Contextual and Non-Contextual Word Embeddings: an in-depth Linguistic Investigation

Alessio Miaschi[⋆ ⋄], **Felice Dell'Orletta**[⋄]
[⋆]Department of Computer Science, University of Pisa
[⋄]ItaliaNLP Lab, Istituto di Linguistica Computazionale "Antonio Zampolli", Pisa
alessio.miaschi@phd.unipi.it, felice.dellorletta@ilc.cnr.it

Abstract

In this paper we present a comparison between the linguistic knowledge encoded in the internal representations of a contextual Language Model (BERT) and a contextual-independent one (Word2vec). We use a wide set of probing tasks, each of which corresponds to a distinct sentence-level feature extracted from different levels of linguistic annotation. We show that, although BERT is capable of understanding the full context of each word in an input sequence, the implicit knowledge encoded in its aggregated sentence representations is still comparable to that of a contextual-independent model. We also find that BERT is able to encode sentence-level properties even within single-word embeddings, obtaining comparable or even superior results than those obtained with sentence representations.

1 Introduction

Distributional word representations (Mikolov et al., 2013) trained on large-scale corpora have rapidly become one of the most prominent component in modern NLP systems. In this context, the recent development of context-dependent embeddings (Peters et al., 2018; Devlin et al., 2019) has shown that such representations are able to achieve state-of-the-art performance in many complex NLP tasks.

However, the introduction of such models made the interpretation of the syntactic and semantic properties learned by their inner representations more complex. Recent studies have begun to study these models in order to understand whether they encode linguistic phenomena even without being explicitly designed to learn such properties (Marvin and Linzen, 2018; Goldberg, 2019; Warstadt et al., 2019). Much of this work focused on the definition of *probing models* trained to predict simple linguistic properties from unsupervised representations. In particular, those work provided evidences that contextualized Neural Language Models (NLMs) are able to capture a wide range of linguistic phenomena (Adi et al., 2016; Perone et al., 2018; Tenney et al., 2019b) and even to organize this information in a hierarchical manner (Belinkov et al., 2017; Lin et al., 2019; Jawahar et al., 2019). Despite this, less study focused on the analysis and the comparison of contextual and non-contextual NLMs according to their ability to encode implicit linguistic properties in their representations.

In this paper we perform a large number of probing experiments to analyze and compare the implicit knowledge stored by a contextual and a non-contextual model within their inner representations. In particular, we define two research questions, aimed at understanding: (i) which is the best method for combining BERT and Word2vec word representations into sentence embeddings and how they differently encode properties related to the linguistic structure of a sentence; (ii) whether such sentence-level knowledge is preserved within BERT single-word representations.

To answer our questions, we rely on a large suite of probing tasks, each of which codifies a particular propriety of a sentence, from very shallow features (such as sentence length and average number of characters per token) to more complex aspects of morphosyntactic and syntactic structure (such as the depth of the whole syntactic tree), thus making them as suitable to assess the implicit knowledge encoded by a NLM at a deep level of granularity.

The remainder of the paper is organized as follows. First we present related work (Sec. 2), then, after briefly presenting our approach (Sec. 3), we describe in more details the data (Sec. 3.1), our set of probing features (Sec. 3.2) and the models used for the experiments (Sec. 3.3). Experiments and results are described in Sec. 4 and 5. To conclude, in Sec. 6 we summarize the main findings of the study.

Proceedings of the 5th Workshop on Representation Learning for NLP (RepL4NLP-2020), pages 110–119
July 9, 2020.

Contributions In this paper: (i) we perform an in-depth study aimed at understanding the linguistic knowledge encoded in a contextual (BERT) and a contextual-independent (Word2vec) Neural Language Model; (ii) we evaluate the best method for obtaining sentence-level representations from BERT and Word2vec according to a wide spectrum of probing tasks; (iii) we compare the results obtained by BERT and Word2vec according to the different combining methods; (iv) we study whether BERT is able to encode sentence-level properties within its single word representations.

