How Do BERT Embeddings Organize Linguistic Knowledge?

Giovanni Puccetti[†◇], **Alessio Miaschi**[⋆◇], **Felice Dell'Orletta**[◇]
[†] Scuola Normale Superiore, Pisa
[⋆]Department of Computer Science, University of Pisa
[◇]Istituto di Linguistica Computazionale "Antonio Zampolli", Pisa
ItaliaNLP Lab – *www.italianlp.it*
giovanni.puccetti@sns.it, alessio.miaschi@phd.unipi.it,
felice.dellorletta@ilc.cnr.it

Abstract

Several studies investigated the linguistic information implicitly encoded in Neural Language Models. Most of these works focused on quantifying the amount and type of information available within their internal representations and across their layers. In line with this scenario, we proposed a different study, based on Lasso regression, aimed at understanding how the information encoded by BERT sentence-level representations is arranged within its hidden units. Using a suite of several probing tasks, we showed the existence of a relationship between the implicit knowledge learned by the model and the number of individual units involved in the encodings of this competence. Moreover, we found that it is possible to identify groups of hidden units more relevant for specific linguistic properties.

1 Introduction

The rise of contextualized word representations (Peters et al., 2018; Devlin et al., 2019) has led to significant improvement in several (if not every) NLP tasks. The main drawback of these approaches, despite the outstanding performances, is the lack of interpretability. In fact, high dimensional representations do not allow for any insight of the type of linguistic properties encoded in these models. Therefore this implicit knowledge can only be determined a posteriori, by designing tasks that require a specific linguistic skill to be tackled (Linzen and Baroni, 2020) or by investigating to what extent certain information is encoded within contextualized internal representations, e.g. defining probing classifier trained to predict a variety of language phenomena (Conneau et al., 2018a; Hewitt and Manning, 2019; Tenney et al., 2019a).

In line with this latter approach and with recent works aimed at investigating how the information is arranged within neural models representations (Baan et al., 2019; Dalvi et al., 2019; Lakretz et al., 2019), we proposed an in-depth investigation aimed at understanding how the information encoded by BERT is arranged within its internal representation. In particular, we defined two research questions, aimed at: (i) investigating the relationship between the sentence-level linguistic knowledge encoded in a pre-trained version of BERT and the number of individual units involved in the encoding of such knowledge; (ii) understanding how these sentence-level properties are organized within the internal representations of BERT, identifying groups of units more relevant for specific linguistic tasks. We defined a suite of probing tasks based on a variable selection approach, in order to identify which units in the internal representations of BERT are involved in the encoding of similar linguistic properties. Specifically, we relied on a wide range of linguistic tasks, which resulted to successfully model different typology of sentence complexity (Brunato et al., 2020), from very simple features (such as sentence length) to more complex properties related to the morphosyntactic and syntactic structure of a sentence (such as the distribution of specific dependency relations).

The paper is organized as follows. In Sec. 2 we present related work, then we describe our approach (Sec. 3), with a focus on the model and the data used for the experiments (Sec. 3.1) and the set of probing tasks (Sec. 3.2). Experiments and results are discussed in Sec. 4 and 5. To conclude, we summarize the main findings of our work in Sec. 6.

2 Related work

In the last few years, a number of recent works have explored the inner mechanism and the linguistic knowledge implicitly encoded in Neural Language Models (NLMs) (Belinkov and Glass, 2019). The most common approach is based on

Proceedings of Deep Learning Inside Out (DeeLIO):
The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, pages 48–57

the development of *probes*, i.e. supervised models trained to predict simple linguistic properties using the contextual word/sentence embeddings of a pre-trained model as training features (Conneau et al., 2018b; Zhang and Bowman, 2018; Miaschi et al., 2020). These latter studies demonstrated that NLMs are able to encode a wide range of linguistic information in a hierarchical manner (Blevins et al., 2018; Jawahar et al., 2019; Tenney et al., 2019b) and even to support the extraction of dependency parse trees (Hewitt and Manning, 2019). For instance, Liu et al. (2019) quantified differences in the transferability of individual layers between different models, showing that higher layers of RNNs (ELMo) are more task-specific (less general), while transformer layers (BERT) do not exhibit this increase in task-specificity.

Other works also investigated the importance of individual neurons within models representations (Qian et al., 2016; Bau et al., 2019; Baan et al., 2019). Dalvi et al. (2019) proposed two methods, *Linguistic Correlations Analysis* and *Cross-model correlation analysis*, to study whether specific dimensions learned by end-to-end neural models are responsible for specific properties. For instance, they showed that open class categories such as verbs and location are much more distributed across the network compared to closed class categories (e.g. coordinating conjunction) and also that the model recognizes a hierarchy of linguistic proprieties and distributes neurons based on it. Lakretz et al. (2019), instead, proposed a detailed study of the inner mechanism of number tracking in LSTMs at single neuron level, showing that long distance number information (from the subject to the verb) is largely managed by two specific units.

Differently from those latter work, our aim was to combine previous approaches based on probes and on the study on individual units in order to propose an in-depth investigation on the organization of linguistic competence within NLM contextualized representations.

