Classifier Probes May Just Learn from Linear Context Features

Jenny Kunz and **Marco Kuhlmann**
Dept. of Computer and Information Science
Linköping University
jenny.kunz@liu.se and marco.kuhlmann@liu.se

Abstract

Classifiers trained on auxiliary probing tasks are a popular tool to analyze the representations learned by neural sentence encoders such as BERT and ELMo. While many authors are aware of the difficulty to distinguish between "extracting the linguistic structure encoded in the representations" and "learning the probing task," the validity of probing methods calls for further research. Using a neighboring word identity prediction task, we show that the token embeddings learned by neural sentence encoders contain a significant amount of information about the exact linear context of the token, and hypothesize that, with such information, learning standard probing tasks may be feasible even without additional linguistic structure. We develop this hypothesis into a framework in which analysis efforts can be scrutinized and argue that, with current models and baselines, conclusions that representations contain linguistic structure are not well-founded. Current probing methodology, such as restricting the classifier's expressiveness or using strong baselines, can help to better estimate the complexity of learning, but not build a foundation for speculations about the nature of the linguistic structure encoded in the learned representations.

1 Introduction

The impressive performance of neural language models such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) has led many authors to believe that these models must have learned linguistic structure, and a whole new research branch concerned with the analysis and interpretation of neural models has emerged, with dedicated areas at many conferences and the BlackboxNLP workshop being established as a large and well-known venue in a short time. An appealing and widespread analysis technique, perhaps due to its simplicity and generalizability, is to use a model's word representations as input to a simple classifier, and train this classifier on an auxiliary task that can be expected to benefit from linguistic knowledge. The argument behind this technique is that, if a simple classifier is enough to learn the task based on word-level representations alone, then it is probable that these representations indeed encode the hypothesized linguistic structure (Hupkes et al., 2018). Many of the papers applying classifier probes suggest that models such as BERT and ELMo encode structure that is similar to well-known syntactic annotations, e.g. dependency trees (Hewitt and Manning, 2019), or even to a traditional NLP pipeline (Tenney et al., 2019a).

A standard assumption in probing is that, if we feed only the representation of one word at a time to the classifier, then no complex inference based on other words in the sentence can happen, and the probe will not be able to learn the classification task well—unless the necessary linguistic information already is encoded in the representation (Lin et al., 2019). Therefore, the argument goes, high performance on a probing task can be taken as evidence that the word-level representations encode linguistic structure. In Section 3, we challenge this view by showing experimentally that the word-level representations learned by ELMo and BERT encode the closer neighborhood of the word in a surprisingly exact way. To explore the consequences of this observation, we formulate the *context-only hypothesis*, which offers an alternative explanation of how a probe could reach high accuracy on the auxiliary task: it could simply make use of a

Proceedings of the 28th International Conference on Computational Linguistics, pages 5136–5146
Barcelona, Spain (Online), December 8-13, 2020

compressed representation of the neighboring words. We propose that any work arguing to have found linguistic structure in word representations should be able to reject this hypothesis.

In Section 4, we examine whether past studies using classifier probes are capable of rejecting our context-only hypothesis, and find that they are not. To argue for the presence of linguistic structure in word embeddings, most of these studies either compare the probe's performance to a baseline, or restrict the probe's expressiveness or the number of training examples. Regarding the first line of argument, we scrutinize proposed baselines and argue that none of them allows us to conclude the presence or absence of linguistic structure in embeddings. We posit that, as long as a model is not fully understood, such baselines probably cannot be found. Regarding the second line, we argue that training restrictions do not provide a reliable setup either, as long as no theoretical ground exists that proves that a certain classifier with a certain set of training data is too weak to learn the probing task. We present experiments showing that the most restricted probes often fail to outperform even uncontextualized baselines. A high variation in very restricted setups also indicates that the results of experiments with restricted training scenarios are hard to interpret, and that their interpretation hence mandates careful testing and error analyses. We speculate that the correct interpretation of such experiments would probably also require full understanding of the model, as well as a clear definition of the distinction between "learning the task" and "extracting linguistic structure", which is lacking for dense embeddings.

We conclude the paper by arguing for an alternative view of the results obtained with classifier probes, namely to regard word representations as containing features that are more or less helpful to learn the probing task.