2 Related Work

In the last few years, several methods have been devised to open the black box and understand the linguistic information encoded in NLMs (Belinkov and Glass, 2019). They range from techniques to examine the activations of individual neurons (Karpathy et al., 2015; Li et al., 2016; Kádár et al., 2017) to more domain specific approaches, such as interpreting attention mechanisms (Raganato and Tiedemann, 2018; Kovaleva et al., 2019; Vig and Belinkov, 2019) or designing specific *probing tasks* that a model can solve only if it captures a precise linguistic phenomenon using the *contextual* word/sentence embeddings of a pre-trained model as training features (Conneau et al., 2018; Zhang and Bowman, 2018; Hewitt and Liang, 2019). These latter studies demonstrated that NLMs are able to encode a wide range of linguistic information in a hierarchical manner (Belinkov et al., 2017; Blevins et al., 2018; Tenney et al., 2019b) and even to support the extraction of dependency parse trees (Hewitt and Manning, 2019). Jawahar et al. (2019) investigated the representations learned at different layers of BERT, showing that lower layer representations are usually better for capturing surface features, while embeddings from higher layers are better for syntactic and semantic properties. Using a suite of probing tasks, Tenney et al. (2019a) found that the linguistic knowledge encoded by BERT through its 12/24 layers follows the traditional NLP pipeline: POS tagging, parsing, NER, semantic roles and then coreference. Liu et al. (2019), instead, 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.

Closer to our study, Adi et al. (2016) proposed a method for analyzing and comparing different sentence representations and different dimensions, exploring the effect of the dimensionality on the resulting representations. In particular, they showed that sentence representations based on averaged Word2vec embeddings are particularly effective and encode a wide amount of information regarding sentence length, while LSTM auto-encoders are very effective at capturing word order and word content. Similarly, but focused on the resolution of specific downstream tasks, Shen et al. (2018) compared a Single Word Embedding-based model (SWEM-based) with existing recurrent and convolutional networks using a suite of 17 NLP datasets, demonstrating that simple pooling operations over SWEM-based representations exhibit comparable or even superior performance in the majority of cases considered. On the contrary, Joshi et al. (2019) showed that, in the context of three different classification problems in health informatics, context-based representations are a better choice than word-based representations to create vectors. Focusing instead on the geometry of the representation space, Ethayarajh (2019) first showed that the contextualized word representations of ELMo, BERT and GPT-2 produce more context specific representations in the upper layers and then proposed a method for creating a new type of static embedding that outperforms GloVe and FastText on many benchmarks, by simply taking the first principal component of contextualized representations in lower layers of BERT.

Differently from those latter work, our aim is to investigate the implicit linguistic knowledge encoded in pre-trained contextual and contextual-independent models both at sentence and word levels.

3 Our Approach

We studied how layer-wise internal representations of BERT encode a wide spectrum of linguistic properties and how such implicit knowledge differs from that learned by a context-independent model such as Word2vec. Following the probing task approach as defined in Conneau et al. (2018), we proposed a suite of 68 probing tasks, each of which corresponds to a distinct linguistic feature capturing raw-text, lexical, morpho-syntactic and syntactic characteristics of a sentence. More specifically, we defined two sets of experiments. The

Level of Annotation	Linguistic Feature	Label
Raw Text	Sentence Length	sent_length
	Word Length	char_per_tok
	Type/Token Ratio for words and lemmas	ttr_form, ttr_lemma
POS tagging	Distibution of UD and language–specific POS	upos_dist_*, xpos_dist_*
	Lexical density	lexical_density
	Inflectional morphology of lexical verbs and auxiliaries (Mood, Number, Person, Tense and VerbForm)	verbs_*, aux_*
Dependency Parsing	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_prepositional_chain_len, prep_dist_*
	Clause length (n. tokens/verbal heads)	avg_token_per_clause
	Order of subject and object	subj_pre, obj_post
	Verb arity and distribution of verbs by arity	avg_verb_edges, verbal_arity_*
	Distribution of verbal heads and verbal roots	verbal_head_dist, verbal_root_perc
	Distribution of dependency relations	dep_dist_*
	Distribution of subordinate and principal clauses	principal_proposition_dist, subordinate_proposition_dist
	Average length of subordination chains and distribution by depth	avg_subordinate_chain_len, subordinate_dist_1
	Relative order of subordinate clauses	subordinate_post

Table 1: Linguistic Features used in the probing tasks.

first consists in evaluating which is the best method for generating sentence-level embeddings using BERT and Word2vec single-word representations. In particular, we defined a simple probing model that takes as input layer-wise BERT and Word2vec combined 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 probing feature. Moreover, we compared the results to understand which model performs better according to different levels of linguistic sophistication.

