Hitting "Probe"rty with Non-Linearity, and More

Avik Pal
14156350
University of Amsterdam
avik.pal@student.uva.nl

Madhura Pawar
14386739
University of Amsterdam
madhura.pawar@student.uva.nl

Abstract

Structural probes (Hewitt and Manning, 2019) learn a linear transformation to find how dependency trees are embedded in the hidden states of language models. This simple design may not allow for full exploitation of the structure of the encoded information. Hence, to investigate the structure of the encoded information to its full extent, we incorporate non-linear structural probes. We reformulate the design of non-linear structural probes introduced by (White et al., 2021) making its design simpler yet effective. We also design a visualization framework that lets us qualitatively assess how strongly two words in a sentence are connected in the predicted dependency tree. We use this technique to understand which non-linear probe variant is good at encoding syntactical information. Additionally, we also use it to qualitatively investigate the structure of dependency trees that BERT encodes in each of its layers. We find that the radial basis function (RBF) is an effective non-linear probe for the BERT model than the linear probe.

1 Introduction

In human languages, the meaning of a sentence is composed hierarchically - small chunks of words together make successively larger chunks. We refer to this tree-structured hierarchy of a sentence as a dependency tree. The paper, (Hewitt and Manning, 2019) introduces "Structural Probes" that tests a simple hypothesis for how dependency trees may be embedded in the hidden states of a language model. The probe in this context is a model that learns a linear transformation such that two words that are syntactically close to one another should also have less distance between their respective contextual representations.

One of the advantages of these structural probes is simplicity. However, this simple design may not allow for full exploitation of the structure of the encoded information. This motivates us to investigate whether the dependency structure is encoded in a non-linear way. (White et al., 2021) introduces three non-linear structural probes - polynomial, radial basis function, and sigmoid. We reformulate these non-linear probe variants. We then apply these probes on contextualized embeddings of BERT and $\text{BERT}_{\text{LARGE}}$ (Devlin et al., 2019). Since different layers of BERT capture different linguistic properties (Rogers et al., 2020), can our probing method tell us how dependency information is encoded in every layer? This question (as far as we know) hasn't been answered yet in the existing literature. We answer this question by applying a non-linear probing variant-RBF across the layers of BERT and $\text{BERT}_{\text{LARGE}}$.

We expect the non-linear variants to perform better than linear probes due to the complex nature of the dependency trees which might be better captured by non-linear probes. We test this hypothesis quantitatively using undirected attachment score (UUAS) (Hewitt and Manning, 2019). Additionally, we test this hypothesis qualitatively by introducing a measure that computes the strength with which the probe predicts dependency between two words.

We find that Radial Basis Function (RBF) probe is more effective than the linear probe for BERT. And we are also able to understand how every layer gradually encodes syntactical information to form a complete dependency tree.

2 Background

Previous works by (Gulordava et al., 2018; Kuncoro et al., 2018; Linzen and Leonard, 2018; Futrell et al., 2019) show that Recurrent Neural Networks (RNNs) were able to learn the syntactic structure of languages within their hidden states without being explicitly trained on them. But these experiments could only show local linguistic phenomena owing to the limited potential of the representations formed by looking at the immediate hidden state in

an iterative manner that loses distant past information. With the advent of the attention mechanism (Bahdanau et al., 2015), the encoders could form better representations by looking at hidden states of all previous steps enabling the model to capture distant information well. This expedited more robust frameworks to quantitatively align with underlying linguistic structures in hidden states.

Structural probing is one such framework that attempts to provide evidence of syntax trees embedded in hidden states using UUAS (Hewitt and Manning, 2019) while also providing experimental setups to visualize syntactic information learning capability between layers. In the paper (White et al., 2021), the authors further kernelize this framework to incorporate non-linear formulations for increased expressivity of the hidden states.

Our work is majorly a reformulation of these kernelized non-linear probes with simpler design and effectiveness.

