Semantic Role Labeling in Portuguese: Improving the State of the Art with Transfer Learning and BERT-based Models

Ana Sofia Medeiros Oliveira
Mestrado em Ciência de Dados
Departamento de Ciência de Computadores
2020

Orientador
Alípio Mário Guedes Jorge, Professor Associado, Faculdade de Ciências

Coorientador
Daniel Alexandre Bouçanova Loureiro, Faculdade de Ciências

U.PORTO

FC FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, ______/______/_________

MSc

Abstract

Semantic role labeling is the natural language processing task of determining "Who did what to whom", "when", "where", "how", etc. In this thesis, we explored state of the art techniques for this task in English and applied them to Portuguese. Using a model architecture with only a pre-trained BERT-based model (a language model), a linear layer, softmax and Viterbi decoding, we improved the state of the art performance in Portuguese by over $15F_1$, using their methodology. Moreover, using a robust methodology designed by us, we compared the usage of monolingual and multilingual pre-trained models for this task, and applied techniques such as cross-lingual transfer learning and transfer learning from dependency parsing to improve our results. We provide an evaluation of the models obtained and a heuristic to choose the most appropriate one for different applications based on the obtained results. We find that using the state of the art techniques in multilingual models, these surpass the Portuguese models for the semantic role labeling task in this language, possibly relieving the need to train monolingual models when the data for a specific language is not abundant.

Resumo

A anotação de papéis semânticos é a tarefa de processamento de linguagem natural que determina "Quem fez o quê a quem", "quando", "onde", "como", etc. Nesta dissertação, explorámos técnicas do estado da arte para esta tarefa em Inglês e aplicámo-las a Português. Usando uma arquitectura com apenas um modelo baseado em BERT pré-treinado (um modelo de linguagem), uma camada linear, softmax e descodificação com o algoritmo Viterbi, melhorámos o estado da arte em Português por mais de $15F_1$, a usar a metodologia existente. Adicionalmente, usando uma metodologia robusta criada por nós, comparámos o uso de modelos pré-treinados monolingues e multilingues para esta tarefa, e aplicámos técnicas como aprendizagem por transferência entre línguas e aprendizagem por transferência de um analisador sintático de dependências para melhorar os nossos resultados. É fornecida uma avaliação dos modelos obtidos e uma heurística para escolher o modelo mais apropriado para diferentes aplicações baseada nos resultados obtidos. Descobrimos que, usando técnicas do estado da arte em modelos multilingues, estes superam os modelos em Português na anotação de papéis semânticos, deixando possivelmente de ser necessário o treino de modelos monolingues quando os dados anotados para uma língua específica não são abundantes.

Contents

List of Tables

List of Figures

Acronyms

NLP	Natural Language Processing
SRL	Semantic Role Labeling
POS	Part-of-Speech
NN	Neural Network
FNN	Feedforward Neural Network
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
MLM	Masked Language Model
NSP	Next Sentence Prediction
UD	Universal Dependencies
CV	Cross-Validation
SIGNLL	Special Interest Group on Natural Language Learning of the Association for Computational Linguistics
CoNLL	SIGNLL Conference on Computational Natural Language Learning

Chapter 1

Introduction

1.1 Motivation

Natural Language Processing (NLP) is a field of Artificial Intelligence that has seen an increased importance in recent years. It is a challenging field: natural languages, i.e. human languages, unlike computer languages, are very ambiguous, owing to polysemy or simply ambiguous phrase construction. Moreover, some natural languages suffer from a lack of annotated resources which complicates the development of automatic computational tools.

Due to its challenges, NLP is an important field of study. In fact, NLP applications are an integral part of our lives, e.g., providing quick translations of sentences (Google Translate) or correcting spelling mistakes in documents and text messages, to name a few.

Semantic role labeling (SRL) is a NLP task that, roughly, attempts to automatically discover "Who did what to whom" and "where", "when", "how", etc. It is usually viewed as an intermediate task and can be useful for applications that perform, for example, question answering [66], information extraction [13] and summarization [36], by providing valuable information about the text's meaning to these systems.

A lot of research has been done in SRL, encouraged by several shared tasks from the SIGNLL Conference on Computational Natural Language Learning (the CoNLL shared tasks), but up until now most of this research has focused on English. It has proved difficult to apply the proposed models to a language that has less annotated resources, such as Portuguese: previous attempts led to a large drop in performance when applying the models to Portuguese due to the reduced size of the labeled data set.

Pre-trained contextualized language models, such as BERT, RoBERTa, etc., may offer a way to improve results without the expensive labeling of more data. By pretraining in unlabeled data, the mentioned models learn the language structure before attempting a specific task using the small amount of labeled data available. The work developed in this thesis revolves around the state-of-the-art pre-trained BERT-based models. We will apply a monolingual model, BERTimbau [68], to improve the current state of the art in semantic role labeling.

Additionally, we will investigate a few different approaches for reaching better performance in this task in Portuguese, namely by using multilingual models, using cross-lingual transfer learning and transfer learning from a model fine-tuned in a syntax-related task. We use syntax in our transfer learning approach due to its known close relation with semantics [26]; in fact, a lot of SRL models proposed for English use syntax.

1.2 Research questions

The thesis's research questions are the following:

1. Do new developments in models for semantic role labeling in English bring improvements to the task in Portuguese?
2. How do the state of the art multilingual language models compare to existing monolingual models for the semantic role labeling task in Portuguese?
3. Does cross-lingual transfer learning from English help the multilingual models' performance in semantic role labeling in Portuguese?
4. Can we improve the results of the SRL task by training the language model on another task first?

1.3 Contributions

Our main contributions with this thesis include:

1. A new state of the art model for Portuguese SRL based on pre-trained BERT models.

2. A new methodology for the evaluation of Portuguese SRL models, based on stratified 10-fold cross validation on journalistic text and evaluation on an out-of-domain opinion data set.

3. A comparison of the performance of multilingual and monolingual models for this task.

4. We show that using data from high-resource languages improves the scores of multilingual models in low-resource languages (when both are annotated similarly).

5. We show that pre-training with dependency parsing can help models identify argument span boundaries in situations where the text has mistakes.

6. We provide an heuristic to help future users of our models choosing the best one for their application.

7. We have made available models based on all the tested pre-trained models and trained with the complete set of available data (PropBank.Br + Buscapé), as well as the code used in this work[1].

1.4 Thesis Layout

In this section, we describe the organization of the rest of the thesis. In Chapter 2, we provide a brief introduction of topics related to semantic role labeling that may be useful for understanding the discussion of this task. In Chapter 3, we summarize the work developed in semantic role labeling, both in English and in Portuguese, and describe the data sets and methodology used in this task. The proposed architecture and methodology are described in detail in Chapter 4 followed by Chapter 5 where we describe the experiments performed and present results and their analysis. Finally, in Chapter 6, we present the conclusions of this work and mention some paths of future work.

[1] https://github.com/asofiaoliveira/srl_bert_pt.

Chapter 2

Background Knowledge

This chapter provides an introduction to some concepts related to semantic role labeling, useful for understanding its discussion.

2.1 Natural Language Processing

Natural Language Processing (NLP) is the area of Artificial Intelligence that deals with the computational processing of human languages and attempts to perform useful tasks with them [27]. In this context, "natural" means naturally evolved, such as all languages spoken by humans, as opposed to formal languages, which have strict syntax and semantics, such as Python. The lack of rigorous rules and the ambiguity in natural languages makes NLP a challenging field.

Natural Language Processing is a broad field that encompasses many tasks, from Part-of-Speech (POS) tagging to Machine Translation, Information Extraction and Semantic Role Labeling.

2.2 Base NLP Tasks

In this section, we describe some of the most relevant tasks for semantic role labeling. These are tasks that are frequently used with SRL models, providing valuable input. We will also use dependency parsing as a pre-training task in our experiments.

2.2.1 POS tagging

Part-Of-Speech (POS) tagging is a NLP task that aims to label each token in a sentence with its POS tag. The tags define the token's syntactic and morphological function in a sentence, for example, noun, verb, pronoun, etc. There are various POS tagsets, each defining a different list of possible POS tags [35].

2.2.2 Parsing

Parsing is the task of finding the syntactic structure of a sentence [35]. Below, constituency parsing, dependency parsing and shallow parsing are briefly presented.

Constituency parsing is based on context free grammars, that specify how sentences can be divided into increasingly smaller blocks of words, called constituents [27]. Constituency parse information is usually represented as a constituency-based parse tree, like the one on the left of Figure 2.1. Note that each node in the tree represents a constituent.

Dependency parsing is based on dependency grammars, where sentences are represented as words and binary relations between them [35]. Information is also usually represented as a tree – dependency-based parse tree – where each node is a word and child nodes are dependent on their parent node. An example tree is presented on the right of Figure 2.1. This example is simplified and does not include the grammatical relations between words.

Shallow parsing provides only partial syntactic information instead of a full tree. An example of shallow parsing is chunking, the task of dividing text into non-overlapping sentence segments (chunks) such that syntactically related words end up in the same phrases (e.g. the noun phrase (NP) "A solução" in Figure 2.1) [63].

2.3 Semantic Role Labeling

Semantic Role Labeling (SRL) is a NLP task that consists of determining "Who did What to Whom, How, When and Where", i.e., identifying in sentences events, their participants and properties of the events [35, 47]. In practice, this translates to finding arguments that bear a semantic role in relation to a predicate. Consider sentence 1.

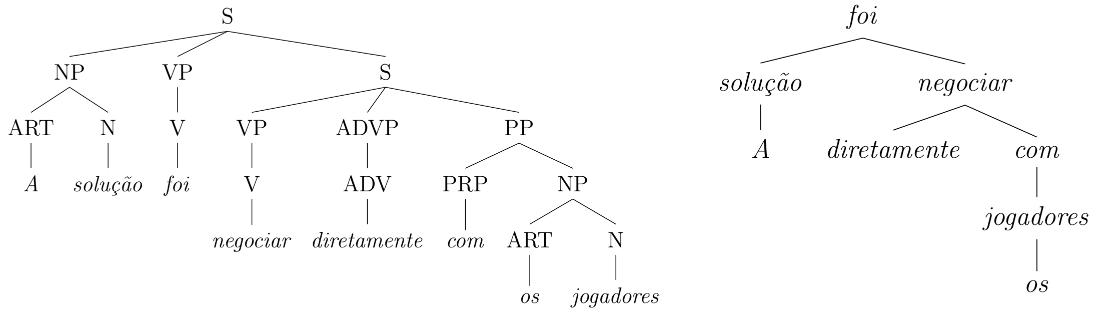


Figure 2.1: Constituency-based (left) and Dependency-based (right) parse trees of the sentence "A solução foi negociar diretamente com os jogadores".

(1) [$_{\text{Agent}}$ John] [$_{\text{Predicate}}$ broke] [$_{\text{Theme}}$ the window].

The predicate determines the event. It says "what" happened [47]. Predicates can be verbs (*decide*), nouns (*decision*), light verbs (*make_ decision*) or adjectives (*nice*). In sentence 1, the predicate is "broke"; the event is that something broke.

The participants are expressed by arguments, which are groups of words, contiguous or not, that add information to the predicate and take on a semantic role in relation to it. Semantic roles are, thus, the role an argument takes in an event, e.g., the agent of the action (for argument "John"). The semantic representation provided by semantic roles can be useful in other tasks, such as information extraction, question answering and machine translation [73].

A proposition is the set of a predicate and its arguments [12], therefore, each sentence has as many propositions as predicates. Note that semantic roles are defined relative to a predicate, so a word can take on different roles for different predicates, but only one role per predicate.

Semantic roles are useful for generalizing across different expressions of the same event [35]. There are many ways to communicate something and while the syntactic structure may change depending on the sentence construction, the semantic meaning stays the same. Consider sentences 1, 2 and 3, an example taken from Gildea and Jurafsky [35].

(2) [$_{\text{Theme}}$ The window] [$_{\text{Predicate}}$ broke].

(3) [$_{\text{Instrument}}$ The rock] [$_{\text{Predicate}}$ broke] [$_{\text{Theme}}$ the window].

In sentences 1 and 3, "the window" is the object of the sentence, while in 2, it is the subject. Nonetheless, in all sentences, "the window" is the thing that broke, the theme of the sentence. As can be seen, arguments can switch positions and become different syntactically, but they always have the same meaning (they are the same semantically).

On the other hand, the subject in sentences 1 and 3 have different roles: "John" is the person who broke the window, the agent, while "the rock" is the instrument used to break the window. Syntactically, however, both are the same.

For machine learning, there needs to be a gold standard data set, i.e., a manually annotated (or automatically annotated and manually revised) data set, so models can learn to make predictions. However, there is no list of agreed upon semantic roles and definitions [47]. Thus, projects aiming to create a SRL data set must define the set of roles to be used (refer to Section 3.1 for examples of such annotation projects).

2.3.1 Types of Semantic Role Labeling

The semantic role labeling task can be formulated in two ways. One is based on constituency-based parsing (span-based SRL) and the other on dependency-based parsing (dependency-based SRL).

In span-based SRL, the objective is to find the word spans, i.e., groups of contiguous words, that constitute arguments of a verb and label them correctly. The spans may or may not be constituents in the parse tree, but due to the close relationship between syntax and semantics, constituents are good argument candidates.

On the other hand, in dependency-based SRL, the objective is to find only the head word of the argument, instead of the whole argument. The difference between these two formulations for the predicate "foi" in sentence "A solução foi negociar diretamente com os jogadores" can be seen in sentences 4 and 5.

(4) $[_{\mathrm{A1}}$ A solução] $[_{\mathrm{Rel}}$ foi] $[_{\mathrm{A2}}$ negociar diretamente com os jogadores]

(5) A $[_{\mathrm{A1}}$ solução] $[_{\mathrm{Rel}}$ foi] $[_{\mathrm{A2}}$ negociar] diretamente com os jogadores

In this dissertation, we focus on span-based semantic role labeling, but in Chapter 3 we give an overview of proposed models for both types.

2.3.2 Semantic Role Labeling Systems

A semantic role labeling system is a set of models (including the SRL model) that takes as input raw text only and outputs the roles for the predicates in the sentence. There can be end-to-end SRL models, where the one model does everything needed. More commonly, several models performing different tasks are combined sequentially to produce the results. We define model as a machine learning algorithm or neural network architecture that has been trained on some data.

In order to deploy a semantic role labeling system, one might need a predicate identification model to give the SRL model the predicates for which to identify roles. Sometimes, systems also have predicate sense disambiguation models. Here, the goal is to discover the appropriate sense for the sentence in a polysemous verb (a verb that can take on different meanings). Both these tasks are sometimes overlooked, and researchers only focus on producing models for the SRL task itself.

2.4 Neural Networks

Neural networks (NNs) are machine learning models of great importance for modern NLP [27].

Neural networks are composed of neural units, such as the one represented in Figure 2.2, which take a set of inputs (x_i) and perform a computation to produce an output (a). In the computation they use a set of weights W_i, a bias term b and a non-linear function f, called the activation function.

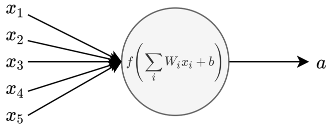


Figure 2.2: Representation of a neural network unit.

Groups of units are called layers and layers combine to form networks.

The simplest type of neural network is the feed-forward (FNN, Figure 2.3). In this network, there is an input layer, an output layer and a varying number of hidden layers between them. The hidden layers learn a representation of the input and the output layer, using this representation, computes an output [35]. In the standard configuration, the units in one layer connect to all units in the following layer. Every connection in the network has its own set of weights.

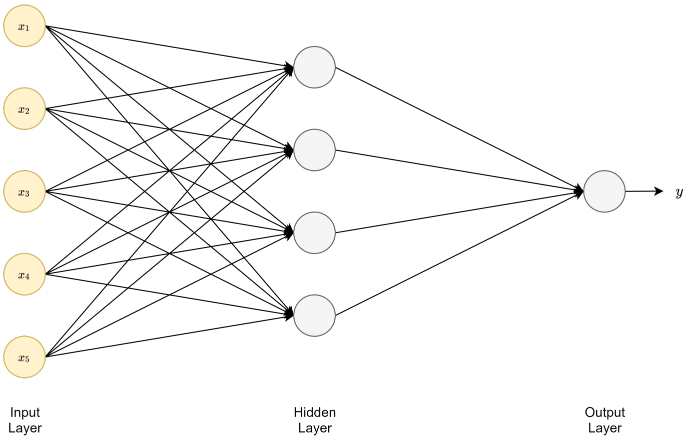


Figure 2.3: Representation of a feed-forward neural network with five inputs, one hidden layer with four units and one output unit.

Neural networks are trainable – once they produce an output, a loss function computes how different the predicted output is from the gold standard output; the network then updates all its parameters W, b in order to achieve a better output.

Another network type, widely used in NLP, is the recurrent neural network (RNN), which includes any neural network that has cycles in its architecture [35]. In this type of network, there can be inputs to a unit from the same layer or from subsequent layers. An example network is presented in Figure 2.4.

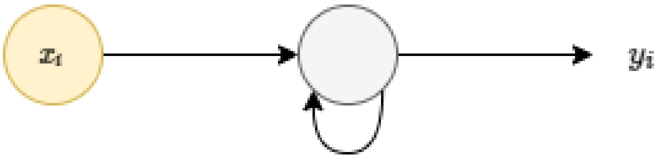


Figure 2.4: Representation of a (simple) recurrent neural network.

RNNs are designed to handle sequences by taking into account the temporal aspect of them: input elements are handled sequentially. Due to the lack of a fixed input layer size, this type of network can handle input sequences of varying length.

Despite being able to handle sequences, standard RNNs have vanishing/exploding gradient problems. The long short-term memory network (LSTM) is a more complex type of RNN that solves the mentioned problem [35], among other advantages. The LSTM is the type of RNN most used in SRL.

2.5 Word Embeddings

Word embeddings are a way to represent words as real-valued vectors. They also bring important semantic information to the NLP tasks that use them [35]. They are based on the distributional hypothesis which states that words that appear in a similar context have a similar meaning. These word representations are, therefore, built based on word occurrences in text and mapped to points in a multidimensional semantic space.

Word embeddings can be learned using unlabeled text in a self-supervised manner and then used in other NLP tasks (transfer learning) [27]. Embeddings can also be learned for sub-words, i.e., parts of words.

The rest of this section briefly introduces some models to obtain word embeddings, namely Word2Vec, GloVe and ELMo. The discussion of BERT is deferred to Section 2.7.2, after the introduction of the attention mechanism.

Word2vec is a class of simple models to learn word embeddings proposed by Mikolov et al. [49]. The idea is to train a classifier to predict how likely a word is to appear in another word's context [35]. There are two model architectures: the Continuous Bag of Words attempts to predict a target word from its context and the Skip-gram attempts to predict the context words from a given target word.