In the second set of experiments, we measured how many sentence-level properties are encoded in single-word representations. To do so, we performed our set of probing tasks using the embeddings extracted from both BERT and Word2vec individual tokens. In particular, we considered the word representations corresponding to the first, last and two internal tokens for each sentence of the UD dataset.

3.1 Data

In order to perform the probing experiments on gold annotated sentences, 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 Probing Features

As previously mentioned, our method is in line with the probing tasks approach defined in Conneau et al. (2018), which aims to capture linguistic information from the representations learned by a NLM. Specifically, in our work, each probing task correspond to predict the value of a specific linguistic feature automatically extracted from the POS tagged and dependency parsed sentences in the English UD dataset. The set of features is based on the ones described in Brunato et al. (2020) and it includes characteristics acquired from raw, morpho-syntactic and syntactic levels of annotation. As described in Brunato et al. (2020), this set of features has been shown to have a highly predictive role when leveraged by traditional learning models on a variety of classification problems, covering different aspects of stylometric and complexity analysis.

As shown in Table 1, these features capture several linguistic phenomena ranging from the average length of words and sentence, to morpho–syntactic information both at the level of POS distribution and about the inflectional properties of verbs. More complex aspects of sentence structure are derived from syntactic annotation and model global and local properties of parsed tree structure, with a focus on subtrees of verbal heads, the order of subjects and objects with respect to the verb, the distribution of UD syntactic relations and features referring to the use of subordination.

3.3 Models

We relied on a pre-trained English version of BERT (BERT-base uncased, 12 layers) for the extraction of the contextual word embeddings. To obtain the representations for our sentence-level tasks we experimented the activation of the first input token (*[CLS]*)[1] and four different combining methods: *Max-pooling*, *Min-pooling*, *Mean* and *Sum*. Each of this four combining methods returns a single $\vec{s}$ vector, such that each s_n is obtained by combining the n^{th} components $w_{1n}, w_{2n}, ..., w_{mn}$ of the embedding of each word in the input sentence.

In order to conduct a comparison of context-based and word-based representations when solving our set of probing tasks, we performed all the probing experiments using also the embeddings extracted from a pre-trained version of Word2vec. In particular, we trained the model on the English Wikipedia dataset (dump of March 2020), resulting in 300-dimensional vectors. In the same manner as BERT's contextual representations, we experimented four combining methods: *Max-pooling*, *Min-pooling*, *Mean* and *Sum*.

We used a linear Support Vector Regression model (LinearSVR) as probing model.

4 Evaluating Sentence Representations

The first set of experiments consists in evaluating which is the best method for combining word-level embeddings into sentence representations in order to understand what kind of implicit linguistic properties are encoded within both contextual and non-contextual representations using different combining methods. To do so, we firstly extracted from each sentence in the UD dataset the corresponding word embeddings using the output of the internal representations of Word2vec and BERT layers (from input layer *-12* to output layer *-1*). Secondly, we computed the sentence-representations according to the different combining strategies defined in 3.3. We then performed our set of 68 probing tasks using the LinearSVR model for each sentence representation. Since the majority of our probing features is correlated to sentence length, we compared probing results with the ones obtained with a baseline computed by measuring the ρ coefficient between the length of the UD sentences and each of the 68 probing features.

Categories	BERT	Word2vec	Baseline
Raw text	**0.65**	0.51	0.37
Morphosyntax	0.49	**0.57**	0.28
Syntax	0.55	**0.56**	0.44
All features	0.53	**0.56**	0.38

Table 2: BERT (average between layers) and Word2vec ρ scores computed by averaging *Max-*, *Min-*, *Mean* and *Sum* scores according to the three linguistic levels of annotations and considering all the probing features (*All features*). Baseline scores are also reported.

Categories	Sum	Min	Max	Mean
Raw text	**0.56**	0.51	0.51	0.46
Morphosyntax	0.59	0.52	0.54	**0.61**
Syntax	**0.61**	0.55	0.55	0.54
All features	**0.60**	0.54	0.55	0.57

Table 3: Word2vec probing scores obtained with the four sentence combining methods.

