Improving Portuguese Semantic Role Labeling with Transformers and Transfer Learning

Sofia Oliveira  
Dept. of Computer Science - FCUP  
University of Porto  
Porto, Portugal  
anasofia.oliveira@protonmail.com

Daniel Loureiro  
Dept. of Computer Science - FCUP  
University of Porto  
Porto, Portugal  
daniel.b.loureiro@inesctec.pt

Alípio Jorge  
Dept. of Computer Science - FCUP  
University of Porto  
Porto, Portugal  
amjorge@fc.up.pt

Abstract—The Natural Language Processing task of determining “Who did what to whom” is called Semantic Role Labeling. For English, recent methods based on Transformer models have allowed for major improvements in this task over the previous state of the art. However, for low resource languages, like Portuguese, currently available semantic role labeling models are hindered by scarce training data. In this paper, we explore a model architecture with only a pre-trained Transformer-based model, a linear layer, softmax and Viterbi decoding. We substantially improve the state-of-the-art performance in Portuguese by over 15 F1. Additionally, we improve semantic role labeling results in Portuguese corpora by exploiting cross-lingual transfer learning using multilingual pre-trained models, and transfer learning from dependency parsing in Portuguese, evaluating the various proposed approaches empirically.

I. INTRODUCTION

Semantic Role Labeling (SRL) is the task of determining “Who did What to Whom, How, When and Where”, i.e., identifying events in sentences, their participants and properties [1], [2]. The predicate (usually a verb) corresponds to the event and the participants are expressed by arguments, which take on a semantic role in relation to a predicate. SRL is commonly viewed as an intermediate task that can be useful for applications such as: question answering [3], [4]; information extraction [5], [6] and summarization [7], [8].

The most prominent Transformer-based Language Models (e.g. BERT [9]) are trained on English text, and have been a key component of state-of-the-art solutions for English SRL [10], [11]. For Portuguese, however, there has not been much work done, and the current state of the art, Falci et al. [12] is based on a model by He et al. [13] that has since been surpassed in English. Our work aims at transferring state-of-the-art techniques proposed for the English language to a lower resource language; Portuguese. We apply a monolingual model, BERTimbau [14], to improve the current state of the art in Portuguese SRL.

Additionally, we investigate further approaches for reaching better performance in this task for 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 [15]; in fact, SRL models have traditionally relied on syntax.

Our main contributions for the state of the art on the SRL task are the following.

- A new model for Portuguese SRL.
- A comparison of the performance of multilingual and monolingual models.
- We show that using data from a high-resource language improves the scores of multilingual models in a low-resource language.
- We show that pre-training with dependency parsing can help models identify argument span boundaries when we have unclean text.
- We share our code and models trained with the complete set of available data (PropBank.Br + Buscapé)[1]

II. RELATED WORK

One of the most prominent early works in automatic Semantic Role Labeling was by Gildea and Jurafsky [15]. They trained a system based on statistical classifiers on a part of the FrameNet dataset [16]. Since then, several models have been proposed, based first on machine learning algorithms and later on different neural network architectures. Currently, the models with the best performance in English SRL include in their architecture a pre-trained Transformer-based model. Examples of this are the works of Shi and Lin [10] and Li et al. [11].

In Portuguese, there have been fewer models proposed, most of them based on techniques proven efficient in English. Alva-Manchego and Rosa [17] used machine learning algorithms with features extracted from the gold parse trees of PropBank.Br v1 to predict arguments. Fonseca and Rosa [18] built a system similar to that of Collobert et al. [19], with no parsing information, but divided the task into two steps – argument identification and classification. Hartmann et al. [20] compared the two previous systems using PropBank.Br versions 1.1 and 2 and automatically parsed trees. They concluded that the model of Alva-Manchego and Rosa performed better.

https://github.com/asofiaoliveira/srl_bert_pt

©2021 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. DOI: 10.1109/DSAA53316.2021.9564238
The most recent model we know for supervised Portuguese SRL is by Falci et al. [12]. It follows the architecture of He et al. [13] and is a 2-layer bi-LSTM using word embeddings and a global inference step with IOB and PropBank constraints.

