Finding Universal Dependency patterns in multilingual BERT's self-attention mechanisms

Jenny Elizabeth Valencia Carmona
STUDENT NUMBER: 2038222

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

Thesis committee:

dr. Grzegorz Chrupała
dr. Martin Atzmüller

Tilburg University
School of Humanities and Digital Sciences
Tilburg, The Netherlands
June 2020

Preface

This thesis marks the final stage for my Master's degree in Data Science & Society at Tilburg University, written from February to June 2020. It is dedicated to my family and friends for their love and support. I would also like to thank my thesis advisor dr. Grzegorz Chrupala for his help and patience. A special and personal thanks to Bancolombia and Colfuturo for allowing me to pursue this dream.

Finding Universal Dependency patterns in multilingual BERT's self-attention mechanisms

Jenny Elizabeth Valencia
Carmona

BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained unsupervised language model, which has shown outstanding results within the state of the art of the NLP. Despite that, the reasons for its extraordinary performance are still unclear and, in general, a complete understanding of the exact way that this model processes language has not been reached yet. Recent work focused on understanding the internal structure of context-based pre-trained models, such as BERT, has conducted a series of visual and quantitative tests, looking for underlying syntactic patterns in different parts of its architecture. Almost all studies so far have been carried out in the English language, and it is unknown if some findings can be extended to other languages. Hence, the purpose of this project is to carry out a cross-lingual analysis, focused on the analysis of the way that BERT captures universal dependency relations in English, German, Italian and Spanish languages. To achieve this, an analysis was performed using attention weights to measure the performance and distance metrics to find and compare the underlying syntactic information in each language. Probing classifiers were also used, seeking to improve the previous metrics adding word information. This study showed that when the different patterns of attention in BERT between languages are compared, there are no specialized heads, but even so, there are similarities in the patterns that are analyzed when compared in a parallel way between languages.

1. Introduction

The use of Deep Learning models has meant a significant improvement in the performance of a wide range of tasks in the area of Natural Language Processing (NLP). This advance has been mostly due to concepts such as transfer learning and pre-trained models (Otter, Medina, and Kalita 2020). In NLP, transfer learning consists of a train a model using an extensive data set and then adapt that model for different tasks using a different data set.

Initially, these models are pre-trained in large text corpus in an unsupervised manner due to the difficulties of finding annotated corpus, and then they are fine-tuned to the specific tasks of small data-sets using supervised models. Improvement of the results observed are attributed when using pre-trained models is because the structure of the induced language model manages to learn a universal representation of language, which translates into improvements in initialization, generalization, adjustment and performance in the target language (Sun et al. 2020).

One of the most popular pre-trained models recently is BERT (Bidirectional Encoder Representations from Transformers). BERT is a model that has shown outstanding results in the state of the art of the NLP and has set a benchmark in eleven language tasks including GLUE, MultiNLI, SQuaADv1.1 and SQuaD v2.0 (Devlin et al. 2019). Subsequently, it has been fine-tuned in various tasks, including classification models, sentiment analysis, machine translation, among others, showing better precision than zero-shot learning models (Tran, Bisazza, and Monz 2018; Sun et al. 2020).

A fair question about this type of model is about the interpretability of the obtained results. Many times, the lack of explainability of models leads to low prediction reliability. In the case of models such as BERT, the search for explainability has been focused on the analysis of how this model manages to represent syntactic and semantic information.

The concept of syntax-aware models (SALMs) is not new. Previously, recurrent neural networks based models (RNNs), like long short-term memory (LSTMs), demonstrated to be capable of capturing semantic and syntactic phenomena not only at the superficial level but also by revealing more profound linguistic knowledge as reported by Linzen, Dupoux, and Goldberg (2016), Shi, Padhi, and Knight (2016) and Gulordava et al. (2018). Later on, models based on transformers (Vaswani et al. 2017) appeared on the scene.

The multi-head attention mechanism is a crucial part of the transformer architecture. Thanks to that, the model is capable to simultaneously focus on different parts of input and calculate the representations of its input and output without using RNNs or some sort of sequential model. In the case of BERT, this is performed in a bidirectional way, also having a deep transforming network that processes long texts efficiently (Vig

and Belinkov 2019). Unlike context-free models like Word2vec or GloVe, BERT generates a representation of embeddings from a single token for each word in the vocabulary that is list based on the other words present in the sentence.

The hypothesis that BERT's attention-heads learn unsupervised syntactic relations has already been addressed in some previous studies in the English language. Several authors have found patterns in various layers, analyzing the weights of attention associated with a couple of words, such as attention to the same word or in specific off-sets. In the case of some syntactic dependencies, it has been possible to observe particular patterns in specific layer-heads, where the weights of attention are concentrated between the syntax-heads and their respective dependent, in some cases, exceeding the baseline associated with the average distance of dependence in each language. These studies start from the assumption that the self-attention weights between two words can be interpreted and that, in this case, in some attention heads, they work indicating the degree of dependency relation between two words. On the contrary, other authors consider that these relations are not underlying BERT and should be an additional step in fine-tuning the model (Sundararaman et al. 2019).

In the case of this project, a greater understanding of the syntactic dependency representations on BERT will be sought, looking to exploit the underlying syntactic information combined with previously annotated texts and analyze how this is captured through the mechanisms of BERT's attention heads from different languages using multilingual BERT (M-BERT). This BERT version takes 100 languages from Wikipedia corpus.

Except for machine translation (MT) research, most research has focused on linguistic analysis in the English language, so this analysis is expected to provide new information in the field. Each language has unique characteristics that make syntactic rules differ from each other, so the findings to date in pre-trained models must consider that their conclusions may only be limited to one language, and that including other languages may lead to better insights in understanding the natural language.

Overall, this project seeks to answer the following questions:

RQ1 - How good is M-BERT capturing syntactic dependency relations in its attention weights in other languages besides English?

RQ2 – Which specific attention heads have the highest accuracy finding syntactic dependency relations using the attention weights of the M-BERT model?

RQ3 – How similar are the attention heads among different languages?

To answer the questions above, a series of experiments will be carried out with the English-BERT and the multilingual version M-BERT using annotated data sets in 4 different languages (Spanish, Italian, English, German) following the visual and quantitative methodology of Clark et al. (2019) partly. The maximum attention weight from attention-heads between words were used to create the inputs of the prediction model, and the baseline choosen the weights corresponding to the common off-set that shows the best overall performance.

Distance and variability measures can also be used to address the underlying similarities between representations in different languages. Probing classifiers will be implemented using a combination of attention weights and pre-trained multilingual word2vec embeddings, so that they would provide answers to the research questions on how the weight information is used, translating them into linguistic information in a more understandable way.

Although in this case, the proposed tests will be used on English-BERT and M-BERT, the results can serve as the basis for the analysis of other pre-trained models.

2. Background

2.1 BERT

BERT is an open-sourced language representation model developed by researchers at Google in 2018. BERT was trained using unlabeled text and massive data sets from Wikipedia and Book Corpus, a data set containing more than 10,000 books of different genres. Much of its popularity is because BERT can be fine-tuned to perform various tasks (Devlin et al. 2019).

BERT has been a breakthrough for NLP compared to traditional models and has served as inspiration for new architectures like GPT-2 (Radford et al. 2019), RoBERTa (Liu et al. 2019) and XLNet (Yang et al. 2020) .

What makes BERT different from OpenAI GPT (a left-to-right transformer) and ELMo (a left-to-right and right-to-left trained LSTM concatenation), is that the model architecture is a deep bi-directional transformer encoder.

Transformers were proposed by Vaswani et al. (2017). Unlike RNNS, transformer forgoes the sequential structure of previous models to change it for a focus on attention. This new structure is more efficient to capture long-distance dependencies because it is capable of look at other positions in the input sequence for contextual information that allows the model a better encoding of each word. Transformers use multiple layers, and each layer contains attention heads.