Evaluation was performed with a 5-cross fold validation and using Spearman correlation score (ρ) between predicted and gold labels as evaluation metric.

Table 2 report average ρ scores aggregating all probing results (*All features*) and according to raw text (*Raw text*), morphosyntactic (*Morphosyntax*) and syntactic (*Syntax*) levels of annotations. Scores are computed by averaging *Max-*, *Min-pooling*, *Mean* and *Sum results*. As a general remark, we notice that the scores obtained by Word2vec and BERT's internal representations outperforms the ones obtained with the correlation baseline, thus showing that both models are capable of implicitly encoding a wide spectrum of linguistic phenomena. Interestingly, we can notice that Word2vec sentence representations outperform BERT ones when considering all the probing features in average.

We report in Table 3 and Figure 1 the probing scores obtained by the two models. For what concerns Word2vec representations, we notice that the *Sum* method prove to be the best one for encoding raw text and syntactic features, while mo-

[1] As suggested in Jawahar et al. (2019), the *[CLS]* token somehow summerizes the information encoded in the input sequence.

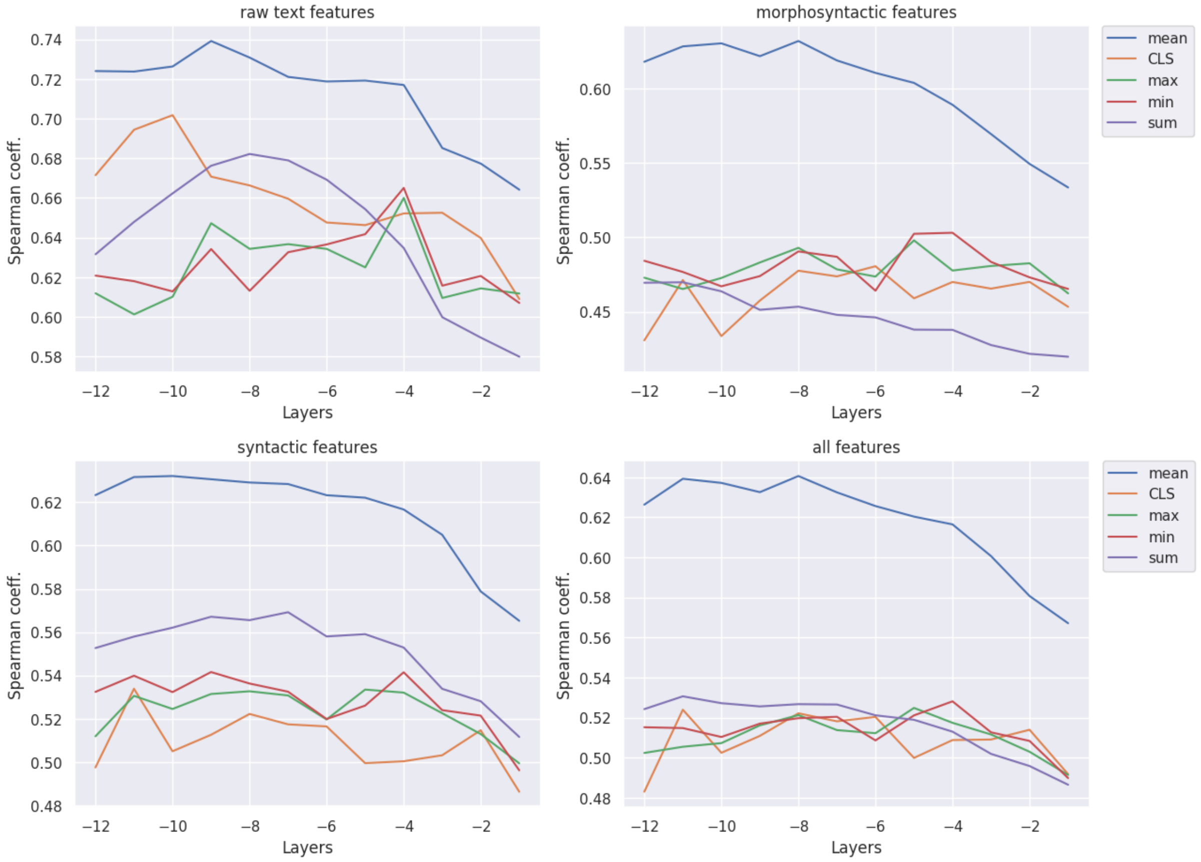

Figure 1: Layerwise ρ scores for the three categories of raw-text, morphosyntactic and syntactic features. Layerwise average results are also reported. Each line in the four plots corresponds to a different aggregating strategy.

