dependency tree-based srl with proper pruning and extensive feature engineering. In: *Proceedings of the Twelfth Conference on Computational Natural Language Learning*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008. (CoNLL '08), p. 253–257. ISBN 978-1-905593-48-4. Disponível em: <http://dl.acm.org/citation.cfm?id=1596324.1596370>. 27

76 MàRQUEZ, L. et al. Semantic role labeling as sequential tagging. In: *Proceedings of the Ninth Conference on Computational Natural Language Learning*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005. (CONLL '05), p. 193–196. Disponível em: <http://dl.acm.org/citation.cfm?id=1706543.1706579>. 27, 30

77 PARK, K.-M.; RIM, H.-C. Maximum entropy based semantic role labeling. In: *Proceedings of the Ninth Conference on Computational Natural Language Learning*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005. (CONLL '05), p. 209–212. Disponível em: <http://dl.acm.org/citation.cfm?id=1706543.1706583>. 27

78 MOSCHITTI, A.; PIGHIN, D.; BASILI, R. Tree kernels for semantic role labeling. *Computational Linguistics*, MIT Press, v. 34, n. 2, p. 193–224, 2008. 27, 28, 29, 39

79 COLLINS, M.; DUFFY, N. New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted perceptron. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 40th annual meeting on association for computational linguistics*. [S.l.], 2002. p. 263–270. 27

80 KAZAMA, J.; TORISAWA, K. Speeding up training with tree kernels for node relation labeling. In: *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005. (HLT '05), p. 137–144. Disponível em: <http://dx.doi.org/10.3115/1220575.1220593>. 27

81 MOSCHITTI, A.; BASILI, R. Verb subcategorization kernels for automatic semantic labeling. In: *Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005. (DeepLA '05), p. 10–17. Disponível em: <http://dl.acm.org/citation.cfm?id=1631850.1631852>. 27

82 CHE, W. et al. Using a hybrid convolution tree kernel for semantic role labeling. *ACM Transactions on Asian Language Information Processing (TALIP)*, ACM, New

York, NY, USA, v. 7, n. 4, p. 13:1–13:23, nov. 2008. ISSN 1530-0226. Disponível em: <http://doi.acm.org/10.1145/1450295.1450298>. 28

83 COLLOBERT, R.; WESTON, J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In: *Proceedings of the 25th International Conference on Machine Learning*. New York, NY, USA: ACM, 2008. (ICML '08), p. 160–167. ISBN 978-1-60558-205-4. Disponível em: <http://doi.acm.org/10.1145/1390156.1390177>. 28

84 ZHOU, J.; XU, W. End-to-end learning of semantic role labeling using recurrent neural networks. In: *ACL (1)*. [S.l.: s.n.], 2015. p. 1127–1137. 28, 40, 45, 49, 55, 59

85 LUND, K.; BURGESS, C. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, Springer, v. 28, n. 2, p. 203–208, 1996. 28

86 MIKOLOV, T. et al. Distributed representations of words and phrases and their compositionality. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2013. p. 3111–3119. 28, 53

87 PENNINGTON, J.; SOCHER, R.; MANNING, C. Glove: Global vectors for word representation. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. [S.l.: s.n.], 2014. p. 1532–1543. 28, 53

88 LEVY, O.; GOLDBERG, Y.; DAGAN, I. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, v. 3, p. 211–225, 2015. 28, 53

89 TRASK, A.; MICHALAK, P.; LIU, J. sense2vec-a fast and accurate method for word sense disambiguation in neural word embeddings. *arXiv preprint arXiv:1511.06388*, 2015. 28, 53

90 BOJANOWSKI, P. et al. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016. 28

91 JR, W. R. F.; MARTIN, J. H. Dependency-based semantic role labeling using convolutional neural networks. In: *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*. [S.l.: s.n.], 2015. p. 279–288. 28, 48

92 FITZGERALD, N. et al. Semantic role labeling with neural network factors. In: *EMNLP*. [S.l.: s.n.], 2015. p. 960–970. 28, 40

