What Makes My Model Perplexed?
A Linguistic Investigation on Neural Language Models Perplexity

Alessio Miaschi[⋆ ⋄], Dominique Brunato[⋄], Felice Dell'Orletta[⋄], Giulia Venturi[⋄]
[⋆]Department of Computer Science, University of Pisa
[⋄]Istituto di Linguistica Computazionale "Antonio Zampolli", Pisa
ItaliaNLP Lab – *www.italianlp.it*
alessio.miaschi@phd.unipi.it, {name.surname}@ilc.cnr.it

Abstract

This paper presents an investigation aimed at studying how the linguistic structure of a sentence affects the perplexity of two of the most popular Neural Language Models (NLMs), BERT and GPT-2. We first compare the sentence–level likelihood computed with BERT and the GPT-2's perplexity showing that the two metrics are correlated. In addition, we exploit linguistic features capturing a wide set of morpho-syntactic and syntactic phenomena showing how they contribute to predict the perplexity of the two NLMs.

1 Introduction and Motivation

Perplexity is one of the most standard metrics to assess the quality of a language model. It is also used in different scenarios, such as to classify formal and colloquial tweets (González, 2015), to detect the boundaries between varieties belonging to the same language family (Gamallo et al., 2017), to identify speech samples produced by subjects with cognitive and/or language diseases e.g. dementia, (Cohen and Pakhomov, 2020) or to assess whether it matches various human behavioural measures, such as gaze duration during reading (Demberg and Keller, 2008; Goodkind and Bicknell, 2018). With the recent success gained by Neural Language Models (NLMs) across a variety of NLP tasks, the notion of perplexity has started being investigated also to dig into issues related to the interpretability of contextual word representations, with the aim of understanding whether there is a relationship between this metric and the grammatical abilities implicitly encoded by a NLM (Gulordava et al., 2018; Marvin and Linzen, 2018; Kuncoro et al., 2019). In this context, Hu et al. (2020) and Warstadt et al. (2020) observed a dissociation between the perplexity of a NLM and its performance on targeted syntactic assessments probing the model's ability to encode a range of subtle syntactic phenomena. These findings seem to be valid for models tested across languages (Mueller et al., 2020).

In this paper, we address this scenario but from a different perspective. Rather than studying the relation between the NLM's perplexity and its linguistic competences assessed on sentences undergoing controlled syntactic modifications, we focus on sentences representative of real usage. Our purpose indeed is to understand which linguistic phenomena of the input sentence may make perplexed a NLM and whether they can effectively predict the assigned perplexity score. To have a in-depth understanding of the relation between linguistic structure and perplexity, we rely on a wide spectrum of linguistic features modeling a variety of phenomena, specifically morpho-syntactic and syntactic ones. As we also intend to evaluate the possible influence of the NLM architecture on this relation, in all our experiments we consider two of the most popular NLMs, a traditional unidirectional one, i.e. GPT-2 (Radford et al., 2019), and a bidirectional model such as BERT (Devlin et al., 2019).

Contributions In this paper: (i) we showed that a sentence-level likelihood computed by masking each word sequentially for the BERT model has a robust correlation with GPT-2's perplexity scores; (ii) we verified whether it is possible to predict NLMs' perplexities using a wide set of linguistic features extracted by a sentence; (iii) we identified the linguistic properties of a sentence that mostly cause perplexity, reporting differences and similarities between the two models.

2 Our Approach

We defined two sets of experiments. The first consists in investigating the relationship between BERT and GPT-2 sentence-level perplexity (*PPL*) scores. To do so, we first computed BERT and GPT-2 *PPL* scores for sentences contained in the English Universal Dependencies (UD) treebank (Nivre et al., 2016) and we assessed their corre-

Proceedings of Deep Learning Inside Out (DeeLIO):
The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, pages 40–47
Online, June 10, 2021. ©2021 Association for Computational Linguistics

lation. In the second set of experiments, we studied whether a simple regression model that takes as input a wide range of linguistic features automatically extracted from each UD sentence is able to predict the two NLMs sentence-level perplexities.

To understand which linguistic phenomena contribute to the prediction of BERT and GPT-2 PPLs, and how these features differ between them, we performed an in-depth investigation training the regression model with one feature at a time.