rophosyntactic properties are better represented averaging all the word embeddings (*Mean*). In general, best results are obtained with probing tasks related to morphosyntactic and syntactic features, like the distribution of POS (e.g. *upos_dist_PRON*, *upos_dist_VERB*) or the maximum depth of the syntactic tree (*parse_depth*). If we look instead at the average ρ scores obtained with BERT layerwise representations (Figure 1), we observe that, differently from Word2vec, best results are the ones related to raw-text features, such as sentence length or Type/Token Ratio. The *Mean* method prove to be the best one for almost all the probing tasks, achieving highest scores in the first five layers. The only exceptions mainly concern some of the linguistic features related to syntactic properties, e.g. the average length of dependency links (*avg_links_len*) or the maximum depth of the syntactic tree (*parse_depth*), for which best scores across layers are obtained with the *Sum* strategy. The *Max-* and *Min-pooling* methods, instead, show a similar trend for almost all the probing features. Interestingly, the representations corresponding to the *[CLS]* token, although considered as a summarization of the entire input sequence, achieve results comparable to those obtained with *Max-* and *Min-pooling* methods. Moreover, it can be noticed that, unlike *Max-* and *Min-pooling*, the representations computed with *Mean* and *Sum* methods tend to lose their average precision in encoding our set of linguistic properties across the 12 layers.

Layers	Mean	Max-pooling	Min-pooling	Sum
-12	.052	-.058	-.038	-.091
-11	.065	-.055	-.038	-.084
-10	.063	-.053	-.043	-.088
-9	.058	-.044	-.036	-.089
-8	.066	-.039	-.034	-.088
-7	.058	-.046	-.033	-.088
-6	.051	-.048	-.045	-.094
-5	.046	-.035	-.032	-.096
-4	.042	-.043	-.025	-.102
-3	.026	-.049	-.041	-.113
-2	.006	-.057	-.045	-.119
-1	-.007	-.069	-.063	-.128

Table 4: Average ρ differences between BERT and Word2vec probing results according to the four embedding-aggregation strategies.