III. DATA

PropBank is a project to annotate text with semantic roles [21]. It divides roles into two types: 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, whereas the numbered roles’ meanings are specific to each predicate sense. They are defined in the predicate’s frame file.

Besides the two mentioned types, there are also continuation roles and reference roles. Continuation roles (C-x, where x is any other role) are used when the argument is non-contiguous. Reference roles (R-x, where x is a non-continuation role) are used on pronouns that refer to other roles present in the sentence.

PropBank.Br [22] is the (Brazilian) Portuguese version of PropBank and it follows the same annotation style as the original, having a similar role set. Unlike the recent versions of PropBank, PropBank.Br only annotates arguments related to verbal predicates and is much smaller in size. There are two versions of this dataset from two different data sources. The first version is based on the Brazilian Portuguese portion of the Bosque corpus, which is a part of the treebank “Floresta Sintáctica” [23]. It contains sentences extracted from the newspaper “Folha de São Paulo” of 1994 which have been manually annotated.

The second version of the dataset contains sentences extracted from the PLN-Br corpus [24], also a journalistic corpus based on “Folha de São Paulo”. The sentences in this version were automatically parsed by PALAVRAS [25]. The annotation project of this second version produced also a smaller annotated corpus, named Buscápé, which is based on a corpus of product reviews [26]. Buscápé contains many misspelled words, poorly constructed sentences, verbs conjugated in the wrong tense and accents missing or in excess. This is very common in texts written on the internet in more casual settings. Thus, it represents an important and realistic benchmark for model testing. The frame files for PropBank.Br are called Verbo-Brasil [27], [28] and were based on PropBank’s frame files.

In some experiments, we also use the English CoNLL-2012 SRL data [29] and the Universal Dependencies (UD) Portuguese data set, UD-Portuguese_Bosque [30]. The former is a data set in English for the SRL task which had to be pre-processed to match the Portuguese data we use in our work. The latter is a dependency parsing data set.

http://www.nilc.icmc.usp.br/portlex/index.php/en/downloads
Available at http://143.107.183.175:12680/verbobrasil/sobre.php?lang=eng
https://universaldependencies.org

IV. MODEL AND EVALUATION METHODOLOGY

The model architecture used in our experiments is presented in Figure 1. It includes a pre-trained transformer-based model on top of which a linear layer, a softmax function and Viterbi decoding are applied. The linear layer’s parameters are randomly initialized and both these and the pre-trained model’s parameters are tuned during training. The Viterbi 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.

Our architecture performs neither predicate identification nor predicate disambiguation. Assuming we know the verbs’ positions in a sentence (included in the datasets for training and testing), the sentence is passed to the SRL architecture one instance at a time. We call instance to the set of the sentence’s tokens and a specific verb. The verb’s position is indicated through the segment embeddings input (which are added to the token embeddings and the position embeddings [9]). This architecture does not use predicate sense disambiguation to make predictions. The identification of arguments and their classification with the appropriate semantic role are done at the same time.

In the rest of this section, we first describe the methodology for comparing our approach with the state of the art for Portuguese (Section IV-A). Then we propose a methodology of our own to compare our architecture in different training conditions (Section IV-B).

A. Methodology for comparison with Portuguese state of the art

We compare our proposed approach using the Portuguese pre-trained transformer-based models with the model of Falci et al. [12] (our baseline). We first pre-process the PropBank.Br v1.1 CoNLL-conll05 dataset using the same steps as the baseline, and then perform 20-fold cross-validation with the same folds. A brief description of the dataset for this comparison is presented in Table 1 (column “Met.1”). After training, the models were evaluated with the script provided in the CoNLL-2005 Shared Task srl-eval.pl.

B. Methodology for comparing our models

We will compare models trained in the Portuguese data (but based on different pre-trained transformer-based models) with models trained in both English and Portuguese and with models pre-trained first on dependency parsing.

To this end, we perform 10-fold cross-validation. We use stratified sampling for multi-label data [31] to create the folds from the complete PropBank.Br dataset (versions 1.1 and 2), using the iterative-stratification package. Each fold is used as a test set once. The remaining folds are again sampled to obtain a validation set of approximately the same size of the test set. In addition to these, the out-of-domain Buscápé corpus was used to test all model runs.