GloVe, introduced by Pennington et al. [55], is a model that uses global counts of word co-occurrence to build word embeddings, i.e., how many times a word appears in the context of the target word in the whole text. This contrasts with word2vec which goes through the text word by word making predictions [27].

ELMo, from Peters et al. [56], is a neural model that finds contextualized word embeddings, i.e. the word representations found by this model are dependent on the entire sentence it appears in. The other models described in this section learn one embedding per word, but polysemous words such as "bat" (which can be an animal or an object) can benefit from context by learning different embeddings for their different meanings.

2.6 Sequence Labeling

Sequence labeling is a type of NLP task where a model receives an input sequence and outputs a label for each element of the input [27]. The usual architecture consists of a word encoder (outputs word representations, such as word embeddings) passed to a sequence encoder (outputs a representation of the sequence), followed by a classification layer. Sequence encoders are typically either based on recurrent neural networks or, more recently, attention-based networks (see Section 2.7).

Span-recognition tasks, e.g. span-based SRL, whose objective is to identify and classify spans of text, can be formulated as a sequence labeling problem using IOB encoding to produce labels for each input word.

IOB encoding is a tagging format introduced by Ramshaw and Marcus [62] that allows the representation of span labels as individual labels for each word. IOB stands for the three base tags, B for the beginning of a span, I for inside a span, and O for outside any span. In SRL, the IOB tags are B-x, I-x and O, where x is a semantic role, i.e., there is a B and an I tag for each possible role. The example from Section 2.3.1 can be seen in Table 2.1 in IOB encoding.

A	solução	foi	negociar	diretamente	com	os	jogadores	.
B-A1	I-A1	B-V	B-A2	I-A2	I-A2	I-A2	I-A2	O

Table 2.1: Semantic role annotation of "A solução foi negociar diretamente com os jogadores" for the verb "foi" in IOB encoding.

2.7 The Attention Mechanism

In the area of machine translation, models initially used RNNs to encode an input sentence into a single vector and predict the whole translation from that vector. This compression led to loss of information in longer sentences. On the other hand, the attention mechanism, first proposed by Bahdanau et al. [4], outputs representation vectors for each input element. This allows the model to focus on different parts of the input sentence for each word being predicted [43]. This mechanism is based on the way a human would translate a sentence, by looking at all of the relevant original words for each word to be translated.

Self-attention is a type of attention mechanism that computes a representation of a sequence by relating its different elements [74].

The rest of this section presents some attention-based models for Natural Language Processing.

2.7.1 Transformer

The Transformer, proposed by Vaswani et al. [74], is a neural network architecture that relies on attention and feed-forward networks.

Unlike RNNs, which are sequential by nature, the Transformer is more parallelizable, allowing faster training. The architecture also captures long-range dependencies more easily. RNNs have difficulty capturing long-range dependencies between tokens, due to the information having to travel a long path (several units) between distant tokens. In contrast, the Transformer architecture has a constant number of operations connecting all inputs.

The Transformer is composed of a stack of N encoders and N decoders. The network does not model token position directly. Hence, positional encoders, indicating the token's position in the sentence, are added to its representation.

This architecture achieved state-of-the-art performance in machine translation tasks [74] and many improvements have since been proposed, among which BERT, introduced below.

2.7.2 BERT

BERT (Bidirectional Encoder Representations from Transformers), proposed by Devlin et al. [16], is a technique for training models for an architecture based on the encoders from the Transformer. A BERT model is a model trained with the BERT technique. The idea is to first pre-train a language model in an unlabeled corpus that can then be applied to multiple tasks. The language structure learned during pre-training helps the model generalize the examples seen in specific tasks.

The model's parameters are, thus, first trained in an unlabeled corpus using two tasks: Masked Language Model (MLM), which aims to predict a masked token using its context (the rest of the sentence), and Next Sentence prediction (NSP), where the model receives as input two sentences and it needs to predict whether or not the second occurs after the first in the corpus. This results in a model that can then be fine-tuned to each specific task using a labeled set or used as features for another model, becoming context-aware word representations.

A BERT model is a deep bidirectional model, since in the MLM objective the model fuses the left and right context to predict the masked words [16].

2.7.3 Other models

RoBERTa (Robustly optimized BERT approach) [42] builds on the BERT technique, changing a few of its design choices. More specifically, the authors investigate the importance of hyperparameter choices and training data size and remove the next sentence prediction pre-training task.

Multilingual BERT[1] and XLM-R [15] are multilingual models trained on monolingual data. They are the multilingual versions of the BERT and RoBERTa models, respectively. Multiligual BERT is trained on the Wikipedia corpora for 104 languages, while XLM-R is trained on a CommonCrawl Corpus in 100 languages.

Multilingual models encode sentences in an embedding space shared by all languages. This allows them to perform zero-shot cross-lingual training, i.e., training in one language and applying the model in another. This is useful for languages with few annotated resources, as it presents an alternative to the expensive data annotation.

[1] https://github.com/google-research/bert/blob/master/multilingual.md

Chapter 3

Semantic Role Labeling

The identification of events and their participants has been studied for a long time. The first known attempt to understand this was by the Indian Panini in his Sanskrit grammar sometime between the 7th and the 4th century BCE. He defined semantic relationships between verb and noun arguments as part of this grammar in a set of rules known as the Karakas [35].

Semantic roles were re-introduced into modern linguistics by Fillmore and Gruber [35]. Since then, many sets of semantic roles have been proposed but there hasn't been one that was well accepted by all linguists [52].

Nevertheless, there are some projects worth mentioning for they resulted in annotated data sets that influenced the surge of automatic approaches to semantic role labeling. The work later developed by Fillmore in case semantics led to the development of the FrameNet data set. The work carried out by Beth Levin in case frame dictionaries led to the development of the VerbNet data set and the PropBank annotation project which was developed by Martha Palmer and colleagues [35].

This chapter focuses on automatic and supervised approaches for semantic role labeling. We begin by describing some relevant projects for the task in Section 3.1. In Section 3.2, we give an overview of proposed models for SRL since the appearance of annotated data sets and in Section 3.3 we discuss the use of syntax in this task. The next section describes the evaluation for the task and mentions some shared tasks that provided a consistent methodology for the comparison of systems. Finally, in Section 3.5, we give an overview of the work developed for this task in the Portuguese language: the data sets available and proposed models.

3.1 English Resources

In this section, we describe PropBank and FrameNet, the two most important projects that built annotated data sets for semantic role labeling, as well as two related projects: NomBank and VerbNet.

3.1.1 FrameNet

FrameNet, introduced in Baker et al. [5], is a project that produced a hand-labeled SRL data set based on the theory of Frame Semantics by Charles J. Fillmore and his colleagues [21]. Verbs, nouns and adjectives are linked to a frame according to shared meaning.

Each frame defines specific semantic roles, called frame elements, and predicates related to the frame, called lexical units. For each lexical unit, there is a set of annotated sentences taken from the British National Corpus, meant to exemplify the possible appearances of frame elements. For example, sentence 6 is an example for the lexical unit "awareness", part of the frame *Awareness.*

(6) The way you move, sit and stand will show [$_{\text{Cognizer}}$ you] [have]^{Supp} a [$_{\text{Degree}}$ greater] [$_{\text{Topic}}$ body] AWARENESS^{Target} and pride .

Frame elements can be core roles, specific to that frame, or non-core roles, which are more general and can be present in all frames. In our *Awareness* example, there are the core roles **Cognizer** and **Topic** and the non-core role **Degree** (*Supp* indicates a support verb[1]). Besides linking related words through frames, the project also links related frames. For example, the frame *Awareness* is linked to *Mental_activity* and *Expectation.*

The FrameNet database has been developed in other languages such as Portuguese, French and Chinese.

3.1.2 PropBank

The Proposition Bank [53], or PropBank, is another project for creating a hand-labeled SRL corpus, originally based on journalistic text from the Penn Treebank [46]

[1] Defined in https://framenet.icsi.berkeley.edu/fndrupal/glossary

– a repository of hand-labeled constituency trees. Initially annotating only verbal predicates, PropBank currently also has annotations for nominal and adjectival predicates as well as complex predicates such as light verbs [8]. It currently also includes other types of text, such as broadcast news and webtext. The most popular data set is the Ontonotes v5.0 [76] that includes different types of text in English, Chinese and Arabic.

There are two types of semantic roles in PropBank: numbered roles A0-A5 and non-numbered roles that represent modifiers or adjuncts, e.g., AM-TMP (temporal modifier) and AM-LOC (location modifier). The AM's are defined globally across predicates, i.e., they have the same meaning for every predicate in the repository. On the other hand, the numbered roles' meaning is specific to each predicate sense and it is defined in the predicate's frame file, even though, in general, A0 and A1 are the Agent and the Patient (the participant who undergoes a change of state) or Theme (the participant most affected by the event), respectively [52, 35].

(7) $[_{\mathrm{A2}}$ This can opener$]$ $[_{Rel}$ opens$]$ $[_{\mathrm{A1}}$ bottles$]$, $[_{\mathrm{AM\text{-}DIS}}$ too$]$!

In sentence 7, we show an example of a sentence taken from the "open" frame file[2]. In this example, the predicate is the verb "open" and it has sense "01", which signifies "(cause to) become open; change of state, free for passage/entry". In the frame file, we can see that A1 means "thing opening" and "A2" instrument. This sentence also includes an AM-DIS – a discourse marker. The meaning of each AM is explained in the current PropBank Annotation Guidelines [7].

PropBank corpora for other languages have been developed. We introduce the Portuguese version of this project in Section 3.5.1

PropBank is more widely used than FrameNet to train semantic role labeling systems, because the latter consists of illustrative sentences for each predicate instead of language-representative annotated text [47]. Furthermore, PropBank is based on the Penn Treebank corpus and therefore has hand-annotated syntactic trees which were fundamental for the development of the first automatic semantic role labelers.

[2] http://verbs.colorado.edu/propbank/framesets-english-aliases/open.html

3.1.3 Other resources

VerbNet [37] organizes verbs into classes according to which semantic roles the verbs allow. It was not used in supervised semantic role labeling due to the lack of an associated annotated corpus [47].

NomBank [48] is an annotation project, similar to PropBank, that annotates the nouns of the Penn TreeBank corpus with semantic roles. Many of the new PropBank noun frame files are taken from the NomBank frame files [8].

3.2 Models

In this section, we present some of the most common algorithms and architectures for the semantic role labeling task. The SRL task can be formulated either as a classification task or a sequence labeling task.

As a classification task, models receive as input information about an argument candidate and a predicate and must decide whether or not the former has a semantic role relative to the latter and which role that is. This is the way SRL is usually formulated in non-neural approaches, which will be described in Section 3.2.1.

As a sequence labeling task, models receive as input the whole sentence and a target predicate and output a IOB label for each word, so that all arguments for a predicate are labeled at the same time. This is in general the way neural approaches treat SRL and we will discuss such approaches in Section 3.2.2.

3.2.1 Non-Neural Methods

Non-neural approaches to SRL follow, in general, an architecture based on three steps: syntactic parsing, argument identification and argument classification.

Argument identification is a binary classification task to distinguishing arguments from non-arguments in a list of candidates. **Argument classification** is a multi-class classification task which aims to label the identified arguments with a semantic role. The classification algorithms used for these steps vary from work to work. Both these classifiers receive as input discrete features generated with the help of a syntactic parse tree.

The syntactic parsers used can be constituency-based, dependency-based or shallow. Constituency parsers have the advantage of reducing the argument candidate list – only constituents are considered as candidate arguments. Nonetheless, Johansson and Nugues [34] showed that dependency parse-based systems, popularized by the CoNLL-2008 shared task, could perform comparably to the constituency parse-based state of the art. On the other hand, producing full syntactic parses is time consuming [52]. For this reason, shallow parses were used as a source of partial syntactic information. However, Punyakanok et al. [59] showed that SRL systems based on shallow parsers yielded results inferior to those based on constituency parsers.

Next, we describe some improvements made to this base system.

Pruning. To deal with the imbalance between non-arguments and arguments in the candidate list, many constituency-based systems introduced a pruning step, specifically the pruning algorithm proposed by Xue and Palmer [79], before argument identification, that discards very unlikely constituents from the argument candidate list [52].

Global Inference. The predictions made by the argument classifier are local, thus there is no certainty that they will make sense when considering the whole sentence. The global inference step combines the local predictions into the most probable global argument structure. There are many ways to combine the predicted roles, for example, Punyakanok et al. [60] used an Integer Linear Program to enforce linguistic and structural constraints. Systems that make global inference are referred to as global, while systems that do not are referred to as local.

System or Model Combination. Another common aspect of non-neural systems is system or model combination, for example, by combining the output of SRL systems based on different syntactic parsers to mitigate errors associated with parsers [52].

The model's use of discrete features has the disadvantage of requiring extensive feature engineering and vast linguistics knowledge to choose an optimal set. Moreover, argument identification and classification require different sets of features and those that help one step may hurt the other [47], doubling the amount of feature engineering needed.

These drawbacks motivated the use of neural networks in SRL, since they can implicitly learn features [14]. The focus, therefore, shifted from coming up with optimal sets of features to designing better networks.

3.2.2 Neural Methods

Nowadays, most NLP tasks use systems based on neural networks [27] and semantic role labeling is no different.

In general, systems using neural networks treat SRL as a sequence labeling task, using word encoders, a sentence encoder and a classifier. They usually perform argument identification and classification at the same time. Some researchers also add a syntax-encoding layer to the architecture, but overall there is an effort to eschew the discrete syntactic features that were popular in non-neural methods.

Word Encoders. Common word encoders for neural networks are pre-trained word embeddings, like GloVe ([73, 32, 40]). Using pre-trained word embeddings helps the system use information from words never seen during training [14]. Some researchers added ELMo word representations, leading to an increase in performance when compared to using only non-contextual word embeddings ([56, 70]). Other forms of word representations are sometimes used, such as character-based representations ([31, 33]) or POS tags embeddings ([44]).

Sentence Encoders. Sentence encoders are neural network architectures, like CNNs, LSTMs or FNNs with attention. Collobert et al. [14] were among the first to apply neural networks to the SRL task. They used a CNN, which allowed them to learn dependencies between words, but only within a fixed-sized window. To achieve a performance comparable to the state of the art, they had to add discrete syntactic features to the model. When comparing a CNN-based SRL system with a LSTM-based one, Zhou and Xu [80] found LSTMs yielded better results.

For this reason, systems based on deep bidirectional LSTMs are more common ([80, 32, 44, 45, 33, 51, 39]). Their ability to preserve both past and future context through many time steps proves useful in SRL due to the long range dependencies commonly found in this task.

Despite their success, LSTMs prove to still have some difficulties in learning long range dependencies between a predicate and its arguments. The model must have a large vector to encode longer sentences, translating to a big memory usage, and it ends up losing information in the longest sentences while wasting memory in the shorter ones [73]. In addition, training LSTM networks is very time consuming, since, due to their sequential nature, no parallelization of computations can be performed [74].

To deal with these problems, researchers started using attention mechanisms in their models ([73, 70, 67, 38]). Attention is highly parallelizable and can model long-range dependencies with a constant number of operations [73].

Recent attention-based models proposed for SRL include the BERT-LSTM model of Shi and Lin [67], the *AllenNLP* BERT-based model and Li et al. [38]'s RoBERTa model.

Global inference and model combination, described in the previous section, are also widely used with neural methods to improve results.

3.3 Syntax

Non-neural approaches relied heavily on syntax, using a syntactic parser to process a sentence to extract features. The use of syntactic features in SRL models stemmed from the close relation between semantic roles and syntax [26].

However, parsing is time-consuming [52] and, in real-world applications, automatic parsers introduce an error which propagates to the semantic role labeling system and limits its performance.

Due to this setback and the difficulty of encoding parse trees in neural networks, many researchers attempted to develop syntax-agnostic models ([14, 80, 32, 44, 51]), achieving state of the art results. The improvements of syntax-agnostic neural SRL models over traditional models were larger in out-of-domain data where traditional models were more conditioned by the errors of parsers [44, 31]. Interestingly, using deep bi-LSTMs, some found that their models seemed to be implicitly learning syntactic structure from the input sentences [80, 44].

The work developed in this setting showed that syntax is not a necessary pre-requisite for good performance in semantic role labeling, contrary to the belief at the time [39]. Nonetheless, syntax can still bring improvements to SRL models and therefore some researchers still choose to include it in their (neural) models ([45, 78, 33]).

3.4 Evaluation

Semantic role labeling systems are evaluated using three measures: precision, recall and F_1 measure. The precision (p) is the proportion of predicted arguments that are correct. Recall (r) is the proportion of true arguments predicted by the system. The F_1 measure is the harmonic mean of precision and recall:

$$F_1 = \frac{2pr}{p + r}. \tag{3.1}$$

These metrics are calculated for each role being predicted, as well as for the total predictions from a model or system (these are called the overall scores).

For span-based SRL, an argument is correct if both its boundaries and its semantic role are correct. Common data sets to evaluate span-based SRL are the data sets from CoNLL-2004 [11], CoNLL-2005 [12] and CoNLL-2012 [57] shared tasks. The first has shallow syntactic information as part of the input and the others have full constituency-based parsing information.

In dependency-based SRL, an argument is correct if the system identified its head word and the correct semantic role. Common data sets are the CoNLL-2008 [71] and CoNLL-2009 [28]. These include dependency-parse trees as part of the input.

All the mentioned data set are based on data sets from the PropBank project. For CoNLL-2004, CoNLL-2005 and CoNLL-2008, only verbal predicates required annotation. CoNLL-2009 and CoNLL-2012 include data sets for languages other than English.

3.5 Semantic Role Labeling in Portuguese

Having reviewed the state of the art in the English language, we now present the work developed in Portuguese for the SRL task. We first describe the resources available in Portuguese for training SRL models and then the models proposed so far, both supervised and semi-supervised.

3.5.1 Resources

Propbank.Br[3] [18] is the (Brazilian) Portuguese version of PropBank and it follows the same annotation style as the original, having a similar role set (the roles existent in Portuguese and their definitions are provided in Appendix A). Unlike the recent

[3] http://143.107.183.175:21380/portlex/index.php/pt/downloads

versions of PropBank, PropBank.Br only annotates arguments related to verbal predicates and is much smaller in size.

There are two versions of this data set from two different data sources. The first version of the data set is based on the Brazilian Portuguese portion of the Bosque corpus, which is a part of the treebank "Floresta Sintá(c)tica" [1]. It contains sentences extracted from the newspaper "Folha de São Paulo" of 1994 which have been hand parsed. We use version 1.1 in our work, which includes predicate and sense annotations that allow us to link verbs to their frame files.

The second version of the data set contains sentences extracted from the PLN-Br corpus [9], also a journalistic corpus based on "Folha de São Paulo". The sentences in this version were automatically parsed by PALAVRAS [6]. The annotation project of this second version produced also a smaller annotated corpus, named Buscapé, which is based on a corpus of product reviews of the same name [29].