An attention head takes the n words of the input sentence into an abstract vector. This vector extract different components: query q_i, key k_i, and value vectors v_i. The head computes attention weights α between all pairs of words as softmax normalized dot products between the query and key vectors. Then, the weighted sum of the value vectors will be the output of each attention head (Clark et al. 2019).

$$\alpha_{ij} = \frac{\exp(q_i^T k_j)}{\sum_{l=1}^{n} \exp(q_i^T k_l)} \qquad o_i = \sum_{j=1}^{n} \alpha_{ij} v_j \tag{1}$$

BERT has twelve attention heads for each layer. Each attention mechanism allows each word in the input sentence to pay attention to any other word. Transformers train many independent attention heads and each head has the possibility focus on a different type of combinations. See Figure 1.

In addition to being bi-directional, BERT is also pre-trained. Its model architecture is first trained on a language modeling target and then adjusted for a subsequent supervised task. Model weights are learned in advance through two parallel unsupervised tasks: Masked language modeling (MLM) and Next sentence prediction (NSP):

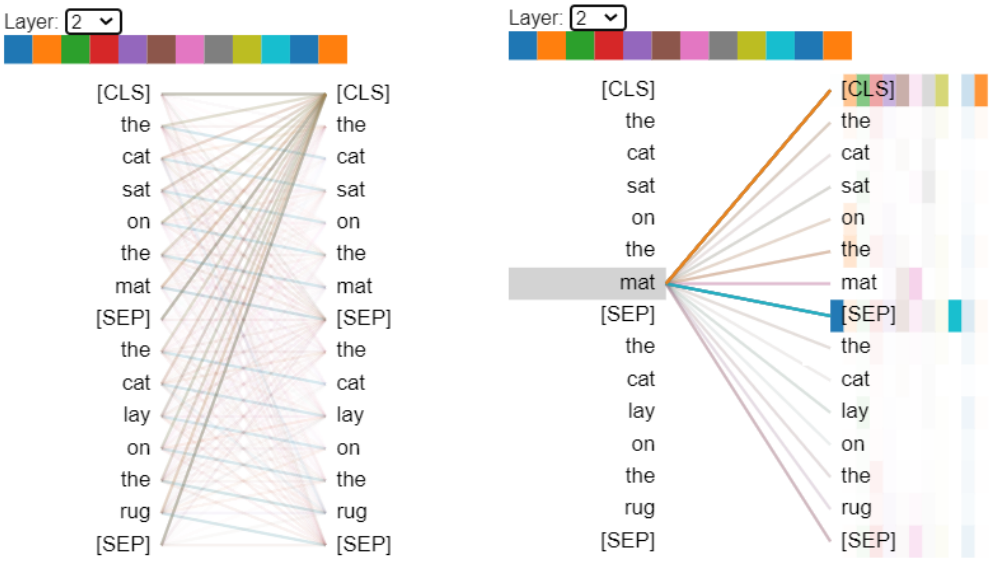


Figure 1: Visualization of attention in BERT using BertViz (Vig 2019). 2019). Left: visualization of attention between all words of the sentence. Right: visualization of attention from selected word only.

- Masked Language Modeling (MLM): It is used to predict a missing word given the left and right context. Before feeding the word sequences into BERT, 15% of the words in each sequence are replaced randomly with a [MASK] token and the model tries to predict the original value of the masked words, based on the information from unmasked words.
- Next Sentence Prediction (NSP): BERT learns to model the relations between sentences predicting if a second sentence follows another.

2.2 Multilingual BERT (M-BERT)

Multilingual BERT (M-BERT) is a 12-layer transformer trained on Wikipedia pages for multiple languages. To balance languages with different size of examples, languages like English were under-sampled.

M-BERT learns in an unsupervised way, which means that the equivalencies between the different languages have not been indicated. M-BERT produces a representa-

tion that seems to generalize well in all languages for a variety of subsequent tasks. The results of this model have been remarkable, especially in zero-shot cross-lingual model transfer tasks (Pires, Schlinger, and Garrette 2019).

Wu and Dredze (2019) related M-BERT's multilingual performance to the number of words shared between the source and target languages. His initial hypothesis is based on explaining the success of M-BERT with the level of language similarity, either lexical or structural similarities (word order or word's frequency), or both.

In contrast, Pires, Schlinger, and Garrette (2019) and Karthikeyan et al. (2020) consider that the success of M-BERT is not due to a lexical overlap between common words between languages, nor to the ability to memorize vocabulary, but instead, to its ability to generalize. They also notice that such generalization does not work equally well in all cases, mainly where topological similarity is not available, for example when the Subject-Verb order is different. However, it is still possible to achieve decent results under such circumstances. Also, it is considered that the use of word embeddings in all the languages that must be assigned to a shared space force the concurrent pieces to be also assigned to a shared space, thus extending the effect to other word embeddings until different languages are close of a shared space.

Kondratyuk and Straka (2019) Compare M-BERT's generalization with how learning a new language can improve a speaker's proficiency in previous languages, i.e. a model that has access to multilingual information can learn generalizations between languages that would not have been possible through a monolingual model.

3. Related Work

Achieve a better understanding of the internal representations of language that is stored in pre-training models is a booming line of research since it seeks to give more significant explanations to the outstanding capabilities of these models that are later used in supervised tasks. These types of works have generally focused on the search for semantic and syntactic structures through the internal architectures of these models.

For the search of syntactic representations, the focus is to search patterns in the attention mechanisms using hierarchical trees or diagnostic classifiers that measure

the accuracy of specific relations. BERT, despite its highly complex, large number of parameters and its unsupervised training with unlabeled data, seems to leave clear traces of linguistic knowledge.

In BERT, each layer-head combination has independent parameter values, for that reason, in each one, it is possible to look for a unique analytical and visual representation that could correspond to a hidden linguistic representation (Vig and Belinkov 2019).

Clark et al. (2019) found that BERT attention's heads exhibit patterns in English, such as attention on unique tokens, or on certain specific off-sets, which may correspond to syntactic notions given the relation between the distance of each word with their syntactic head.

Kovaleva et al. (2019) conducted a series of experiments, using a subset of GLUE tasks and a set of handcrafted features, he extracted self-attention weights heads and determining token by token which target tokens were providing more information. He found five patterns in the different heads. Then, he disables redundant heads seeking to validate the importance of the relations, and he did not find detriment in its results.

Htut et al. (2019) took the maximum attention weight and calculated the maximum spanning tree looking to extract dependency relations. The author found attention heads in BERT and RoBERTa that can perform significantly better than their baselines. Individual attention heads were evaluated for syntax incorporation of syntactic structures through fine-tuning on a semantics-oriented task, finding as a result that this encourages useful long-distance dependencies. However, it slightly degrades the performance in other shorter-distance dependency types.

Voita et al. (2019) carried out a study of the contribution made by individual attention heads in transformers models. They found that the heads related to linguistic representations were consistent and played different roles: some heads referred to positional, syntactic, and others paid attention to rare words. To identify essential heads, they used a layer-wise relevance propagation-based method. Finally, a pruning was performed on the heads using a method based on stochastic gates observing that specialized

heads were last to be pruned, managing to remove most of the heads without affecting performance.

Another approach to search for a syntactic structure is to use the syntactic information contained in contextualized embeddings instead of using only attention weights. Goldberg (2019), Tenney et al. (2019) and Reif et al. (2019) suggested that context embeddings encoded dependency parse trees geometrically and provided evidence that it could work quantitatively using the intensity of the tokens in a sentence. In the same way, Hewitt and Manning (2019) provided evidence for the existence of syntactic knowledge through syntax trees searching, which was embedded in the representations of BERT's word's space as linear transformations. They found representations that were not found in baselines, supporting the underlying learning of the model. Meanwhile, Mareček and Rosa (2018) derived distribution of individual words through an aggregation of self-attention over layers and constructed various types of syntax trees based on that information (example: constituency trees, undirected trees, dependency trees) and implemented them in several sentences sampled from Penn Tree Bank.