Available at https://www.cs.upc.edu/~srcnll/soft.html
Available at https://github.com/trent-b/iterative-stratification
The XML files of the three datasets (PropBank.Br v1.1, PropBank.Br v2 and Buscapé) were all pre-processed in the same fashion. We eliminated instances 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 and for the corpora PropBank.Br v2 and Buscapé, we removed any instances with flags "WRONGSUBCORPUS", "LATER" or "REEXAMINE", since, according to the guide, these indicate something wrong with the sentence that prevents its annotation.

In contrast to previous works, we only annotated an argument with a "C-" role if the respective words were non-contiguous to the original argument. 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.

As in the previous methodology, the models were evaluated with the official script from the CoNLL 2005 Shared Task. Table I presents a brief description of both datasets in this methodology (columns “Met.2 CV” and “Met.2 Buscapé”), as well as of the CoNLL-2012 dataset.

V. Experiments

We will now run experiments that try to answer the following research questions:

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 Portuguese SRL?

3) Does cross-lingual transfer learning from English help the multilingual models’ performance in Portuguese SRL?

4) Can we improve the results of the SRL task by training the language model on dependency parsing previously?

5) Is the Portuguese data useful or can we rely on models trained with English only?

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. Li et al. [11] 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 [32].

On the other hand, syntax has often been used to boost SRL performance in proposed architectures. It is therefore relevant to see if it can also help Transformer-based architectures, particularly in low-resource languages. We integrate syntax using intermediate-task fine-tuning, which has at times led to improvements in models’ scores for other tasks [33].
VI. RESULTS AND ANALYSIS

The models were implemented in Python using the package AllenNLP v1.0.0rc3 \cite{allenai-allennlp-models}, the (cased) mBERT, XLM-R\textsubscript{base} and XLM-R\textsubscript{large} models trained by Huggingface Transformers v2.9.0 \cite{xlm-r-large} and the Portuguese BERT models (base and large), which will be referred to as brBERT, trained by \textit{neuralmind-ai} \cite{brbert} and built on PyTorch v1.5.0 \cite{pytorch}. In each experiment, all models are trained with the same hyperparameters. With the exception of number of epochs, batch size and learning rate, all hyperparameters used followed the defaults in AllenNLP’s English BERT SRL training configuration.\footnote{https://github.com/allenai/allennlp-models/blob/v1.0.0rc3/training_config/syntax/bert_base_srl.onnet}

A. Comparison with Portuguese state of the art

As mentioned, work in Portuguese SRL has been scarce. From the models mentioned in Section \ref{sec:related_work} most are trained only on PropBank v1.1 and achieve less than 70F\textsubscript{1} in cross-validation. The exception is Alva-Manchego’s model, trained in PropBank v2 and tested in PropBank v1.1 by Hartmann et al. \cite{alva-manchego}. However, we compare only to the work of Falci et al. \cite{Falci2016}, since they describe their methodology in detail, allowing us to compare the models trained in the exact same data.

We ran both brBERT\textsubscript{base} and brBERT\textsubscript{large} on the PropBank.Br v1.1 corpus pre-processed as described in section \ref{sec:experiments} with a batch size of 16 and a learning rate of 4 \times 10^{-5} for up to 100 epochs with early stopping after 10 epochs without improvement. The parameters are based on AllenNLP’s model configuration, but due to memory constraints, the batch size had to be reduced (from 32 to 16); the learning rate was reduced as well (from 3 \times 10^{-5} to 4 \times 10^{-5}) due to the fact that some folds’ models were overfitting on no argument predictions. In Table \ref{tab:related_work} we present the results for the best model of all folds and the average F\textsubscript{1} across folds\footnote{Note that throughout the text, whenever average results are mentioned, we mean averaged across folds.} since these were the metrics reported by our baseline. The values for the baseline are taken from Falci et al. \cite{Falci2016}.