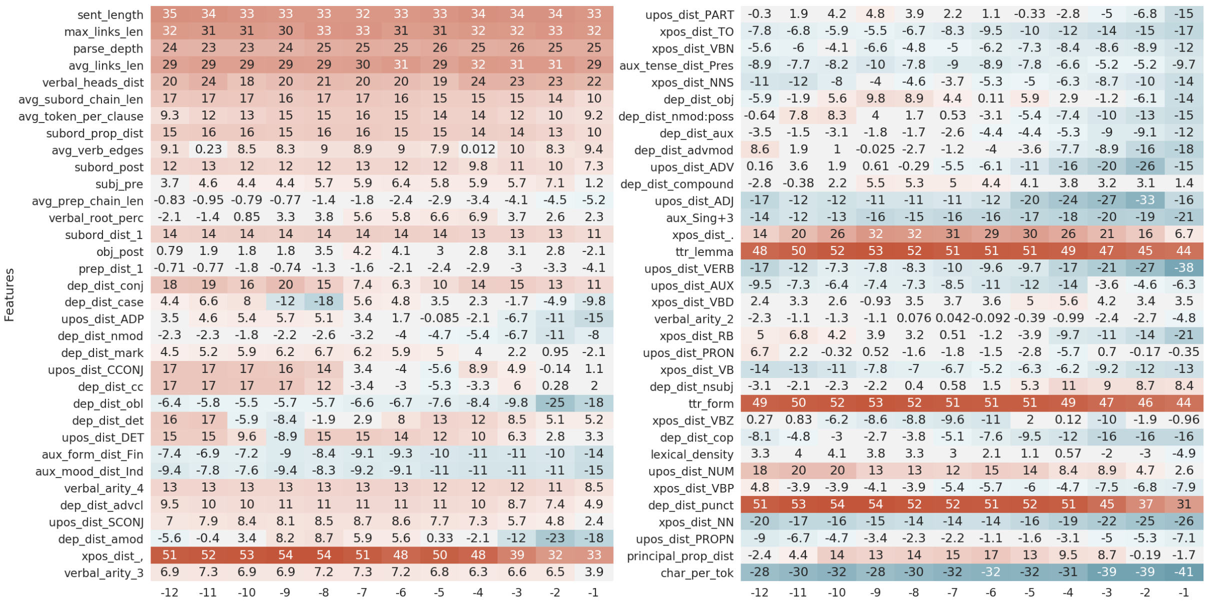

Figure 2: Differences between BERT and Word2vec scores (multiplied by 100) for all the 68 probing features (ranked by correlation with sentence length), obtained with the *Mean* aggregation strategy. BERT scores are reported for all the 12 layers. Positive (*red*) and negative (*blue*) cells correspond to scores for which BERT outperforms Word2vec and vice versa.

In order to investigate more in depth how the linguistic knowledge encoded by BERT across its layers differs from that learned by Word2vec, we report in Table 4 average ρ differences between the two models according to the four combining strategies. As a general remark, we can notice that, regardless of the aggregation strategy taken into account, BERT and Word2vec sentence representations achieve quite similar results on average. Hence, although BERT is capable of understanding the full context of each word in an input sequence, the amount of linguistic knowledge implicitly encoded in its aggregated sentence representations is still comparable to that which can be achieved with a non-contextual language model.

In Figure 2 we report instead the differences between BERT and Word2vec scores for all the 68 probing features (ordered by correlation with sentence length). For the comparison, we used the representations obtained with the *Mean* combining method. As a first remark, we notice that there is a clear distinction in terms of ρ scores between features better predicted by BERT and Word2vec. In fact, features most related to syntactic properties (left heatmap) are those for which BERT results are generally higher with respect to those obtained with Word2vec. This result demonstrates that BERT, unlike a non-contextual language model as Word2vec, is able to encode information within its representations that involves the entire input sequence, thus making more simple to solve probing tasks that refer to syntatic characteristics.

Focusing instead on the right heatmap, we observe that Word2vec non-contextual representations are still capable of encoding a wide spectrum of linguistic properties with higher ρ values compared to BERT ones, especially if we consider scores closer to BERT's output layers (from *-4* to *-1*). This is particularly evident for morphosyntactic features related to the distribution of POS categories (*xpos_dist_**, *upos_dist_**), most likely because non-contextual representations tend to encode properties related to single tokens rather than syntactic relations between them.

5 Evaluating Word Representations

Once we have probed the linguistic knowledge encoded by BERT and Word2vec using different strategies for computing sentence embeddings, we investigated how much information about the structure of a sentence is encoded within single-word contextual representations. For doing so, we performed our sentence-level probing tasks using a single BERT word embedding for each sentence in the UD dataset. We tested four different words, corresponding to the first, the last and two internal tokens for each sentence in the UD dataset. In