2.1 Linguistic Features

The set of considered linguistic features is based on the ones described in Brunato et al. (2020) which are acquired from raw, morpho-syntactic and syntactic levels of annotation for a total of 78 features that can be categorised in 9 groups corresponding to different linguistic phenomena. A summary of the linguistic features is reported in Table 1, while the whole list is provided in Appendix A.

As shown in Table, these features model linguistic phenomena ranging from raw text one, to morpho–syntactic information and inflectional properties of verbs, to more complex aspects of sentence structure modeling global and local properties of the whole parsed tree and of specific subtrees, such as the order of subjects and objects with respect to the verb, the distribution of UD syntactic relations, also including features referring to the use of subordination and to the structure of verbal predicates.

All these features have been shown to play a highly predictive role when leveraged by traditional learning models on a variety of classification problems, also including the development of probes as reported by Miaschi et al. (2020), who showed that these features can be effectively used to profile the knowledge encoded in the language representations of a pretrained NLM.

2.2 Models and Data

For our experiments, we rely on the pre-trained version of the two NLMs previously defined. BERT (Devlin et al., 2019) is a Transformer-based masked language model, pretrained on BookCorpus (Zhu et al., 2015) and English Wikipedia. GPT-2 (Radford et al., 2018) is a large transformer-based language model trained using the language modeling task (LM) on 8 million documents for a total of 40 GB of text.

Linguistic Feature
Raw Text Properties
Sentence Length
Word Length
Vocabulary Richness
Type/Token Ratio for words and lemmas
Morphosyntactic information
Distibution of UD and language–specific POS
Lexical density
Inflectional morphology
Inflectional morphology of auxiliary verbs
Verbal Predicate Structure
Distribution of verbal heads and verbal roots
Verb arity and distribution of verbs by arity
Global and Local Syntactic Tree Structures
Depth of the whole syntactic tree
Average length of dependency links and of the longest link
Average length of prepositional chains and distribution by depth
Clause length
Relative order of elements
Order of subject and object
Syntactic Relations
Distribution of dependency relations
Use of Subordination
Distribution of subordinate and principal clauses
Average length of subordination chains and distribution by depth
Relative order of subordinate clauses

Table 1: Linguistic Features used in the experiments.

We first computed GPT-2's sentence-level perplexities by dividing the sum of all sub-word conditional log-probabilities by the total number of words for each sentence in the UD dataset. On the other hand, since BERT masked language modeling task does not allow to compute well-formed probability distributions over sentences, we measure BERT sentence-level likelihood by masking each word sequentially and computing the probability as follows:

$$p(S) \approx \prod_{i=1}^{k} p(w_i|context)$$

where *context*, given the deep bidirectionality of the model, corresponds to $w_1, ..., w_{i-1}, w_{i+1}, ..., w_k$. The perplexity is then computed as follows:

$$PPL_S = e^{(\frac{p(S)}{N})}$$

where *N* correspond to the length of sentence *S*. In order to uniform the terminology, in what follows we will refer to the BERT sentence-level likelihood as perplexity.

In order to evaluate our approach on gold annotated sentences, we relied on the three English Universal Dependencies (UD) treebanks: the English version of ParTUT (Sanguinetti and Bosco, 2015), the UD version of the GUM corpus (Zeldes, 2017) and of the English Web Treebank (EWT) (Silveira et al., 2014). Overall, the final dataset consists of 22,505 sentences.

Lengths	ρ score	# samples
All	0.63	22,505
n=10	0.66	847
n=15	0.60	793
n=20	0.64	643
n=25	0.53	422
n=30	0.54	277

Table 2: Spearman correlations between BERT and GPT-2 perplexities computed for all UD sentences (*All*) and sentences with fixed-length n.