3 Approach

3.1 Structural Probe

The goal of the structural probe is to see whether the syntactic distance between any two words can be approximated by a learned, linear distance function:

$$d_B(h_i^l, h_j^l) = ||Bh_i^l - Bh_j^l||_2 \quad (1)$$

where h_i^l and h_j^l are word embeddings from the l^{th} layer of the language model for i^{th} and j^{th} word where $h_i^l, h_j^l \in \mathbb{R}^d$ and $B \in \mathbb{R}^{d \times m}$ is a linear projection matrix. To learn this probe, (Hewitt and Manning, 2019) minimize the following objective with respect to B through gradient descent:

$$\min_B \sum_l \frac{1}{|s^l|^2} \sum_{i,j} |d_T^l(h_i^l, h_j^l) - d_B^l(h_i^l, h_j^l)^2| \quad (2)$$

where $|s^l|$ is the length of the sentence and $d_T^l(h_i^l, h_j^l)$ is the actual distance between words i and j in the dependency tree of sentence s. This minimizes the difference between the syntactic distances obtained from the dependency tree and the distance between the two vectors under our learned transformation. From now on we refer to this linear design of structural probes as "linear probes".

3.2 Non-Linear Probes

(White et al., 2021) finds the distance between two embeddings as follows:

$$||\phi(Bh_i^l, Bh_i^l) - 2\phi(Bh_i^l, Bh_j^l) + \phi(Bh_j^l, Bh_j^l)||_2 \quad (3)$$

where ϕ is the non-linear function that takes as input the linear transformation of two representations. We modify the kernel to make it a function that takes as input just one linear transformation of the representation and computes distance as follows:

$$d_B(h_i^l, h_j^l) = ||\phi(Bh_i^l) - \phi(Bh_j^l)||_2 \quad (4)$$

We design our probe in such a way so that it has a simple design while factoring in the effect of applying non-linearity on the embeddings. Firstly we use the polynomial function as follows:

$$\phi_{\text{poly}}(Bh_i^l) = (Bh_i^l + c)^d \quad (5)$$

where $d \in \mathbb{Z}_+$ and $c \in \mathbb{R}_{\geq}0$. Next, we consider the radial-basis function (RBF), defined as follows:

$$\phi_{\text{rbf}}(Bh_i^l) = \exp(-\frac{||Bh_i^l||^2}{2\sigma^2}) \quad (6)$$

where σ acts like a scaling factor. Lastly, we implement the sigmoid function:

$$\phi_{\text{rbf}}(Bh_i^l) = \tanh(aBh_i^l + b) \quad (7)$$

where a and b are scalar-valued tuning parameters.

3.3 Tree Distance Evaluation Metric

As our quantitative metric, We use Unlabeled Undirected Attachment Score (UUAS) which is the percent of edges predicted correctly by the probe against the actual dependency tree (gold tree). UUAS score merely conveys whether there exists an edge or not between two word embeddings. It would be helpful to also know how strongly the probe thinks two words are connected in a sentence's predicted dependency tree. Inspired by the qualitative visualization introduced by (Coenen et al., 2019), we implement a similar visualization framework to see how strongly the probe predicts connectivity between h_i^l and h_j^l:

$$\text{strength}(h_i^l, h_j^l) = \frac{d_B(h_i^l, h_j^l)}{d_T(h_i^l, h_j^l)} \quad (8)$$

This metric can help us visualize qualitatively how different probes encode the dependency trees from the contextual representations. Also, this measure can help us gain insight into how each layer encodes dependency tree syntaxes.