93 CHE, W. et al. A cascaded syntactic and semantic dependency parsing system. In: *Proceedings of the Twelfth Conference on Computational Natural Language Learning*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008. (CoNLL '08), p. 238–242. ISBN 978-1-905593-48-4. Disponível em: <http://dl.acm.org/citation.cfm?id=1596324.1596367>. 28

94 JOHANSSON, R.; NUGUES, P. Dependency-based syntactic-semantic analysis with propbank and nombank. In: *Proceedings of the Twelfth Conference on Computational Natural Language Learning*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008. (CoNLL '08), p. 183–187. ISBN 978-1-905593-48-4. Disponível em: <http://dl.acm.org/citation.cfm?id=1596324.1596355>. 28, 40, 45

95 DAS, D.; SMITH, N. A. Semi-supervised frame-semantic parsing for unknown predicates. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011. (HLT '11), p. 1435–1444. ISBN 978-1-932432-87-9. Disponível em: <http://dl.acm.org/citation.cfm?id=2002472.2002648>. 29

96 WATANABE, Y. et al. A pipeline approach for syntactic and semantic dependency parsing. In: *Proceedings of the Twelfth Conference on Computational Natural Language Learning*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008. (CoNLL '08), p. 228–232. ISBN 978-1-905593-48-4. Disponível em: <http://dl.acm.org/citation.cfm?id=1596324.1596365>. 29

97 YURET, D.; YATBAZ, M. A.; URAL, A. E. Discriminative vs. generative approaches in semantic role labeling. In: *Proceedings of the Twelfth Conference on Computational Natural Language Learning*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008. (CoNLL '08), p. 223–227. ISBN 978-1-905593-48-4. Disponível em: <http://dl.acm.org/citation.cfm?id=1596324.1596364>. 29

98 CIARAMITA, M. et al. Desrl: A linear-time semantic role labeling system. In: *Proceedings of the Twelfth Conference on Computational Natural Language Learning*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008. (CoNLL '08), p. 258–262. ISBN 978-1-905593-48-4. Disponível em: <http://dl.acm.org/citation.cfm?id=1596324.1596371>. 29

99 ZHAO, H. et al. Multilingual dependency learning: A huge feature engineering method to semantic dependency parsing. In: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. (CoNLL '09), p. 55–60. ISBN 978-1-932432-29-9. Disponível em: <http://dl.acm.org/citation.cfm?id=1596409.1596418>. 29

100 WATANABE, Y.; ASAHARA, M.; MATSUMOTO, Y. Multilingual syntactic-semantic dependency parsing with three-stage approximate max-margin linear models. In: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. (CoNLL '09), p. 114–119. ISBN 978-1-932432-29-9. Disponível em: <http://dl.acm.org/citation.cfm?id=1596409.1596429>. 29

101 PRADHAN, S. et al. Support vector learning for semantic argument classification. *Machine Learning*, Springer, v. 60, n. 1-3, p. 11–39, 2005. 29, 30, 39

102 WANG, H.-L.; ZHOU, G.-D. Semantic role labeling of chinese nominal predicates with dependency-driven constituent parse tree structure. *Journal of Computer Science and Technology*, Springer Science & Business Media, v. 28, n. 6, p. 1117, 2013. 29, 39

103 PUNYAKANOK, V. et al. Semantic role labeling via integer linear programming inference. In: *Proceedings of the 20th International Conference on Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2004. (COLING '04). Disponível em: <https://doi.org/10.3115/1220355.1220552>. 30

104 TOUTANOVA, K.; HAGHIGHI, A.; MANNING, C. D. Joint learning improves semantic role labeling. In: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005. (ACL '05), p. 589–596. Disponível em: <https://doi.org/10.3115/1219840.1219913>. 30

105 KOOMEN, P. et al. Generalized inference with multiple semantic role labeling systems. In: *Proceedings of the Ninth Conference on Computational Natural Language Learning*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005. (CONLL '05), p. 181–184. Disponível em: <http://dl.acm.org/citation.cfm?id=1706543.1706576>. 30