The first version is available both in the CoNLL format (format from the CoNLL shared tasks) and in XML, while the second version is available only in XML.

The frame files for PropBank.Br are called Verbo-Brasil[4] [17, 19] and were based on PropBank's frame files.

There is also a Brazilian Portuguese version of FrameNet called FrameNet Brasil[5] and CINTIL-PropBank, also a Portuguese version of PropBank. Since all previous work in Portuguese SRL used PropBank.Br and since annotation differs somewhat for both FrameNet and CINTIL-PropBank, we will use PropBank.Br in this thesis.

3.5.2 Models

In the Portuguese language, few models have been proposed for (automatic) semantic role labeling. Here, we present some of the most important work in this area. We first describe supervised learning methods for Portuguese SRL.

Sequeira et al. [65] propose a SRL model that uses machine learning methods to predict verbs, A0 and A1. They also introduce a data set, BosqueUE, based on the European Portuguese portion of Bosque 8.0, of Floresta Sintá(c)tica [1]. However, we could not find this data set.

[4] Available at http://143.107.183.175:12680/verbobrasil/sobre.php?lang=pt-br

[5] https://www.ufjf.br/framenetbr/

Fonseca et al. [22] describe a system based on that of Collobert et al. [14] to perform several NLP tasks but they do not present any results.

Alva-Manchego and Rosa [2] proposed a benchmark for SRL in Brazilian Portuguese. Their model was based on traditional methods: they used machine learning models with features extracted from the gold parse trees of PropBank version 1 to predict arguments. Furthermore, they created a CoNLL formatted version of this data set, so that future work could use it to make comparable systems.

Fonseca and Rosa [23] built a system similar to that of Collobert et al. [14], but divided the task into two steps – argument identification and classification. Their system did not use parsing information, only word representations of the input.

Hartmann et al. [30] compared the Alva-Manchego and Rosa [2] and Fonseca and Rosa [23] systems using PropBank.Br versions 1.1 and 2. To make a fairer comparison, in the former model, they used automatically parsed trees instead of gold-standard. They found that even using automatically generated syntactic trees, the model of Alva-Manchego and Rosa performed better than that of Fonseca and Rosa.

The most recent model (we could find) for supervised Portuguese SRL is by Falci et al. [20] and follows the architecture of He et al. [32]. It is a 2-layer bi-LSTM model that uses word embeddings and a global inference step with IOB and PropBank constraints.

In terms of semi-supervised work, some work has been done in the Portuguese language, due to the scarcity of annotated data [10]. Alva-Manchego and Rosa [3] describe a proposal for a system using maximum entropy models, without presenting results. Carneiro et al. [10] compare several semi-supervised models for SRL, but only for three verbs: "give", "do" and "say", obtaining results better than the supervised methods of that time.

Chapter 4

Model and Methodology

In the previous chapter, we presented an overview of the work done for semantic role labeling both in English and in Portuguese. It can be noted that there is much less work done in the Portuguese language and that the most recent model proposed ([20]) is based on an a model for English that has been since surpassed. In this chapter, we describe an approach for Portuguese based on the state of the art developments in English.

Our architecture is a pre-trained BERT-based model, with a classifier on top which uses Viterbi decoding. We call pre-trained BERT-based model to any model pre-trained using either BERT or techniques built upon BERT, e.g., RoBERTa or XLM-R.

We first detail our proposed architecture in Section 4.1, followed by an outline of the evaluation procedure to compare to a previous model for Portuguese in Section 4.2. Finally, in Section 4.3, we describe a methodology for a systematic empirical evaluation that enables the comparison between approaches. This methodology will be helpful in the appraisal of the several state of the art techniques we will employ in an attempt to improve the performance of our model.

4.1 The Architecture

The SRL architecture we will use in the experiments is based on *AllenNLP*'s, implemented in their package with the same name [25], which they claim achieves state of the art performance among single models (as opposed to ensembles) for English SRL[1].

[1] https://demo.allennlp.org/semantic-role-labeling

The model architecture is presented in Figure 4.1. It includes a pre-trained BERT-based model on top of which a linear layer, a softmax function and Viterbi decoding are applied. In this section, we will explain each of these components more in depth.

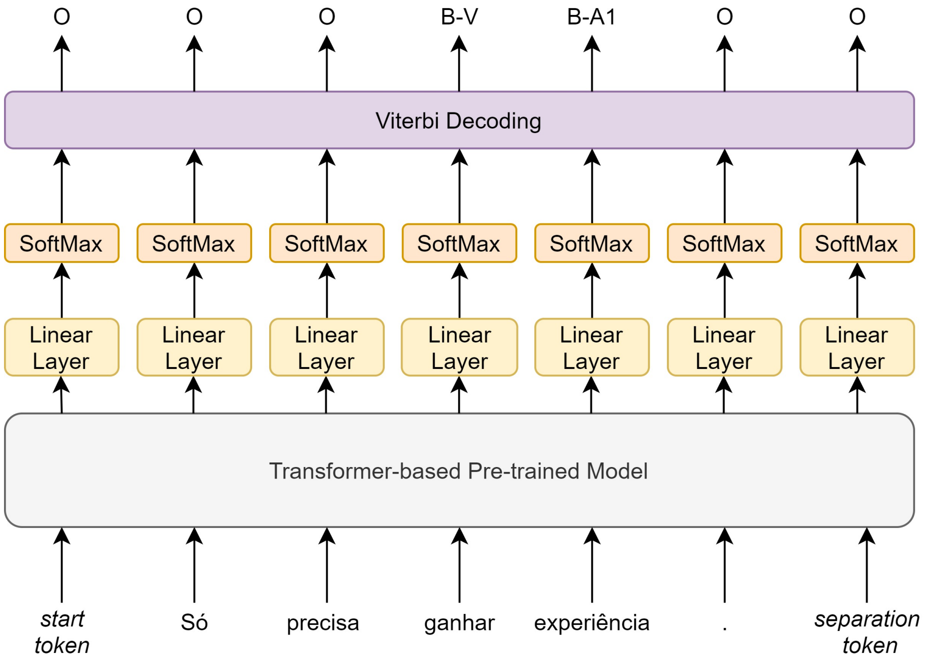


Figure 4.1: Architecture of the models used in this thesis. The model is shown here predicting the argument structure for the verb "ganhar" in the sentence "Só precisa ganhar experiência."

We detail in Section 4.1.1 the way input propositions are processed to be passed to the pre-trained model. In Section 4.1.2, we explain the architecture of a transformer encoder, the basis of a BERT model, and in Section 4.1.3 give a few details about the pre-trained BERT-based models we use in this thesis. Finally in Section 4.1.4, we explain the linear layer, softmax and Viterbi decoding applied to the representations output by the pre-trained model.

4.1.1 Network Inputs

Each input proposition is first tokenized, resulting in sub-words we will refer to simply as tokens, and special tokens are added to the start and end of it.

The inputs to the pre-trained BERT-based model are the sum of three vectors:

- token embeddings – learned embeddings for each token of the input sentence;

- position embeddings – the Transformer does not model token position directly, so learned embeddings representing the positions of each token in the sentence are added to the inputs;
- token type embeddings – this is used as a way to pass the proposition's predicate position to the model.

4.1.2 The Transformer Encoder

The Transformer encoder consists of a stack of L encoder layers, each comprised of two sub-layers. In Figure 4.2 we present one encoder layer. As can be seen in the representation, the inputs to the layer, of size H, (which can either be the inputs to the system or the output of the previous layer) passes first through a multi-head attention sub-layer and then through a fully connected feed forward network sub-layer and outputs a vector of size H.

All sub-layers have a residual connection around them (the input to the sub-layer is summed to its output) and layer normalization after, so the input to the following sub-layer is $LayerNorm(x+SubLayer(x))$. Additionally, dropout [69] and label smoothing [72] are applied for regularization.

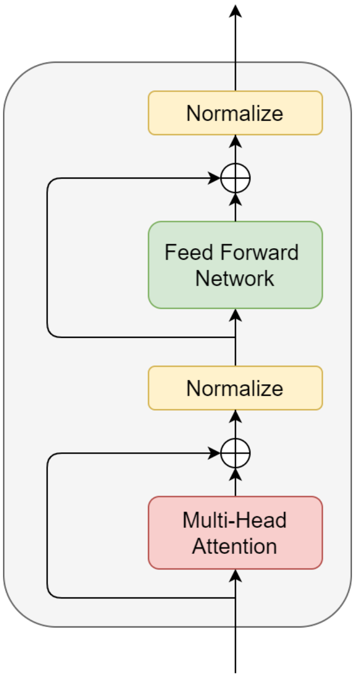


Figure 4.2: A layer from a Transformer encoder, drawn after Vaswani et al. [74].

4.1.2.1 The Attention Mechanism

Transformers use an attention mechanism called scaled dot-product attention. There are three inputs: a query (q) that represents the target token, and a set of key (K)-value (V) pairs associated with all input tokens. The calculation is summarized in Eq. 4.1

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{4.1}$$

First, we compute the compatibility between query and keys by calculating the dot-products between these vectors and dividing them by $\sqrt{d_K}$ (the dimensionality of the query and of each key). Next, we apply a softmax function and obtain weights (the bigger the weights, the more important the associated key-value pair for a target token) and finally multiply the weights with the values.

In practice, the representations for all tokens are calculated at the same time, using a matrix Q containing the queries for all inputs.

4.1.2.2 Multi-Head Attention

Multi-head attention is a technique that allows the model to learn different representations of the same input concurrently without increasing the computational cost.

This technique runs A attention heads, i.e., scaled dot product attention functions, in parallel, each with a smaller projected version of the original Q, K and V vectors. The outputs of the A functions are then concatenated and passed through a linear layer (a NN layer with no activation function).

The calculations are summarised in Eq. 4.2.

$$\begin{aligned} MultiHead(Q, K, V) &= Concat(head_1, ..., head_h)W^O \\ \text{where}\, head_i &= Attention(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{4.2}$$

In practice, Q, K and V are all the same vector, the input to the encoder layer.

4.1.2.3 Fully Connected Feed Forward Network

Another component in the encoder layer is a fully connected feed forward network with one hidden layer with d_{ff} units and a ReLU activation function (in all our models $d_{ff} = 4H$, where H is the size of both the FNN's input and output). This network is independently applied to each representation returned by the previous sub-layer.

4.1.3 Pre-trained Models Specifications

A BERT-based model is a transformer encoder pre-trained in a specific set of tasks. Having described its underlying architecture, we now give some details about each model's pre-training parameters. The models presented here are the pre-trained models we will use in our experiments.

The BERTimbau models (henceforth **brBERT**), by Souza et al. [68], are Portuguese models pre-trained with the BERT technique in the brWac corpus [75] with a vocabulary size of 30k tokens and the usual training objectives – MLM and NSP.

The multilingual cased BERT model, henceforth **mBERT**[2] is a multilingual version of the Devlin et al. [16]'s BERT. It is trained in monolingual Wikipedia data for 104 languages, including Portuguese. The model is pre-trained using the two usual training objectives, MLM and NSP, with no indication of the languages being processed. A shared vocabulary with a size of 110k tokens is created with a sampling of the Wikipedia data set that makes high-resource languages under-sampled and low-resource languages over-sampled.

XLM-R are multilingual versions of the RoBERTa model introduced in Conneau et al. [15]. They are pre-trained in monolingual, clean CommonCrawl data for 100 languages, one of which being Portuguese. The pre-training data for these models is larger than for the others. The only training objective used is the MLM objective using the monolingual data, with no indication about the languages of each training sentence and no cross-lingual data. A shared vocabulary with a size of 250k tokens is created in a similar manner to mBERT's vocabulary.

In Table 4.1, we present an overview of the pre-trained model's parameters by order of increasing number of parameters. Multilingual models naturally have more parameters than their monolingual counterparts due to their larger vocabulary. For more details

[2] https://github.com/google-research/bert/blob/master/multilingual.md

about the models and their pre-training, the reader is referred to Souza et al. [68], the mBERT github[2] and Conneau et al. [15].

Model	L	H	A	#Parameters	Corpus	Tokenizer
BERT$_{base}$	12	768	12	110M	brWaC	WordPiece
mBERT	12	768	12	172M	Wikipedia	WordPiece
XLM-R$_{base}$	12	768	12	270M	CommonCrawl	SentencePiece
BERT$_{large}$	24	1024	24	340M	brWaC	WordPiece
XLM-R$_{large}$	24	1024	16	550M	CommonCrawl	SentencePiece

Table 4.1: The number of encoding layers (L), the hidden size (H), the number of attention heads per layer (A), the total number of parameters for the transformer encoder (#Parameters), the corpus used for pre-training the models, and the tokenizer each model uses for each pre-trained model used in this thesis.

4.1.4 Linear Layer, Softmax and Viterbi

In our architecture, Figure 4.1, after obtaining representations from the pre-trained BERT-based model, the representation vectors corresponding to each input token are independently passed to the same randomly initialized linear layer and softmax.

Softmax is a function that normalizes a vector and turns it into a probability distribution [35]. The function is in Eq. 4.3 for a vector v of dimensionality d.

$$\text{softmax}(v) = \frac{e^{v_i}}{\sum_{j=1}^{d} e^{v_j}}. \tag{4.3}$$

The result is a set of probabilities for all possible labels for each input token. The probabilities for the whole sentence are given to a Viterbi decoding algorithm, a decoding algorithm that finds the most likely tag sequence [35]; it is a form of global inference, described in Section 3.2.1. This algorithm is conditioned on the output having to be a valid IOB sequence, i.e., "I-x" tags must be preceded either by "B-x" or "I-x", where x is any semantic role label.

4.2 Methodology for Comparison with Portuguese State of the Art

We compare the described architecture using the Portuguese pre-trained models with the model proposed by Falci et al. [20], which we will refer to as our baseline. We compare only to this work because it is the only one that details their methodology. We first pre-process the PropBank.Br v1.1 CoNLL-format data set[3] using the same steps as the baseline, described below, and then perform 20-fold cross-validation with the same data set divisions. The pre-processing steps were the following:

1. Eliminated all propositions that had words with more than one label. Each word in a sentence can only have one semantic role relating to each predicate, so these were considered annotation mistakes.

2. Re-joined all words formed by prepositional contractions which were separated by the annotators, since word contractions such as "do" ("de" + "o") are always used in Portuguese, even in formal writing.

3. Separated groups of words such as "em_termos_de" into individual words. These words were concatenated by the annotators of the corpus due to being either expressions or proper names. In a real setting, such words would not be concatenated and so the model needs to know how to use them separately.

4. Removed all propositions that had a continuation role, e.g. "C-A0".

The first column of Table 4.2 has the number of times each role appears in the pre-processed data set. After training, the models were evaluated with the script provided in the CoNLL-2005 Shared Task *srl-eval.pl*[4]. The results are the precision, recall and F_1 measure for each role and the overall scores for these metrics.

4.3 Methodology for Comparing our Models

To compare our own models with each other, we chose to perform 10-fold cross-validation (CV). Hence, we used stratified sampling for multi-label data [64] to create

[3] Available at http://143.107.183.175:21380/portlex/index.php/en/downloadsingl

[4] Available at https://www.cs.upc.edu/ srlconll/soft.html and through the *allennlp_ models* package.

the folds from the complete PropBank.Br data set (versions 1.1 and 2), using the *iterative-stratification*[5] package. Each fold produced is used as a test set once and the remaining folds are again sampled to obtain a validation set of approximately the same size of the test set. The division of data in each model run is, therefore, 80% for the training set, 10% for validation set and 10% for the test set. In addition to these, the out-of-domain Buscapé corpus is used to test all model runs. In other words, each fold's model is evaluated in the appropriate test set from the PropBank.Br CV folds and in the Buscapé set.

Buscapé constitutes a more difficult and yet at times more adequate data set. Looking at the sentences it includes, there are plenty of misspelled words, poorly constructed sentences, verbs conjugated in the wrong tense and accents missing or in excess. This, however, is very common in texts written on the internet in more casual settings. Thus, it may be an important benchmark to have if we want a model to be used in these situations.

In regards to data pre-processing, we used the XML files of all three data sets (PropBank.Br v1.1, PropBank.Br v2 and Buscapé)[6] in order to pre-process them in the same way. As in the previous methodology, we eliminated propositions with more than one label for a word, we separated expressions joined with "_" and re-joined words formed by prepositional contractions. Additionally, we removed arguments labeled as "AM-MED" or "AM-PIN" because there is no mention of these labels in the annotation guides[7] and for the corpora PropBank v2 and Buscapé, we removed any propositions with flags "WRONGSUBCORPUS", "LATER" or "REEXAMINE", since, according to the guide, these indicate something wrong with the sentence that prevents its annotation.

The way continuation arguments were annotated differed from previous works. In our pre-processing, we only annotated an argument with a "C-" role if the words that constituted it were non-contiguous to the original argument, whereas the CoNLL file for PropBank v1.1 separates arguments if they are not in the same node in the constituency-based syntactic tree, even if they are contiguous.

After pre-processing, the data sets to be used had the role counts presented in the

[5] Available at https://github.com/trent-b/iterative-stratification

[6] Available at http://143.107.183.175:21380/portlex/index.php/en/downloadsingl

[7] The annotation guide for PropBank v1.1 is available at http://143.107.183.175:21380/portlex/images/arquivos/propbank-br/propbank.br%20tutorial.pdf and for PropBank v2 at http://www.nilc.icmc.usp.br/semanticnlp/includes/projects/propbankbr-/files/MANUAL%20DE%20ANOTACAO%20DO%20PROPBANK%20v5.pdf

columns "Met. 2" of Table 4.2.

As in the previous methodology, the models were evaluated with the official script from the CoNLL 2005 Shared Task.

Semantic role	Met. 1	Met. 2	
	PropBank v1.1	PropBank	Buscapé
A0	2891	6274	258
A1	5061	11680	452
A2	1290	2991	142
A3	139	296	6
A4	111	165	11
A5	1	1	0
AM-ADV	346	675	38
AM-ASP	0	179	38
AM-CAU	141	293	28
AM-COM	0	27	1
AM-DIR	13	26	0
AM-DIS	288	662	37
AM-EXP	0	2	1
AM-EXT	74	206	19
AM-GOL	0	30	3
AM-LOC	672	1444	39
AM-MNR	384	969	64
AM-MOD	0	251	23
AM-NEG	322	769	77
AM-NSE	0	39	14
AM-PAS	0	285	7
AM-PRD	169	360	9
AM-PRP	143	373	19
AM-REC	8	9	0
AM-TML	0	12	14
AM-TMP	1082	2441	64
V	5600	13665	709

Table 4.2: Number of appearances of each semantic role in the pre-processed data sets to be used to compare with previous results in Portuguese (Met. 1) and among our own models (Met. 2).