Sundararaman et al. (2019) and Im and Cho (2017), believe that syntactic or semantic relations are not captured in an underlying form, or that they are found in a weak way. Therefore, it is necessary to previously train the models and induce the syntax trees on them. Currey and Heafield (2019) considered that patterns found cannot necessarily be interpreted as explainability, so in order to seek a better performance of the models, they consider that the linguistic structures must be "added" to achieve better performance. Raganato and Tiedemann (2018) indicated that Transformers models might not learn syntactic structures in the same way as sequential models and call for explicitly adding this type of information. Gulordava et al. (2018) focus too on including stimuli, that also emphasizes linguistic structure over other behaviours.

From another perspective, the location of the type of knowledge extracted in this type of model is generally associated with specific layers. Lin, Tan, and Frank (2019) considers that BERT encodes positional information about words in its lower layers and hierarchical information in upper layers. Jawahar, Sagot, and Seddah (2019) suggests that the lower layers of the Transformer decoder handle modelling the language, while the last layers handle the input sentence. Peters et al. (2018) found that the lower layers

of a language model encode a higher local syntax, while the higher layers capture complex semantics.

Although there has been previous work on the analysis of attention heads in BERT, the present work seeks to be the first to look for such dependency relations learned by BERT's attention heads in languages other than English, focusing on the similarities and differences found in each attention head according to the dependency relations analyzed. Structures of the different languages will be used to analyze if the patterns established in English BERT may correspond to fortuitous results or if, on the contrary, this type of behaviour is extended or related with inputs in a cross-lingual analysis, and therefore with differences in its syntax to the English language.

4. Methods

This work will seek evidence of the existence of syntactic patterns in multiple languages through the analysis of Universal Dependencies relations, based on self-attention weights through the attention in a multi-language scenario through a series of experiments.

4.1 Maximum attention weights extraction(MAX)

This work starts from the hypothesis of using the attention head weights in BERT as a measure of the relation between two pairs of tokens corresponding to the same sentence. For this thesis, the underlying relation sought is the syntactic dependency relation, as expressed in the Universal Dependencies data set.

The maximum attention weights extraction methodology will be used to estimate the syntactic head of each word (MAX) (Htut et al. 2019).

For a relation (w_i, w_j) between word w_i and w_j relations for all sentences are extracted if $j = argmax W[i]$ for each row i in attention matrix W.

This simple method allows calculating syntactic dependency relations without needing to form a hierarchical tree. For each given sentence, using a previously anno-

tated corpus, self-attention weights from BERT will be extracted for each word, that is, the attention paid by each of the words to others and themselves in different directions will be calculated. The highest of these weights will indicate which word is given more attention and will be considered the predicted syntactic head of the particular word and will be compared with the actual syntactic head, according to the original annotated text.

The baseline chosen for the analysis of dependency relations will be the most common positional off-set between two words. For example, the determiner: 'det' relation usually occurs one post-word off-set. Variations in the off-set between -3 and +3 will be included looking if it is possible to find patterns related to long-distance dependency patterns.

4.2 Probing classifiers

Using the methodology use by Clark et al. (2019), two probing classifiers will be implemented. First, using attention weights only. Then, using attention weights combined with word embeddings. The data source will be separated like this: 80% to training and 20% for validation.

Probing classifiers work as classification models, in this case, each attention map is taken as input to the model, seeking to measure the model's ability to capture the knowledge of syntactic relations, which are distributed across multiple attention heads. Each dependency is evaluated exclusively, using a softmax classifier and then a global indicator is calculated.

The first probing classifier works by calculating the probability per word, that is, that another word in the same sentence corresponds to its syntax head. This result is accomplished through a linear combination of attention weights. See Equation 2.

$$p(i|j) \propto \exp\left(\sum_{k=1}^{n} w_k \alpha_{ij}^k + u_k \alpha_{ji}^k\right) \tag{2}$$

The second probing classifier uses, in addition to attention heads weights, the particular information of each word, that is, the multilingual pre-trained Word2Vect embeddings. See Equation 3.

$$p(i|j) \propto \exp\left(\sum_{k=1}^{n} W_{k,:}(v_i \oplus v_j)\alpha_{ij}^{k} + U_{k,:}(v_i \oplus v_j)\alpha_{ji}^{k}\right) \tag{3}$$

Where u denotes word2vec embeddings and $\oplus$ denotes concatenation.

5. Experimental Setup

This chapter describes the experiments performed that seek to answer the research questions previously presented in the introduction. The technical aspects of the implementation of the English-BERT and the M-BERT model will be explained using the proposed languages and the methods used for their interpretation, evaluation and contrast.

5.1 Software

All steps following post-extraction weight analysis were implemented using Python in Google Colab.

The following pre-trained models will be used [1]:

BERT-BASE uncased: 12-layer, 768-hidden, 12-heads, 110M parameters. Trained on lower-cased English text.

BERT-BASE -multilingual-uncased 12-layer, 768-hidden, 12-heads, 110M parameters. Trained on lower-cased text in the top 102 languages with the largest Wikipedias.

BERT-English and multilingual BERT are based on the Tensor Flow implementation.

1 https://huggingface.co/transformers/pretrained_models.html

The Base version was selected to each model due to its higher performance and interpretability as it has fewer attention heads to analyze.

5.2 Data

Universal Dependencies (UD) [2] is a cross-linguistically consistent treebank annotation of grammar across different languages. It is widely used in multilingual parsing research problems, using unified annotations between several languages, allowing comparison. The bases corresponding to Parallel Universal Dependencies (PUD) treebanks, created for the Conference on Computational Natural Language Learning 2017 (CoNLL2017), will be used explicitly for four relatively different languages belonging to two families. The languages used in these experiments were chosen, considering their similarity in each family. Spanish and Italian for Romance, English and German for Germanic. PUD data-sets contain treebank annotations that use basic Stanford Style dependencies. In the data-sets, all data contains the same 1000 sentences in their respective language, ordered in the same way, taken from news and Wikipedia and which were annotated morphologically and syntactically by Google according to Google universal annotation guidelines to be subsequently validated.

Word Embeddings: The embed_tweets_multi_300M_52D pretrained embeddings model was used.[3] This is a word embedding pre-trained model with Word2Vec on 300 million multilingual Tweets using 52 dimensions.

5.3 Feature extraction

Each language featured contains an unlabeled version and a parsed version. The latter appears in the CoNLL-U format. In this format, annotations are encoded in raw text files where blank lines mark sentence boundaries.

Table 1 shows the number of dependencies found in the training text for each language. For subsequent analyzes, only those with at least 100 samples in all languages

2 http://universaldependencies.org/
3 http://spinningbytes.ch/resources/word embeddings

rel	English	German	Italian	Spanish
advcl	293	220	250	175
advmod	852	1.103	777	843
aux	410	365	462	375
case	2.499	2.053	3.443	3.698
cc	574	724	589	565
ccomp	135	169	137	149
conj	634	842	662	652
cop	316	274	298	289
det	2.047	2.736	3.751	3486
mark	555	459	506	275
nmod	1.076	1.101	1.876	1.803
nsubj	1.393	1.481	1.126	1.189
nsubj:pa	239	207	167	165
nummod	254	226	202	192
obj	876	895	849	790
obl	1.237	1.344	1.591	1.550
punct	2.451	2.771	2.303	2.292
xcomp	271	190	249	356

Table 1: Frecuency distribution of Parallel Universal Dependencies (PUD) by language for dependencies amod, det, aux and nmod.