106 MORANTE, R.; DAELEMANS, W.; ASCH, V. V. A combined memory-based semantic role labeler of english. In: *Proceedings of the Twelfth Conference on Computational Natural Language Learning*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008. (CoNLL '08), p. 208–212. ISBN 978-1-905593-48-4. Disponível em: <http://dl.acm.org/citation.cfm?id=1596324.1596361>. 30

107 ZHUANG, T.; ZONG, C. A minimum error weighting combination strategy for chinese semantic role labeling. In: *Proceedings of the 23rd International Conference on Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. (COLING '10), p. 1362–1370. Disponível em: <http://dl.acm.org/citation.cfm?id=1873781.1873934>. 30

108 SUTTON, C.; MCCALLUM, A. Joint parsing and semantic role labeling. In: *Proceedings of the Ninth Conference on Computational Natural Language Learning*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005. (CONLL '05), p. 225–228. Disponível em: <http://dl.acm.org/citation.cfm?id=1706543.1706587>. 30

109 YI, S.-t.; PALMER, M. The integration of syntactic parsing and semantic role labeling. In: *Proceedings of the Ninth Conference on Computational Natural Language Learning*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005. (CONLL '05), p. 237–240. Disponível em: <http://dl.acm.org/citation.cfm?id=1706543.1706590>. 30

110 SAMUELSSON, Y. et al. Mixing and blending syntactic and semantic dependencies. In: *Proceedings of the Twelfth Conference on Computational Natural Language Learning*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008. (CoNLL '08), p. 248–252. ISBN 978-1-905593-48-4. Disponível em: <http://dl.acm.org/citation.cfm?id=1596324.1596369>. 30

111 HENDERSON, J. et al. A latent variable model of synchronous parsing for syntactic and semantic dependencies. In: *Proceedings of the Twelfth Conference on Computational Natural Language Learning*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008. (CoNLL '08), p. 178–182. ISBN 978-1-905593-48-4. Disponível em: <http://dl.acm.org/citation.cfm?id=1596324.1596354>. 30

112 JOHANSSON, R. Statistical bistratal dependency parsing. In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. (EMNLP '09), p. 561–569. ISBN 978-1-932432-62-6. Disponível em: <http://dl.acm.org/citation.cfm?id=1699571.1699586>. 30

113 BAEZA-YATES, R.; RIBEIRO-NETO, B. et al. *Modern information retrieval*. [S.l.]: ACM press New York, 1999. v. 463. 31

114 CARRERAS, X.; MàRQUEZ, L. Introduction to the conll-2004 shared task: Semantic role labeling. In: NG, H. T.; RILOFF, E. (Ed.). *HLT-NAACL 2004 Workshop: Eighth Conference on Computational Natural Language Learning (CoNLL-2004)*. Boston, Massachusetts, USA: Association for Computational Linguistics, 2004. p. 89–97. Disponível em: <http://acl.ldc.upenn.edu/W/W04/W04-2412.bib>. 31

115 CARRERAS, X.; MÀRQUEZ, L. Introduction to the conll-2005 shared task: Semantic role labeling. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the Ninth Conference on Computational Natural Language Learning*. [S.l.], 2005. p. 152–164. 31

116 HAJIČ, J. et al. The conll-2009 shared task: Syntactic and semantic dependencies in multiple languages. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*. [S.l.], 2009. p. 1–18. 31, 32, 36

117 LEVY, Y.; ELLIS, T. J. A systems approach to conduct an effective literature review in support of information systems research. *Informing Science: International Journal of an Emerging Transdiscipline*, Informing Science Institute, v. 9, n. 1, p. 181–212, 2006. 32

118 MOHER, D. et al. Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *Annals of internal medicine*, Am Coll Physicians, v. 151, n. 4, p. 264–269, 2009. 32

119 PRADHAN, S. S. et al. Shallow semantic parsing using support vector machines. In: *HLT-NAACL*. [S.l.: s.n.], 2004. p. 233–240. 33