3 Approach

To study how the information used by BERT to implicitly encode linguistic properties is arranged within its internal representations, we relied on a variable selection approach based on Lasso regression (Tibshirani, 1996), which aims at keeping as few non-zero coefficients as possible when solving specific regression tasks. Our aim was to identify which weights within sentence-level BERT internal representations can be set to zero, in order to understand the relationship between hidden units and linguistic competence and whether the information needed to perform similar linguistic tasks is encoded in similar positions. We relied on a suite of 68 sentence-level probing tasks, each of which corresponds to a specific linguistic feature capturing characteristics of a sentence at different levels of granularity. In particular, we defined a Lasso regression model that takes as input layer-wise BERT representations for each sentence of a gold standard Universal Dependencies (UD) (Nivre et al., 2016) English dataset and predicts the actual value of a given sentence-level feature. Lasso regression consists in adding an L_1 penalization to the usual ordinary least square loss. To do so, one of the most relevant parameters is λ, which tunes how relevant the L_1 penalization is for the loss function. We performed a grid search with cross validation for each feature-layer pair, in order to identify the best suited value for λ according to each task. Specifically, our goal was to find the most suited value for seeking the best performance when having as few non-zero coefficients as possible.

3.1 Model and data

We used a pre-trained version of BERT (BERT-base uncased, 12 layers). In order to obtain the representations for our sentence-level tasks we experimented with the activation of the first input token (*[CLS]*) and the mean of all the word embeddings for each sentence (*Mean-pooling*).

With regard to the data used for the regression experiments, we relied on the Universal Dependencies (UD) English dataset. The dataset includes three UD English treebanks: UD_English-ParTUT, a conversion of a multilingual parallel treebank consisting of a variety of text genres, including talks, legal texts and Wikipedia articles (Sanguinetti and Bosco, 2015); the Universal Dependencies version annotation from the GUM corpus (Zeldes, 2017); the English Web Treebank (EWT), a gold standard universal dependencies corpus for English (Silveira et al., 2014). Overall, the final dataset consists of 23,943 sentences.

3.2 Linguistic features

As already mentioned, we defined a suite of probing tasks relying on a wide set of sentence-level linguistic features automatically extracted from the parsed sentences in the UD dataset. The set of

Level of Annotation	Linguistic Feature	Label
	Raw Text Properties	
Raw Text	Sentence Length	sent_length
	Word Length	char_per_tok
	Vocabulary Richness	
Vocabulary	Type/Token Ratio for words and lemmas	ttr_form, ttr_lemma
	Morphosyntactic information	
	Distibution of UD and language–specific POS	upos_dist_*, xpos_dist_*
POS tagging	Lexical density	lexical_density
	Inflectional morphology	
	Inflectional morphology of lexical verbs and auxiliaries	xpos_VB-VBD-VBP-VBZ, aux_*
	Verbal Predicate Structure	
	Distribution of verbal heads and verbal roots	verbal_head_dist, verbal_root_perc
	Verb arity and distribution of verbs by arity	avg_verb_edges, verbal_arity_*
	Global and Local Parsed Tree Structures	
	Depth of the whole syntactic tree	parse_depth
	Average length of dependency links and of the longest link	avg_links_len, max_links_len
	Average length of prepositional chains and distribution by depth	avg_prep_chain_len, prep_dist_*
Dependency Parsing	Clause length	avg_token_per_clause
	Order of elements	
	Order of subject and object	subj_pre, obj_post
	Syntactic Relations	
	Distribution of dependency relations	dep_dist_*
	Use of Subordination	
	Distribution of subordinate and principal clauses	principal_prop_dist, subordinate_prop_dist
	Average length of subordination chains and distribution by depth	avg_subord_chain_len, subordinate_dist_1
	Relative order of subordinate clauses	subordinate_post

Table 1: Linguistic Features used in the experiments.

features is based on the ones described in Brunato et al. (2020) which are acquired from raw, morpho-syntactic and syntactic levels of annotation and can be categorised in 9 groups corresponding to different linguistic phenomena. As shown in Table 1, these features model linguistic phenomena ranging from raw text one, to morpho–syntactic information and inflectional properties of verbs, to more complex aspects of sentence structure modeling global and local properties of the whole parsed tree and of specific subtrees, such as the order of subjects and objects with respect to the verb, the distribution of UD syntactic relations, also including features referring to the use of subordination and to the structure of verbal predicates.

4 Linguistic competence and BERT units

As a first analysis, we investigated the relationship between the implicit linguistic properties encoded in the internal representations of BERT and the number of individual units involved in the encoding of these properties. Figure 1 and 2 report layerwise R^2 results for all the probing tasks along with the number of non-zero coefficients obtained with the sentence representations computed with the *[CLS]* token and the *Mean-pooling* strategy respectively. As a first remark, we can notice that the *Mean-pooling* method proved to be the best one for almost all the probing features across the 12 layers. Moreover, in line with Hewitt and Manning (2019), we noticed that there is high variability among different tasks, whereas less variation occurs among the model layers. In general, we observe that best scores are related to features belonging to raw text and vocabulary proprieties, such as sentence length and Type/Token Ratio. Nevertheless, BERT representations implicitly encode information also related to more complex syntactic features, such as the order of the subject (*subj_pre*) or the distribution of several dependency relations (e.g. *dep_dist_det*, *dep_dist_punct*). Interestingly, the knowledge about POS differs when we consider more granular distinctions. For instance, within the broad categories of verbs and nouns, worse predictions were obtained by sub-specific classes of verbs based on tense, person and mood features (see especially past participle, *xpos_dist_VBN*). Similarly, within the verb predicate structure properties, we observe that lower R^2 scores were obtained by features related to sub-categorization information about verbal predicates, such as the distribution of verbs by arity (*verbal_arity_**).