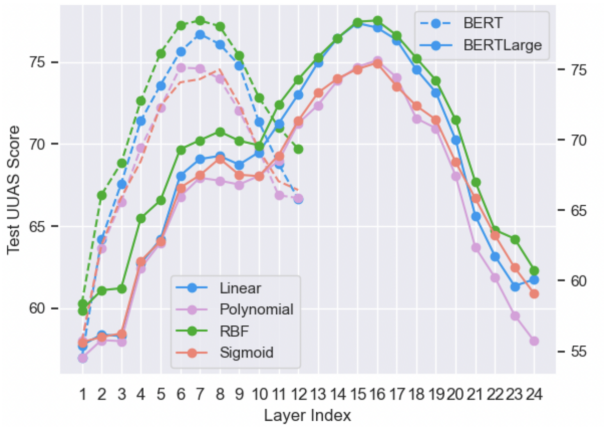


Figure 1: UUAS across BERT and $\text{BERT}_{\text{LARGE}}$ layers

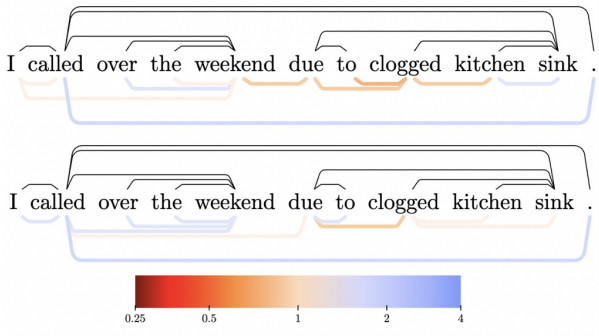


Figure 2: Dependency trees for BERT Layer 12 by Linear Probe (above) and RBF Probe (below). Edges in black are the gold trees.

4 Experiments

4.1 Dataset

We use the Universal Dependencies English Web Treebank (UD-EWT) dataset (Silveira et al., 2014). The EWT trees were hand-corrected to Universal Dependencies which have a CoNLL-U format (Buchholz and Marsi, 2006). We keep the original train-dev-test split of 75.5-12-12.5 provided by (Silveira et al., 2014).

4.2 Models and Setup

We use our probes to analyze the dependency structure present in the hidden states of BERT and $\text{BERT}_{\text{LARGE}}$. We take a contextualized representation of the words in the sentence while taking the mean over all sub-word representations. We keep hyperparameters for Equation 5: $c = 0, d = 2$, Equation 6: $\sigma = 1$, and Equation 7: $a = 1, b = 0$. We train our probes for 200 epochs with early stopping enabled and an initial learning rate of 0.001 with a reduction factor of 0.5 on plateauing. The code implementation of our project is here[1].

5 Results and Discussion

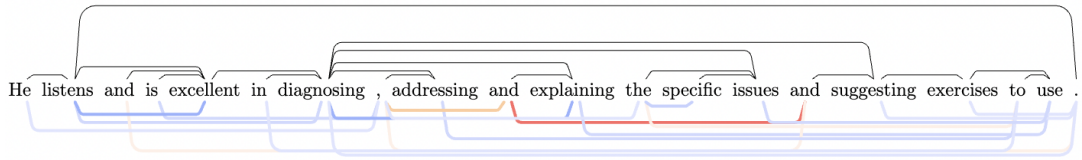
We initially implemented the equations for non-linear probes introduced by (White et al., 2021). However, the design of their probes wasn't clear as they don't mention how they find the distance between two embeddings. Because of this, we weren't able to reproduce their work. Hence, based on our intuition and understanding of structural probes we came up with Equations 4-7.

We observe in Figure 1 that for BERT and $\text{BERT}_{\text{LARGE}}$, **RBF Probe has consistently high UUAS scores** across all layers compared to the other two non-linear probes (Polynomial and Sigmoid). (White et al., 2021) provides reasoning for this by showing a resemblance of RBF's structure with that of BERT's attention mechanism. BERT and $\text{BERT}_{\text{LARGE}}$'s UUAS score for the last layer is 69.71 and 60.77 respectively for RBF. The UUAS score's trend for a linear probe for both these models runs along with RBF's score with 66.65 and 60.10 for BERT and $\text{BERT}_{\text{LARGE}}$ respectively. Then the question arises, can we conclude that RBF Probe is better than Linear Probe?

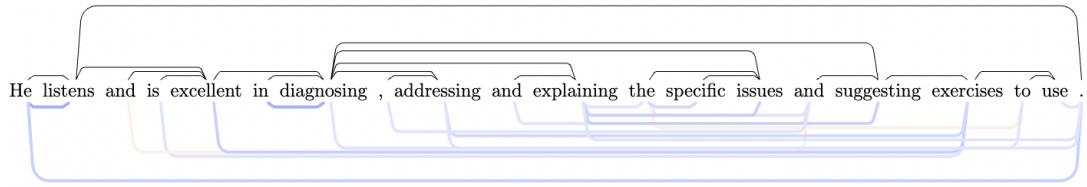
To investigate this further, we utilize the visualization technique we introduced in Section 3.3. The strength between two words (Equation 8) is shown by the gradient in Figure 2. The edges predicted by probes having lesser strength are orange in color while the edges having greater strength have bluish color. We see in Figure 2 that linear probes predict more edge dependencies which are (1) "false negatives" (*weekend — due, to — clogged*), (2) "true positives" but with a lower strength (*I — called, the — weekend*). While the RBF probe comparatively predicts lesser edges but with higher strength and more correctness. We did the qualitative visualizations for multiple sentences at random and found similar patterns for them too. However, as future work, we can think of incorporating the strength with which two edges are connected in the calculation of UUAS scores.