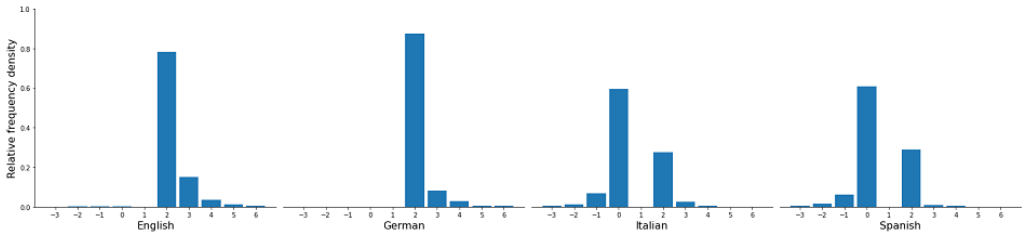
were chosen. Figure 2 presents the distribution of the distances between the word and its syntactic head for four dependency relations. The similarity of the distribution is evident for the relations corresponding to the same family.

The following features were chosen:

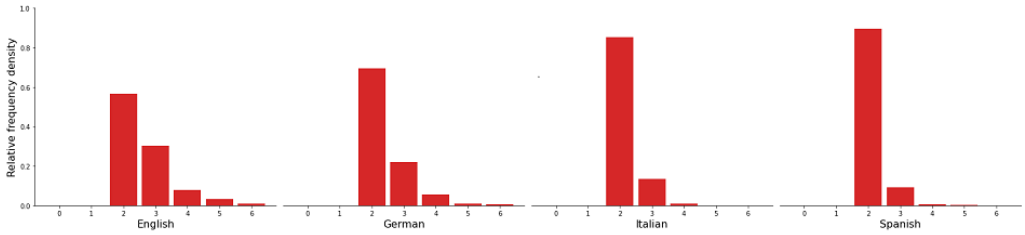
- Column 2 - FORM: Word form or punctuation symbol.
- Column 7 - HEAD: Head of the current word, which is either a value of ID or zero (0).
- Column 8 - DEPREL: Universal dependency relation to the HEAD (root if HEAD = 0).

5.4 Training

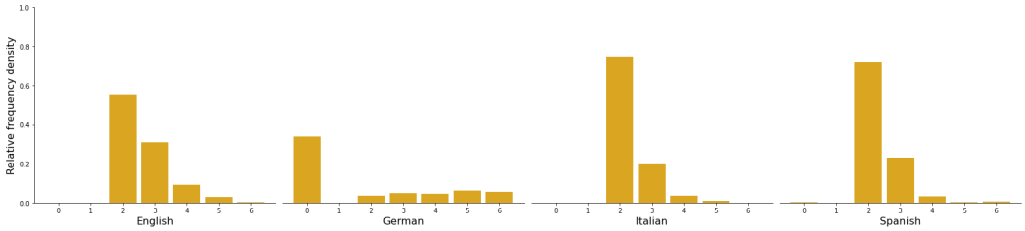
Following Htut et al. (2019), attention weights for each language were extracted using the maximum attention weights extraction (MAX) method. Each example was adjusted



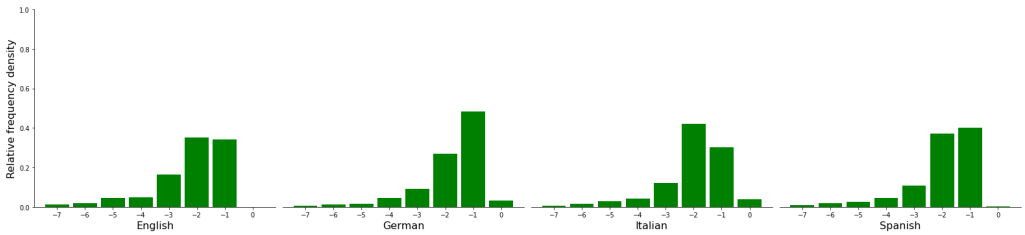
(a) amod



(b) det



(c) aux



(d) nunmod

Figure 2: Relative frequency of a word's distances to its syntactic head for the relations amod, det, aux and nummod.

to the BERT format, adding [CLS] / [SEP] to each sentence. The prediction of the syntactic head was made for each word. Likewise, the prediction for baseline was calculated using the weights corresponding to the attention values between -3 and +3 off-sets.

Using the probing classifier method introduced by Clark et al. (2019), the attention head values in BERT were used as inputs within a classification model. This time, instead of predicting a particular dependency relationship, It is evaluated how good BERT is by predicting overall the total dependencies of a text. Two probing tasks were implemented, the first using only the attention heads, and the second using word embeddings.

5.5 Evaluation

For the evaluation of predictions, the evaluation measures will be:

Accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

Where TP, FN, FP and TN represent the number of true positives, false negatives, false positives and true negatives, respectively.

F1 score: precision (PR) and recall (RE) are defined as:

$$PR = \frac{TP}{TP + FP} \qquad RE = \frac{TP}{TP + FN} \tag{5}$$

where TP is the number of true positives, FP the number of false positives and FN the number of false negatives.

Since in this work a multiclass classifier is being considered, the $F1$ score for each class is defined as:

$$f1_c = \frac{2PR_cRE_c}{PR_c + RE_c} \tag{6}$$

where PR_c and RE_c are the precision and recall of the class c.

the overall $F1$ score is defined as

$$F1 = \frac{1}{n}\sum_{c \in A} f1_c \tag{7}$$

where A is the set of all classes.

Distance Metrics:

To compare the 144 (layers x heads) x 4 (languages) attention heads of different languages, the cosine distance and MSE score will be used.

Cosine Distance Cosine distance is define as

$$cos\_dis(X, Y) = 1 - \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2}\sqrt{\sum_{i=1}^{n} Y_i^2}} \tag{8}$$

Mean squared error (MSE)

$$MSE(X, Y) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} (X_i - Y_i)^2 \tag{9}$$

In the case of the comparison with English-BERT(English_o), since each pre-trained model has been initialized randomly, the same attention function may be located in a

head with a different location in the multilingual model. The previously used MSE and Cosine Distance metrics are not invariant to the permutation. Therefore, the metrics will be used using the total permutations of the vectors and calculating the minimum distance between the heads in each layer, including all the possible permutations of the heads in the corresponding layer of the other language. Finally, the minimum difference per layer is calculated, obtaining a vector that will later be summarized using an average.

5.6 Reproducibility

Code is available at: https://github.com/eliza166/mbertparsing

6. Results

In this section, the results of the comparative experiments in the four chosen languages will be presented.

In the first part, attention heads were calculated with the weights that generated maximum precision using CoNLL. 2017 data-set. Table 2 is sorted by dependency name and displays results between languages in parallel. For this part, only relations with several samples equal to or greater than 100 were considered. As in previous analysis in the English language, there is evidence of some dependency relations in specific attention heads, which do not necessarily coincide between languages. However, most relations between languages in the same family are quite similar.

In most cases, the accuracy calculated based on the attention weights usually shows a considerable improvement compared to the accuracy calculated based on the off-set. Other cases presented, show slight values differences between the value calculated by the accuracy and the value calculated using off-set, possibly indicating that BERT would be learning the distance to the word head, instead of a hierarchy relation.

In the second part, its sought to compare the capture capacity of semantic dependencies, first visually. See figures 5, 6, 7, 8 and the similarity between different attention heads in different languages using cosine distance and MSE metrics. See tables 5, 6, 7, 8.