3 A Linguistic Investigation on Perplexity

As a first step, we assessed whether there is a relationship between the perplexity of a traditional NLM and of a masked NLM. We thus calculated BERT and GPT-2 perplexity scores for each UD sentence and measured the correlation between them. Since *PPL* scores are highly affected by the length of the input sequence, we computed ρ correlation coefficients also considering groups of sentences with fixed length. Specifically, we relied on Spearman correlation because we were interested in measuring how the variations in perplexity scores relate each other, rather than focusing on the actual *PPL* values. Results are reported in Table 2. As we can notice, even considering samples with fixed length, the two NLMs' perplexities exhibit moderate to substantial correlation (with $p < 0.001$), thus showing that BERT an GPT-2 do not diverge excessively in their ability of predicting the likelihood of the input sentences. Moreover, this allows us to confirm that, although the deep bidirectional structure of BERT does not permit to compute a well-formed probability distribution over a sentence (see Section 2.2), this metric could be considered as a valid approximation of the perplexity computed with a unidirectional NLM.

Once established the correlation between the perplexities of the two NLMs, we performed a second experiment to investigate (i) if the considered set of linguistic features plays a role in predicting their perplexity and (ii) which are the features that contribute more to the prediction task. To do so, we trained a LinearSVR model that predicts perplexity's scores using our set of linguistic properties as input features. Since most of them refer to syntactic properties of sentence that are strongly correlated with its length, we considered as a baseline a SVR model that takes sentence length as input and outputs BERT/GPT-2 sentence's perplexity. Regression results deriving by considering both the whole set (*All*) and each of the 9 groups of linguistic features separately are reported in Figure 1. As a general remark, for the whole UD dataset, we can observe that the results considering both all and the 9 groups of linguistic features outperform the results obtained by the baseline, i.e. ρ=0.38 for BERT and 0.22 for GPT-2 respectively. This demonstrates that the considered features are able to model aspects involved in NLM's perplexity that go beyond the simple length of sentence. This is particularly the case of GPT-2, suggesting that the probability assigned to a sentence by a traditional NLM is more explainable in terms of linguistic phenomena mainly affecting morpho-syntactic and syntactic structure. Consequently, the baseline score is higher for BERT. If we consider the scores obtained for each group of sentences with fixed length, we can see that higher scores are obtained for groups containing shorter sentences, for both NLMs. This is quite expected since in these sentences the possible output space is smaller for almost all features,

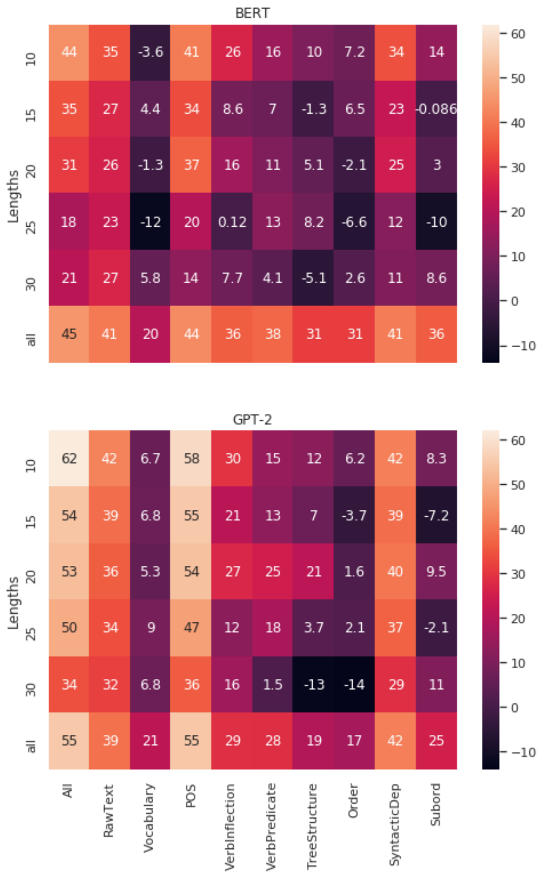

Figure 1: BERT and GPT-2 ρ scores (multiplied by 100) obtained with the LinearSVR model using linguistic features, for the whole UD dataset and groups of sentences with fixed length.