2 Preliminaries

We start the paper by introducing our notation and general experimental setup.

2.1 Neural Sentence Encoders

A neural sentence encoder can be viewed as a parametrized function R that maps a sentence $s = w_{1:n}$ to a sequence of d-dimensional embedding vectors $\boldsymbol{w}_{1:n}$, one vector $\boldsymbol{w}_i \in \mathbb{R}^d$ for each token w_i in s. We denote this token embedding by $R(w_i)$, leaving the sentence s implicit. We study two neural sentence encoders:

ELMo ELMo (Peters et al., 2018) was arguably the first contextualized word representation that was widely adopted in the NLP community. It implements the function R as a bidirectional LSTM (Hochreiter and Schmidhuber, 1997) consisting of a forward and a backward language model that predict the next (respectively the previous) token conditioned on the LSTM accumulation of the preceding (respectively the future) tokens, jointly maximizing the log likelihood of both directions. The ELMo model that we use for our experiments has a character-based word representation layer with 512 dimensions, and 2 bi-LSTM hidden layers with 1024 units each. The token representation $R(w_i)$ is the weighted sum of the 3 layers. We use the original pre-trained version available at TensorFlow Hub.

BERT BERT (Devlin et al., 2019) is based on a Transformer model's encoder (Vaswani et al., 2017), which contextualizes the word representations with self-attention and fully connected layers. BERT is pre-trained with two objectives: The Masked Language Model randomly replaces some input tokens with a special MASK token, with the objective to predict the vocabulary id of the original token at that position. This approach naturally includes both left and right context. The Next Sentence Prediction objective makes BERT learn the relationship of two sentences by predicting if the second sentence is following the first one in the original document or not. For our experiments, we use the original English BERT-Base model published by Google, with 12 layers and 16 attention heads per layer, accessed via the HuggingFace Transformers library (Wolf et al., 2019).

2.2 Probes and Probing Tasks

By a *classifier probe* or simply *probe*, we mean a classifier that is trained on pairs (x_i, y_i) consisting of representations x_i and labels y_i, with the intention of revealing what linguistic structure is encoded in the representations. Like several other works on probing, we target probing tasks related to syntactic

dependency parsing based on the English Web Treebank data set from the Universal Dependencies (UD) project (Nivre et al., 2018). The probes in our experiments are simple feed-forward networks with one hidden layer consisting of 64 units and a final softmax layer. We train and evaluate on 5 random seeds and shuffles of the data and report the sample mean. We implement all our probes in PyTorch (Paszke et al., 2017) and optimize them with Adam (Kingma and Ba, 2015) with an initial learning rate of 0.001. The code behind the experiments is available at `https://github.com/jekunz/probing`.

3 Extracting Linguistic Structure, One Embedding at a Time

The purpose of a probe is to reveal what linguistic structure is encoded in word representations. There is little doubt that these representations contain useful information; but whether this information takes the form of traditional relational information between words, such as syntactic or semantic structure, is quite a different issue.

3.1 Extracting Structure vs. Learning a Task: A Continuum

In the context of probing, we need to have a methodology that allows us to answer the question whether a word representation encodes *linguistic structure*, or whether it merely encodes *some features* useful for learning the probing task. This question cannot be satisfactorily answered. For dense embeddings, there is likely to be a continuum between two extreme scenarios:

1. The embeddings contain no useful information at all, and the probing task is learned from scratch.
2. The embeddings contain a direct, human-interpretable representation of linguistic information.

Many works in probing motivate their methodology like Lin et al. (2019), who argue that because they feed one embedding at a time, it is not possible for the probe to "compute complex contingencies among the representations of multiple words". We claim that there is no safe ground for this assumption. Neural sentence encoders employ non-local mechanisms such as recurrent neural networks and attention, and are therefore well able to learn representations of words in their sentential context. It seems reasonable to assume then, that a probe with access to representations computed with the help of these mechanisms should, at least in principle, be able to learn any task which can be learned by end-to-end-systems that include the encoders as architectural components. We hypothesize that BERT's masked language model objective in particular may make the learned representations memorize words in the immediate neighborhood, as the states associated with these words are what is used for predictions during pre-training.