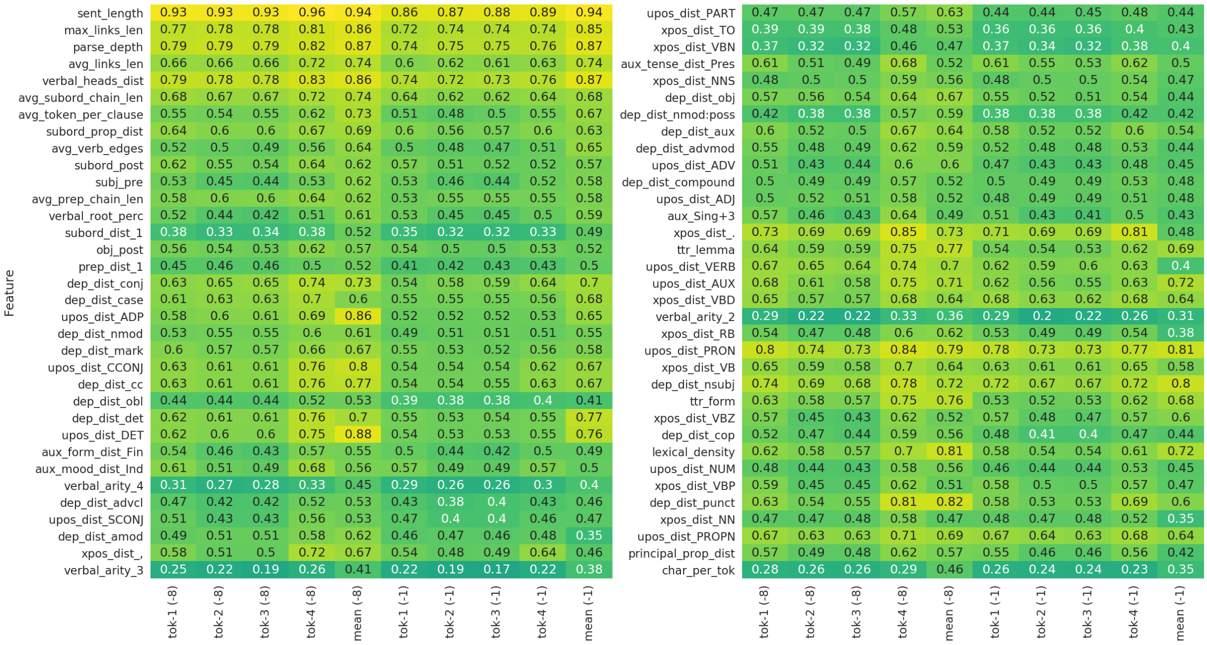

Figure 3: Probing scores obtained by BERT word (*tok_**) and sentence (*mean*) representations extracted from layers *-1* and *-8*. Sentence embeddings are computed using the *Mean* method.

Embeddings	Raw	Morphoyntax	Syntax	All
BERT-1 (-8)	0.62	0.57	0.55	0.57
BERT-2 (-8)	0.59	0.53	0.53	0.53
BERT-3 (-8)	0.59	0.52	0.52	0.53
BERT-4 (-8)	0.65	**0.66**	**0.62**	**0.64**
BERT-1 (-1)	0.55	0.55	0.51	0.53
BERT-2 (-1)	0.54	0.51	0.49	0.50
BERT-3 (-1)	0.54	0.51	0.49	0.50
BERT-4 (-1)	0.59	0.57	0.53	0.55
[CLS] (-8)	**0.66**	0.47	0.52	0.51
[CLS] (-1)	0.61	0.45	0.49	0.48
Word2vec-1	0.26	0.26	0.22	0.24
Word2vec-2	0.17	0.21	0.18	0.19
Word2vec-3	0.17	0.19	0.17	0.18
Word2vec-4	0.13	0.15	0.12	0.13

Table 5: Average ρ scores obtained by BERT and Word2vec according to word representations corresponding to the first, the last and two internal tokens of each input sentence. Results are computed according to the three linguistic levels of annotation and considering all the probing features (*All*). Average scores obtained with the *[CLS]* token are also reported.

particular, we extracted the embeddings from the output layer (*-1*) and from the layer that achieved best results in the previous experiments (*-8*). We used probing scores obtained with Word2vec embeddings for the same tokens as baseline. In Table 5 we report average ρ scores obtained by BERT (*BERT-**) and Word2vec (*Word2vec-**) according to word-level representations extracted from the four tokens mentioned above. Results were computed aggregating all probing results (*All*) and according to raw text (*Raw*), morphosyntactic (*Morphosyntax*) and syntatic (*Syntax*) levels of annotation. For comparison, we also report average scores obtained with the *[CLS]* token.

As a first remark, we can clearly notice that even with a single-word embedding BERT is able to encode a wide spectrum of sentence-level linguistic properties. This result allows us to highlight the main potential of contextual representations, i.e. the capability of capturing linguistic phenomena that refer to the entire input sequence within single-word representations. An interesting observation is that, except for the raw text features, for which the best scores are achieved using *[CLS]*, higher performance are obtained with the embeddings corresponding to *BERT-4*, i.e. the last token of each sentence. This result seems to indicate that *[CLS]*, although being used for classification predictions, does not necessarily correspond to the most linguistically informative token within each input sequence.

Comparing the results with those achieved using Word2vec word embeddings, we notice that BERT scores greatly outperform Word2vec for all the probing tasks. This is a straightforward result and can be easily explained by the fact that the lack of contextual knowledge does not allow single-word representations to encode information that are related to the structure of the whole sentence.