Furthermore, to investigate how dependency tree syntaxes are encoded in every layer, we applied all four probing variants to BERT and $\text{BERT}_{\text{LARGE}}$. However, since the RBF probe's performance is promising, we discuss the results of probing with this variant for BERT. Out of the 12 layers, we select three particular layers to discuss because of

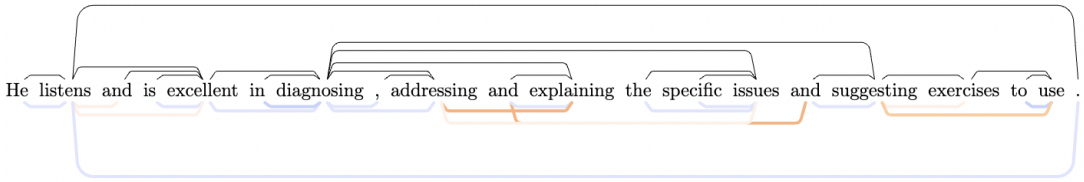
[1] https://github.com/madhurapawaruva/nlp2-probing-lms



(a) Layer 3 (UUAS = 68.842)



(b) Layer 6 (UUAS = 77.269)



(c) Layer 12 (UUAS = 69.71)

Figure 3: Dependency trees from predicted squared distances on BERT by RBF Probe for layers $\in \{3, 6, 12\}$

their functional position in the BERT model (lower, middle, and last layer) and also because of their high UUAS scores. The lower layers of the BERT try to capture linear word order (Rogers et al., 2020) and this is affirmed by our observation of Figure 3a as every other word is spanning out and connecting to every other word in its neighborhood. We see words from the first sub-part of the sentence (*He listens and is excellent is diagnosing*), connecting with words from the second sub-part of the sentence (*addressing and explaining*) which then is getting connected with the last sub-part of the sentence.

Figure 3b shows the dependency tree of mid-layer 6 where many edges with high strength are predicted. The mid layers convey the most syntactic information as per (Rogers et al., 2020). That might be true as we see stronger subject-verb agreement (*he-listens*) but we also see wrong subject-verb pairs being predicted (*diagnosing-exercises*). Layer 6 overall forms a lot of correct and incorrect dependencies and the high UUAS score can be attributed to the fact that due to so many connections being formed, some connections are predicted correctly but they might've been just formed "by chance" and do not have any meaning. **The UUAS score thus needs to also penalize such false-positive dependencies.**

Finally, we see that Figure 3c has very selective connections and the correct connections are predicted with high strength while the incorrect with low strength. This affirms with the observation in (Rogers et al., 2020) where the final layers are more task-specific.

We also try to analyze whether the context is important for language models to form representations with syntax dependencies. We do this by running the same experiments but passing individual words of the sentence independently to obtain non-contextualized representations for them. We find that this degrades the performance and contexts are indeed important. Detailed discussion in Appendix A.

6 Conclusion

In conclusion, we see that non-linear probes like RBF are better at describing the knowledge its embeddings have about linguistic properties like syntax trees. Our qualitative experiments make us realize that UUAS is not enough indicator of how good a probing design is, and incorporating the strength of predicted edges and false positives may make the UUAS more robust. Also, we see that layer-wise BERT gains knowledge functionally but our understanding of this is still in its infancy and more rigorous experiments are needed. This we would like to address in our future work.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 149–164, New York City. Association for Computational Linguistics.

Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda B. Viégas, and Martin Wattenberg. 2019. Visualizing and measuring the geometry of bert. *ArXiv*, abs/1906.02715.

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. Association for Computational Linguistics.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. 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 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. 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 1195–1205, New Orleans, Louisiana. 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. Association for Computational Linguistics.