3.2 Neighboring Word Identity Probes

To assess the extent to which the sentential context of a token is encoded in pre-trained representations, we use a neighboring word identity prediction task, as first suggested by Zhang and Bowman (2018). In this task, given a single token representation $R(w_i)$ as input, the probe predicts the identity of the word type of the token w_{i+k}, for some fixed offset $k \in \mathbb{Z}$. We interpret the performance on this task as a conservative estimate of the utility of the information encoded in $R(w_i)$ about the exact linear neighborhood of w_i. It is a conservative estimate because, even if a probe is unable to predict the *exact* neighbor of w_i, it may still be able to recover a distributionally similar word, say, a different inflectional form, which for many tasks would be almost as useful as an exact match.

Experimental Setup Using the general setup described in Section 2.2, we train and evaluate seven different classifiers, one classifier for each offset $-3 \leq k \leq +3$. Our data set gives us 43,124 training examples and a vocabulary (set of output labels) of size 8,282. Note that for $k = 0$, the classifier uses the token embedding to predict the identity of its own word type. We interpret the performance on this specific task as an upper bound for the performances on the tasks with $|k| > 1$.

3.2.1 Results: BERT

For BERT, the accuracy on the neighboring word identity task across different layers is shown in Figure 1. Because the scores for all but the immediate neighbors ($|k| = 1$) differ only little between the left context (negative k) and the right context (positive k), we conflate them into one mean score for $\pm k$ to increase the

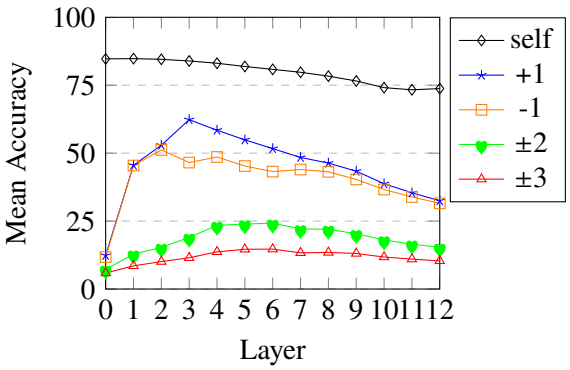

Figure 1: Neighboring word identity probes: Results for BERT

readability of the plot. For the immediate neighbors, there is a performance gap especially in the middle layers, and we therefore report both accuracies separately.

While the prediction accuracy for all words but the center word is only between 6% and 12% in layer 0, it increases in the next layers, suggesting that the word representations available in these layers have memorized configurations of neighboring words in the training set. The accuracy for the next word tops at 62.3% in layer 3, while that for the previous word reaches its maximum in layer 2 with 51.2%. For words two steps steps away, the top result is in layer 6 with 24.2%, and for words three steps away it is 14.7% in layer 5. For higher layers, the performance at all offsets is dropping again. The word itself can be best recovered from layer 0, and accuracy drops steadily until layer 11.

Our results clearly show that the linear context of a word can be recovered from the word representation to a substantial extent, especially when it comes to the word's direct neighbors. When comparing results to the baseline accuracies for layer 0 (representing an uncontextualized language model) and to the accuracies for recovering the word itself (as an upper bound), we see that, although the encoding of the linear context is clearly not a lossless compression, it is still quite informative. Not surprisingly, the context information is the noisier the further away the neighboring words are; but even for words three steps away ($k = \pm 3$) the word representations in higher layers are still clearly more predictive than the representation in the uncontextualized layer—for example, we see 14.7% accuracy in layer 4, compared to 5.9% in layer 0.

Clark et al. (2019) find that one of BERT's attention heads in layer 3 specifically attends to the next token. The results in Figure 1 are consistent with this observation, as in layer 3 we see a clear peak in the accuracies for the prediction of the next token.

3.2.2 Results: ELMo

The results for ELMo can be found in Table 1. ELMo always has the highest performance in layer 1, except for the prediction of the word itself, which is best predicted by the word embedding layer. Layer 2 seems to correspond to BERT's late layers in that the performance drops for all context words.