The BERT models give better results than the bi-LSTM model of the baseline. The improvements on the average F\textsubscript{1} are of 15.85 for brBERT\textsubscript{base} and 16.91 for brBERT\textsubscript{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, however, likely due to the fewer resources used in Portuguese, which hindered the LSTM-based model.

| Met.1 | Met.2 CV | Met.2 Buscapé | CoNLL-2012 |
|-------|---------|---------------|------------|
| #Instances | 5600 | 13665 | 709 | 316155 |
| #Annotated Args | 13135 | 30459 | 1364 | 659297 |
| #Semantic Roles | 18 | 26 | 23 | 21 |

B. Comparing monolingual to multilingual models

In this section, we perform a more robust evaluation of the different pre-trained Transformer-based models by following a different methodology, described in Section \ref{sec:experiments} which uses more data and different pre-processing steps. We also experiment with alternative techniques for further improvement, particularly cross-lingual transfer learning from English.

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 \times 10^{-5} for up to 100 epochs with an early stopping after 10 epochs without improvement. The proposed architecture with the different pre-trained models was run in three scenarios:

1) Fine-tuning only with Portuguese data;
2) Fine-tuning first with pre-processed CoNLL-2012 data for 5 epochs, followed by fine-tuning with Portuguese data (only for multilingual models, represented by a superscript “+En”, e.g., XLM-R\textsuperscript{En}).
3) Intermediate task fine-tuning in dependency parsing for 10 epochs, before fine-tuning in SRL (for the best models from the previous scenarios, represented by a superscript “+UD”).

1) Overall Results: Table \ref{tab:results} presents the average precision, recall and F\textsubscript{1} measure of the 10 models (one for each fold) in their respective test sets and in the Buscapé set.

Note that the values of the monolingual brBERT models are smaller than the ones reported in the previous section because those values were from the validation set and these are from test sets. Moreover, the results in this section are evaluated in more data, with more challenging roles (scarcer data) and include continuation arguments, which were removed for the previous section.

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 \cite{XLM-R}. 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 dataset. This is likely due to this dataset being more difficult, as previously discussed.

2) 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.

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 2 we report the average error for each model in each dataset decomposed into error from argument identification ("Arg Id") and from argument classification ("Arg Class").

Several things are noticeable in this figure. To start with, the error from the non-identification of arguments is much larger than the error from the mislabeling of arguments for all models and both datasets. Hence, our models are better at attributing semantic roles than at identifying argument spans. Secondly, errors are higher in Buscapé. We cannot blame the drop in performance on one of these sub-tasks – both are worse. Lastly, the difference in $F_1$ between models is due both to differences in the identification and the classification of arguments.

3) How do the state of the art multilingual language models compare to existing monolingual models for Portuguese SRL?: When training only with Portuguese data, brBERT$_{large}$ and XLM-R$_{large}$ have similar scores in the test sets, but the monolingual model outperforms the multilingual by approximately 1$F_1$ in the out-of-domain dataset. 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 brBERT$_{base}$ and XLM-R$_{base}$ in the test set than in the Buscapé set, but the difference is still relatively small (approximately 1$F_1$). The mBERT model performs the worst: over 3$F_1$ points below brBERT$_{base}$.

4) Does cross-lingual transfer learning from English help the multilingual models' performance in Portuguese SRL?: When including cross-lingual transfer learning, all multilingual models get an increase in their scores. The less powerful the model, the larger this increase. In this scenario, XLM-R$_{large}$ achieves the best score in the average of the test folds, but is still slightly behind brBERT$_{large}$ in Buscapé. On the other hand, XLM-R$_{base}$ improves enough to become on par with brBERT$_{base}$ in both datasets. As before, mBERT underperforms compared to all other models.

5) Can we improve the results of the SRL task by training the language model on dependency parsing first?: Using the UD dataset did not have a positive impact in the performance of Portuguese SRL for the monolingual BERT$_{large}$ model. For the multilingual XLM-R$_{large}$ and XLM-R$_{En}$$_{large}$ models, fine-tuning first with UD boosts the performance in the Buscapé dataset. In fact, with the UD data, we improve the best results obtained before on Buscapé.