	English_o			English			German			Italian			Spanish		
rel	**Acc**	**F1**	**B/L**	**Acc**	**F1**	**B/L**	**Acc**	**F1**	**B/L**	**Acc**	**F1**	**B/L**	**Acc**	**F1**	**B/L**
advcl	29.4	30.6	8.2	38.2	38.5	8.2	27.3	27.5	9.5	33.2	34.3	6.4	33.1	33.5	5.7
advmod	56.7	53.4	48.6	61.3	58.9	48.6	55.5	52.5	38.9	62.3	60.1	39.8	60.9	57.7	39.0
amod	88.9	88.1	78.3	83.2	81.1	78.3	87.4	85.7	87.3	87.5	86.0	59.6	89.7	88.5	60.7
aux	77.6	48.5	55.4	80.2	45.8	55.4	83.8	55.7	34.0	87.9	58.1	74.7	83.2	53.6	72.0
case	82.8	76.2	36.2	77.8	79.0	36.2	76.8	82.8	46.4	78.1	87.2	56.0	82.4	83.4	47.5
cc	59.2	78.0	43.4	60.5	71.4	43.4	57.5	72.1	37.2	61.1	71.5	35.7	67.1	75.5	41.1
ccomp	51.1	57.4	14.8	28.9	58.2	14.8	21.9	54.0	3.0	37.2	58.7	8.8	38.9	64.7	18.1
conj	54.1	51.7	27.8	53.5	28.3	27.8	43.8	21.9	25.8	48.8	36.6	19.2	49.4	37.4	22.2
cop	68.4	51.0	35.4	73.4	51.9	35.4	80.7	43.8	23.7	76.8	47.4	39.9	83.0	46.9	46.4
det	95.2	67.3	56.7	87.5	71.8	56.7	91.7	79.7	69.5	87.2	75.6	85.3	90.2	83.0	89.7
mark	67.6	94.0	53.7	62.5	84.2	53.7	67.1	88.1	31.2	67.0	81.6	49.8	51.3	84.9	20.4
nmod	38.5	65.8	35.2	45.2	60.4	35.2	55.1	62.3	48.3	58.8	64.7	42.1	57.1	48.7	40.1
nsubj	54.2	34.2	39.0	56.5	42.7	39.0	52.5	51.5	23.2	57.4	53.3	24.7	59.5	53.5	25.6
nsubj:pa	59.4	48.9	42.7	56.1	52.3	42.7	55.6	47.4	13.5	57.5	53.9	28.1	51.5	56.3	23.0
nummod	72.8	57.9	57.5	78.0	53.8	57.5	69.5	55.5	62.8	86.6	58.3	86.6	95.8	50.7	95.8
obj	85.8	73.9	39.4	83.4	75.9	39.4	78.8	67.8	15.5	82.4	86.8	51.9	84.4	96.5	53.9
obl	30.9	84.4	21.9	40.9	82.0	21.9	47.7	75.9	26.6	52.7	81.5	27.0	47.3	83.8	32.1
punct	25.7	27.6	9.5	21.3	36.4	9.5	29.3	44.7	15.4	24.2	46.9	13.6	25.0	42.1	12.0
xcomp	69.7	68.3	46.5	65.7	64.7	46.5	50.0	50.7	17.4	75.9	75.6	44.2	45.8	43.3	32.6

Table 2: Best accuracies and F1 scores of English-BERT (English_o) and M-BERT using MAX method for Parallel Universal Dependencies (PUD) data. Baseline accuracies are calculated by choosing the best accuracy by taking off-set weights within the range of -3 and +3 for each word

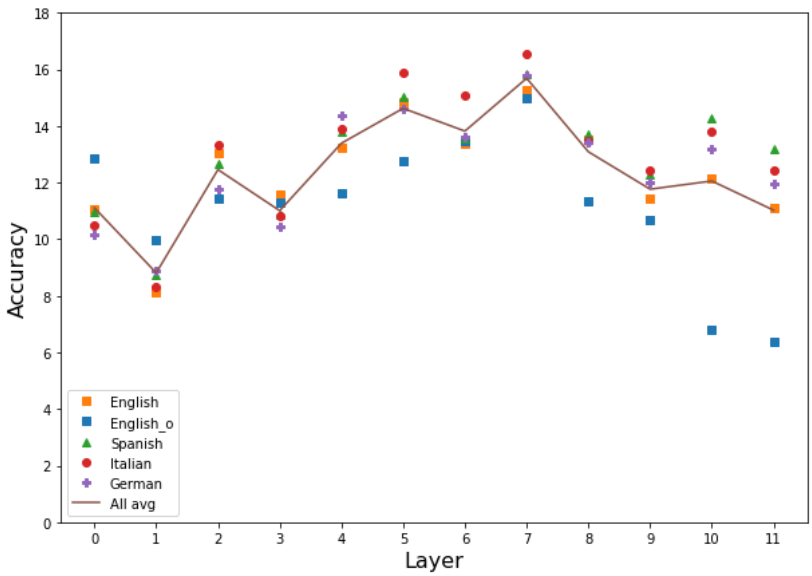


Figure 3: Best global accuracy per layer for each language. The solid line represents the average accuracy for all the relations found in the text.

Figure 5 shows the similarity between the average of the accuracy score for each head between M-BERT Spanish- M-BERT Italian and M-BERT English- M-BERT Ger-

man, respectively. In heat maps, the dark colour reflects higher accuracies, which leads to the conclusion that generally, regardless of the language, the layers located between positions 1-7 tend to encode this type of relation better. The last layer does not reflect significant values for this task using M-BERT.

Contrary to the initial hypothesis, which was established that individual attention heads focused on the recognition of certain linguistic relations, the attention head 6-1 seems to recognize more than one dependency relation. The hypothesis of the trend towards higher attention weights in the initial layers independent of the language is supported.

In the case of the 'amod' relation, both Spanish and Italian data-sets have the best results in -1 fixed position and English and German in +1. See Table 3.

In the case of the relation 'det', the four languages share the same off-set for their baseline (+1), Figure 6 allows to verify the similarities between the Spanish, Italian and German languages, although these forms do not seem to be replicated in the figure of English.

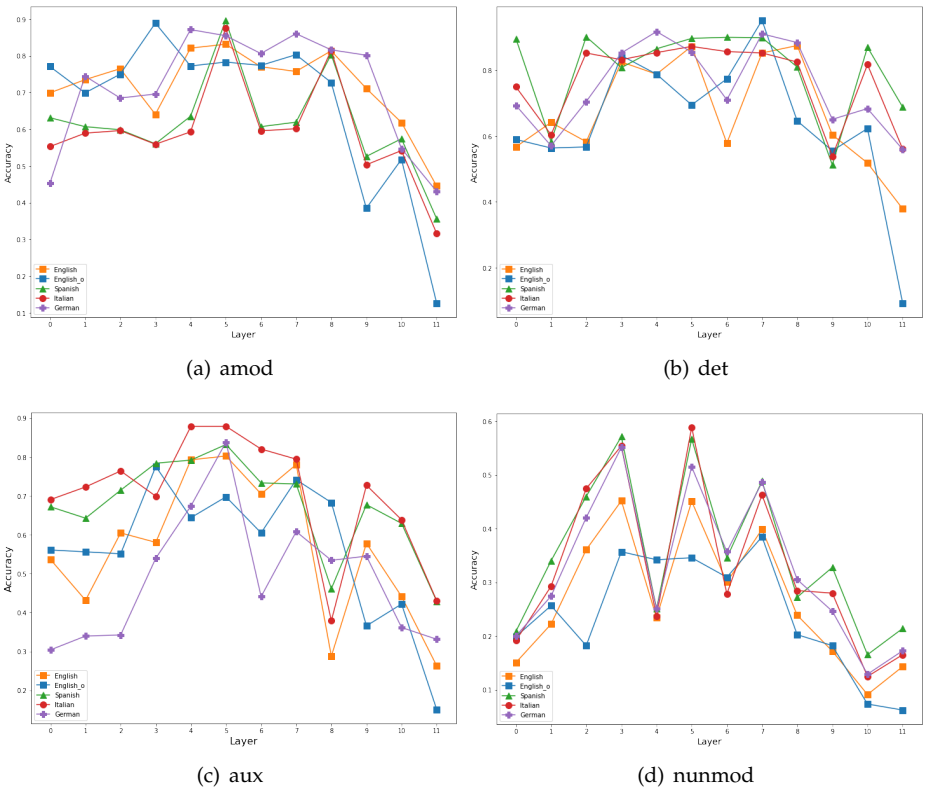
In the third part, cosine similarity was calculated between the heads of the different languages in specific relations, starting from the average of the accuracy in each head. Table 5, for example, shows the distance between the same attention heads for a specific relation. For relation amod, English and German function in a similar way to how M-BERT represents this relation in Spanish and Italian. On the other hand, the Spanish, Italian and English languages seem to coincide and are distant from the same relation represented in German.