ELMo's best layer generally underperforms BERT's best layers in predicting neighboring words. These results raise the question of whether BERT's superior performance e.g. in syntax probes may be due not to its better modeling of linguistic structure, but due to its better modeling of the exact context. Especially

	self	−1	+1	±2	±3
BERT (best)	84.786	**51.180**	**62.322**	**24.188**	**14.685**
Embedding	**84.971**	11.294	12.083	7.594	5.634
Layer 1	81.458	43.851	43.699	21.494	12.784
Layer 2	72.933	34.865	33.711	15.920	10.086
Weighted Sum	81.261	38.552	37.650	17.496	10.395

Table 1: ELMo Results: Word Identity

the direct neighbors seem to have a notably more exact representation in BERT.

4 A Framework for the Analysis of Probing Experiments

The experiments that we presented in the previous section show that pre-trained word representations encode the sentential context of a token to a significant extent. But how can we know whether they encode any specific linguistic structure? In this section we provide a framework for the theoretical analysis of a probing experiment with regards to this question.

4.1 The Context-Only Hypothesis

Two common lines of reasoning when arguing that a probing experiment *does* show the presence of linguistic structure in word representations are the use of strong baselines and restrictions on expressiveness and learning scenarios. Similar to other areas of NLP, probing baselines should be designed in such a way that performance above the baseline can be taken as evidence that the word representations encode linguistic structure. The reasoning behind restrictions on expressiveness and learning is exemplified by Peters et al. (2018), who write: "As the linear classifier adds only a small amount of model capacity, this is direct test of the biLM's [bidirectional language model's] representations." Other approaches to limiting a probe's learning capabilities include restricting the number of hidden units or training set size.

We will review the validity of baselines and restrictions on learning relative to the following hypothesis, which serves a role similar to that of a null hypothesis in statistical significance testing:

The Context-Only Hypothesis: The only information that the classifier probe uses to learn the auxiliary task T is information about the identity of the neighboring words of each w_i.

In the preceding section, we have shown that this hypothesis is not trivially false, as information about neighboring words appears to be present in the word representations to a high extent. We argue that a probing study that claims that word representations encode linguistic structure should be able to reject the context-only hypothesis—that is, it should show that the probe uses information that goes beyond information about the identity of neighboring words, such as syntactic or semantic structure.

4.2 Review of Baselines under the Hypothesis

The most widespread instrument to test if the representation $R(w_i)$ really encodes linguistic information that is specifically useful for the probing task T is to compare the performance of the probe as it operates on $R(w_i)$ with its performance as it operates on some baseline representation that can be assumed *not* to contain information useful for T. We identify four categories of baselines in previous work: random input embeddings, random targets, uncontextualized word embeddings, and contextualized embeddings where the contextualization process is not learned.

Random Word Embeddings In this baseline, each word type is assigned a random embedding. This baseline is used, e.g., by Zhang and Bowman (2018). As random embeddings contain no structural information at all, this baseline is obviously weak with regard to the context-only hypothesis. It only tests the model's capacities to memorize random inputs, but discards even similarities of words that have been shown useful already in earlier uncontextualized word embeddings.

Random Targets (Control Tasks) This baseline was proposed by Hewitt and Liang (2019). Each word type is assigned an output randomly using a deterministic function. As this baseline cannot meaningfully use any pre-existing information, it is obviously weak with regard to the hypothesis—it only tests the model's capacities to memorize. As the targets are randomly assigned, no information from the embeddings can be useful, which makes this baseline comparable to the random word embeddings baseline with respect to the context-only hypothesis.

Uncontextualized Embeddings This baseline is often the uncontextualized part of the model, e.g. the embedding layer in ELMo (*word embedding*) or BERT (*BERT-0*). As the context-only hypothesis assumes explicit information about neighboring words, strong performance relative to this baseline is not enough to disprove the hypothesis.

Learning-Free Contextualized Embeddings Some works include baselines where the word representations are enriched with sentence context in the form of, e.g., randomly initialized LSTMs (Zhang and Bowman, 2018) or, similarly, ELMo with all layers except for the character-based embedding layer randomized (Hewitt and Manning, 2019; Tenney et al., 2019b). These baselines are arguably stronger than the previous ones as they have the potential to include the same information about the context as the representations we are seeking to probe. However, we cannot assume the representations in these baselines to be "$R(w_i)$ minus a representation of T". The $R(w_i)$ can still contain more useful information than the baseline that are not related to T. They could, for example, represent a better compression of the context; they could model different types of context (as we have seen in Section 3.2, which context is modeled even varies largely across layers of the same model); or, under a more restricted definition of a representation of T, they could include features or relations in higher layers that are useful, but still have no connection to T other than being features useful for T. As long as we do not have full understanding of the information encoded in R, we do not know enough about the relation of $R(w_i)$ and the baseline and cannot reject the context-only hypothesis.