Focusing instead on the relationship between R^2 scores and number of non-zero coefficients, we can notice that although best scores are achieved at lower layers (between layers 12 and 8 for both configurations), the highest number of non-zero coefficients occurs instead at layers closer to the output. This is particularly evident for the results achieved using the *[CLS]* token, for which we observe a continuous increase across the 12 layers in the number of units used by the the probing models.

For both configurations, features more related to the structure of the whole syntactic tree are

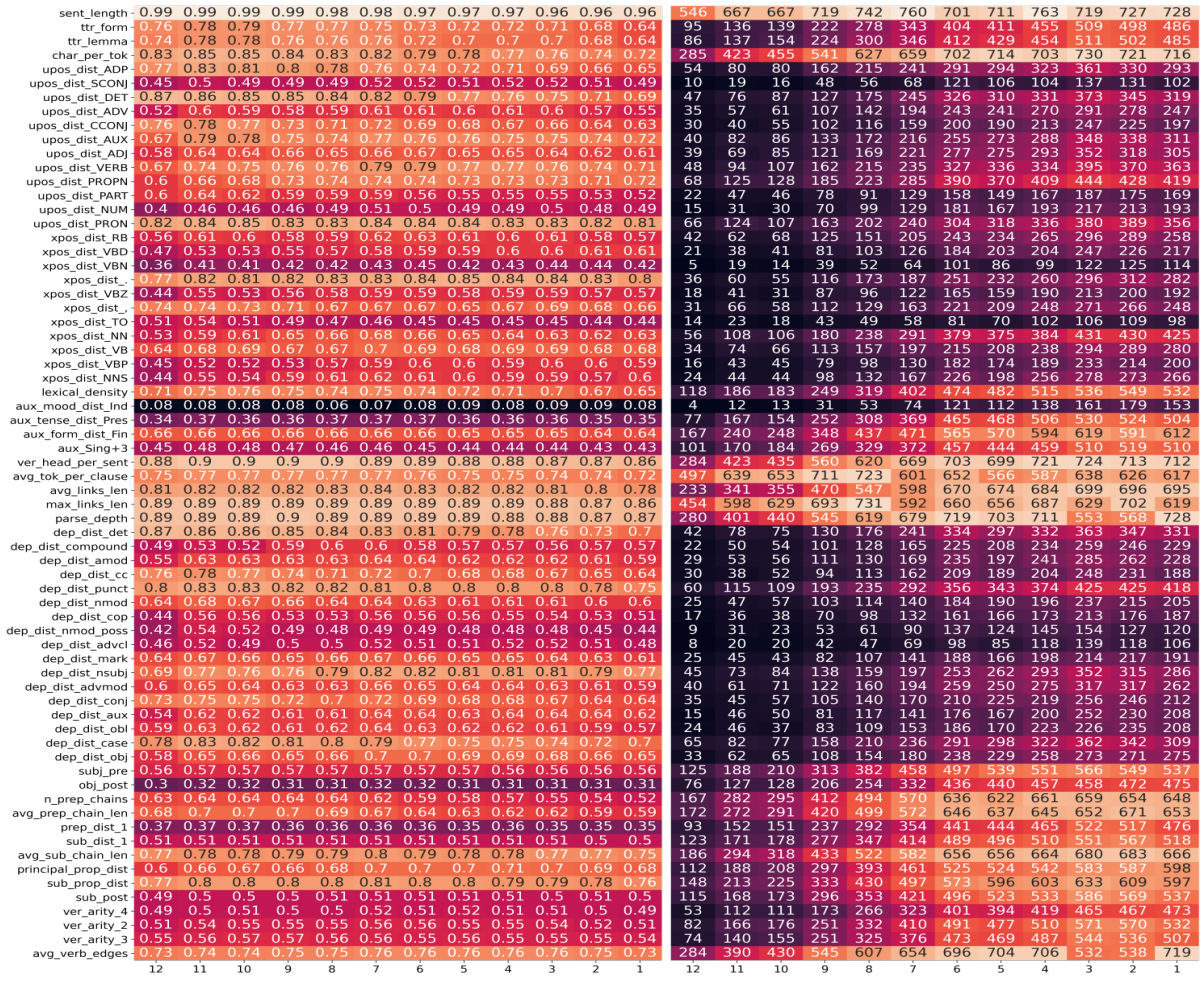


Figure 1: Layerwise R^2 results for all the probing tasks (*left heatmap*) along with the number of non-zero coefficients (*right heatmap*) obtained with the sentence representations computed using the *[CLS]* token.

those for which less units were set to zero during regression (e.g. *max_links_len*, *parse_depth*, *n_prepositional_chains*), while properties belonging to word–based properties (i.e. features related to POS and dependency labels) were predicted relying on less units. Moreover, we can clearly notice that features related to specific POS and dependency relationships are also those that gained less units through the 12 layers (e. g. *xpos_dist_.*, *xpos_dist_AUX*). On the contrary, features belonging to the structure of the syntactic tree tend to acquire more non-zero units as the output layer is approached. This is particularly evident for the linguistic features predicted using sentence representations computed using the *[CLS]* token (e.g. *subj_pre*, *parse_depth*, *n_prepositional_chains*). We believe this is due to the fact that the interdependence between different units in each representation tend to increase across layers, thus making the information less localized especially for those features that belong to the whole structure of the syntactic tree. This is coherent with the fact that using the *Mean-pooling* strategy a higher number of non-zero coefficients was preserved also in the very first input layers, suggesting that this strategy increases the interdependence between each unit and makes the extraction of localized information more complex.