Chapter 5

Experiments and Results

In this chapter, we describe the experiments run in the project, specify the implementation details and present the results. We will run several experiments, to answer the following questions:

1. Do new developments in models for semantic role labeling in English bring improvements to the task in Portuguese? (Section 5.2)

2. How do the state of the art multilingual language models compare to existing monolingual models for the semantic role labeling task in Portuguese? (Section 5.3.4)

3. Does cross-lingual transfer learning from English help the multilingual models' performance in semantic role labeling in Portuguese? (Section 5.3.5)

4. Is it useful at all to use the Portuguese data or can we rely on models trained with English data only? (Section 5.3.6)

5. Can we improve the results of the SRL task by training the language model on another task first? (Section 5.3.7)

In Section 5.3.8, we present the results of a statistical significance test of the models from Section 5.3. In Section 5.3.9, we detail the process of choosing an appropriate model for an application.

5.1 Implementation Details

The architecture was implemented in Python using the package *AllenNLP* [25], the mBERT and XLM-R models trained by *Hugginface Transformers* [77] and the Portuguese BERT model trained by *neuralmind-ai* [68] and built on PyTorch [54].

In each experiment, all models are trained with the same hyperparameters, since the focus of this work is to compare different approaches in the same settings. With the exception of number of epochs, batch size and learning rate, all hyperparameters used followed the values set for *AllenNLP*'s English BERT SRL system[1]: embedding dropout of 0.1, optimizer *huggingface_ adamw* without bias correction and a slanted triangular learning rate scheduler.

The code used for these experiments, the tool to choose the best model for a specific application and the models trained on all the data are available in https://github.com/asofiaoliveira/srl_

5.2 Comparison with Portuguese state of the art

We ran both brBERT$_{\text{base}}$ and brBERT$_{\text{large}}$ on the PropBank.Br v1.1 corpus preprocessed as described in Section 4.2 with a batch size of 16 and a learning rate of 4×10^{-5} for up to one hundred epochs with early stopping after ten epochs without improvement. The parameters are based on *AllenNLP*'s model configuration, but due to memory constraints, the batch size had to be lowered (from 32 to 16); the learning rate was then lowered as well (from 5×10^{-5} to 4×10^{-5}) due to the fact that some folds' models were getting stuck on predicting no arguments.

In Table 5.1, we present the results for the best model out of all folds and the average F_1 across folds[2], since these were the metrics reported by our baseline. The values for the baseline are taken from Falci et al. [20].

The BERT models give better results than the bi-LSTM model of the baseline. The improvements are of $15.85F_1$ and $16.91F_1$ for the brBERT$_{\text{base}}$ and brBERT$_{\text{large}}$ models, respectively. This was expected due to the superior performance of BERT models in the English language. The difference is larger than the one observed in English,

[1] Parameters detailed in https://github.com/allenai/allennlp-models/blob/v1.0.0rc3/training_config/syntax/bert_base_srl.jsonnet

[2] Note that throughout the text, whenever average results are mentioned, we mean averaged across folds.

Model	**Best Model**			**Average F_1**
	p (%)	r (%)	F_1	
Baseline	67.62	68.75	68.18	65.63
brBERT$_{\text{base}}$	84.24	85.37	84.80	81.48
brBERT$_{\text{large}}$	85.98	84.53	85.25	82.54

Table 5.1: Comparison of our proposed BERT-based model with the bi-LSTM-based baseline on 20-fold cross-validation. The results for the baseline were taken from Falci et al. [20].

however, likely due to the fewer resources used in Portuguese, which hindered the LSTM-based model.

5.3 Comparing our Models

In the previous section, we compared our proposed architecture to an existing model proposed for Portuguese following its own evaluation methodology. The resulting models are simple (just a language model with a FFN on top) and follow the state of the art of the English language: a single monolingual BERT-based model, trained in the available data for our language.

In this section, we perform a more robust evaluation of the different pre-trained BERT-based models by following a different methodology, described in Section 4.3, which uses more data and different pre-processing steps. We also apply some state of the art techniques, such as cross-lingual transfer learning.

5.3.1 Motivation and Implementation Details

The advantage of multilingual models is that they can learn the task in other languages and boost performance in Portuguese. Compared to English, Portuguese has much less SRL annotated data, i.e. Portuguese is a low-resource language, while English is a high-resource language. Li et al. [38] showed that less data leads to poorer performance of models, even with powerful models as the ones being tested – a drop from $86.47 F_1$ to $75.96 F_1$ with the RoBERTa model when using only 3% of the CoNLL-2012 data. Since the annotation of more data is expensive, one way to attempt to mitigate this impact is to train a SRL multilingual model in other languages and use the trained model parameters as a starting point for training in Portuguese. This type of cross-lingual transfer learning, where information from high-resource languages is used in a low-resource language, has already been proven useful for other tasks [41].

We ran all of the models in this section with a batch size of 4 (due to memory constraints) and a learning rate of 1×10^{-5} for up to one hundred epochs with an early stopping after ten epochs without improvement. The proposed architecture with the different pre-trained models was run in two scenarios:

1. Fine-tuning only with Portuguese data;
2. Fine-tuning first with pre-processed CoNLL-2012 data for five epochs, followed by fine-tuning with Portuguese data (only for multilingual models, represented by a superscript "+En", e.g., $\text{brBERT}^{\text{+En}}_{\text{base}}$).

The CoNLL-2012 [57] data had to be pre-processed to match Portuguese data: some instances of the data set were removed, due to their size and to keep the batch size at 4; all roles that do not exist in the Portuguese data set were removed from the data and reference numbered arguments (denoted with "R-Ax" where x is a number) were replaced with the non-reference role label ("Ax") and the original argument annotations for these roles (the annotations of arguments that had the respective "Ax") were eliminated.

In the following subsections, we describe the obtained results, first looking at their overall performance (Section 5.3.2) and then analysing more detailed results. We first compare the models using only training scenario number 1 (Section 5.3.4), then we present the differences brought by training with English data and compare models from both scenarios (Section 5.3.5). In Section 5.3.6, we present the results of fine-tuning the multilingual models in English and testing it in Portuguese. In Section 5.3.7, we analyze the impact on fine-tuning first in dependency parsing for the best pre-trained models in previous sections. In Section 5.3.8, we present the results of a statistical significance test for the models in the section. Finally, in Section 5.3.9, we give some concluding remarks and advice on choosing the best model.

5.3.2 Overall Results

Table 5.2 presents the average overall precision, recall and F_1 measure of the 10 models (one for each fold) in their respective test sets and in the Buscapé set. The average results for each semantic role for all models can be found in Appendix B.

Model	**Average of Test Folds**				**Average of Buscapé**			
	p (%)	r (%)	F_1	δF_1	p (%)	r (%)	F_1	δF_1
$\text{brBERT}_{\text{base}}$	75.78	76.83	76.30		74.00	72.68	73.33	
$\text{brBERT}_{\text{large}}$	76.65	78.20	77.42		**75.58**	**74.14**	**74.85**	
$\text{XLM-R}_{\text{base}}$	74.42	76.04	75.22		73.12	72.54	72.82	
$\text{XLM-R}^{\text{+En}}_{\text{base}}$	76.09	76.93	76.50	1.28	74.29	73.22	73.74	0.92
$\text{XLM-R}_{\text{large}}$	76.74	78.47	77.59		74.36	73.34	73.84	
$\text{XLM-R}^{\text{+En}}_{\text{large}}$	**77.71**	**78.75**	**78.22**	0.63	75.36	73.77	74.55	0.71
mBERT	72.34	73.20	72.76		67.10	66.70	66.89	
$\text{mBERT}^{\text{+En}}$	74.22	75.56	74.88	2.12	69.41	68.98	69.19	2.3

Table 5.2: Average of results of each model in the test set and Buscapé set.

Note that the values of the monolingual brBERT models are smaller than the ones reported in the Section 5.2 because the previous results are from validation sets and these are from test sets. Moreover, these are evaluated in more data, with more "difficult" roles (roles with few appearances) and include continuation arguments, which were removed for the previous section (see all the differences in Sections 4.2 and 4.3).

As expected, large models have a superior performance with respect to their base counterparts. XLM-R performs better than mBERT, which had already been reported for other tasks ([15]). Moreover, all models have a drop in all measures in the out-of-domain Buscapé set when compared to the test folds of the PropBank data set. This is likely due to this data set being more difficult, as discussed in Section 4.3. The performance drop is larger in recall than in precision, meaning the model is having more difficulty in predicting the correct arguments.

When training only with Portuguese data, $\text{brBERT}_{\text{large}}$ and $\text{XLM-R}_{\text{large}}$ have similar scores in the test sets, but the monolingual model outperforms the multilingual by approximately $1F_1$ in the out-of-domain data set. We believe that the monolingual model, having been pre-trained only in Portuguese data, has learned a better language structure, compared to XLM-R, and, therefore, can better understand the Buscapé data, despite the errors it contains. As for the base models, there is a larger difference between $\text{brBERT}_{\text{base}}$ and $\text{XLM-R}_{\text{base}}$ in the test set than in the Buscapé set, but the difference is still relatively small (approximately $1F_1$). The mBERT model performs the worst: over $3F_1$ points below the monolingual base model.

When including cross-lingual transfer learning, all multilingual models get an increase in their scores. The less powerful the model, the larger is this increase. In this scenario, XLM-R$_{\text{large}}$ achieves the best score in the average of the test folds, but is still slightly behind brBERT$_{\text{large}}$ in Buscapé. On the other hand, XLM-R$_{\text{base}}$ improves enough to become on par with brBERT$_{\text{base}}$ in both data sets. As before, mBERT underperforms compared to all other models.

5.3.3 Argument Identification and Classification

The SRL task can be seen as two sub-tasks: argument identification and argument classification. The performance in the classification, and, therefore, the scores obtained in SRL, are constrained by argument identification, since only correctly identified spans can be checked for the correct label. Table 5.3 reports the average precision, recall and F_1 measure for argument identification.

Model	**Average of Test Folds**			**Average of Buscapé**		
	p (%)	r (%)	F_1	p (%)	r (%)	F_1
brBERT$_{\text{base}}$	83.29	84.4	83.84	82.55	81.33	81.93
brBERT$_{\text{large}}$	83.53	85.22	84.36	83.2	81.88	82.53
XLM-R$_{\text{base}}$	82.19	83.91	83.04	82.32	81.86	82.08
+CoNLL-2012	83.35	84.21	83.78	82.69	81.61	82.14
XLM-R$_{\text{large}}$	83.68	85.56	84.6	82.55	81.66	82.1
+CoNLL-2012	84.34	85.45	84.89	83.26	81.76	82.5
mBERT	81.72	82.67	82.19	79.68	79.32	79.49
+CoNLL-2012	82.44	83.88	83.15	80.24	79.88	80.05

Table 5.3: Average of argument identification of each model in the test set and Buscapé set.

Additionally, the total error of the model can be decomposed in the error in identifying argument spans and the error in classifying the correctly identified spans. In Figure 5.1, we report the average error for each model in each data set decomposed into error from argument identification ("Arg Id") and from argument classification ("Arg Class").

Several things are noticeable in this figure. Firstly, the error from the non-identification of arguments is much larger than the error from the mislabeling of arguments for all models and both data sets. Hence, our models are better at attributing semantic roles

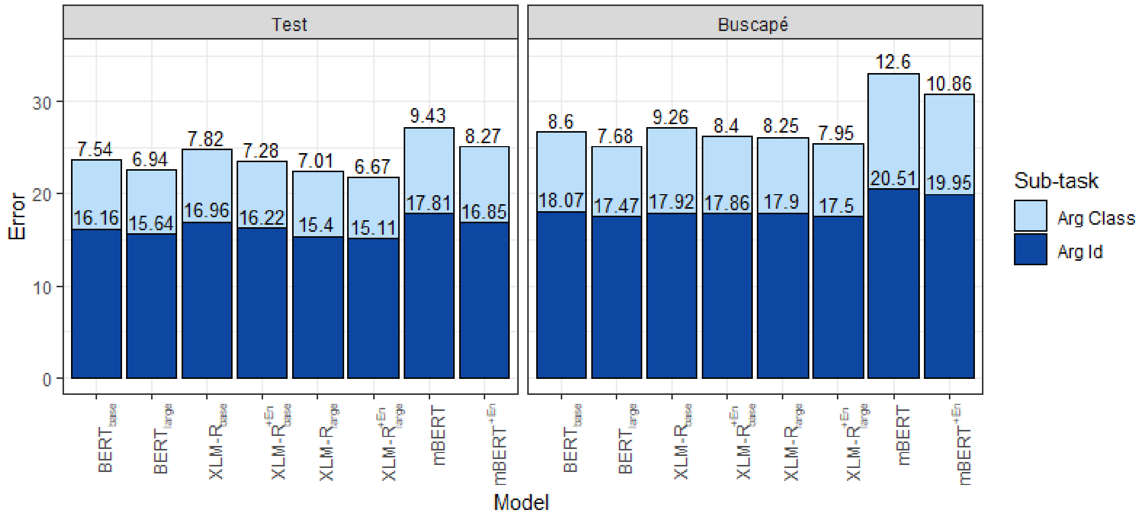


Figure 5.1: Average error in F_1 from argument identification ("Arg Id") and argument classification ("Arg Class") for each model.

than at identifying argument spans. Secondly, all errors are larger in Buscapé, so we cannot attribute the drop in performance to one of these sub-tasks – both are worse. Lastly, the difference in F_1 measure between models is due both to differences in the identification and the classification of arguments.

The results reported so far refer to the overall measures, i.e., the compound scores from all semantic roles. We now look at the results for each semantic role, to understand if the differences in the overall measures come from a difference in the types of roles identified or just poorer general performance. Additionally, it is important to know if the models can identify some roles which may be more important for applications; for example, for information extraction, temporal and location modifiers as well as numbered roles will be more important than auxiliary verbs.

5.3.4 Comparing the Models Trained Only in Portuguese Data

How do the state of the art multilingual language models compare to existing monolingual models for the semantic role labeling task in Portuguese?

In Figures 5.2 and 5.3, we present the results of SRL (argument identification + argument classification) and of argument identification for base and large models, respectively. Performance is similar for each role, so less powerful models are not losing performance because they do not predict certain classes, but because they are worse in most of them.

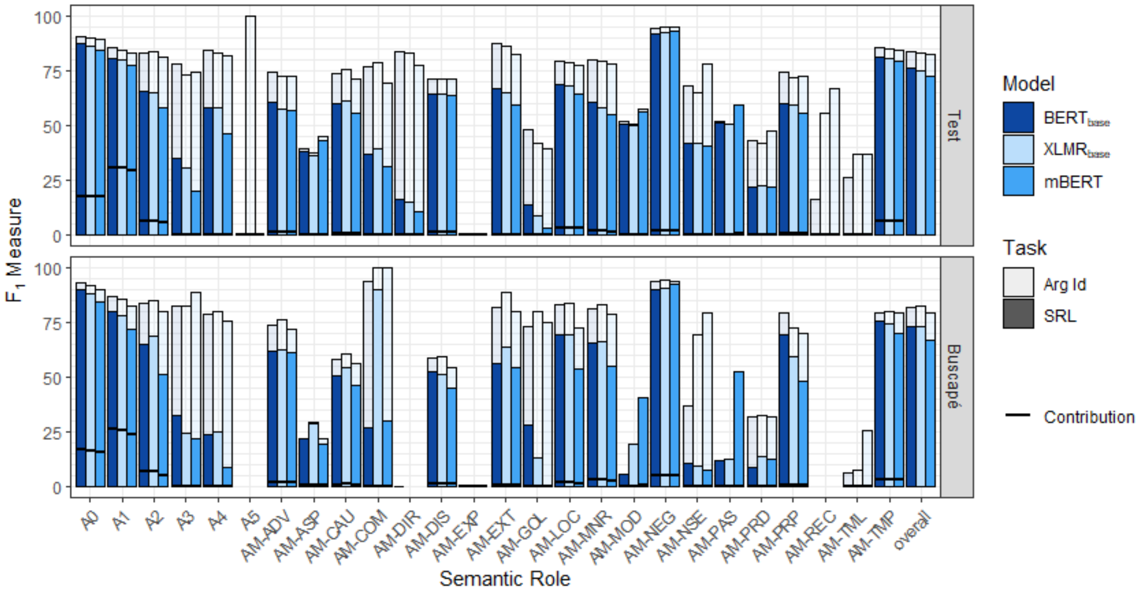


Figure 5.2: Bars – Average SRL (opaque) and argument identification (more transparent) results for the base models in both the test sets and the Buscapé set. Horizontal black lines – Contribution of each role to the overall score.

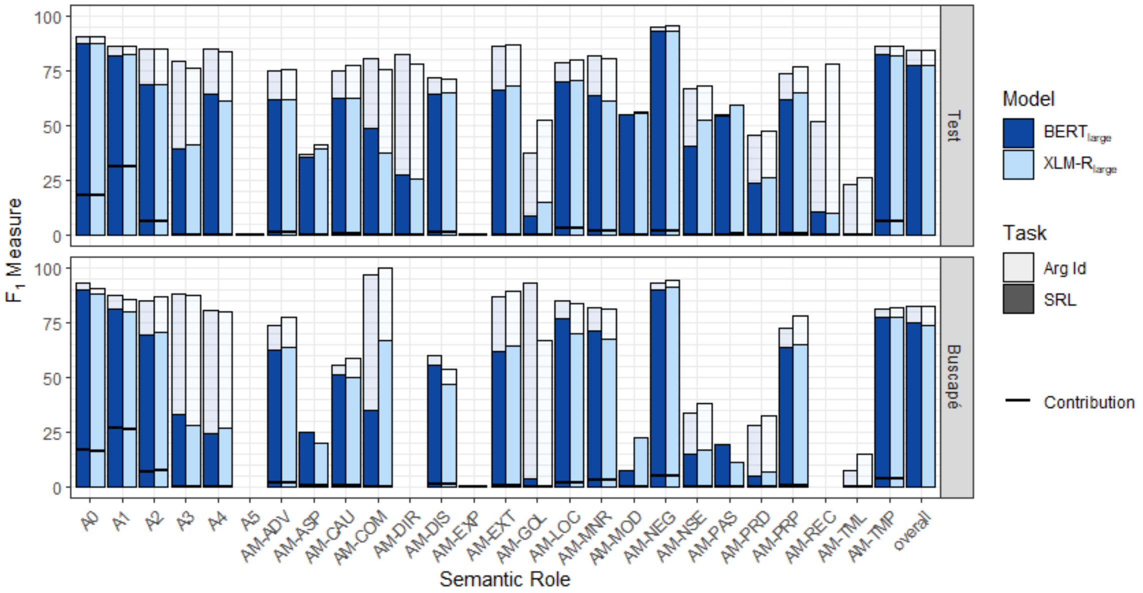


Figure 5.3: Bars – Average SRL (opaque) and argument identification (more transparent) results for the large models in both the PropBank CV test sets and the Buscapé set. Horizontal black lines – Contribution of each role to the overall score.