4.3 Review of Model and Training Restrictions under the Hypothesis

The idea behind putting explicit restrictions on probes is that a less expressive probe or a probe that is only trained on a small number of samples can only access information that is easily available, and will therefore only be able to succeed in learning the auxiliary task if the linguistic structure needed to solve the task was encoded already during pre-training.

One approach to restricting a probe is to restrict its classifier architecture, e.g. using a linear model or only few hidden units. This idea was most systematically studied by Hewitt and Liang (2019), but other works are also arguing for simple design, e.g. Lin et al. (2019) and Alain and Bengio (2016). However, the justification for the choice of expressiveness has been the empirical comparison against a baseline—in Hewitt and Liang's case the control tasks described in the previous subsection. As we already pointed out, this approach has the problem that we do not know how the baseline relates to the representation even for strong baselines, as long as we do not have full understanding of the representation. To the best of our knowledge, there have not been attempts to provide a theoretical proof that it is not possible to learn an auxiliary task T with a specific set of embeddings and design of the probe, and this is probably difficult because a definition of the distinction between learning the task vs. learning to decode is lacking.

Restricting the size of the data or the number of training steps has also been explored by some works. Zhang and Bowman (2018) and Hewitt and Liang (2019) limit the data set by taking only a portion of it for training. Talmor et al. (2019) take an even more restricted approach: Their proposed metric WS, based on work by Blier and Ollivier (2018), is a weighted sum over the scores at different training steps, giving higher weight to earlier training steps and starting as early as at 64 training steps. However, no theoretical reasoning why a certain number of data points or training steps is too small to learn the task is provided. WS is motivated by the work of Yogatama et al. (2019), who use a related setup for the evaluation of models in generalizing to new tasks. Our own attempts at finding grounding for training restrictions in learning theory, particularly for lower bounds on the learning of certain tasks, were not successful. There are several "rules of thumb" for how many training examples are necessary to learn a task, e.g. the Vapnik-Chervonenkis (VC) dimension or Foley's rule (Priddy and Keller, 2005); but they are not provable lower bounds but meant for practitioners who look after a suitable data size for good performance. We argue that there will be a large variation depending on the number of classes, the complexity of the task, etc., making general rules probably an inappropriate way of addressing this problem.

4.4 Empirical Analysis of Training Restrictions

In the following experiments, we test the behavior of restricted probes with two types of modifications: decreasing the number of hidden units and the number of training examples, similar in spirit to Talmor et al. (2019)'s WS metric. Like several other works on probing, we target syntactic dependency parsing based on the English Web Treebank data set from the Universal Dependencies (UD) project (Nivre et al., 2018). Following Tenney et al. (2019b) and Hewitt and Manning (2019), we use an edge probing setup and two specific tasks:

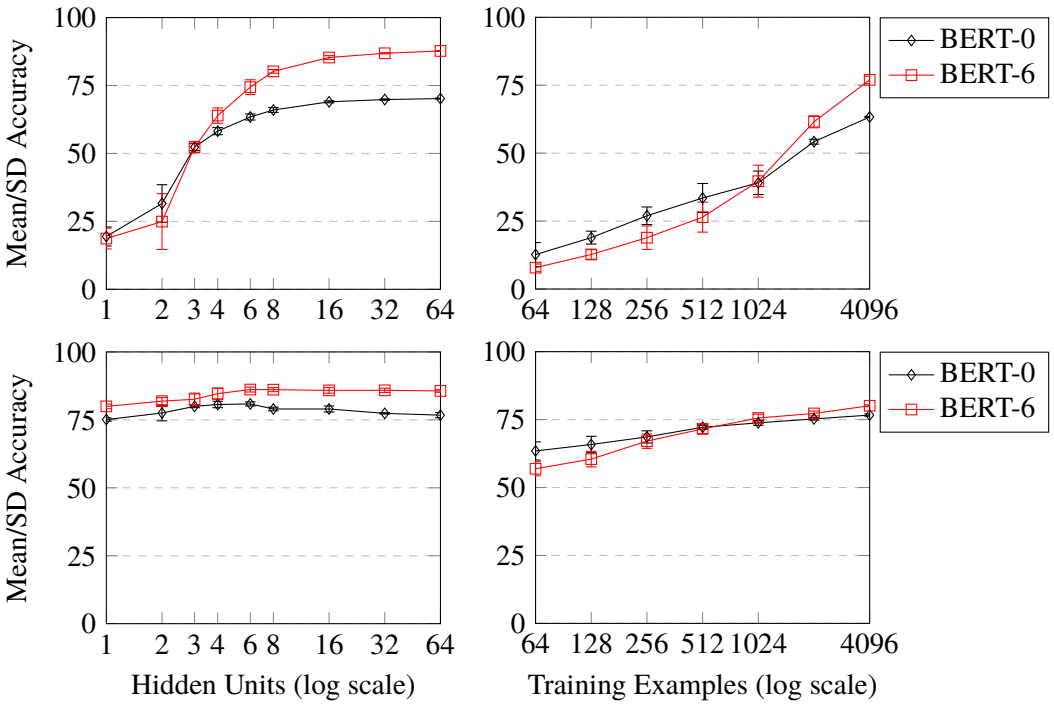

Figure 2: Restrictions on BERT-based models for label (above) and head-dependent pair (below) prediction, trained with representations from the uncontextualized layer BERT-0 and on BERT-6.

- Dependency Label Prediction, where for a given head–dependent pair (w_i, w_j), the objective is to determine the type of the dependency edge.

- Dependency Edge Prediction, where for a pair of words (w_i, w_j), the objective is to determine whether w_i is the syntactic head of w_j. The negative samples are reduced randomly so that both labels have the same number of training examples, i.e. 50% for each.

Experimental Setup We use the general setup described in Section 2.2. We decrease the number of hidden units and the number of training examples systematically by halving: from 64 hidden units down to just 1 hidden unit, and from 4,096 training examples down to 64. As the evaluation measure for both tasks, we use the prediction accuracy for all labels. We restrict the data size to the dependencies in 400 sentences, resulting in about 5,000 dependencies.[1] In addition to the sample mean accuracy, we also report one sample standard deviation with Bessel's correction.

4.4.1 Results: BERT

The results for the restricted BERT models are shown in Figure 2. We show the performances of the layer that performs best at the two probing task (BERT-6; red curve) and the performances of the uncontextualized layer (BERT-0; black curve). When comparing the two curves, we find that, for the most restricted models in terms of the number of hidden units, the differences in performance are very small. They increase in favor of BERT-6 with increasing number of hidden units especially for the dependency labels task. For this task we see a larger increase with more complex probes probably because having fewer parameters than output labels prevents the model from reaching a good performance. Restricting the number of training examples shows a consistent pattern: For very restricted numbers, up to 1,024 for the labels and up to 512 for the pairs task, BERT-0 reaches a higher accuracy than BERT-6. More training examples allow BERT-6 to outperform the uncontextualized baseline.

[1]This size was determined in preliminary experiments aimed at reducing training time and saving resources while at the same time not losing much in accuracy.

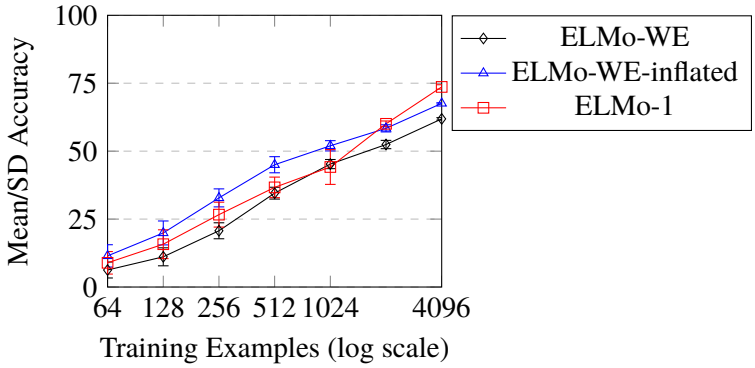

Figure 3: Restrictions on ELMo-based models for label prediction, trained with representations from the uncontextualized layer ELMo-WE and the best performing layer BERT-1, as well as an "inflated" version of ELMo-WE that has the same dimensionality as ELMo-1.