In order to focus more closely on the relationship between R^2 scores and non-zero units, we reported in Figures 3a and 3b average R^2 scores versus average number of non-zero coefficients, along with the line of best fit, for each layer and according to the *[CLS]* token and to the *Mean-pooling* strategy respectively. Interestingly, for both *[CLS]* and *Mean-pooling* representations, R^2 scores tend to improve as the number of non-zero coefficients increases. Moreover, when considering sentence representations computed with the *[CLS]* token, this behaviour becomes more pronounced as the output

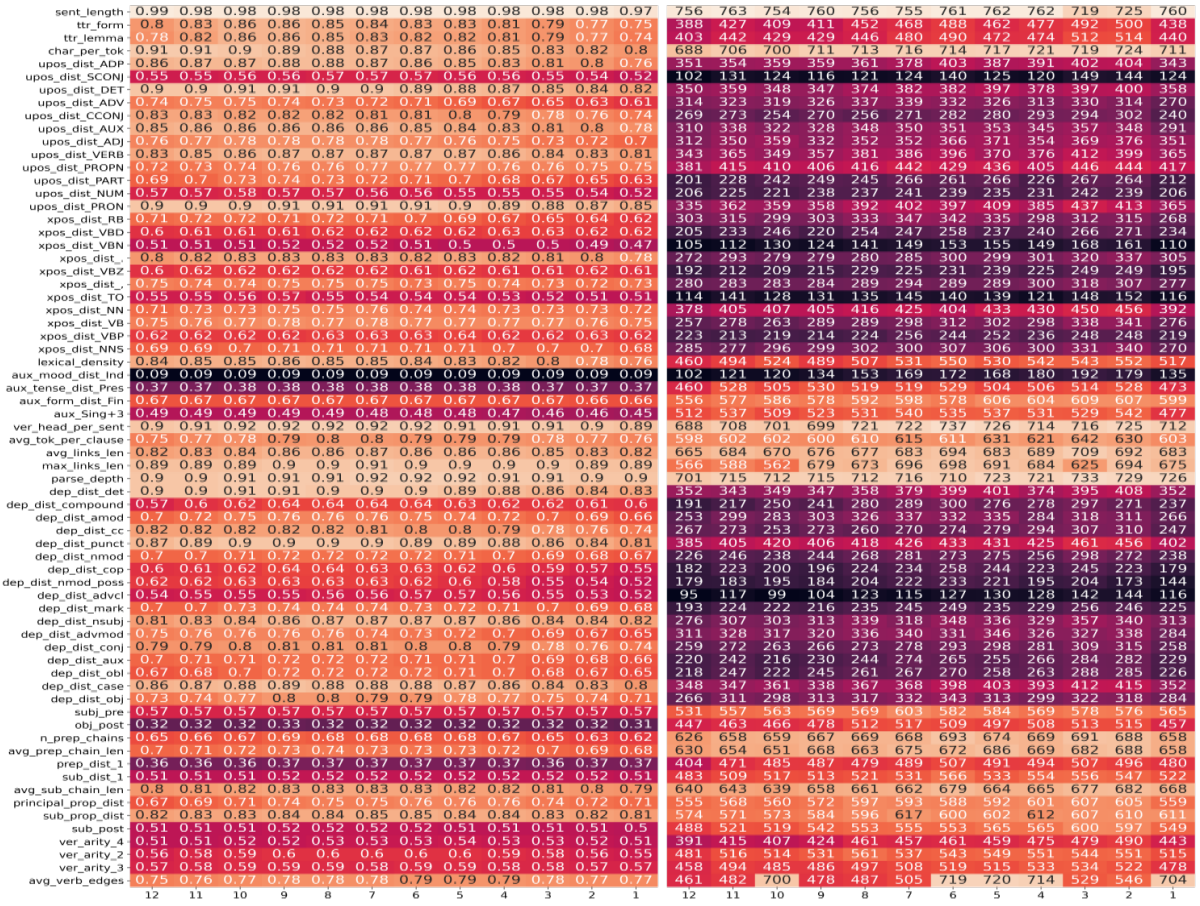
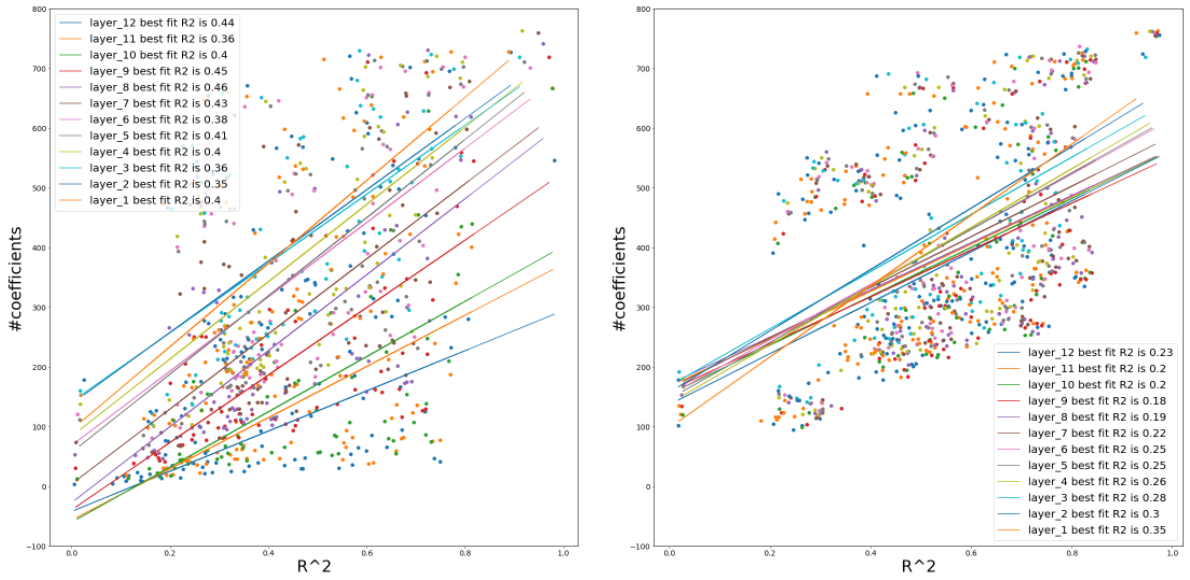