It is interesting to observe in both figures the low performance of numbered roles A3 and A4. The arguments are being identified ("Arg Id" is high) but most are mislabeled. Looking at the confusion matrices (not displayed), we see these arguments are mostly labeled either with another numbered role or with a modifier whose definition may appear in a numbered role (e.g., for the verb "vir" (come), A3 and A4 are the place where (something) comes from and the place (something) goes; these could be confused with AM-LOC, the location modifier, since they both indicate locations). The low performance is very likely due to the definitions of these semantic roles being flexible, i.e., different for each verb, combined with the low number of appearances. The model can determine the presence of an argument but not which role it corresponds to for that verb.

Recall that the overall performance is computed using all argument occurrences. Thus, for example, in Figure 5.2, we have the XLM-R$_{\text{base}}$ model outperforming brBERT$_{\text{base}}$ in Buscapé in several roles (AM-COM, AM-PRD, A4, A2, AM-CAU, AM-ASP, etc.) but these roles count little to the overall performance, since they have relatively few appearances.

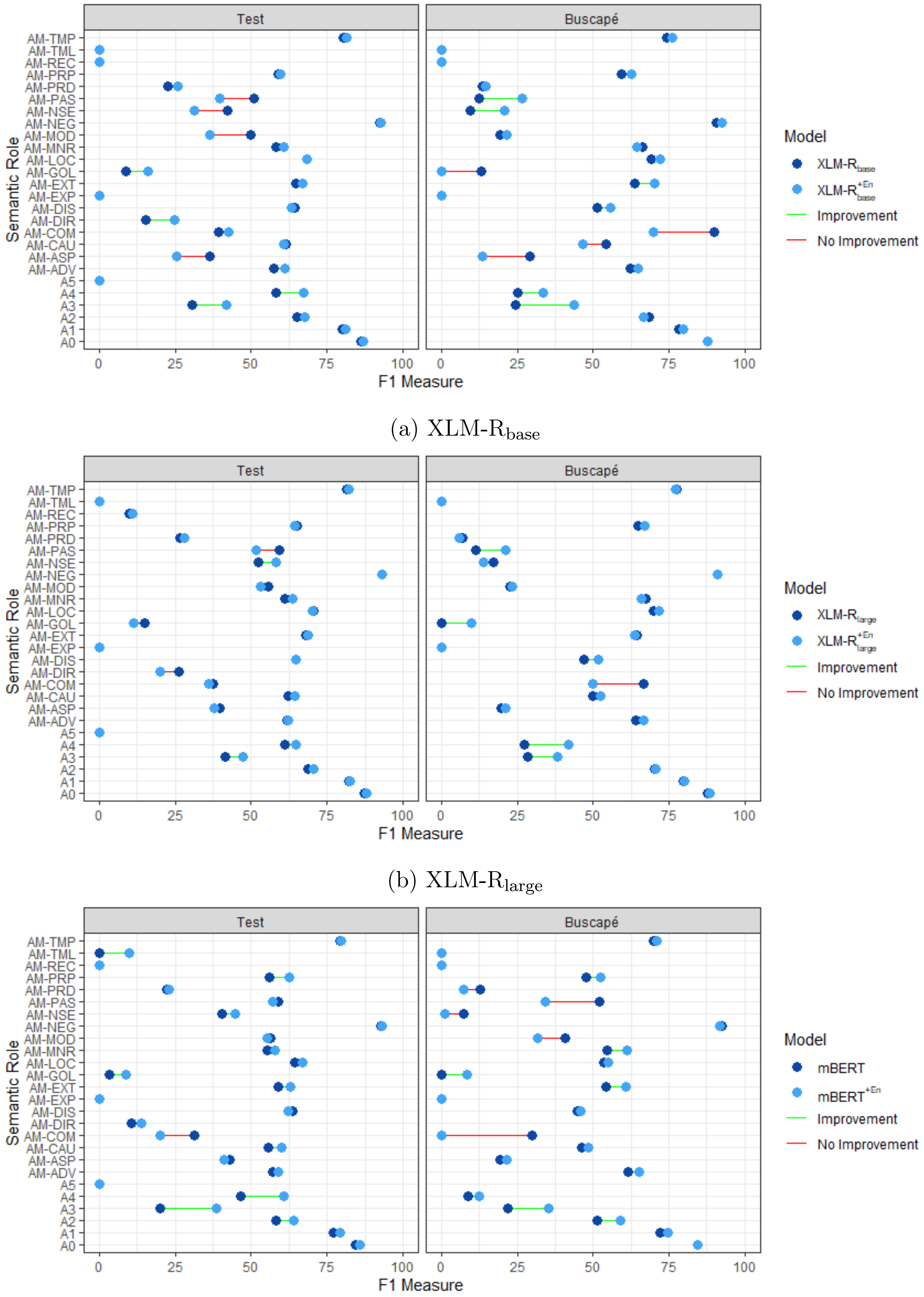
To illustrate this, we include in the figures, as black horizontal lines, the SRL results weighted by the proportion of appearances of each role in each corpus. This represents an approximate contribution of each role to the overall score. Evidently, the model's scores are mostly determined by the numbered roles A0, A1 and A2, which explains why they have such high scores – they are seen the most and they count the most, so the model learns them best. This is an important information. If we are interested in roles other than these three, the overall metrics may not be the best to distinguish between these models.

5.3.5 Cross-Lingual Transfer Learning

Does cross-lingual transfer learning help the multilingual models' performance in semantic role labeling in Portuguese?

Let us now examine the effect that using the additional English data has on the results per role. In Figure 5.4 we can see the changes in F_1 measure of the additional training for each model.

Evidently, the extra data improves the models' performance in some semantic roles and hurts it in others. Nonetheless, the overall score increases because it performs



(a) XLM-R$_{\text{base}}$

(b) XLM-R$_{\text{large}}$

(c) mBERT

Figure 5.4: Comparison between average F_1 measures of the SRL task for the multilingual models trained only in Portuguese and trained both in English and Portuguese. The line between the models is green if the training with English brought an improvement and red otherwise.

better in the roles with more appearances. Interestingly, all models are better in the mentioned difficult numbered roles, A3 and A4. This is probably due to the models seeing more data and, therefore, seeing more verbs and learning which numbered roles they take (since the Portuguese frame files are based on those of PropBank, verbs with the same meaning have the same set of numbered roles).

Comparing the "augmented" multilingual models to the monolingual models (Figures 5.5 and 5.6), we see that the improvement in the difficult numbered roles makes the multilingual XLM-R models better in these roles than their monolingual counterparts. As noted when referring to the overall results in Table 5.2, the model pairs brBERT$_{\text{base}}$/XLM-R$_{\text{base}}$ and brBERT$_{\text{large}}$/XLM-R$_{\text{large}}$ have very similar performance.

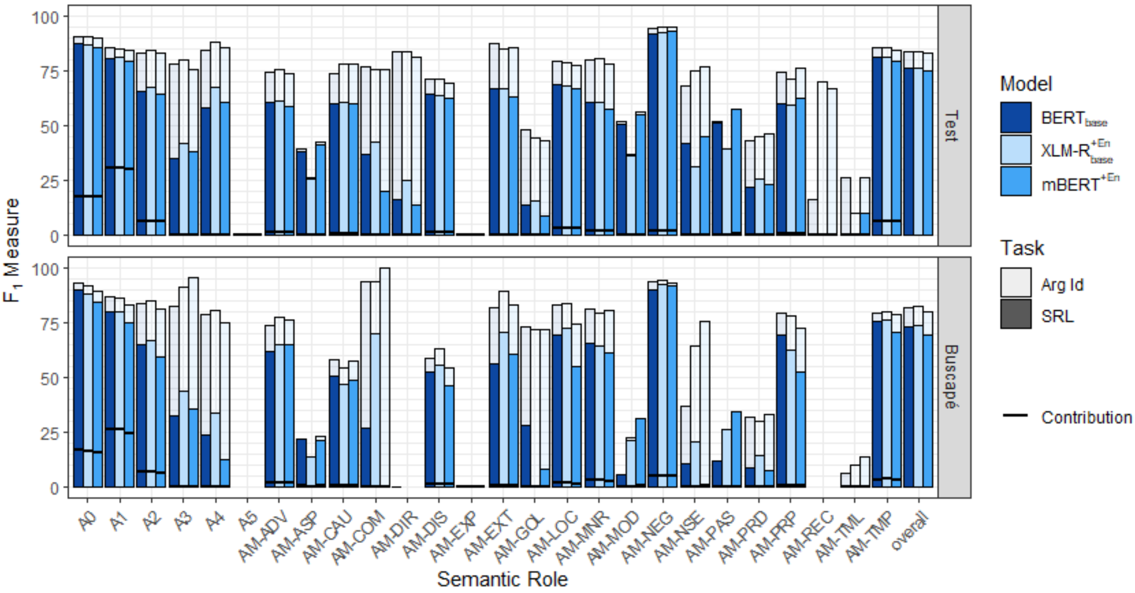


Figure 5.5: Bars – Average SRL (opaque) and argument identification (more transparent) results for brBERT$_{\text{base}}$, XLM-R$^{\text{+En}}_{\text{base}}$ and mBERT$^{\text{+En}}$ in both the test sets and the Buscapé set. Horizontal black lines – Contribution of each role to the overall score.

5.3.6 Zero-shot Cross-lingual Transfer Learning

Is it useful at all to use the Portuguese data or can we rely on models trained with English data only?

We provide in Table 5.4 the zero-shot transfer learning results from the three multilingual models trained in the CoNLL-2012 data, i.e., the models were fine-tuned in

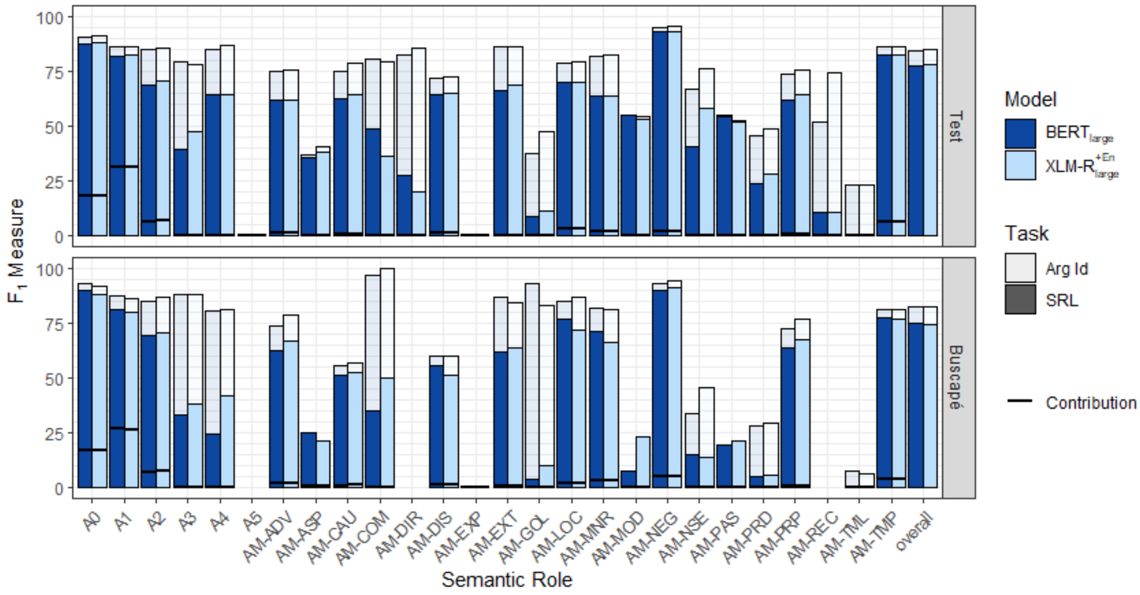


Figure 5.6: Bars – Average SRL (opaque) and argument identification (more transparent) results for brBERT$_{\text{large}}$ and XLM-R$^{\text{+En}}_{\text{large}}$ in both the test sets and the Buscapé set. Horizontal black lines – Contribution of each role to the overall score.

English and tested in Portuguese.

Model	**Average of Test Folds**			**Buscapé**		
	p (%)	r (%)	F_1	p (%)	r (%)	F_1
mBERT	60.73	65.59	63.07	57.83	59.31	58.56
XLM-R$_{\text{base}}$	63.58	69.90	66.59	63.56	67.01	65.24
XLM-R$_{\text{large}}$	64.64	70.85	67.60	63.05	66.94	64.94

Table 5.4: Results for zero-shot cross-lingual transfer learning. The three models were trained on the pre-processed CoNLL-2012. The results are the average of the 10 folds and the result in the Buscapé corpus.

Despite the large drop in F_1 measure between zero-shot cross-lingual transfer learning (Table 5.4) and training only with Portuguese data – between 9 and 10 F_1 points –, it is encouraging to see that the former still performs reasonably well in the Portuguese data. It means languages with no annotated SRL data can use multilingual models trained for this task in English and obtain acceptable results. However, the gains of having own language resources are evident.

5.3.7 Transfer Learning from Syntax

Can we improve the results of the SRL task by training the language model on another task first?

In this section, we study the impact of pre-training with syntax on the best models from Table 5.2.

In Chapter 3, we saw that syntax has often been used to boost SRL performance in proposed architectures. It would therefore be interesting to see if it can also help BERT-based architectures, particularly in low-resource languages where performance is worse.

To that end, we perform intermediate task fine-tuning in dependency parsing before fine-tuning in SRL. In other words, we will fine-tune a pre-trained BERT-based model in a dependency parsing task, and then further fine-tune it in SRL (note that only "pre-trained BERT-based model"'s weights are fine-tuned in the intermediate task; the model's linear layer is still randomly initialized). This type of transfer learning from different tasks has at times led to improvements in models' scores for other tasks [58].

We use the Universal Dependencies[3] (UD) Portuguese data set, UD-Portuguese_Bosque [61], based on the Bosque data set, from Floresta Sintá(c)tica. The UD project provides a consistent treebank annotation across languages. The data set includes the annotation of syntactic dependencies, POS tags and morphological features. We train a model that receives as input the tokenized sentence and predicts the syntactic dependencies between the words. This data set was chosen because there was already a data set reader and model architecture implemented for UD in the *AllenNLP* package, facilitating our work.

Table 5.5 reports the results obtained with this double fine-tuning. The fine-tuning in UD was run for 10 epochs with a learning rate of 1×10^{-5} and a batch size of 4. The SRL models were run in the same conditions and in the same data as the models from the previous section. We will refer to these models using a superscript "+UD".

From the table, we can infer that double fine-tuning using the UD data set did not have a positive impact in the performance of the monolingual $\text{BERT}_{\text{large}}$ model. For the multilingual $\text{XLM-R}_{\text{large}}$ and $\text{XLM-R}^{\text{+En}}_{\text{large}}$ models, fine-tuning first with UD boosts the performance in the Buscapé data set. In fact, with the UD data, we improve the

[3] https://universaldependencies.org

Model	Average of Test Folds				Average of Buscapé			
	p (%)	r (%)	F_1	δF_1	p (%)	r (%)	F_1	δF_1
brBERT$_{\text{large}}$	76.65	78.20	77.42		75.58	74.14	74.85	
brBERT$^{\text{+UD}}_{\text{large}}$	76.90	78.19	77.53	0.11	75.25	73.75	74.49	-0.36
XLM-R$_{\text{large}}$	76.74	78.47	77.59		74.36	73.34	73.84	
XLM-R$^{\text{+UD}}_{\text{large}}$	77.00	78.40	77.69	0.10	75.77	74.08	74.91	1.07
XLM-R$^{\text{+En}}_{\text{large}}$	**77.71**	**78.75**	**78.22**		75.36	73.77	74.55	
XLM-R$^{\text{+En+UD}}_{\text{large}}$	77.38	78.57	77.97	-0.25	**75.69**	**74.44**	**75.05**	0.50

Table 5.5: Average of results of the best models from the previous section in the test set and Buscapé set when using double fine-tuning.

best results obtained before on Buscapé.

In Figure 5.7, we show the F_1 measure of XLM-R$^{\text{+UD}}_{\text{large}}$ and XLM-R$^{\text{+En+UD}}_{\text{large}}$ in the Buscapé set and the results obtained by the best model in this data set in the previous section, BERT$_{\text{large}}$.

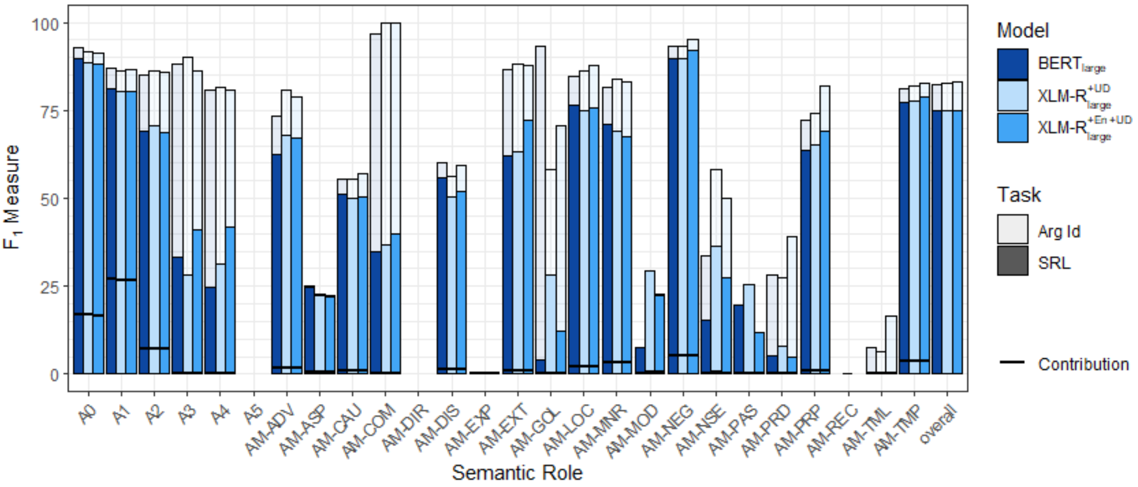


Figure 5.7: Bars – Results of BERT$_{\text{large}}$, XLM-R$^{\text{+UD}}_{\text{large}}$ and XLM-R$^{\text{+En+UD}}_{\text{large}}$ in the Buscapé data set. Horizontal black lines – Contribution of each role to the overall score.

Again, using English data for SRL training increases the scores of roles A3 and A4. The overall performance of the three models ends up being quite similar, but we can see that BERT$_{\text{large}}$ is better in A0 and A1, the most common roles, while XLM-R$^{\text{+En+UD}}_{\text{large}}$ is better in the more uncommon numbered roles and in some important modifiers, such as temporal and negation.