120 KITCHENHAM, B. Procedures for performing systematic reviews. *Keele, UK, Keele University*, v. 33, n. 2004, p. 1–26, 2004. 35

121 DESCHACHT, K.; MOENS, M.-F. Semi-supervised semantic role labeling using the latent words language model. In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. (EMNLP '09), p. 21–29. ISBN 978-1-932432-59-6. Disponível em: <http://dl.acm.org/citation.cfm?id=1699510.1699514>. 38, 40

122 BOXWELL, S. A.; MEHAY, D.; BREW, C. Brutus: A semantic role labeling system incorporating ccg, cfg, and dependency features. In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. (ACL '09), p. 37–45. ISBN 978-1-932432-45-9. Disponível em: <http://dl.acm.org/citation.cfm?id=1687878.1687885>. 39

123 SIEGEL, S.; CASTELLAN, N. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, 1988. (McGraw-Hill international editions. Statistics series). ISBN 9780070573574. Disponível em: <https://books.google.com.br/books?id=bq3uAAAAMAAJ>. 39

124 CHEN, Z.; JI, H. Language specific issue and feature exploration in chinese event extraction. In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. (NAACL-Short '09), p. 209–212. Disponível em: <http://dl.acm.org/citation.cfm?id=1620853.1620910>. 39

125 XUE, N. Semantic role labeling of nominalized predicates in chinese. In: *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006. (HLT-NAACL '06), p. 431–438. Disponível em: <http://dx.doi.org/10.3115/1220835.1220890>. 39

126 LIM, S.; LEE, C.; RA, D. Dependency-based semantic role labeling using sequence labeling with a structural {SVM}. *Pattern Recognition Letters*, v. 34, n. 6, p. 696 – 702, 2013. ISSN 0167-8655. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0167865513000354>. 40

127 BJöRKELUND, A.; HAFDELL, L.; NUGUES, P. Multilingual semantic role labeling. In: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. (CoNLL '09), p. 43–48. ISBN 978-1-932432-29-9. Disponível em: <http://dl.acm.org/citation.cfm?id=1596409.1596416>. 40

128 ZHAO, H.; CHEN, W.; KIT, C. Semantic dependency parsing of nombank and propbank: An efficient integrated approach via a large-scale feature selection. In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. (EMNLP '09), p. 30–39. ISBN 978-1-932432-59-6. Disponível em: <http://dl.acm.org/citation.cfm?id=1699510.1699515>. 40

129 ZHAO, H. et al. Multilingual dependency learning: Exploiting rich features for tagging syntactic and semantic dependencies. In: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. (CoNLL '09), p. 61–66. ISBN 978-1-932432-29-9. Disponível em: <http://dl.acm.org/citation.cfm?id=1596409.1596419>. 40

130 TOUTANOVA, K.; HAGHIGHI, A.; MANNING, C. D. A global joint model for semantic role labeling. *Computational Linguistics*, MIT Press, v. 34, n. 2, p. 161–191, 2008. 40

131 HE, L. et al. Deep semantic role labeling: What works and what's next. In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. [S.l.: s.n.], 2017. 45, 49, 55, 58

132 PRADHAN, S. et al. Semantic role labeling using different syntactic views. In: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005. (ACL '05), p. 581–588. Disponível em: <https://doi.org/10.3115/1219840.1219912>. 45, 57

133 TÄCKSTRÖM, O.; GANCHEV, K.; DAS, D. Efficient inference and structured learning for semantic role labeling. *Transactions of the Association for Computational Linguistics*, v. 3, p. 29–41, 2015. 45, 57

134 SUTSKEVER, I.; VINYALS, O.; LE, Q. V. Sequence to sequence learning with neural networks. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2014. p. 3104–3112. 45

135 GRAVES, A.; MOHAMED, A.-r.; HINTON, G. Speech recognition with deep recurrent neural networks. In: IEEE. *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*. [S.l.], 2013. p. 6645–6649. 45, 55