Figure 2: Layerwise R^2 results for all the probing tasks (*left heatmap*) along with the number of non-zero coefficients (*right heatmap*) obtained with the sentence representations computed with the *Mean-pooling* strategy.

layer is reached. This is in line with what we already noticed, namely that the interdependence between different units tend to increase across layers, especially when taking into account representations extracted without using a mean-pooling strategy.

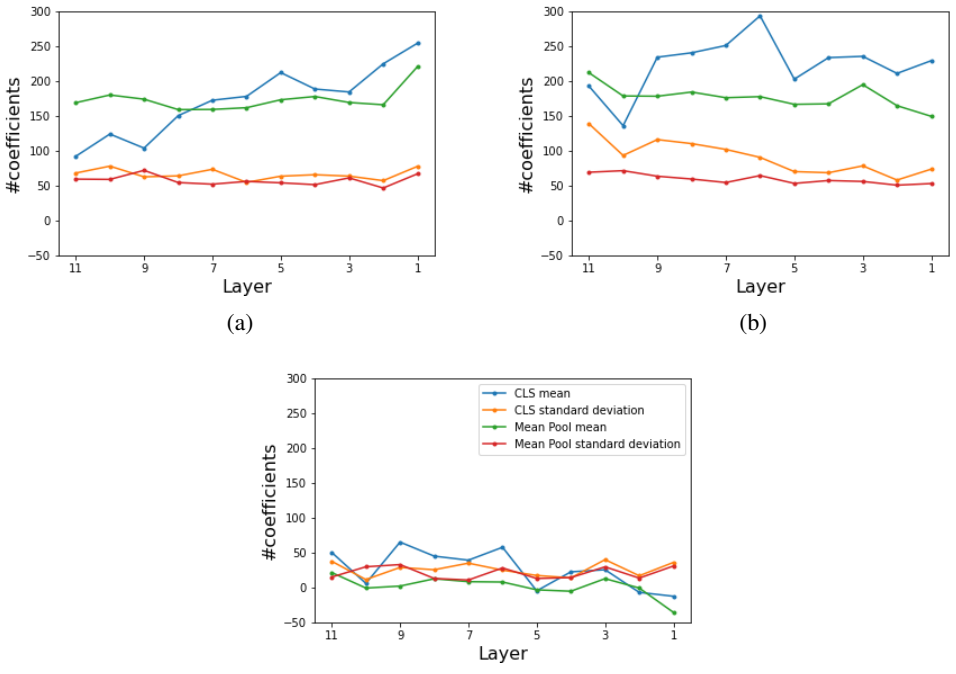
In order to investigate more in depth the behaviour of BERT hidden units when solving the probing tasks, we focused more closely at how the different units in the internal representations are kept and lost across subsequent layers. Figure 4 reports the average number of non-zero coefficients in a layer that are set to zero in the following one (4a), the average number of zero coefficients in a layer that are set to non-zero in the following one (4b) and the average value of the difference between the number of non-zero coefficients at pairs of consecutive layers (4c). As it can be observed, there is high coherence between each layer and its subsequent one, meaning that the variation in the number of selected coefficient is stable (4c). However, the first two plots also show that there is a higher variation when considering non-zero coefficients in the same positions between pairs of layers. This underlines the fact that the information is not localized within BERT's internal representations, since the algorithm shows a degree of freedom in which units can be zeroed and which cannot.

In Figure 5 we report instead how many times each individual unit in the *[CLS]* (5a) and *Mean-pooling* (5b) internal representations has been kept non-zero when solving the 68 probing tasks for all the 12 BERT layers (816 regression task). In general, we can observe that the regression tasks performed using sentence-level representations obtained with the *Mean-pooling* strategy tend to use more hidden units with respect to the *[CLS]* ones. It is also interesting to notice that there is a highly irregular unit (number 308) that has been kept different from zero in a number of tasks and layers much higher than the average. This could suggest that this unit is particularly relevant for encoding almost all the linguistic properties devised in our probing tasks.