Additionally, both represented models that were fine-tuned in UD are in general better at argument identification than BERT$_{\text{large}}$, leading us to believe that the use of syntax may help identify spans in more difficult data. This agrees with previous work, such as Strubell et al. [70], which found that syntax helped with span boundary identification. However, the differences are small, so we cannot be certain such differences are not just accidental and based on the little test data available for this task.

5.3.8 Statistical Significance

To study the statistical significance of these results, we have conducted a Friedman test [24] with all models (except the zero-shot cross-lingual models) for

- the total 20 F_1 scores (10 test folds, 10 Buscapé results);
- the scores of the test folds;
- the scores in the Buscapé set.

We reject the null hypothesis of the results being from the same distribution for all tests run, so there are differences in the obtained models.

We show in Figures 5.8 and 5.9 the results of the post-hoc Nemenyi test [50]. We can see in Figure 5.9 that XLM-R$^{\text{+En+UD}}_{\text{large}}$ achieves the best scores on average, but it is not statistically different from all other large models. Looking separately at the results for each data set (Figure 5.9), we see that the best model is not statistically different from even the best base models.

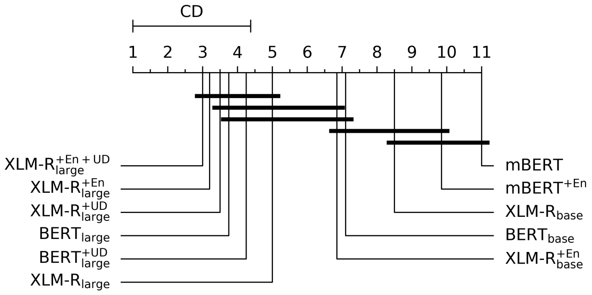
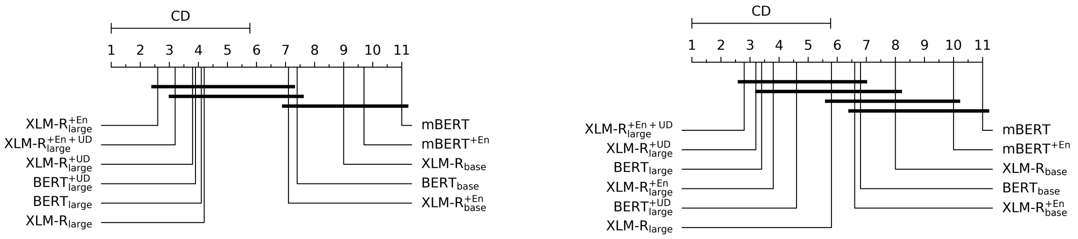


Figure 5.8: Comparison of the models using the Nemenyi post-hoc test with all obtained results. Models that are not statistically different (with $\alpha = 0.05$) are connected.



(a) Test Folds

(b) Buscapé

Figure 5.9: Comparison of the models using the Nemenyi post-hoc test in each set of results (PropBank.Br CV test folds – left; Buscapé – right). Models that are not statistically different (with $\alpha = 0.05$) are connected.

Note that since we are using 10-fold cross-validation, the assumption of independence of the samples for the Friedman test is not guaranteed, so we must take this analysis of statistical significance with caution.

5.3.9 Choosing the Best Model

All things considered, choosing the best model is not an easy task. It depends on many factors regarding the intended application. In this section, we provide an heuristic for choosing a model for an application, based on the obtained results. The decision diagram in Figure 5.10 summarises our heuristic.

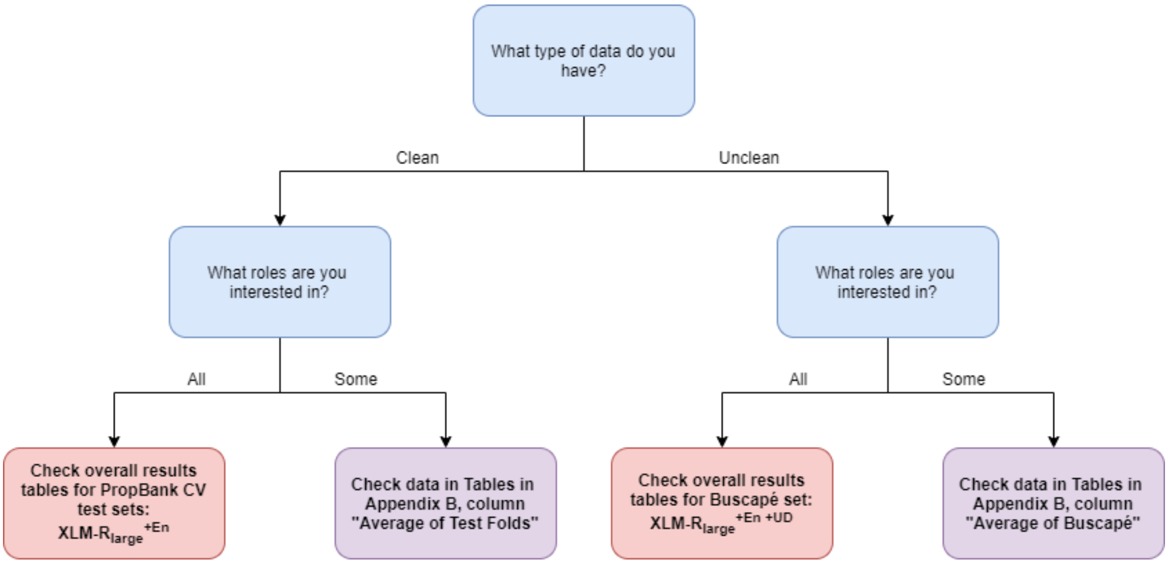

What type of data do you have?
Clean
Unclean
What roles are you interested in?
What roles are you interested in?
All
Some
All
Some
Check overall results tables for PropBank CV test sets: XLM-R$_{large}$$^{+En}$
Check data in Tables in Appendix B, column "Average of Test Folds"
Check overall results tables for Buscapé set: XLM-R$_{large}$$^{+En +UD}$
Check data in Tables in Appendix B, column "Average of Buscapé"

Figure 5.10: Heuristic for choosing the most appropriate model in different situations.

First of all, it is important to determine the type of data involved. If the model is to be applied to text of a more formal variety, where there is some certainty that sentences will be properly structured and words properly spelled ("clean" data, e.g. journalistic text), it is better to look at the scores achieved in the PropBank CV test sets. On the other hand, if one is dealing with text from online sources where there is no guarantee of the mentioned constraints, or it is certain they are not met ("unclean" data), it is better to look at the scores from the Buscapé set.

As mentioned, the overall scores of the models may not be the most appropriate to distinguish between them in all situations. However, the distribution of roles in these data sets is likely to be representative of that found in the Portuguese language. Therefore, if one is interested in the best model for the language in general, they can choose the one with the highest F_1 in the relevant data set. Recall that in the previous section we showed that there is no statistical difference between many models, so one may choose any of the similar models for an application (similar models for each set are reported in Figure 5.9).

If, on the other hand, one is interested instead only in a subset of roles, it is best to choose the best model in those roles by evaluating the presented figures (or the results tables in Appendix B). For example, if we are interested in the best model for A0, A1, AM-LOC, AM-TMP and AM-NEG, the best models will be XLM-$\text{R}^{\text{+En}}_{\text{large}}$ and $\text{BERT}_{\text{large}}$ for properly written data and not, respectively. We will make available a tool that automatically determines the best model for a scenario (an implementation of the heuristic) and computes the F_1 measure for a subset of roles, when necessary.

The results obtained allow us to compare the models, as intended. However, these scores may not correspond to the model's actual performance for two reasons. Firstly, there are semantic roles that were only annotated in the second version of PropBank, therefore the results for these are not reliable, as the data is inconsistent.

On the other hand, upon inspecting the resulting predictions and gold labels, there seem to be some poorly annotated sentences. For example, when predicting the arguments for the verb "retirar" (withdraw) in the sentence "Os EUA devem retirar suas tropas da Somália até março." (The USA should withdraw their troops from Somalia until March.), the gold labels say that "suas tropas da Somália" (their troops from Somalia) corresponds to A1 – "entidade ou coisa sendo retirada" (entity or thing being withdrawn). However, we believe that A1 should be "suas tropas" (their troops) and A2 – "local de onde foram retiradas" (place from where they were withdrawn) –

should be "da Somália" (from Somalia)[4].

This is just one example, and without a complete revision of the data, it is hard to determine many such mistakes. This is merely a caveat to warn possible users not to take the obtained scores as ground truth.

[4] The sentence could also be translated as "The USA should withdraw their Somalian troops until March", and this ambiguity is likely where the confusion comes from, but this sentence makes less sense.

Chapter 6

Conclusion

This thesis aimed to apply state of the art research in English to the task of semantic role labeling in Portuguese. To this end, we used an architecture that has achieved state of the art performance in English, comprised of a pre-trained BERT-based model and a classifier with Viterbi decoding.

We first showed that this architecture achieves better results than the previous state of the art in Portuguese SRL on an existing benchmark, using monolingual pre-trained models in this language. We then systematically compared models based on monolingual and on multilingual pre-trained models, using the data available in Portuguese. We studied the effects of training with English data on multilingual models. Finally, we investigated the use of dependency parsing for language model pre-training.

The used architecture achieves a new state of the art for SRL in Portuguese, improving previous results by over $15 F_1$ points. Regarding the comparison, we found that, with the techniques employed, the multilingual model XLM-R$_{\text{large}}$ could achieve better results than the monolingual BERT$_{\text{large}}$, despite having less attention heads per encoding layer (refer to Table 4.1), and, therefore, less power to learn the language structure. This suggests that monolingual models may become unnecessary when powerful multilingual models are available, at least for low-resource languages, such as Portuguese. Additionally, we presented an heuristic for choosing the most appropriate model for different applications that may be useful for practitioners and researchers.

We found that the largest percentage of errors in all models comes from the non-identification of arguments, instead of their misclassification. Thus, we suggest that future research focus on this sub-task. We showed that dependency parsing helped in span identification. However, it is possible that constituency parsing would yield even better results, since arguments for a predicate are commonly constituents in the constituency parse tree.

Another interesting possibility for future work would be to apply the constraints from Li et al. [38], which were reported to help their RoBERTa model when trained with less data.

We consider the most important line of future work, however, to be the improvement of the Portuguese data sets, by harmonising the role set across versions and by having all of the data sets manually revised once again to eliminate any annotation errors that may exist.

References

[1] Susana Afonso, Eckhard Bick, Renato Haber, and Diana Santos. Floresta sintá(c)tica: A treebank for portuguese. In *Proceedings of the Third International Conference on Language Resources and Evaluation, LREC 2002, May 29-31, 2002, Las Palmas, Canary Islands, Spain*, 2002.

[2] Fernando Emilio Alva-Manchego and João Luís Garcia Rosa. Semantic role labeling for brazilian portuguese: A benchmark. In *Advances in Artificial Intelligence - IBERAMIA 2012 - 13th Ibero-American Conference on AI, Cartagena de Indias, Colombia, November 13-16, 2012. Proceedings*, pages 481–490, 2012.

[3] Fernando Emilio Alva-Manchego and João Luís Garcia Rosa. Towards semi-supervised brazilian portuguese semantic role labeling: Building a benchmark. In *Computational Processing of the Portuguese Language - 10th International Conference, PROPOR 2012, Coimbra, Portugal, April 17-20, 2012. Proceedings*, pages 210–217, 2012.

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[5] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The berkeley framenet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, COLING-ACL '98, August 10-14, 1998, Université de Montréal, Montréal, Quebec, Canada. Proceedings of the Conference.*, pages 86–90, 1998.

[6] Eckhard Bick. *The Parsing System "PALAVRAS": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework.* PhD thesis, University of Aarhus, Denmark, Aarhus University Press, 2000.

[7] Claire Bonial, Julia Bonn, Kathryn Conger, Jena Hwang, Martha Palmer, and Nicholas Reese. English propbank annotation guidelines. Center for Computational Language and Education Research, Institute of Cognitive Science, University of Colorado at Boulder, 2015.

[8] Claire Bonial, Julia Bonn, Kathryn Conger, Jena D. Hwang, and Martha Palmer. Propbank: Semantics of new predicate types. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014*, pages 3013–3019. European Language Resources Association (ELRA), 2014.

[9] Mírian Bruckschen, Fernando Muniz, José Guilherme C. de Souza, Juliana Thiesen Fuchs, Kleber Infante, Marcelo Muniz, Patrícia Nunes Gonçalves, Renata Vieira, and Sandra Aluísio. Anotação linguística em xml do corpus PLN-BR. Technical report, University of São Paulo, 2008.

[10] Murillo G. Carneiro, Thiago Henrique Cupertino, Liang Zhao, and João Luís Garcia Rosa. Semi-supervised semantic role labeling for brazilian portuguese. *JIDM*, 8(2):117–130, 2017.

[11] Xavier Carreras and Lluís Màrquez. Introduction to the conll-2004 shared task: Semantic role labeling. In *Proceedings of the Eighth Conference on Computational Natural Language Learning, CoNLL 2004, Held in cooperation with HLT-NAACL 2004, Boston, Massachusetts, USA, May 6-7, 2004*, pages 89–97, 2004.

[12] Xavier Carreras and Lluís Màrquez. Introduction to the conll-2005 shared task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning, CoNLL 2005, Ann Arbor, Michigan, USA, June 29-30, 2005*, pages 152–164, 2005.

[13] Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. An analysis of open information extraction based on semantic role labeling. In Mark A. Musen and Óscar Corcho, editors, *Proceedings of the 6th International Conference on Knowledge Capture (K-CAP 2011), June 26-29, 2011, Banff, Alberta, Canada*, pages 113–120. ACM, 2011.

[14] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel P. Kuksa. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537, 2011.

[15] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116, 2019.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[17] Magali Sanches Duran and Sandra M. Aluísio. Automatic generation of a lexical resource to support semantic role labeling in portuguese. In Martha Palmer, Gemma Boleda, and Paolo Rosso, editors, *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics, *SEM 2015, June 4-5, 2015, Denver, Colorado, USA*, pages 216–221. The *SEM 2015 Organizing Committee, 2015.

[18] Magali Sanches Duran and Sandra Maria Aluísio. Propbank-br: a brazilian treebank annotated with semantic role labels. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 1862–1867, 2012.

[19] Magali Sanches Duran, Jhonata Pereira Martins, and Sandra Maria Aluísio. Um repositório de verbos para a anotação de papéis semânticos disponível na web (a verb repository for semantic role labeling available in the web) [in Portuguese]. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, 2013.

[20] Daniel Henrique Mourão Falci, Marco Antônio Calijorne Soares, Wladmir Cardoso Brandão, and Fernando Silva Parreiras. Using recurrent neural networks for semantic role labeling in portuguese. In *Progress in Artificial Intelligence, 19th EPIA Conference on Artificial Intelligence, EPIA 2019, Vila Real, Portugal, September 3-6, 2019, Proceedings, Part II*, pages 682–694, 2019.

[21] Charles J. Fillmore, Christopher R. Johnson, and Miriam R.L. Petruck. Background to Framenet. *International Journal of Lexicography*, 16(3):235–250, 09 2003.

[22] Erick Rocha Fonseca and João Luís Garcia Rosa. An architecture for semantic role labeling on portuguese. In *Computational Processing of the Portuguese Language - 10th International Conference, PROPOR 2012, Coimbra, Portugal, April 17-20, 2012. Proceedings*, pages 204–209, 2012.

[23] Erick Rocha Fonseca and João Luís Garcia Rosa. A two-step convolutional neural network approach for semantic role labeling. In *The 2013 International Joint Conference on Neural Networks, IJCNN 2013, Dallas, TX, USA, August 4-9, 2013*, pages 1–7, 2013.

[24] Milton Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701, 1937.

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

[26] Daniel Gildea and Daniel Jurafsky. Automatic labeling of semantic roles. In *38th Annual Meeting of the Association for Computational Linguistics, Hong Kong, China, October 1-8, 2000.*, 2000.

[27] Masato Hagiwara. *Real-World Natural Language Processing*. Manning Publications, 2019. Manning Early Access Program (MEAP).

[28] Jan Hajic, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Stepánek, Pavel Stranák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. The conll-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task, CoNLL 2009, Boulder, Colorado, USA, June 4, 2009*, pages 1–18, 2009.

[29] Nathan Hartmann, Lucas Avanço, Pedro Balage, Magali Duran, Maria das Graças Volpe Nunes, Thiago Pardo, and Sandra Aluísio. A large corpus of product reviews in Portuguese: Tackling out-of-vocabulary words. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3865–3871, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

[30] Nathan Siegle Hartmann, Magali Sanches Duran, and Sandra Maria Aluísio. Automatic semantic role labeling on non-revised syntactic trees of journalistic texts. *CoRR*, abs/1704.03016, 2017.

[31] Luheng He, Kenton Lee, Omer Levy, and Luke Zettlemoyer. Jointly predicting predicates and arguments in neural semantic role labeling. In *Proceedings of*

the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers, pages 364–369, 2018.

[32] Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 473–483, 2017.

[33] Shexia He, Zuchao Li, Hai Zhao, and Hongxiao Bai. Syntax for semantic role labeling, to be, or not to be. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2061–2071, 2018.

[34] Richard Johansson and Pierre Nugues. Dependency-based semantic role labeling of propbank. In *2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu, Hawaii, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 69–78, 2008.

[35] Daniel Jurafsky and James H. Martin. Speech and language processing. 3rd edition draft, 2019.

[36] Atif Khan, Naomie Salim, and Yogan Jaya Kumar. A framework for multi-document abstractive summarization based on semantic role labelling. *Applied Soft Computing*, 30:737 – 747, 2015.

[37] Karin Kipper Schuler. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon.* PhD thesis, University of Pennsylvania, 2005.

[38] Tao Li, Parth Anand Jawale, Martha Palmer, and Vivek Srikumar. Structured tuning for semantic role labeling, 2020.

[39] Zuchao Li, Shexia He, Jiaxun Cai, Zhuosheng Zhang, Hai Zhao, Gongshen Liu, Linlin Li, and Luo Si. A unified syntax-aware framework for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2401–2411, 2018.

[40] Zuchao Li, Shexia He, Hai Zhao, Yiqing Zhang, Zhuosheng Zhang, Xi Zhou, and Xiang Zhou. Dependency or span, end-to-end uniform semantic role labeling.

In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6730–6737. AAAI Press, 2019.

[41] Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. Choosing transfer languages for cross-lingual learning, 2019.

[42] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[43] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1412–1421, 2015.

[44] Diego Marcheggiani, Anton Frolov, and Ivan Titov. A simple and accurate syntax-agnostic neural model for dependency-based semantic role labeling. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), Vancouver, Canada, August 3-4, 2017*, pages 411–420, 2017.

[45] Diego Marcheggiani and Ivan Titov. Encoding sentences with graph convolutional networks for semantic role labeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1506–1515, 2017.

[46] Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of english: The penn treebank. *Comput. Linguist.*, 19(2):313–330, June 1993.