Adhiguna Kuncoro, Chris Dyer, John Hale, Dani Yogatama, Stephen Clark, and Phil Blunsom. 2018. LSTMs can learn syntax-sensitive dependencies well, but modeling structure makes them better. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1436, Melbourne, Australia. Association for Computational Linguistics.

Tal Linzen and Brian Leonard. 2018. Distinct patterns of syntactic agreement errors in recurrent networks and humans. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society, CogSci 2018, Madison, WI, USA, July 25-28, 2018*. cognitivesciencesociety.org.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

Jennifer C. White, Tiago Pimentel, Naomi Saphra, and Ryan Cotterell. 2021. A non-linear structural probe. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 132–138. Association for Computational Linguistics.

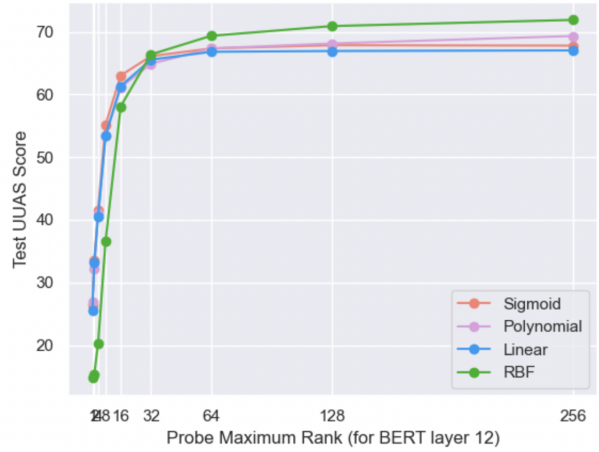


Figure 4: UUAS Scores for varying rank of matrix B where rank $\in \{1, 2, 4, 8, 16, 32, 64, 128, 256\}$

A Non-contextualized representations

Figure 5 shows the performance of probes when word representations are taken independent of the sentence. From Section 5, we see qualitatively that syntactic dependencies are more prominent in deeper layers. For BERT, the trend of Figure 5a is decreasing for all probes. This shows that dependency structures between words are captured much better (from Figure 1) within the context of a sentence.

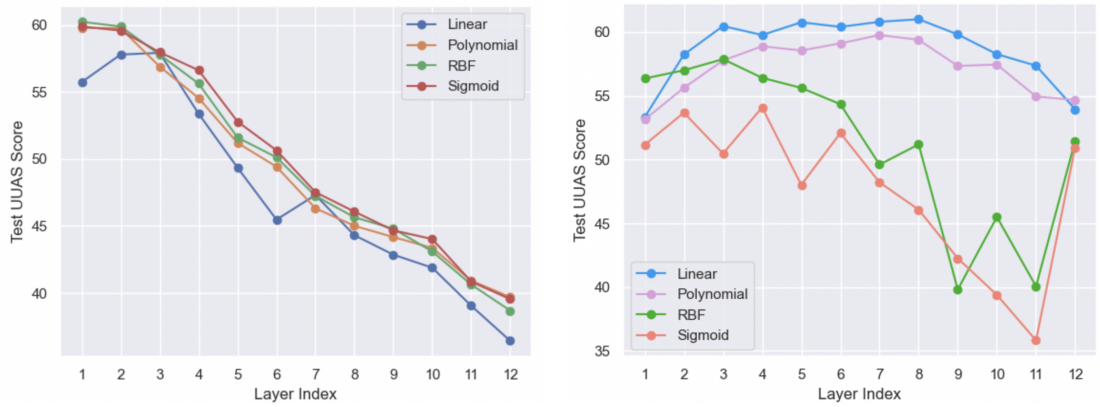
For GPT2 though, results are more arbitrary. From Figure 5b, we observe that linear and polynomial probes have moderately high UUAS scores which are somewhat maintained across layers but UUAS scores from RBF and sigmoid probes are quite haphazard. Since GPT2 has been pre-trained on an autoregressive task of predicting the next token, it's only given the context of previous tokens which is a limited view than what BERT has. This might be the reason that the words have better non-contextualized representations. We couldn't provide intuition on why RBF and sigmoid probes don't work in this case and leave it for future work.