Language	Accuracy	offset	Acu. Offset	Layer	Head	Direction
German	87.4	1	87.33	2	6	head<-dep
Italian	87.5	-1	59.57	5	10	dep->head
English	83.2	1	78.29	5	10	dep->head
Spanish	89.7	-1	60.70	5	10	dep->head
English_o	88.9	1	78.29	5	9	dep->head

Table 3: Accuracy using differents off-sets values for amod relation

Language	Relation	Accuracy	Direction
English	compound	91.4	dep->head
German	fixed	85.7	dep->head
German	amod	87.3	head<-dep
Italian	compound	91.7	dep->head
Italian	fixed	87.1	dep->head
Italian	det	85.3	head<-dep
Italian	aux:pass	83.3	head<-dep
Italian	nummod	86.6	head<-dep
Italian	det:poss	90.0	head<-dep
Spanish	flat:nam	87.1	dep->head
Spanish	det	89.7	head<-dep
Spanish	aux:pass	91.5	head<-dep
Spanish	nummod	95.8	head<-dep

Table 4: Accuracies in Layer 6 Head 1



(a) amod (b) det

(c) aux (d) nunmod

Figure 4: Best accuracies per layer for each language for relations a) amod, b) det, c) aux, d) nummod.

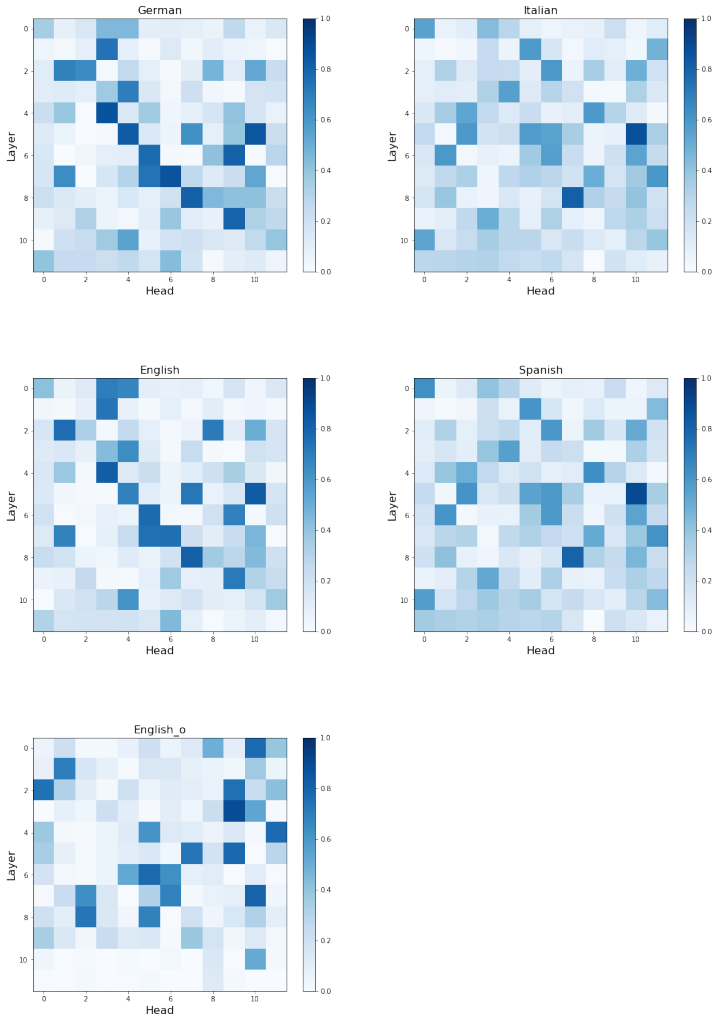


Figure 5: Attention heatmaps visualization for the 'amod' relation using the best accuracy for each head-layer combination. Darker colors correspond to the best scores.English-German and Spanish-Italian couples share visual patterns of recognition of the syntactic dependency 'amod'

amod	**en-en_o**	**en-it**	**en-es**	**en-de**	**it-de**	**es-de**	**es-it**
MSE	0.0891	0.0466	0.0489	0.0033	0.0571	0.0592	0.0006
cosine	0.0341	0.2919	0.2933	0.0129	0.2968	0.3002	0.0030

Table 5: Cosine Distance and MSE metrics to assess M-BERT differences in their ability to predict syntactic dependencies per head of attention for the amod relation. The lower the value, the greater the similarity.

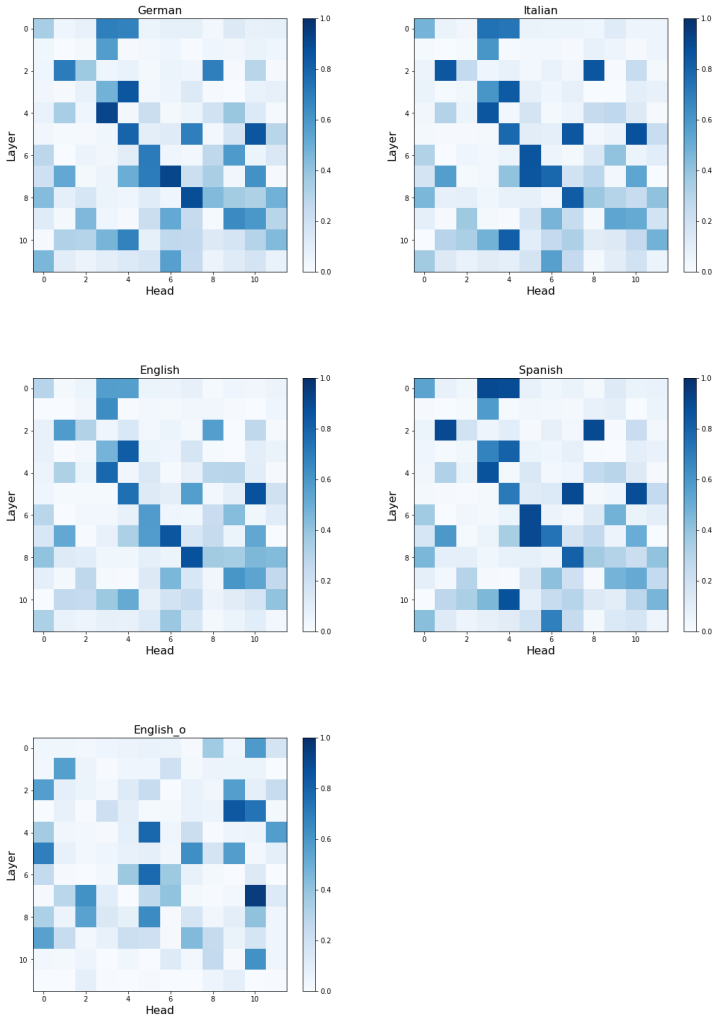


Figure 6: Attention heatmaps visualization for the 'det' relation using the best accuracy for each head-layer combination. German-Italian-Spanish show a similar visual pattern.

det	**en-en_o**	**en-it**	**en-es**	**en-de**	**it-de**	**es-de**	**es-it**
MSE	0.0698	0.0060	0.0079	0.0022	0.0031	0.0055	0.0008
cosine	0.0403	0.0231	0.0352	0.0118	0.0143	0.0242	0.0046

Table 6: Cosine and MSE metrics to assess M-BERT differences in their ability to predict syntactic 'det' dependence