[47] Lluís Màrquez, Xavier Carreras, Kenneth C. Litkowski, and Suzanne Stevenson. Semantic role labeling: An introduction to the special issue. *Computational Linguistics*, 34(2):145–159, 2008.

[48] Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. The NomBank project: An interim

report. In *Proceedings of the Workshop Frontiers in Corpus Annotation at HLT-NAACL 2004*, pages 24–31, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics.

[49] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

[50] Peter Nemenyi. *Distribution-free multiple comparisons.* PhD thesis, Princeton University, 1963.

[51] Hiroki Ouchi, Hiroyuki Shindo, and Yuji Matsumoto. A span selection model for semantic role labeling. *CoRR*, abs/1810.02245, 2018.

[52] Martha Palmer, Daniel Gildea, and Nianwen Xue. *Semantic Role Labeling.* Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2010.

[53] Martha Palmer, Paul R. Kingsbury, and Daniel Gildea. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106, 2005.

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

[55] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[56] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,*

NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers), pages 2227–2237, 2018.

[57] Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. Towards robust linguistic analysis using ontonotes. In Julia Hockenmaier and Sebastian Riedel, editors, *Proceedings of the Seventeenth Conference on Computational Natural Language Learning, CoNLL 2013, Sofia, Bulgaria, August 8-9, 2013*, pages 143–152. ACL, 2013.

[58] Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. Intermediate-task transfer learning with pretrained models for natural language understanding: When and why does it work? *CoRR*, abs/2005.00628, 2020.

[59] Vasin Punyakanok, Dan Roth, and Wen-tau Yih. The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics*, 34(2):257–287, 2008.

[60] Vasin Punyakanok, Dan Roth, Wen-tau Yih, and Dav Zimak. Semantic role labeling via integer linear programming inference. In *COLING 2004, 20th International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2004, Geneva, Switzerland*, 2004.

[61] Alexandre Rademaker, Fabricio Chalub, Livy Real, Cláudia Freitas, Eckhard Bick, and Valeria de Paiva. Universal dependencies for portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling)*, pages 197–206, Pisa, Italy, September 2017.

[62] Lance A. Ramshaw and Mitchell P. Marcus. Text chunking using transformation-based learning. *CoRR*, cmp-lg/9505040, 1995.

[63] Erik F. Tjong Kim Sang and Sabine Buchholz. Introduction to the conll-2000 shared task chunking. In *Fourth Conference on Computational Natural Language Learning, CoNLL 2000, and the Second Learning Language in Logic Workshop, LLL 2000, Held in cooperation with ICGI-2000, Lisbon, Portugal, September 13-14, 2000*, pages 127–132, 2000.

[64] Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis P. Vlahavas. On the stratification of multi-label data. In Dimitrios Gunopulos, Thomas Hofmann, Donato Malerba, and Michalis Vazirgiannis, editors, *Machine Learning and*

Knowledge Discovery in Databases - European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, Proceedings, Part III, volume 6913 of *Lecture Notes in Computer Science*, pages 145–158. Springer, 2011.

[65] João Sequeira, Teresa Gonçalves, and Paulo Quaresma. Semantic role labeling for portuguese - A preliminary approach -. In *Computational Processing of the Portuguese Language - 10th International Conference, PROPOR 2012, Coimbra, Portugal, April 17-20, 2012. Proceedings*, pages 193–203, 2012.

[66] Dan Shen and Mirella Lapata. Using semantic roles to improve question answering. In Jason Eisner, editor, *EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic*, pages 12–21. ACL, 2007.

[67] Peng Shi and Jimmy Lin. Simple BERT models for relation extraction and semantic role labeling. *CoRR*, abs/1904.05255, 2019.

[68] Fabio Souza, Rodrigo Nogueira, and Roberto Lotufo. Portuguese named entity recognition using bert-crf. *arXiv preprint arXiv:1909.10649*, 2019.

[69] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, 2014.

[70] Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. Linguistically-informed self-attention for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 5027–5038, 2018.

[71] Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. The conll 2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning, CoNLL 2008, Manchester, UK, August 16-17, 2008*, pages 159–177, 2008.

[72] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015.

[73] Zhixing Tan, Mingxuan Wang, Jun Xie, Yidong Chen, and Xiaodong Shi. Deep semantic role labeling with self-attention. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4929–4936, 2018.

[74] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.

[75] Jorge A. Wagner Filho, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. The brWaC corpus: A new open resource for Brazilian Portuguese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[76] Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Nianwen Xue, Martha Palmer, Jena D. Hwang, Claire Bonial, Jinho Choi, Aous Mansouri, Maha Foster, Abdel aati Hawwary, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, and Ann Houston. OntoNotes Release 5.0 LDC2013T19. Web Download. Philadelphia: Linguistic Data Consortium, 2013.

[77] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.

[78] Qingrong Xia, Zhenghua Li, Min Zhang, Meishan Zhang, Guohong Fu, Rui Wang, and Luo Si. Syntax-aware neural semantic role labeling. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7305–7313, 2019.

[79] Nianwen Xue and Martha Palmer. Calibrating features for semantic role labeling. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain*, pages 88–94, 2004.

[80] Jie Zhou and Wei Xu. End-to-end learning of semantic role labeling using recurrent neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1127–1137, 2015.

Appendix A

Definitions of PropBank.Br Roles

Role Label	**Definition**
A0	agent or causer
A1	patient, experiencer or theme
A2 / A3 / A4 / A5	defined in the frame files of each verb
AM-ADV	adverbial
AM-ASP[1]	aspect auxiliary verb
AM-CAU	cause
AM-COM[1]	comitative
AM-DIR	direction
AM-DIS	discourse
AM-EXT	extension
AM-EXP[1]	expletive
AM-GOL[1]	goal
AM-LOC	locative
AM-MNR	manner
AM-MOD[1]	modal auxiliary verb
AM-NEG	negation
AM-NSE[1]	non-argumental reflexive pronoun
AM-PAS[1]	passive voice auxiliary verb
AM-PRP[2]	purpose
AM-PRD	secondary predication
AM-REC	reciprocal
AM-TML[1]	temporal auxiliary verb
AM-TMP	time

Table A.1: Role definition for all roles in PropBankBr.

[1] Only exists in PropBank v2 and Buscapé

[2] In PropBank v1.1 this argument has the label "AM-PNC" meaning "purpose not cause". The notation was changed to "AM-PRP" in PropBank v2 and we adopt this label in this dissertation.

Appendix B

Detailed results

This appendix contains the results per role for the average of scores in the PropBank CV test folds and in the Buscapé set for all models mentioned in Section 5.3.

Semantic Role	**Average of Test Folds**			**Average of Buscapé**		
	P (%)	**R** (%)	$\mathbf{F}_1$	**P** (%)	**R** (%)	$\mathbf{F}_1$
A0	86.69	87.77	87.21	86.40	93.26	89.68
A1	79.85	81.90	80.86	77.83	82.57	80.12
A2	65.19	66.22	65.66	65.96	63.80	64.81
A3	38.40	33.16	35.07	26.22	45.00	32.74
A4	58.80	59.15	58.41	27.83	20.91	23.75
A5	0.00	0.00	0.00	-	-	-
AM-ADV	57.19	64.77	60.67	58.78	65.00	61.55
AM-ASP	50.98	31.93	38.47	78.90	13.16	21.78
AM-CAU	58.21	62.61	60.05	45.37	57.50	50.59
AM-COM	39.00	38.33	36.94	25.00	30.00	26.67
AM-DIR	15.83	16.67	16.19	0.00	0.00	0.00
AM-DIS	65.09	64.19	64.43	71.49	41.89	52.39
AM-EXP	0.00	0.00	0.00	0.00	0.00	0.00
AM-EXT	69.37	64.95	66.82	49.73	65.26	56.32
AM-GOL	19.00	13.33	14.00	50.00	20.00	28.00
AM-LOC	68.88	69.07	68.91	69.05	70.26	69.49
AM-MNR	59.72	61.22	60.39	65.39	66.09	65.72
AM-MOD	64.03	43.40	50.53	41.67	3.04	5.61
AM-NEG	91.24	92.86	92.02	91.48	88.83	90.11
AM-NSE	46.07	45.83	42.06	33.81	7.14	10.95
AM-PAS	62.10	45.97	51.47	21.67	8.57	12.08
AM-PRD	22.40	22.36	22.15	7.72	10.00	8.55
AM-PRP	60.28	61.20	60.14	69.51	70.00	69.20
AM-REC	0.00	0.00	0.00	0.00	0.00	0.00
AM-TML	0.00	0.00	0.00	0.00	0.00	0.00
AM-TMP	79.67	83.07	81.31	69.49	82.34	75.36
overall	75.78	76.83	76.30	74.00	72.68	73.33

Table B.1: Averaged results per role for model BERT$_{base}$.

Semantic Role	Average of Test Folds			Average of Buscapé		
	P (%)	**R** (%)	$\mathbf{F}_1$	**P** (%)	**R** (%)	$\mathbf{F}_1$
A0	87.03	88.12	87.55	86.06	93.60	89.66
A1	81.18	82.64	81.91	79.13	83.50	81.25
A2	68.80	68.53	68.64	73.39	65.42	69.13
A3	42.66	38.55	39.73	24.33	55.00	33.27
A4	64.74	65.92	64.30	28.92	21.82	24.72
A5	0.00	0.00	0.00	-	-	-
AM-ADV	57.09	68.07	61.95	60.58	65.26	62.56
AM-ASP	40.04	36.50	35.80	85.05	15.26	24.82
AM-CAU	62.87	62.99	62.56	46.24	57.14	51.00
AM-COM	57.33	48.33	49.10	33.33	40.00	35.00
AM-DIR	30.00	28.33	27.33	0.00	0.00	0.00
AM-DIS	62.17	66.61	64.09	72.43	45.68	55.87
AM-EXP	0.00	0.00	0.00	0.00	0.00	0.00
AM-EXT	66.45	66.98	66.42	57.37	67.89	62.08
AM-GOL	7.83	10.00	8.69	5.00	3.33	4.00
AM-LOC	67.92	72.05	69.84	73.98	79.23	76.49
AM-MNR	64.15	63.47	63.73	68.46	74.06	71.02
AM-MOD	56.88	53.75	54.77	37.50	4.35	7.64
AM-NEG	91.72	94.68	93.15	90.84	89.09	89.93
AM-NSE	46.50	43.33	40.97	41.81	10.00	15.16
AM-PAS	57.21	53.79	54.56	29.88	15.71	19.61
AM-PRD	24.68	23.36	23.78	5.06	5.56	5.28
AM-PRP	61.25	64.09	62.06	64.33	64.21	63.64
AM-REC	11.11	11.11	11.11	-	-	-
AM-TML	0.00	0.00	0.00	0.00	0.00	0.00
AM-TMP	80.73	83.73	82.18	71.59	84.06	77.32
overall	76.65	78.21	77.42	75.58	74.14	74.85

Table B.2: Averaged results per role for model BERT$_{large}$.

Semantic Role	Average of Test Folds			Average of Buscapé		
	P (%)	**R** (%)	$\mathbf{F_1}$	**P** (%)	**R** (%)	$\mathbf{F_1}$
A0	85.74	86.72	86.22	83.53	92.71	87.85
A1	79.26	80.53	79.89	76.29	80.15	78.17
A2	64.87	65.42	65.11	70.05	67.25	68.55
A3	29.52	33.48	30.66	17.53	45.00	24.59
A4	56.15	61.58	58.41	27.61	23.64	25.33
A5	0.00	0.00	0.00	-	-	-
AM-ADV	54.16	61.94	57.56	59.75	65.79	62.31
AM-ASP	48.05	33.17	36.52	89.84	18.16	29.03
AM-CAU	59.96	63.69	61.41	50.22	58.93	54.09
AM-COM	46.83	40.00	39.17	90.00	90.00	90.00
AM-DIR	19.50	15.00	15.36	0.00	0.00	0.00
AM-DIS	62.89	66.33	64.42	63.43	44.05	51.21
AM-EXP	0.00	0.00	0.00	0.00	0.00	0.00
AM-EXT	64.09	66.07	64.79	58.93	70.00	63.73
AM-GOL	6.85	13.33	8.90	20.00	10.00	13.00
AM-LOC	67.13	69.62	68.30	70.29	68.46	69.25
AM-MNR	57.45	60.27	58.38	63.04	70.00	66.19
AM-MOD	56.85	48.94	49.90	66.83	11.74	19.24
AM-NEG	91.34	93.63	92.44	91.62	90.00	90.78
AM-NSE	50.50	40.83	42.25	40.00	5.71	9.61
AM-PAS	61.78	46.23	50.82	16.83	10.00	12.45
AM-PRD	22.50	23.28	22.63	12.29	16.67	13.72
AM-PRP	58.86	60.09	59.13	66.22	54.21	59.27
AM-REC	0.00	0.00	0.00	0.00	0.00	0.00
AM-TML	0.00	0.00	0.00	0.00	0.00	0.00
AM-TMP	78.60	82.42	80.45	69.68	79.84	74.36
overall	74.42	76.04	75.22	73.12	72.54	72.82

Table B.3: Averaged results per role for model XLM-R$_{base}$.

Semantic Role	**Average of Test Folds**			**Average of Buscapé**		
	P (%)	**R** (%)	$\mathbf{F}_1$	**P** (%)	**R** (%)	$\mathbf{F}_1$
A0	86.42	87.79	87.08	82.84	93.41	87.78
A1	80.22	81.71	80.95	77.03	82.77	79.79
A2	68.84	66.86	67.73	73.85	60.99	66.71
A3	43.26	41.30	41.86	34.41	61.67	43.70
A4	66.32	69.56	67.36	34.56	32.73	33.57
A5	0.00	0.00	0.00	-	-	-
AM-ADV	58.12	64.93	61.01	64.24	65.79	64.91
AM-ASP	34.31	24.22	25.48	50.24	8.42	13.58
AM-CAU	60.93	61.29	60.68	42.23	52.14	46.62
AM-COM	48.33	40.00	42.43	61.67	90.00	70.00
AM-DIR	27.83	26.67	24.83	-	-	-
AM-DIS	62.32	65.39	63.48	72.88	45.41	55.54
AM-EXP	0.00	0.00	0.00	0.00	0.00	0.00
AM-EXT	69.08	65.45	66.89	62.37	81.58	70.34
AM-GOL	14.72	20.00	15.90	0.00	0.00	0.00
AM-LOC	65.75	71.56	68.43	68.88	76.41	72.21
AM-MNR	59.80	62.36	60.94	62.19	66.87	64.37
AM-MOD	48.86	33.80	36.60	71.57	13.04	21.44
AM-NEG	91.35	94.03	92.61	94.10	90.78	92.40
AM-NSE	37.50	31.67	31.50	56.50	13.57	20.72
AM-PAS	52.12	37.08	39.62	32.21	22.86	26.47
AM-PRD	27.98	25.17	26.06	16.61	14.44	14.69
AM-PRP	57.85	62.22	59.69	57.82	68.42	62.63
AM-REC	0.00	0.00	0.00	0.00	0.00	0.00
AM-TML	0.00	0.00	0.00	0.00	0.00	0.00
AM-TMP	80.85	82.21	81.51	73.52	78.75	76.01
overall	76.09	76.93	76.50	74.29	73.22	73.74

Table B.4: Averaged results per role for model XLM-$\mathrm{R}_{\mathrm{base}}^{\mathrm{+En}}$.

Semantic Role	**Average of Test Folds**			**Average of Buscapé**		
	P (%)	**R** (%)	$\mathbf{F}_1$	**P** (%)	**R** (%)	$\mathbf{F}_1$
A0	86.98	87.98	87.47	83.08	93.18	87.82
A1	81.35	83.25	82.29	77.00	82.79	79.78
A2	69.24	68.26	68.68	75.74	65.70	70.24
A3	43.63	41.24	41.54	20.65	46.67	28.28
A4	59.67	65.33	61.28	31.02	24.55	27.21
A5	0.00	0.00	0.00	-	-	-
AM-ADV	57.86	66.70	61.71	62.48	65.79	63.99
AM-ASP	52.78	38.37	39.76	70.78	11.84	19.78
AM-CAU	62.15	64.03	62.36	46.39	54.64	49.97
AM-COM	32.39	46.67	37.40	65.00	70.00	66.67
AM-DIR	31.00	30.00	26.07	0.00	0.00	0.00
AM-DIS	65.10	64.81	64.79	67.50	36.49	47.10
AM-EXP	0.00	0.00	0.00	0.00	0.00	0.00
AM-EXT	70.29	66.55	68.14	57.85	72.63	64.25
AM-GOL	14.50	16.67	14.91	0.00	0.00	0.00
AM-LOC	68.37	72.95	70.55	68.87	71.03	69.73
AM-MNR	60.62	61.73	61.04	63.37	71.72	67.18
AM-MOD	58.72	56.20	55.56	58.28	14.78	22.49
AM-NEG	91.41	95.06	93.18	92.08	90.00	91.00
AM-NSE	66.83	54.17	52.55	38.24	11.43	17.06
AM-PAS	63.21	58.57	59.34	16.67	8.57	11.27
AM-PRD	28.35	26.08	26.69	7.54	6.67	6.95
AM-PRP	63.40	68.13	65.29	67.20	63.68	64.68
AM-REC	10.00	10.00	10.00	0.00	0.00	0.00
AM-TML	0.00	0.00	0.00	0.00	0.00	0.00
AM-TMP	80.03	83.15	81.54	72.75	82.81	77.40
overall	76.74	78.47	77.59	74.36	73.34	73.85

Table B.5: Averaged results per role for model XLM-R$_{large}$.

Semantic Role	**Average of Test Folds**			**Average of Buscapé**		
	P (%)	**R** (%)	$\mathbf{F}_1$	**P** (%)	**R** (%)	$\mathbf{F}_1$
A0	86.85	88.94	87.87	83.90	93.18	88.28
A1	81.75	83.19	82.46	76.89	83.76	80.18
A2	71.18	70.10	70.62	76.66	65.28	70.46
A3	48.06	47.70	47.43	29.11	56.67	38.17
A4	64.43	65.44	64.67	45.95	39.09	41.86
A5	0.00	0.00	0.00	-	-	-
AM-ADV	58.23	66.73	62.08	64.10	69.74	66.59
AM-ASP	46.75	33.07	38.03	66.01	13.16	21.18
AM-CAU	63.34	66.43	64.45	48.08	57.50	52.28
AM-COM	40.00	40.00	36.10	50.00	50.00	50.00
AM-DIR	21.19	30.00	20.16	-	-	-
AM-DIS	67.30	63.09	64.88	78.65	38.92	51.53
AM-EXP	0.00	0.00	0.00	0.00	0.00	0.00
AM-EXT	72.73	65.55	68.61	58.33	71.05	63.87
AM-GOL	11.00	13.33	11.50	20.00	6.67	10.00
AM-LOC	67.97	72.72	70.22	71.49	72.31	71.55
AM-MNR	62.93	64.83	63.83	63.60	68.91	66.06
AM-MOD	59.96	49.02	53.09	88.33	13.91	23.39
AM-NEG	91.85	94.55	93.16	92.52	89.74	91.10
AM-NSE	70.83	51.67	58.28	46.33	8.57	13.81
AM-PAS	62.60	46.64	51.65	25.79	18.57	21.33
AM-PRD	30.05	27.17	28.13	5.30	6.67	5.77
AM-PRP	63.81	65.46	64.46	69.92	64.74	67.13
AM-REC	11.11	11.11	11.11	-	-	-
AM-TML	0.00	0.00	0.00	0.00	0.00	0.00
AM-TMP	80.93	83.61	82.22	75.15	79.06	77.03
overall	77.71	78.75	78.22	75.36	73.77	74.55

Table B.6: Averaged results per role for model XLM-$\mathrm{R}^{\text{+En}}_{\text{large}}$.