(a) (b)

Figure 3: Average R^2 scores versus average number of non-zero coefficients, along with the line of best fit, for each layer and according to *[CLS]* (a) and *Mean-pooling* (b) strategy.



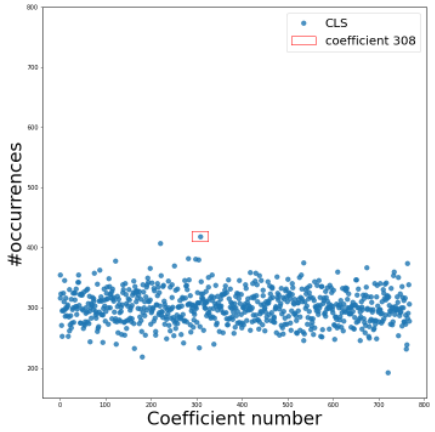
(a) (b)

(c)

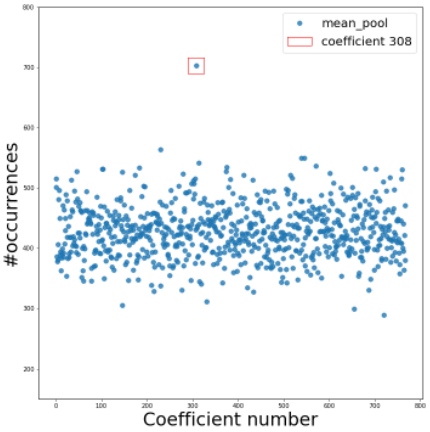
Figure 4: In (a) the average number of non-zero coefficients in a layer that are set to zero in the following one (*average number of dropped coefficients*), in (b) the average number of zero coefficients in a layer that are set to non-zero in the following one (*average number of gained coefficients*) and in (c) the value of the difference between the number of non-zero coefficients at pairs of consecutive layers (*average number of changed coefficients*).

5 Is information linguistically arranged within BERT representations?

Once we have investigated the relationship between the linguistic knowledge implicitly encoded by BERT and the number of individual units involved in it, we verified whether we can identify groups of units particularly relevant for specific probing tasks. To this end, we clustered the 68 probing features according to the weights assigned by the regression



(a)



(b)

Figure 5: Number of times in which each BERT individual unit (computed with *[CLS]* token in (a) and with *Mean-pooing* aggregation strategy in (b)) has been kept as non-zero when solving all the probing tasks for all the 12 layers.

models to each BERT hidden unit. Specifically, we perform hierarchical clustering using correlation distance as distance metric. Figure 6 and 7 report the hierarchical clustering obtained with the *[CLS]* and *Mean-pooling* internal representations at layers 12, 8 and 1. We chose layers 12 and 1 in order to study differences of the clustering of linguistic features taking into account the representations that were more distant and more closer to the language modeling task respectively, while layer 8 was chosen since it was the layer after which BERT's representations tend to lose their precision in encoding our set of linguistic properties.

As a general remark, we can notice that, despite some variations, the linguistic features are organized in a similar manner across the tree layers and for both the configuration. This is to say that, despite the number of non-zero coefficients varies significantly between layers and according to the strategy for extracting the internal representations, the way in which linguistic properties are arranged within BERT embeddings is quite consistent. This suggests that there is a coherent organization of linguistic features according to non-zero coefficients that is independent from the layer and the aggregation techniques taken into account.

Focusing on specific groups of features, we observe that, even if the traditional division with respect to the linguistic annotation levels (see Table 1) has not been completely maintained, it is possible to identify different clusters of features referable to the same linguistic phenomena for all the 3 layers taken into account and for both configurations. In particular, we can clearly observe groups of features related to the length of dependency links and prepositional chains (e.g. *max_links_len*, *avg_links_len*, *n_prepositional_chains*), to vocabulary richness (*ttr_form*, *ttr_lemma*), to properties related to verbal predicate structure and inflectional morphology of auxiliaries (e.g. *xpos_dist_VBD*, *xpos_dist_VBN* *aux_form_dist_Fin*, *aux_tense_dist_pres*) and to the use of punctuation (*xpos_dist_.*, *xpos_dist_,*, *dep_dist_punct*) and subordination (e.g. *subordinate_dist_1*, *subordinate_post*). Interestingly enough, BERT representations also tend to put together features related to each other but not necessarily belonging to the same linguistic macrocategory. This is the case, for instance, of characteristics corresponding to functional properties (e.g. *upos_dist_ADP*, *dep_dist_det*).

6 Conclusions

In this paper we proposed an in-depth investigation aimed at understanding how BERT embeddings encode and organize linguistic competence. Relying on a variable selection approach applied on a suite of 68 probing tasks, we showed the existence of a relationship between the implicit linguistic knowledge encoded by the NLM and the number of individual units involved in the encoding of this knowledge. We found that, according to the strategy for obtaining sentence-level representations, the amount of hidden units devised to encode linguistic properties varies differently across BERT layers: while the number of non-zero units used in the *Mean-pooling* strategy remains more or less constant across layers, the *[CLS]* representations show a continuous increase in the number of