Since the latter results demonstrated that BERT is capable of encoding many sentence-level properties within its single word representations, as a last analysis, we decided to compare these results with the ones obtained using sentence embeddings. In particular, Figure 3 reports probing scores obtained by BERT single word (*tok_**) and *Mean* sentence representations (*sent*) extracted from the output layer (*-1*) and from the layer that achieved best results in average (*-8*).

As already mentioned, for many of these probing tasks, word embeddings performance is comparable to that obtained with the aggregated sentence representations. Nevertheless, there are several cases in which the difference between performance is particularly significant. Interestingly, we can notice that aggregated sentence representations are generally better for predicting properties belonging to the left heatmap, i.e. to the group of features more related to syntactic properties. This is particularly noticeable for the average number of tokens per clause (*avg_token_per_clause*) or the distribution of subordinate chains by length (*subord_dist*), for which we observe an improvement from word-level to sentence-level representations of more than .10 ρ points. On the contrary, probing features belonging to the right heatmap, therefore more close to raw text and morphosyntactic properties, are generally better predicted using single word embeddings, especially when considering the inner representations corresponding to the last token in each sentence (*tok_4*). The property most affected by the difference in scores between word- and sentence-level embeddings is the the distribution of periods (*xpos_dist_.*).

Focusing instead on differences in performance between the two considered layers, we can notice that regardless of the method used to predict each feature, the representations learned by BERT tend to lose their precision in encoding our set of linguistic properties, most likely because the model is storing task-specific information (Masked Language Modeling task) at the expense of its ability to encode general knowledge about the language.

6 Conclusion

In this paper we studied the linguistic knowledge implicitly encoded in the internal representations of a contextual Language Model (BERT) and a contextual-independent one (Word2vec). Using a suite of 68 probing tasks and testing different methods for combining word embeddings into sentence representations, we showed that BERT and Word2vec encode a wide set of sentence-level linguistic properties in a similar manner. Nevertheless, we found that for Word2vec the best method for obtaining sentence representations is the *Sum*, while BERT is more effective when averaging all the single-word representations (*Mean* method). Moreover, we showed that BERT is able in storing features that are mainly related to raw text and syntactic properties, while Word2vec is good at predicting morphosyntactic characteristics.

Finally, we showed that BERT is able to encode sentence-level linguistic phenomena even within single-word embeddings, exhibiting comparable or even superior performance than those obtained with aggregated sentence representations. Moreover, we found that, at least for morphosyntactic and syntactic characteristics, the most informative word representation is the one that correspond to the last token of each input sequence and not, as might be expected, to the *[CLS]* special token.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *arXiv preprint arXiv:1608.04207*.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2017. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10.

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, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. 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.

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.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743.

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, Djamé Seddah, Samuel Unicomb, Gerardo Iñiguez, Márton Karsai, Yannick Léo, Márton Karsai, Carlos Sarraute, Éric Fleury, et al. 2019. What does bert learn about the structure of language? In *57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy*.

Aditya Joshi, Sarvnaz Karimi, Ross Sparks, Cecile Paris, and C Raina MacIntyre. 2019. A comparison of word-based and context-based representations for classification problems in health informatics. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 135–141, Florence, Italy. Association for Computational Linguistics.

Akos Kádár, Grzegorz Chrupała, and Afra Alishahi. 2017. Representation of linguistic form and function in recurrent neural networks. *Computational Linguistics*, 43(4):761–780.

Andrej Karpathy, Justin Johnson, and Li Fei-Fei. 2015. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4365–4374, Hong Kong, China. Association for Computational Linguistics.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in nlp. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy. Association for Computational Linguistics.

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.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

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.

Christian S Perone, Roberto Silveira, and Thomas S Paula. 2018. Evaluation of sentence embeddings in downstream and linguistic probing tasks. *arXiv preprint arXiv:1806.06259*.

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.

Alessandro Raganato 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*. 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.

Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao, and Lawrence Carin. 2018. Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 440–450, Melbourne, Australia. Association for Computational Linguistics.

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*.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Florence, Italy. Association for Computational Linguistics.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, et al. 2019. Investigating bert's knowledge of language: Five analysis methods with npis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2870–2880.

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.