Additionally, both models that were fine-tuned in UD are in general better at argument identification in Buscapé than BERT$_{large}$, 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. [38], 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.

6) 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 IV 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 English and tested in Portuguese.

Despite the large drop in $F_1$ measure between zero-shot cross-lingual transfer learning and training only with Portuguese data – between 9 and 10 $F_1$ points –, it is encouraging to see that the former still perform 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
that sentences will be properly structured and words properly spelled (“clean” data, e.g. journalistic text), it is better to consider the performance 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 requirements, or it is certain they are not met (“unclean” data), it is better to look at the scores from the Buscape set.

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 datasets 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 dataset. If, on the other hand, one is interested only in a subset of roles, it is best to choose the best model for those roles according to the results per roles (an implementation of this heuristic is available with the code for this article).

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. 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 issues with the 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 remove their troops from Somalia until March, the gold labels indicate that “suas tropas da Somália” (their troops from Somalia) corresponds to A1 – “entidade ou coisa sendo retirada” (entity or thing being removed). However, we believe that A1 should be “suas tropas” (their troops) and A2 – “local de onde foram retiradas” (place from where they were removed) – should be “da Somália” (from Somalia). We believe this is due to the usage of automatically annotated parse trees in version 2, which constrained the SRL annotations – in our example, “suas tropas” did not appear as a constituent in the tree, therefore the annotator could not properly classify it. Nonetheless, this dataset is extremely useful in such a low-resource language. We mention this merely as a caveat to warn possible users not to take the obtained scores as fully representative of the model’s performance.

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. Our transfer learning experiment with 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. [11], 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 datasets, by harmonising the role set across versions and by having all of the datasets manually revised once again to eliminate any annotation errors that may exist.

ACKNOWLEDGMENTS

This work was developed in Project Text2Story, financed by the ERDF – European Regional Development Fund through the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 and by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within project PTDC/CCI-COM/31857/2017 (NORTE-01-0145-FEDER-03185).

Daniel Loureiro is supported by the European Union and Fundação para a Ciência e Tecnologia through research grant DFA/BDE/9028/2020.

REFERENCES

[1] D. Jurafsky and J. H. Martin, “Speech and language processing,” 2019, 3rd edition draft.

[2] L. Márquez, X. Carreras, K. C. Litkowski, and S. Stevenson, “Semantic role labeling: An introduction to the special issue,” Computational Linguistics, vol. 34, no. 2, pp. 145–159, 2008. [Online]. Available: https://doi.org/10.1162/coll.2008.34.2.145

[3] D. Shen and M. Lapata, “Using semantic roles to improve question answering,” in 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, J. Eisner, Ed. ACL, 2007, pp. 12–21. [Online]. Available: https://www.aclweb.org/anthology/D07-1002/
[33] Y. Pruksachatkun, J. Phang, H. Liu, P. M. Htut, X. Zhang, R. Y. Pang, C. Vania, K. Kann, and S. R. Bowman, “Intermediate-task transfer learning with pretrained models for natural language understanding: When and why does it work?” CoRR, vol. abs/2005.00628, 2020. [Online]. Available: https://arxiv.org/abs/2005.00628

[34] M. Gardner, J. Grus, M. Neumann, O. Tatford, P. Dasigi, N. F. Liu, M. Peters, M. Schmitz, and L. S. Zettlemoyer, “AllenNLP: A deep semantic natural language processing platform,” 2017.

[35] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Raault, R. Louf, M. Funtowicz, and J. Brew, “Huggingface’s transformers: State-of-the-art natural language processing,” ArXiv, vol. abs/1910.03771, 2019.

[36] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “Pytorch: An imperative style, high-performance deep learning library,” in Advances in Neural Information Processing Systems 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[37] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, “Unsupervised cross-lingual representation learning at scale,” CoRR, vol. abs/1911.02116, 2019. [Online]. Available: http://arxiv.org/abs/1911.02116

[38] E. Strubell, P. Verga, D. Andor, D. Weiss, and A. 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, 2018, pp. 5027–5038. [Online]. Available: https://www.aclweb.org/anthology/D18-1548/