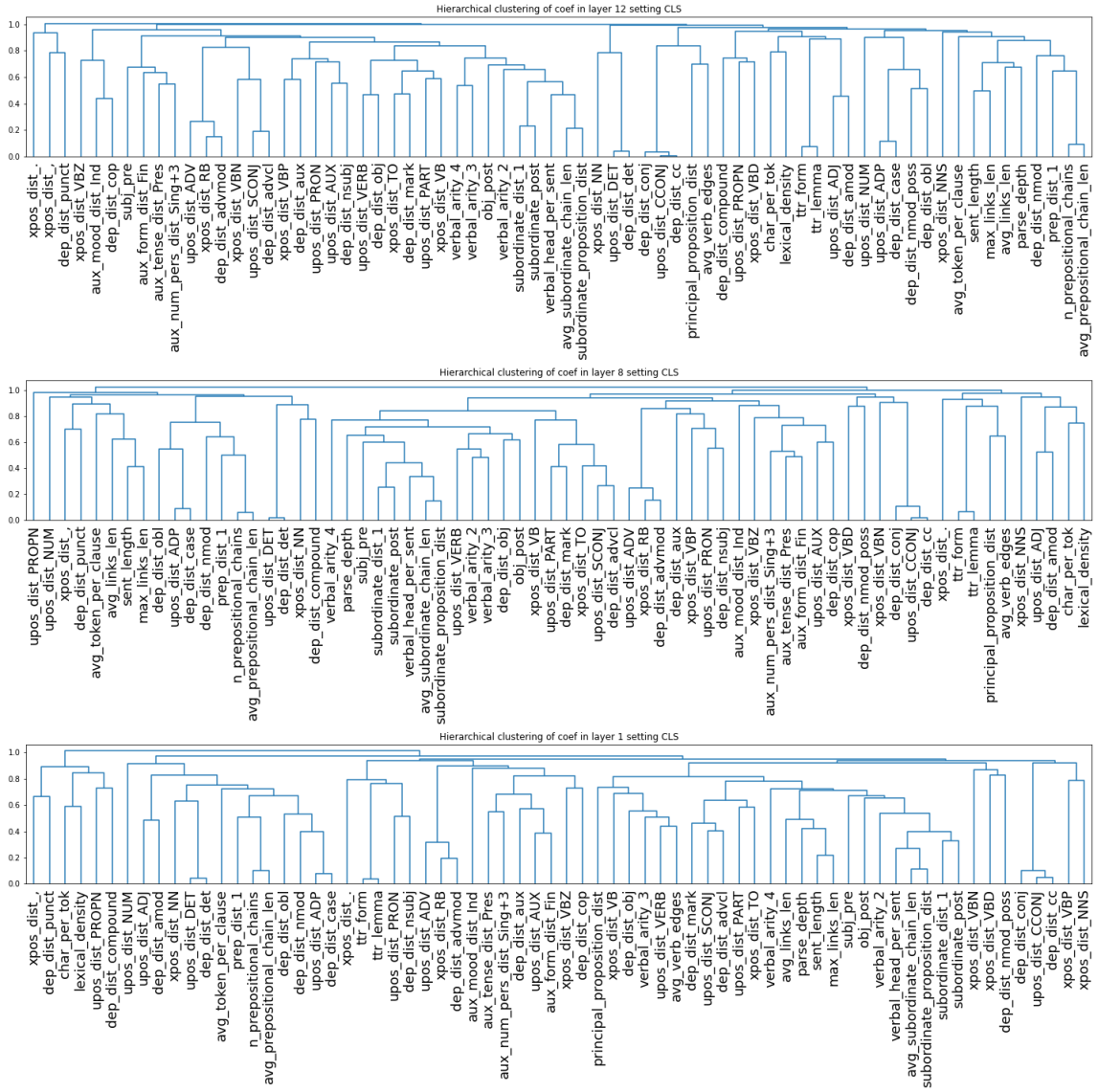


Figure 6: From top to bottom, the hierarchical clustering for the *[CLS]* setting of all the tasks respectively at layers 12, 8 and 1.

used coefficients. Moreover, we noticed that this behaviour is particularly significant for linguistic properties related to the whole structure of the syntactic tree, while features belonging to POS and dependency tags tend to acquire less non-zero units across layers.

Finally, we found that it is possible to identify groups of units more relevant for specific linguistic tasks. In particular, we showed that clustering our set of sentence-level properties according to the weights assigned by the regression models to each BERT unit we can identify clusters of features referable to the same linguistic phenomena and this, despite some variations, is true for both the configurations and for all the BERT layers.

References

Joris Baan, Jana Leible, Mitja Nikolaus, David Rau, Dennis Ulmer, Tim Baumgärtner, Dieuwke Hupkes, and Elia Bruni. 2019. On the realization of compositionality in neural networks. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 127–137, Florence, Italy. Association for Computational Linguistics.

Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2019. Identifying and controlling important neurons in neural machine translation. In *International Conference on Learning Representations*.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey.