Semantic Role	Average of Test Folds			Average of Buscapé		
	P (%)	**R** (%)	$\mathbf{F}_1$	**P** (%)	**R** (%)	$\mathbf{F}_1$
A0	83.78	85.00	84.38	78.85	90.93	84.41
A1	76.58	77.97	77.27	68.43	75.84	71.93
A2	59.80	57.06	58.27	53.66	49.86	51.48
A3	24.87	17.98	20.14	21.82	26.67	21.91
A4	52.49	44.04	46.43	15.11	6.36	8.77
A5	0.00	0.00	0.00	-	-	-
AM-ADV	54.13	60.94	57.17	57.29	67.11	61.35
AM-ASP	44.36	43.17	43.05	58.12	11.84	19.32
AM-CAU	57.17	55.83	55.75	44.60	48.57	46.30
AM-COM	35.00	31.67	31.43	30.00	30.00	30.00
AM-DIR	10.00	11.67	10.71	0.00	0.00	0.00
AM-DIS	64.77	62.99	63.71	67.70	33.78	44.79
AM-EXP	0.00	0.00	0.00	0.00	0.00	0.00
AM-EXT	59.17	59.71	59.13	50.13	60.53	54.39
AM-GOL	3.33	3.33	3.33	0.00	0.00	0.00
AM-LOC	62.02	66.84	64.25	54.10	53.85	53.52
AM-MNR	55.33	55.55	55.29	54.80	55.16	54.79
AM-MOD	63.36	54.58	56.54	93.57	26.52	40.93
AM-NEG	91.01	94.94	92.89	93.19	91.69	92.41
AM-NSE	45.83	38.33	40.52	21.67	5.00	7.51
AM-PAS	60.78	59.42	59.13	67.83	44.29	52.23
AM-PRD	24.56	20.84	22.18	11.39	14.44	12.64
AM-PRP	53.71	59.52	55.94	55.00	44.21	47.89
AM-REC	0.00	0.00	0.00	0.00	0.00	0.00
AM-TML	0.00	0.00	0.00	0.00	0.00	0.00
AM-TMP	77.76	80.69	79.15	69.71	70.16	69.82
overall	72.34	73.21	72.76	67.10	66.70	66.89

Table B.7: Averaged results per role for model mBERT.

Semantic Role	Average of Test Folds			Average of Buscapé		
	P (%)	**R** (%)	$\mathbf{F}_1$	**P** (%)	**R** (%)	$\mathbf{F}_1$
A0	84.88	86.62	85.73	78.31	91.47	84.34
A1	78.61	79.80	79.20	71.00	78.63	74.61
A2	65.34	63.11	64.17	62.00	56.62	59.15
A3	39.62	38.23	38.53	32.31	43.33	35.35
A4	63.66	59.30	60.71	17.67	10.00	12.62
A5	0.00	0.00	0.00	-	-	-
AM-ADV	57.03	61.64	59.07	62.75	68.16	65.23
AM-ASP	55.70	37.52	41.11	68.50	13.16	21.46
AM-CAU	62.30	59.28	60.14	45.84	52.14	48.61
AM-COM	19.17	21.67	20.24	0.00	0.00	0.00
AM-DIR	12.83	16.67	14.05	-	-	-
AM-DIS	60.89	64.05	62.35	70.54	34.59	46.03
AM-EXP	0.00	0.00	0.00	0.00	0.00	0.00
AM-EXT	64.75	62.26	62.86	56.85	65.79	60.78
AM-GOL	8.33	10.00	8.89	13.33	6.67	8.33
AM-LOC	64.53	69.53	66.86	55.69	54.36	54.92
AM-MNR	55.73	60.41	57.82	58.18	64.69	61.21
AM-MOD	60.75	52.57	55.30	78.48	20.00	31.59
AM-NEG	91.83	94.16	92.94	93.32	90.26	91.75
AM-NSE	45.95	45.83	44.85	10.00	0.71	1.33
AM-PAS	61.14	54.42	57.30	53.17	27.14	34.25
AM-PRD	22.22	24.72	23.11	7.23	7.78	7.31
AM-PRP	61.94	64.42	62.46	54.31	51.05	52.35
AM-REC	0.00	0.00	0.00	0.00	0.00	0.00
AM-TML	10.00	10.00	10.00	0.00	0.00	0.00
AM-TMP	77.90	81.51	79.61	70.90	70.78	70.80
overall	74.23	75.56	74.88	69.41	68.98	69.19

Table B.8: Averaged results per role for model mBERT$^{+\mathrm{En}}$.

Semantic Role	**Average of Test Folds**			**Buscapé**		
	P (%)	**R** (%)	$\mathbf{F}_1$	**P** (%)	**R** (%)	$\mathbf{F}_1$
A0	64.27	83.26	72.54	70.69	90.70	79.46
A1	72.16	77.33	74.65	66.03	76.99	71.09
A2	62.11	53.31	57.36	64.71	46.48	54.10
A3	33.45	26.44	29.25	50.00	50.00	50.00
A4	59.61	53.12	55.87	50.00	45.45	47.62
A5	0.00	0.00	0.00	-	-	-
AM-ADV	27.62	52.88	36.24	29.63	42.11	34.78
AM-ASP	0.00	0.00	0.00	0.00	0.00	0.00
AM-CAU	63.50	43.47	51.30	46.43	46.43	46.43
AM-COM	45.00	26.67	31.33	100.00	100.00	100.00
AM-DIR	9.62	35.00	14.84	0.00	0.00	0.00
AM-DIS	48.57	38.66	42.93	72.73	21.62	33.33
AM-EXP	0.00	0.00	0.00	0.00	0.00	0.00
AM-EXT	71.14	56.81	61.97	56.00	73.68	63.64
AM-GOL	0.00	0.00	0.00	0.00	0.00	0.00
AM-LOC	58.68	57.08	57.81	63.64	71.79	67.47
AM-MNR	53.99	61.54	57.41	58.33	65.62	61.76
AM-MOD	18.56	78.91	30.00	62.07	78.26	69.23
AM-NEG	88.02	84.80	86.35	90.00	81.82	85.71
AM-NSE	0.00	0.00	0.00	0.00	0.00	0.00
AM-PAS	0.00	0.00	0.00	0.00	0.00	0.00
AM-PRD	15.44	6.71	9.31	11.11	11.11	11.11
AM-PRP	58.46	59.80	59.00	42.11	42.11	42.11
AM-REC	0.00	0.00	0.00	-	-	-
AM-TML	0.00	0.00	0.00	0.00	0.00	0.00
AM-TMP	69.88	77.28	73.38	47.92	71.87	57.50
overall	63.58	69.90	66.59	63.56	67.01	65.24

Table B.9: Results per role for zero-shot cross-lingual transfer learning for model XLM-R$_{base}$.

Semantic Role	**Average of Test Folds**			**Buscapé**		
	P (%)	**R** (%)	$\mathbf{F}_1$	**P** (%)	**R** (%)	$\mathbf{F}_1$
A0	64.00	83.45	72.44	68.75	89.53	77.78
A1	73.58	78.27	75.85	67.16	80.09	73.06
A2	64.37	56.49	60.15	63.06	49.30	55.34
A3	35.58	33.90	34.40	27.27	50.00	35.29
A4	65.81	54.37	59.27	22.22	18.18	20.00
A5	0.00	0.00	0.00	-	-	-
AM-ADV	26.84	51.07	35.15	27.27	39.47	32.26
AM-ASP	0.00	0.00	0.00	0.00	0.00	0.00
AM-CAU	64.28	49.26	55.11	41.67	35.71	38.46
AM-COM	23.33	21.67	22.33	100.00	100.00	100.00
AM-DIR	9.57	30.00	14.00	0.00	0.00	0.00
AM-DIS	50.28	38.39	43.46	50.00	21.62	30.19
AM-EXP	0.00	0.00	0.00	0.00	0.00	0.00
AM-EXT	75.73	54.81	63.27	76.47	68.42	72.22
AM-GOL	0.00	0.00	0.00	0.00	0.00	0.00
AM-LOC	61.41	57.08	59.11	64.10	64.10	64.10
AM-MNR	56.61	63.49	59.80	58.46	59.37	58.91
AM-MOD	19.38	78.91	31.07	61.54	69.57	65.31
AM-NEG	88.94	84.67	86.72	92.54	80.52	86.11
AM-NSE	0.00	0.00	0.00	0.00	0.00	0.00
AM-PAS	0.00	0.00	0.00	0.00	0.00	0.00
AM-PRD	21.00	10.26	13.75	0.00	0.00	0.00
AM-PRP	58.06	54.70	56.12	52.94	47.37	50.00
AM-REC	0.00	0.00	0.00	-	-	-
AM-TML	0.00	0.00	0.00	0.00	0.00	0.00
AM-TMP	70.80	78.89	74.61	49.48	75.00	59.63
overall	64.64	70.85	67.60	63.05	66.94	64.94

Table B.10: Results per role for zero-shot cross-lingual transfer learning for model XLM-R_{large}.

Semantic Role	**Average of Test Folds**			**Buscapé**		
	P (%)	**R** (%)	$\mathbf{F}_1$	**P** (%)	**R** (%)	$\mathbf{F}_1$
A0	61.86	80.96	70.13	63.25	86.05	72.91
A1	70.06	74.20	72.07	59.77	69.03	64.07
A2	56.73	48.13	52.05	47.06	28.17	35.24
A3	34.21	15.24	20.72	0.00	0.00	0.00
A4	56.42	38.64	44.83	0.00	0.00	0.00
A5	0.00	0.00	0.00	-	-	-
AM-ADV	23.79	50.16	32.23	34.09	39.47	36.59
AM-ASP	0.00	0.00	0.00	0.00	0.00	0.00
AM-CAU	66.58	37.00	47.06	46.15	42.86	44.44
AM-COM	40.00	28.33	32.33	0.00	0.00	0.00
AM-DIR	5.41	15.00	7.49	0.00	0.00	0.00
AM-DIS	39.32	40.36	39.78	50.00	32.43	39.34
AM-EXP	0.00	0.00	0.00	0.00	0.00	0.00
AM-EXT	61.14	40.24	47.36	47.62	52.63	50.00
AM-GOL	0.00	0.00	0.00	0.00	0.00	0.00
AM-LOC	53.41	51.40	52.33	63.64	53.85	58.33
AM-MNR	50.11	54.40	52.10	52.05	59.37	55.47
AM-MOD	18.90	70.54	29.76	61.90	56.52	59.09
AM-NEG	88.22	81.29	84.58	88.73	81.82	85.14
AM-NSE	0.00	0.00	0.00	0.00	0.00	0.00
AM-PAS	0.00	0.00	0.00	0.00	0.00	0.00
AM-PRD	16.68	6.11	8.89	11.11	11.11	11.11
AM-PRP	52.05	56.86	54.13	31.03	47.37	37.50
AM-REC	0.00	0.00	0.00	-	-	-
AM-TML	0.00	0.00	0.00	0.00	0.00	0.00
AM-TMP	67.59	64.48	65.95	51.90	64.06	57.34
overall	60.73	65.59	63.07	57.83	59.31	58.56

Table B.11: Results per role for zero-shot cross-lingual transfer learning for model mBERT.

Semantic Role	**Average of Test Folds**			**Average of Buscapé**		
	P (%)	**R** (%)	$\mathbf{F_1}$	**P** (%)	**R** (%)	$\mathbf{F_1}$
A0	87.40	88.30	87.84	85.75	94.11	89.72
A1	81.46	82.80	82.12	77.76	83.14	80.35
A2	67.60	68.73	68.10	72.15	62.04	66.64
A3	37.64	36.90	36.56	19.12	48.33	27.06
A4	62.95	65.26	63.10	32.03	23.64	26.64
A5	0.00	0.00	0.00	-	-	-
AM-ADV	56.53	64.24	60.03	62.11	67.63	64.68
AM-ASP	44.89	42.09	41.98	64.74	16.32	24.22
AM-CAU	62.94	66.77	64.41	51.18	58.21	54.21
AM-COM	48.69	48.33	46.86	70.00	80.00	73.33
AM-DIR	17.33	15.00	14.00	0.00	0.00	0.00
AM-DIS	63.30	67.22	65.08	71.91	40.27	51.36
AM-EXP	0.00	0.00	0.00	0.00	0.00	0.00
AM-EXT	67.42	67.52	67.30	60.16	68.95	64.00
AM-GOL	11.67	10.00	10.67	10.00	3.33	5.00
AM-LOC	69.10	73.09	70.92	71.00	81.79	75.92
AM-MNR	64.29	61.64	62.75	66.95	69.69	68.21
AM-MOD	56.51	47.82	50.25	57.11	13.04	20.65
AM-NEG	91.91	94.81	93.29	93.09	88.96	90.97
AM-NSE	50.36	50.83	46.91	60.00	14.29	22.00
AM-PAS	59.17	52.68	55.02	25.00	11.43	15.49
AM-PRD	26.24	26.03	25.90	4.83	7.78	5.91
AM-PRP	59.52	62.20	60.30	60.97	62.63	61.35
AM-REC	20.00	20.00	20.00	0.00	0.00	0.00
AM-TML	0.00	0.00	0.00	0.00	0.00	0.00
AM-TMP	81.40	82.99	82.14	74.26	83.12	78.42
overall	76.90	78.19	77.53	75.25	73.75	74.49

Table B.12: Averaged results per role for model $\text{BERT}^{\text{+UD}}_{\text{large}}$.

Semantic Role	Average of Test Folds			Average of Buscapé		
	P (%)	**R** (%)	$\mathbf{F}_1$	**P** (%)	**R** (%)	$\mathbf{F}_1$
A0	86.70	87.77	87.22	84.45	92.98	88.50
A1	81.70	82.54	82.12	78.53	82.68	80.54
A2	68.37	68.90	68.59	74.35	67.32	70.60
A3	49.63	37.20	41.87	22.01	41.67	28.07
A4	59.39	65.18	61.20	35.56	28.18	31.24
A5	0.00	0.00	0.00	-	-	-
AM-ADV	57.59	63.02	59.96	67.00	69.74	68.03
AM-ASP	48.43	48.20	46.11	72.53	13.95	22.19
AM-CAU	64.44	64.03	63.84	47.02	53.57	50.01
AM-COM	19.00	25.00	21.43	35.00	40.00	36.67
AM-DIR	30.00	11.67	16.67	-	-	-
AM-DIS	66.05	68.61	67.19	74.24	38.65	50.25
AM-EXP	0.00	0.00	0.00	0.00	0.00	0.00
AM-EXT	68.04	63.57	65.36	56.29	73.16	63.36
AM-GOL	18.17	20.00	18.08	53.33	20.00	28.33
AM-LOC	68.76	73.02	70.70	73.62	77.18	74.93
AM-MNR	62.14	63.41	62.59	67.81	70.94	69.16
AM-MOD	64.23	62.15	60.90	72.82	19.13	29.33
AM-NEG	91.60	94.03	92.76	90.27	88.96	89.60
AM-NSE	51.19	51.67	47.47	69.53	26.43	36.26
AM-PAS	59.16	69.14	62.92	38.50	20.00	25.42
AM-PRD	29.42	25.62	26.82	8.71	7.78	7.88
AM-PRP	59.73	66.00	62.51	67.15	64.21	65.18
AM-REC	10.00	10.00	10.00	0.00	0.00	0.00
AM-TML	0.00	0.00	0.00	0.00	0.00	0.00
AM-TMP	81.66	83.98	82.79	73.10	83.12	77.73
overall	77.00	78.40	77.69	75.77	74.08	74.91

Table B.13: Averaged results per role for model XLM-$\mathrm{R}_{\mathrm{large}}^{\mathrm{+UD}}$.

Semantic Role	**Average of Test Folds**			**Average of Buscapé**		
	P (%)	**R** (%)	**F$_1$**	**P** (%)	**R** (%)	**F$_1$**
A0	87.02	88.52	87.75	83.67	93.18	88.14
A1	81.38	83.03	82.19	77.05	84.38	80.54
A2	70.22	70.97	70.51	72.96	65.00	68.57
A3	47.98	43.99	45.57	34.59	51.67	41.06
A4	62.42	63.53	62.50	44.33	40.00	42.00
A5	0.00	0.00	0.00	-	-	-
AM-ADV	57.98	64.90	61.02	64.36	70.53	67.11
AM-ASP	49.05	34.22	39.39	75.65	13.42	21.78
AM-CAU	60.45	60.95	60.32	47.46	53.93	50.43
AM-COM	30.17	33.33	29.86	40.00	40.00	40.00
AM-DIR	25.83	26.67	24.52	-	-	-
AM-DIS	63.88	64.17	63.76	81.84	38.65	51.90
AM-EXP	0.00	0.00	0.00	0.00	0.00	0.00
AM-EXT	69.43	65.05	66.95	65.66	80.53	72.19
AM-GOL	15.25	16.67	14.82	18.33	10.00	12.33
AM-LOC	67.99	72.09	69.87	77.00	74.87	75.54
AM-MNR	64.38	65.43	64.76	65.46	69.84	67.51
AM-MOD	57.86	47.75	50.50	61.53	14.35	22.27
AM-NEG	92.49	93.38	92.89	93.82	90.26	91.99
AM-NSE	66.42	43.33	47.05	72.67	18.57	27.33
AM-PAS	61.88	55.49	57.44	19.17	8.57	11.68
AM-PRD	31.32	25.90	27.73	5.95	5.56	4.85
AM-PRP	62.31	64.39	63.04	67.38	71.05	68.90
AM-REC	0.00	0.00	0.00	0.00	0.00	0.00
AM-TML	0.00	0.00	0.00	0.00	0.00	0.00
AM-TMP	82.04	83.81	82.89	75.85	82.03	78.77
overall	77.38	78.57	77.97	75.69	74.44	75.05

Table B.14: Averaged results per role for model XLM-R$_{\text{large}}^{\text{+En+UD}}$.