B More Visualizations of Dependency Trees

Figure 7 shows various dependency trees for $\text{BERT}_{\text{LARGE}}$. We do not observe gradual learning of syntax trees as we did for BERT for these particular layers. Due to time constraint, we couldn't generate visualizations for all of the 24 layers. And it might be the case that these particular layers we visualize might not depict the representative learnings. As part of future work, we aim to investigate whether RBF exhibits gradual learning of syntax tree for $\text{BERT}_{\text{LARGE}}$ too.

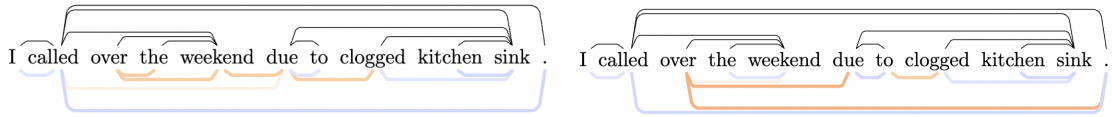
C Analysis of transformation rank

We train our probes of varying d, that is, specifying a matrix $B \in \mathbb{R}^{d \times m}$. As shown in Figure 4, increasing d beyond 64 or 128 leads to no further gains in UUAS.



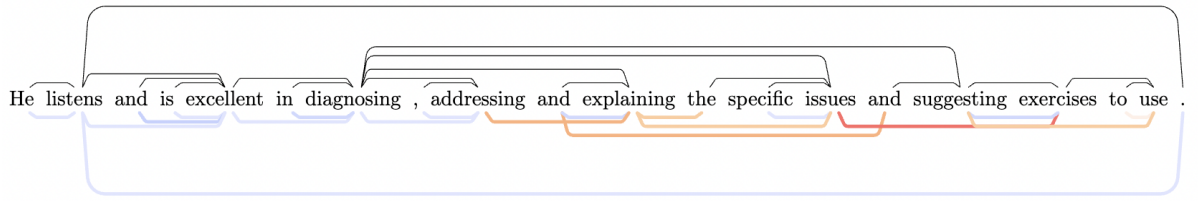
(a) BERT (b) GPT2

Figure 5: UUAS Scores for non-contextualized representations

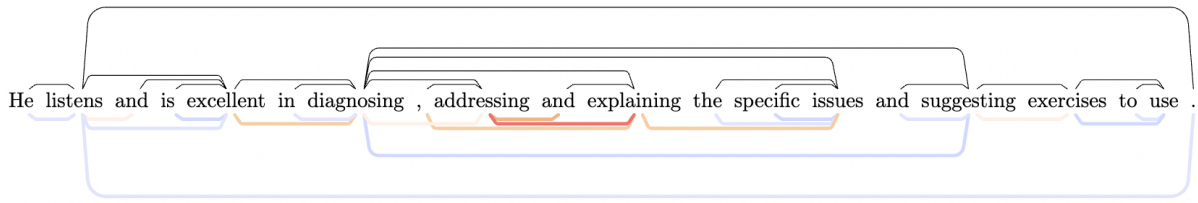


(a) Polynomial (b) Sigmoid

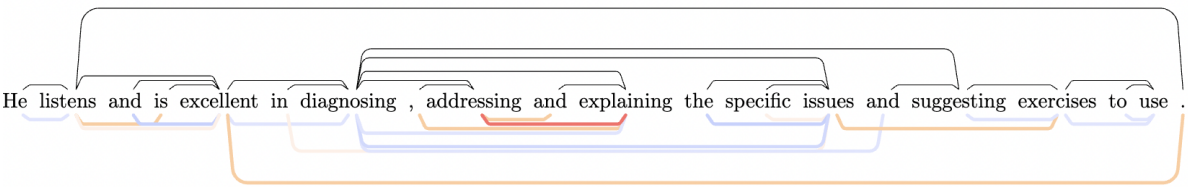
Figure 6: Dependency trees for BERT Layer 12



(a) Layer 3



(b) Layer 16



(c) Layer 24

Figure 7: Dependency trees from predicted squared distances on $\text{BERT}_{\text{LARGE}}$ by RBF Probe for layers $\in \{3, 16, 24\}$