Our results suggest that it is little meaningful to restrict probes without setting the results into a richer context. In very restricted setups, less expressive models fail to improve over non-contextualized baselines in many cases and tend to have a high standard deviation, which makes the results less reliable and significance testing a necessity when comparing models or data sets.

If the task has a better encoding in the representation, we would expect that it should be easier to discover as well, i.e. the more we restrict the probe, the larger the difference in performance in favor of the contextualized representation should be. However, this is not confirmed in our results: In both tasks, the baseline even performs better in the most restricted training setups. We assume that the probe heavily relies on simple clues at the earliest training stages. A brief informal analysis of the results on the labels task revealed that the baseline models are faster at learning the simplest relations, e.g. the *det* relation between nouns and their determiners. This may be the case because in the uncontextualized layers, the words themselves have a clearer and better memorizable representation (e.g. of words like the article *the*), which makes the memorization and learning of simple clues faster and easier.

4.4.2 Results: ELMo

The results with ELMo embeddings are shown in Figure 3. They show the same general trends as the BERT results, with one interesting property: The ELMo character-based word embedding layer (ELMo-WE) has lower performance than ELMo-1 in all evaluation steps. However, "inflating" the character-based embeddings to the same dimensionality as ELMo-1 by simply concatenating them with themselves makes the word embedding layer surpass the performance of ELMo-1 up to step 2,048, although the performance after all training of inflated and non-inflated ELMo-WE is about the same (72.73% without vs. 72.91% with inflation). The performance after all training for ELMo-1 is 85.74%, and for ELMo-2 it is 82.64%. It appears to be crucial to control for the number of parameters, even if not trainable, when evaluating in a limited training setup.

4.4.3 Summary of the Results

In summary, probes need a certain expressiveness or training regime to improve over non-contextualized baselines. But if they need it to discover the representation of the task or to be able to learn the task with the context included in the representations is, as we argue, not obvious at all. With the uncontextualized layer as the baseline, we, in contrast to Hewitt and Liang (2019) using random targets, do not find that restricting models leads to a larger difference in results in favor of the contextualized representations. Instead, we observe the contrary or similar performance. We therefore propose that using one restricted model does not lead to meaningful results. Plotting successively restricted models against the accuracy can, however, help getting a picture of the probe's behavior and approximate the ease of learning the auxiliary probing task T with the word embeddings $R(w_i)$.

4.5 The Pipeline Argument

There is a finding from probing studies that makes it appear reasonable to believe that neural sentence encoders learn tasks in a similar way as humans do. More specifically, they exhibit a pipeline pattern when probing, with lower layers having being the most relevant layers for low-level syntactic tasks, while the middle layers contain more useful information for tasks that involve the detection of relations between words in sentence-level syntax.

One of the most influential voices for this pipeline hypothesis is Tenney et al. (2019a). They probe on all layers and discover that, on eight tasks including part-of-speech tagging, parsing, named entity recognition, semantic role labeling and coreference, the importance of layers on the performance corresponds to a classical NLP pipeline, with the lowest-level tasks having a better performance on the lowest layers. Peters et al. (2018), who probe their ELMo model for part-of-speech tagging, also suggest that consistencies with previous pipelined multi-task learning approaches support this finding: They find that part-of-speech tags are best recovered from layer 1. Belinkov et al. (2017a) find that lower layers in neural machine translation (NMT) systems better capture word structure, and suggest that higher layers have more information about word *meaning*. Shi et al. (2016) suggest that representations from lower layers in NMT are better at "learning" local syntax, while those from higher layers are better at more global syntax tasks. Belinkov et al. (2017b) find that representations from lower layers in neural machine translation systems get better results at part-of-speech tagging, while those from higher layers lead to a better performance on word-level semantic tagging, especially for the tags that are most semantic, like discourse functions and noun concepts.