136 KUMAR, A. et al. Ask me anything: Dynamic memory networks for natural language processing. In: *International Conference on Machine Learning*. [S.l.: s.n.], 2016. p. 1378–1387. 45

137 BONIAL, C. et al. English propbank annotation guidelines. *Center for Computational Language and Education Research Institute of Cognitive Science University of Colorado at Boulder*, 2012. 48

138 GOLDBERG, Y. A primer on neural network models for natural language processing. *J. Artif. Intell. Res.(JAIR)*, v. 57, p. 345–420, 2016. 48

139 WANG, Z. et al. Chinese semantic role labeling with bidirectional recurrent neural networks. In: *EMNLP*. [S.l.: s.n.], 2015. p. 1626–1631. 49, 55

140 HARTMANN, N. S.; DURAN, M. S.; ALUÍSIO, S. M. Automatic semantic role labeling on non-revised syntactic trees of journalistic texts. In: SPRINGER. *International Conference on Computational Processing of the Portuguese Language*. [S.l.], 2016. p. 202–212. 49

141 ALVA-MANCHEGO, F. E.; ROSA, J. L. G. Towards semi-supervised brazilian portuguese semantic role labeling: building a benchmark. In: SPRINGER. *International Conference on Computational Processing of the Portuguese Language*. [S.l.], 2012. p. 210–217. 50

142 CARNEIRO, M. G. et al. Semi-supervised semantic role labeling for brazilian portuguese. *Journal of Information and Data Management*, v. 8, n. 2, p. 117, 2017. 50

143 HARRIS, Z. S. Distributional structure. *Word*, Taylor & Francis, v. 10, n. 2-3, p. 146–162, 1954. 53

144 LANDAUER, T. K.; DUMAIS, S. T. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, American Psychological Association, v. 104, n. 2, p. 211, 1997. 53

145 BOJANOWSKI, P. et al. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016. 53

146 BENGIO, Y. et al. A neural probabilistic language model. *Journal of machine learning research*, v. 3, n. Feb, p. 1137–1155, 2003. 53

147 ŘEHŮŘEK, R.; SOJKA, P. Software Framework for Topic Modelling with Large Corpora. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, 2010. p. 45–50. <http://is.muni.cz/publication/884893/en>. 53

148 BIRD, S.; KLEIN, E.; LOPER, E. *Natural language processing with Python: analyzing text with the natural language toolkit*. [S.l.]: " O'Reilly Media, Inc.", 2009. 53

149 MELAMUD, O. et al. The role of context types and dimensionality in learning word embeddings. In: *Proceedings of NAACL-HLT*. [S.l.: s.n.], 2016. p. 1030–1040. 53, 60

150 BENGIO, Y.; SIMARD, P.; FRASCONI, P. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, IEEE, v. 5, n. 2, p. 157–166, 1994. 54

151 HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. *Neural computation*, MIT Press, v. 9, n. 8, p. 1735–1780, 1997. 54

152 GRAVES, A.; SCHMIDHUBER, J. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, Elsevier, v. 18, n. 5, p. 602–610, 2005. 55

153 HUANG, Z.; XU, W.; YU, K. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015. 55

154 TAI, K. S.; SOCHER, R.; MANNING, C. D. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*, 2015. 55

155 LEE, K.; LEWIS, M.; ZETTLEMOYER, L. Global neural ccg parsing with optimality guarantees. *arXiv preprint arXiv:1607.01432*, 2016. 57

156 CHOLLET, F. et al. *Keras*. [S.l.]: GitHub, 2015. <https://github.com/fchollet/keras>. 58

157 ABADI, M. et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. 2015. Software available from tensorflow.org. Disponível em: <https://www.tensorflow.org/>. 58

158 KINGMA, D.; BA, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 59

159 SRIVASTAVA, N. et al. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, v. 15, n. 1, p. 1929–1958, 2014. 59

160 GAL, Y.; GHAHRAMANI, Z. A theoretically grounded application of dropout in recurrent neural networks. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2016. p. 1019–1027. 59