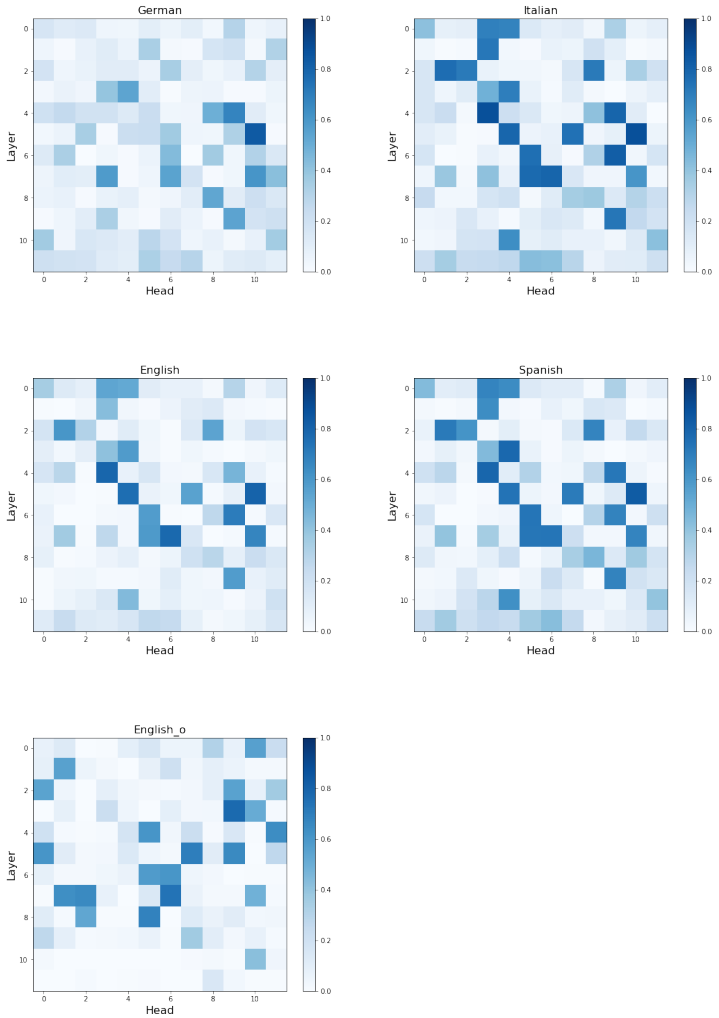


Figure 7: Attention heatmaps visualization for the 'aux' relation using the best accuracy for each head-layer combination. German-Italian-Spanish show a similar visual pattern

aux	**en-en_o**	**en-it**	**en-es**	**en-de**	**it-de**	**es-de**	**es-it**
MSE	0.0619	0.0032	0.0023	0.0316	0.0425	0.0374	0.0014
cosine	0.0415	0.0173	0.0183	0.3360	0.3261	0.3196	0.0076

Table 7: Table Distances for Relation aux

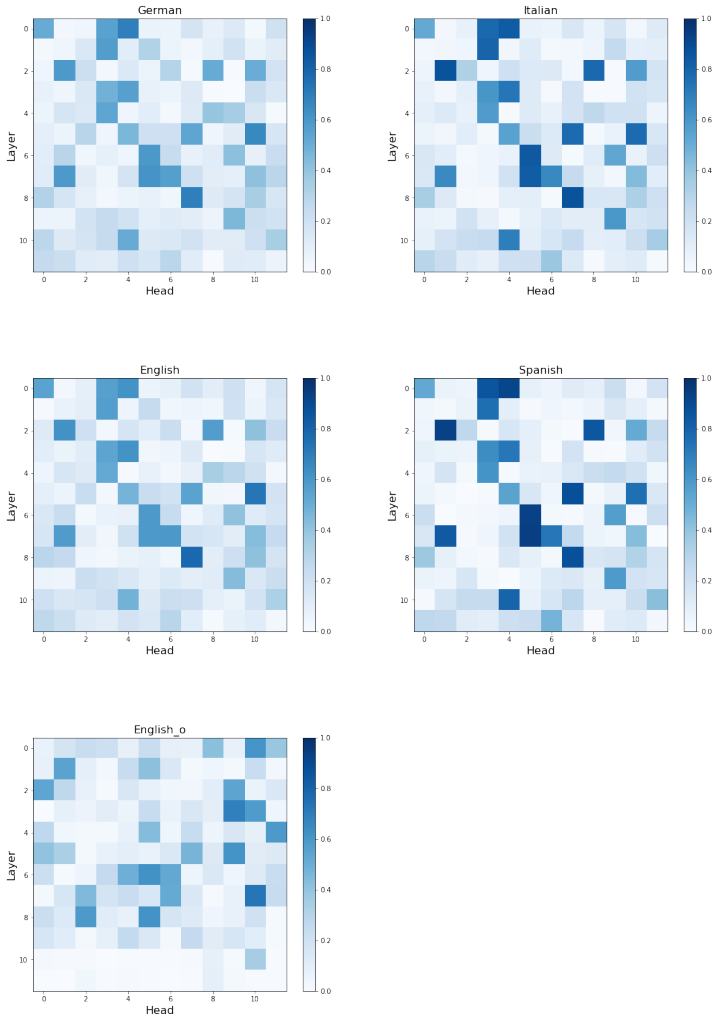


Figure 8: Attention heatmaps visualization for the 'nummod' relation using the best accuracy for each head-layer combination. English-German and Spanish-Italian couples share visual patterns of recognition of this syntactic dependency.

nunmod	**en-en_o**	**en-it**	**en-es**	**en-de**	**it-de**	**es-de**	**es-it**
MSE	0.0049	0.0063	0.0123	0.0014	0.0076	0.0139	0.0023
cosine	0.0308	0.0270	0.0507	0.0100	0.0342	0.0599	0.0102

Table 8: Cosine and MSE metrics to assess M-BERT differences in their ability to predict 'nummod'syntactic dependencies

Table 9 shows the result of implementing probing classifiers using only attention and attention with word embeddings information. The results are similar in both cases for all languages.

In this part, all dependency relationships included training the model, resulting in that the overall performance when all attention maps are used is the almost the same regardless of the language chosen to extract the attention weights from BERT.

	Only Attention	**Attention + words**
English-BERT	57.8	69.3
M-BERT-English	57.1	68.5
M-BERT-German	55.7	67.9
M-BERT-Italian	58.6	69.6
M-BERT-Spanish	56.8	69.1

Table 9: Probing task using two different methods: Attention only: A combination of the attention head weights were used to predict the model result. Attention + words: In addition to the attention weights, the information of each word represented in word2vec multilingual embeddings was used as a model feature.

7. Discussion

The issue of M-BERT's intrinsic syntactic capabilities has been addressed in multiple ways in this work. The objective of this study was to expand the answers to the questions previously analyzed by other authors with data-sets in English, in a context other than machine translation, such as syntax parsing.

This section discusses the results of the experiments carried out in this work. Three research questions were established to achieve the objective of this thesis:

RQ1 - How good is M-BERT capturing syntactic dependency relations in its attention weights in other languages besides English?

In the table 2, corresponding to accuracy and F1 scores, the results show that M-BERT, using English data-set, can calculate dependency relations with results even better than English-BERT (English-o) in some cases (advmod, aux, cc, cop, nmod, nsubj, obl). This result is satisfying, considering that the linguistic information that M-BERT

can interpret is the result of generalization between several languages and that despite being an unsupervised multilingual pre-trained model surpasses a model trained in a specific language.

Regarding the other languages, there present also good performances, although not necessarily in the same proportion for which English-BERT initially stands out. Present experiment provides evidence that languages from the same families tend to present similar results (Italian-Spanish, English-German), although not in all cases. For Xcomp, for example, M-BERT using Italian vocabulary weights achieves an accuracy of 75.9 compared to 45.8 in Spanish, reflecting that despite there being a generalization between languages and particular similarities, M-BERT can capture relevant specific dependency relations in a particular language diminishing their performance in another similar language.