Looking at our results in Figure 1 and Table 1, it is tempting to offer an alternative interpretation of these pipeline results: For local tasks like part-of-speech tagging, information about the word itself and about its direct neighbors is most crucial; therefore the performance is highest that are best at memorizing these neighboring words—the lower layers of BERT or ELMo's layer 1. For more global tasks like syntactic parsing and semantic tasks on the other hand, a broader context is important for good performance, so that the middle layers, which contain the most concrete information about words that are more than one step away, are the most successful point for prediction.

5 Conclusion

Being a very young field of research, probing and interpretation of language representations still lacks a solid theoretical foundation of its methodology, and despite some progress, the inner workings of models such as BERT and ELMo remain largely obscure. We have shown that a simple classifier trained on these models' representations can reconstruct the words in the context of single tokens with surprisingly high accuracy. Based on this observation, we propose to evaluate probes with regard to the context-only hypothesis: That the probing classifier learns the task based on linear context information alone. With this hypothesis in mind, many attempts to discern "learning the probing task" from "extracting linguistic structure" are not reliable in theory and in practice. Looking at random, uncontextualized and contextualized baselines, we have shown that none of them qualifies as a reliable ground for general conclusions about the linguistic knowledge that may be used to solve any sentence-level syntactic task. For the usage of probing model restrictions, we have argued that we are lacking too much knowledge about the representations and about their interplay with the probe's learning capabilities to reason about the presence of some specific feature from the results. A closer look at the learning curves evaluated at small numbers of training steps revealed that conclusions from probes with restricted training regimes have to be handled and interpreted with extreme care. In the most restricted setups, many of the probes fail to outperform even uncontextualized baselines in syntactic dependency pair and label prediction, and it remains theoretically and empirically unclear what setup to choose for a meaningful and reliable probe.

In very recent work, Pimentel et al. (2020) take an information-theoretic perspective on probing, concluding that the difference between learning a task and identifying linguistic structure is non-existent. They argue against any restrictions on the expressiveness of the probe, stating that there is no information gain induced by them. We argue, in line with them, that valid conclusions using restricted probing classifiers can to this point only relate to the ease of learning with the features present in the representations. We propose that, as probing classifiers' simplicity is still appealing, future work should aim at providing

theoretically well-motivated setups for this idea, as for example explored by Voita and Titov (2020) and Yogatama et al. (2019).

We believe that approximating the ease of learning the task with the respective embeddings of the baseline and the representations in question e.g. with plots can be insightful to get an impression of the learning process, but interpretations remain difficult without an extensive error analysis, and comparisons of embeddings or data sets are especially hard due to differences in dimensionalities, numbers and distributions of labels.

Acknowledgements

We thank our group at Linköping University for helpful discussions of this work, and the three anonymous reviewers for their interesting and constructive feedback. This work was partially supported by the Center for Industrial Information Technology (CENIIT) under grant 15.02.

References

Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017a. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada, July. Association for Computational Linguistics.

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2017b. 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, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.

Léonard Blier and Yann Ollivier. 2018. The description length of deep learning models. In *Advances in Neural Information Processing Systems*, pages 2216–2226.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy, August. Association for Computational Linguistics.

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, Minneapolis, Minnesota, June. Association for Computational Linguistics.

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, Hong Kong, China, November. Association for Computational Linguistics.

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, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. 3rd International Conference for Learning Representations.

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, August. Association for Computational Linguistics.

Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, John Bauer, Sandra Bellato, Kepa Bengoetxea, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Rogier Blokland, Victoria Bobicev, Carl Börstell, et al. 2018. Universal dependencies 2.2. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch.

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, New Orleans, Louisiana, June. Association for Computational Linguistics.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. *arXiv preprint arXiv:2004.03061*.

Kevin L Priddy and Paul E Keller. 2005. *Artificial neural networks: An introduction*, volume 68. SPIE press.

Xing Shi, 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, Austin, Texas, November. Association for Computational Linguistics.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2019. oLMpics–On what language model pre-training captures. *arXiv preprint arXiv:1912.13283*.

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, July. 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*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. *arXiv preprint arXiv:2003.12298*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, page abs/1910.03771.

Dani Yogatama, Cyprien de Masson d'Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, et al. 2019. Learning and evaluating general linguistic intelligence. *arXiv preprint arXiv:1901.11373*.

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, Brussels, Belgium, November. Association for Computational Linguistics.