161 CARRERAS, X.; MÀRQUEZ, L. Introduction to the conll-2005 shared task: Semantic role labeling. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the Ninth Conference on Computational Natural Language Learning*. [S.l.], 2005. p. 152–164. 59

162 FUSHIKI, T. Estimation of prediction error by using k-fold cross-validation. *Statistics and Computing*, Springer, v. 21, n. 2, p. 137–146, 2011. 59

Appendix

APPENDIX A – Systematic literature review - support

Figure 14 – The histogram of F_1-score distribution

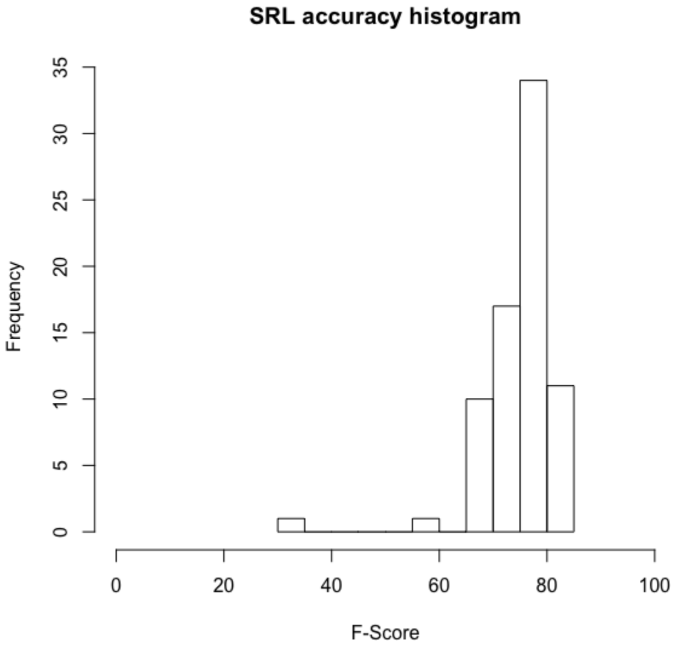
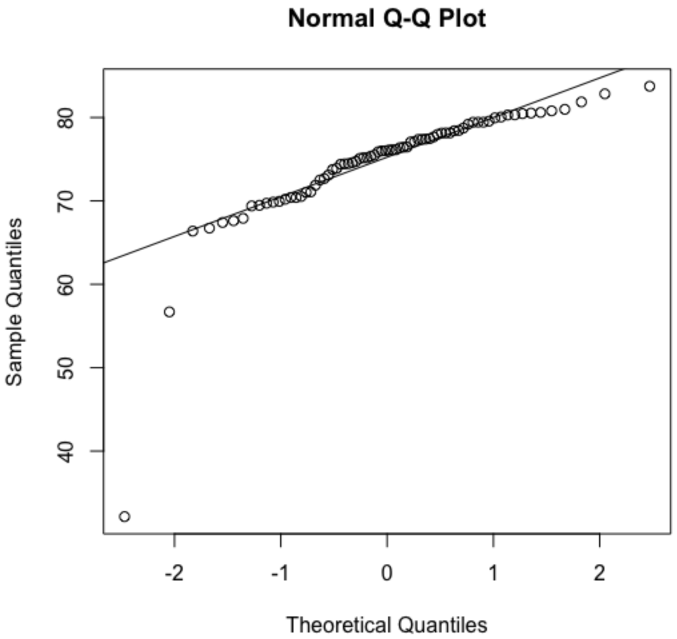


Table 18 – Syntactic combination strategies : Comparable experiments

	Dependency	Constituency	Shallow	Word Embeddings
Dependency		1		1
Constituency	1		6	
Shallow		6		1
Word Embeddings	1		1	

Source: Own Author

Figure 15 – F_1-score distribution from comparable experiments



The Q-Q plot exhibits negatively skewed distribution ≈ -3.42. The Shapiro-Wilk normality test corroborate this result while it rejects the null hypothesis that the sample came from a normally distributed population ($W = 0.71446$, $p\text{-}value = 1.063 \times 10^{-10}$)

Source: Own Author