In general, according to figure 3, even taking into account the respective variations of the different languages, the overall performance of M-BERT for each language seems to be very similar, regardless of the way where BERT has been pre-trained. In contrast, English-BERT does not seem to recognize enough syntactic dependencies in its last two layers.

RQ2 – Which specific attention heads have the highest accuracy finding syntactic dependency relations using the attention weights of the M-BERT model?

In the graphic and quantitative results presented in the previous section, several cases can be seen:

- There are specific heads of attention that seem to represent well the same dependency relation in all languages studied.
- There are specific attention heads that work well for some similar languages.
- Furthermore, there are attention heads that differ in all languages.

In general, most attention heads seem not to be good at recognizing these types of relations, and many of these do not even reach the baseline level based on the off-set.

From this, it can be thought that a good accuracy in one language in a specific head is not reflected in its performance in another language, but in cases where the languages belong to the same family, most of the times, the performance was similar.

Most of the dependencies that were recognized in this work are usually short-distance relations, and the most common off-set number seems to be related to the place in the M-BERT architecture where it could be coded according to its weight.

Figure 4 shows that layer 6 and 7, presents outstanding results in at least four relations analyzed. For other types of relations, none of the possible combinations of weights in the attention heads manages to overcome the baseline value, raising the question of if these relations are stored in another location, or if the combination of multiple attention mechanisms like was use in probing classifier experiment is necessary or if M-BERT fails to interpret this type of information within its internal structure.

It should also be kept in mind that when M-BERT is used instead of English-BERT or other monolingual models, a particular type of loss of information will be produced, trying to adapt itself in some way to capture a better generalization in many languages.

It would be interesting to include in the future how the analysis of the variation in the results works when using specific pre-trained models for each language instead of M-BERT. For example, in Clark et al. (2019), the 'det' ratio reaches an accuracy of 94.3 compared to 87.5 using M-BERT in this work. Although the differences may occur for many reasons, for example, the data-sets used in this thesis had fewer examples than that used in his paper: The Wall Street Journal portion of the Penn Treebank.

RQ3 – How similar are the attention heads among different languages?

Cosine distance and MSE were used in this work to validate this statement since it is possible to find similarities between some attention heads when they are used to check which of them could recognize dependency relations through their weights. However, not necessarily the relations found seem to have more strength between those languages that share a family, but rather, with languages that share the same distances to its syntactic head for that specific relation.

Another relevant finding is that for all the languages analyzed, there could be more than one attention head with outstanding results in the prediction of the model, even corresponding to different layers, but in general, one attention head does not imply that the same head in another language can generate the same results. See Table 4.

8. Conclusion

The main objective of this work was to answer the research questions of how good M-BERT is capturing syntactic dependency relations using its attention weights in other languages besides English; and if there are specific patterns and heads within the M-BERT architecture in charge of executing said task.

The accuracy and F1 metrics of each relation in each one of the layers were compared in a parallel way. Similar accuracies were found through the same layer-head positions between languages, especially between languages from the same families, but this result may vary according to the dependency relation analyzed. Another relevant finding is that for all the languages analyzed, there could be more than one attention head with outstanding results in the prediction of the model, even corresponding to different layers, but in general, one attention head does not imply that the same head in another language can generate the same results.

Since M-BERT is a pre-trained language, its internal structure can hold underlying linguistic information that may be the reason for the model's performance, which is notable when it is used to compare dependence relations, even when is compared to monolingual models like BERT-English.

References

Clark, Kevin, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What Does BERT Look At? An Analysis of BERT's Attention. *arXiv preprint arXiv:1906.04341*.

Currey, Anna and Kenneth Heafield. 2019. Incorporating Source Syntax into Transformer-Based Neural Machine Translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 24–33, Association for Computational Linguistics, Florence, Italy.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Goldberg, Yoav. 2019. Assessing BERT's Syntactic Abilities. *arXiv preprint arXiv:1901.05287*.

Gulordava, Kristina, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. *arXiv preprint arXiv:1803.11138*.

Hewitt, John and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Association for Computational Linguistics, Minneapolis, Minnesota.

Htut, Phu Mon, Jason Phang, Shikha Bordia, and Samuel R. Bowman. 2019. Do Attention Heads in BERT Track Syntactic Dependencies? *arXiv preprint arXiv:1911.12246*.

Im, Jinbae and Sungzoon Cho. 2017. Distance-based Self-Attention Network for Natural Language Inference. *arXiv preprint arXiv:1712.02047*.

Jawahar, Ganesh, Benoît Sagot, and Djamé Seddah. 2019. What Does BERT Learn about the Structure of Language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Association for Computational Linguistics, Florence, Italy.

Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-Lingual Ability of Multilingual BERT: An Empirical Study. *arXiv preprint arXiv:1912.07840*.

Kondratyuk, Dan and Milan Straka. 2019. 75 Languages, 1 Model: Parsing Universal Dependencies Universally. *arXiv preprint arXiv:1904.02099*.

Kovaleva, Olga, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the Dark Secrets of BERT. *arXiv preprint arXiv:1908.08593*.

Lin, Yongjie, Yi Chern Tan, and Robert Frank. 2019. Open Sesame: Getting Inside BERT's Linguistic Knowledge. *arXiv preprint arXiv:1906.01698*.

Linzen, Tal, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 arXiv preprint*.

Mareček, David and Rudolf Rosa. 2018. Extracting Syntactic Trees from Transformer Encoder Self-Attentions. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 347–349, Association for Computational Linguistics, Brussels, Belgium.

Otter, Daniel W., Julian R. Medina, and Jugal K. Kalita. 2020. A Survey of the Usages of Deep Learning for Natural Language Processing. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–21.

Peters, Matthew E., Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018. Dissecting Contextual Word Embeddings: Architecture and Representation. *arXiv preprint arXiv:1808.08949*.

Pires, Telmo, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Association for Computational Linguistics, Florence, Italy.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Raganato, Alessandro and Jörg Tiedemann. 2018. An Analysis of Encoder Representations in Transformer-Based Machine Translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 287–297, Association for Computational Linguistics, Brussels, Belgium.

Reif, Emily, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and Measuring the Geometry of BERT. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., pages 8594–8603.

Shi, Xing, Inkit Padhi, and Kevin Knight. 2016. Does String-Based Neural MT Learn Source Syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534, Association for Computational Linguistics, Austin, Texas.

Sun, Chi, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2020. How to Fine-Tune BERT for Text Classification? *arXiv preprint arXiv:1905.05583*.

Sundararaman, Dhanasekar, Vivek Subramanian, Guoyin Wang, Shijing Si, Dinghan Shen, Dong Wang, and Lawrence Carin. 2019. Syntax-Infused Transformer and BERT models for Machine Translation and Natural Language Understanding. *arXiv preprint arXiv:1911.06156*.

Tenney, Ian, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.

Tran, Ke, Arianna Bisazza, and Christof Monz. 2018. The Importance of Being Recurrent for Modeling Hierarchical Structure. *arXiv preprint arXiv:1803.03585*.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv preprint arXiv:1706.03762*.

Vig, Jesse. 2019. A Multiscale Visualization of Attention in the Transformer Model. *arXiv preprint arXiv:1906.05714*.

Vig, Jesse and Yonatan Belinkov. 2019. Analyzing the Structure of Attention in a Transformer Language Model. *arXiv preprint arXiv:1906.04284*.

Voita, Elena, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned. *arXiv preprint arXiv:1905.09418*.

Wu, Shijie and Mark Dredze. 2019. Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT. *arXiv preprint arXiv:1904.09077*.

Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv:1906.08237 arXiv preprint*.