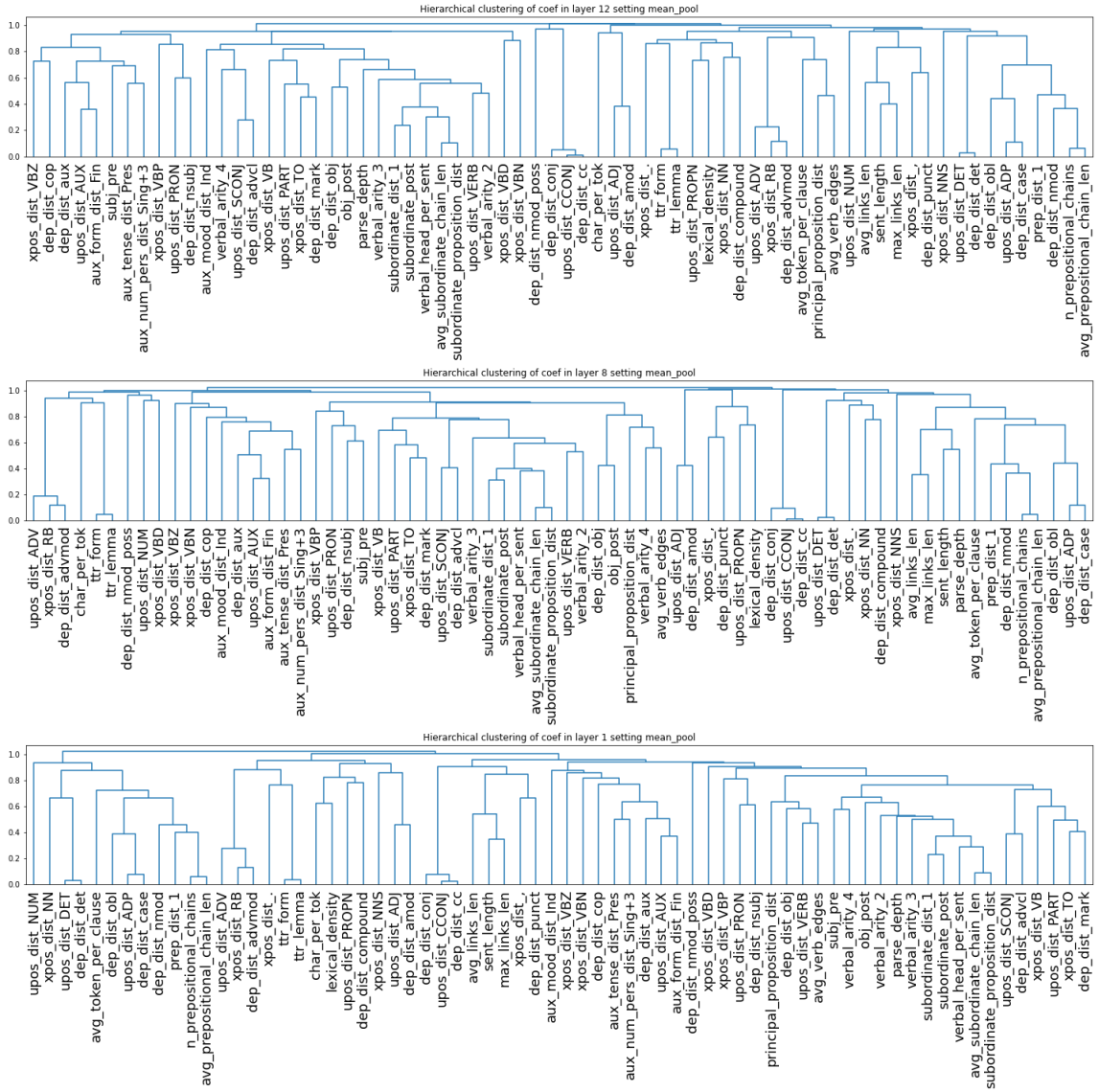


Figure 7: From top to bottom, the hierarchical clustering for the *Mean-pooling* setting of all the tasks respectively at layers 12, 8 and 1.

Transactions of the Association for Computational Linguistics, 7:49–72.

Terra Blevins, Omer Levy, and Luke Zettlemoyer. 2018. Deep rnns encode soft hierarchical syntax. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 14–19.

Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2020. Profiling-ud: a tool for linguistic profiling of texts. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7147–7153, Marseille, France. European Language Resources Association.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018a. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018b. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James Glass. 2019. What is one grain of sand in the desert? analyzing individual neurons in deep nlp models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6309–6317.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186.

John Hewitt 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.

Ganesh Jawahar, 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, Florence, Italy. Association for Computational Linguistics.

Yair Lakretz, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. 2019. The emergence of number and syntax units in LSTM language models. 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 11–20, Minneapolis, Minnesota. Association for Computational Linguistics.

Tal Linzen and Marco Baroni. 2020. Syntactic structure from deep learning. *CoRR*, abs/2004.10827.

Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew E Peters, and Noah A Smith. 2019. Linguistic knowledge and transferability of contextual 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 1073–1094.

Alessio Miaschi, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2020. Linguistic profiling of a neural language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 745–756, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Peng Qian, Xipeng Qiu, and Xuanjing Huang. 2016. Analyzing linguistic knowledge in sequential model of sentence. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 826–835, Austin, Texas. Association for Computational Linguistics.

Manuela Sanguinetti and Cristina Bosco. 2015. Parttut: The turin university parallel treebank. In *Harmonization and Development of Resources and Tools for Italian Natural Language Processing within the PARLI Project*, pages 51–69. Springer.

Natalia Silveira, Timothy Dozat, Marie-Catherine De Marneffe, Samuel R Bowman, Miriam Connor, John Bauer, and Christopher D Manning. 2014. A gold standard dependency corpus for english. In *LREC*, pages 2897–2904.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

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

Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Kelly Zhang and Samuel Bowman. 2018. Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 359–361.