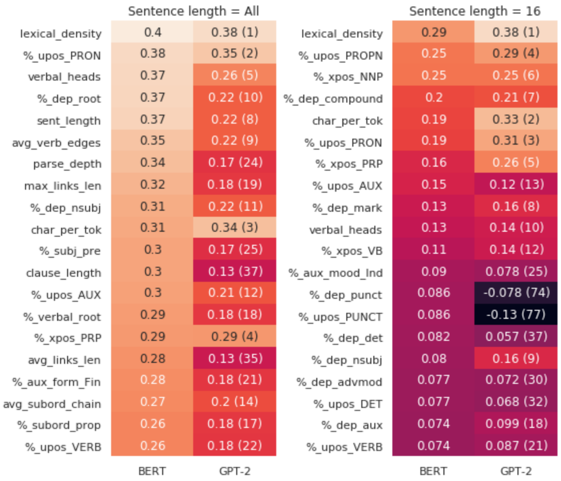

Figure 2: BERT and GPT-2 ρ scores obtained with the LinearSVR model, for the whole UD dataset and 16 token-long sentences. Scores are reported for the 20 top-ranked features for BERT. Numbers in brackets correspond to the relative in the GPT-2 ranking.

thus making them more predictive. Also in this case, the impact of the linguistic features is always higher for the prediction of GPT-2's perplexity.

A more in-depth analysis of these results shows that the distribution of the morpho-syntactic characteristics of a sentence (*POS*) and of the syntactic dependency relations (*SyntacticDep*) are the two most predictive sources of linguistic information. As Figure 1 reports, this holds for the two NLM models and it remains constant throughout all the groups of sentences with fixed lengths. Interestingly, if we consider the whole set of sentences, the effect of the morpho-syntactic information on the prediction of GPT-2's perplexity is exactly the same of that of the whole set of linguistic features. For some sentence lengths (15, 20, 30) the scores obtained using only this type of information outperform even those obtained considering the whole set of features. Note that this last remark is true also in the prediction of BERT's perplexity. As expected the other most predictive group is the one (*RawText*) that includes the length of sentence.

3.1 Focus on the contribution of individual features

To investigate more in depth which linguistic phenomena are more involved in the perplexity of the two models, we trained the LinearSVR model using each individual feature at a time. This was done for both the whole dataset and the subset of sentences (i.e. 758 sentences) having a length of 16 tokens, which corresponds to the mean sentence length of the UD dataset. A subset of results is reported in Figure 2, while the whole results are provided in Appendix B. As we can see in the left-side of the heatmap, the two models share many features in the first ten positions, thus showing that the two NLM architectures are made perplexed by similar linguistic characteristics of a sentence. In particular, for both of them, the two most predictive features correspond to the lexical density and the presence of pronouns confirming the highly predictive power of morpho-syntactic information. They are followed by features related to the presence of verbs and to their internal structure (i.e. *verbal_heads* and *avg_verb_edges*), and, as it was expected, by the length of the sentence. Despite these similarities, we can see that the scores obtained by the regression model to predict BERT's perplexity are on average higher than GPT-2's scores. Considering that we obtained higher scores using all (or groups of) features in the prediction of GPT-2' perplexity (see Figure 1), this latter result may suggest that the interaction among features is less relevant in the prediction of BERT's perplexity. Differences among the two models concern features that are highly sensitive to sentence length, which result to be more predictive of BERT's perplexity. This is the case of syntactic features capturing global and local aspects of sentence structure, i.e. the depth of the whole syntactic tree (*parse_depth*), the maximum length of dependency links (*max_links_len*) and the length of verbal clauses (*clause_length*). Also, the canonical order of nuclear sentence elements such as pre-verbal subjects contribute more to predict BERT's than GPT-2's perplexity. Instead, the distribution of proper nouns (*%_upos_PROPN*), in particular in their singular form (*%_xpos_NNP*), the length of token (*char_per_tok*) and vocabulary richness are more predictive of GPT-2's perplexity. Although we cannot say from ranking results whether features highly ranked are positively or negatively correlated with perplexity, we can hypothesize that knowing the distribution of tokens belonging to open lexical categories (e.g. proper nouns vs determiners) make the perplexity easier to identify.

The right-side heatmap shows the top-ranked features used to predict the two models perplexity for sentences 16-token long. As expected, when sentence length is controlled, the role of other features less related to length becomes predominant.

In particular, morpho-syntactic information is still highly predictive for the two models, with lexical parts-of-speech showing to be relevant not only for GPT-2's but also of BERT's perplexity.

4 Conclusion

In this paper we proposed an investigation of the linguistic phenomena characterizing the perplexity of a undirectional and a bidirectional Neural Language Model, GPT-2 and BERT. We first reported robust correlations between GPT-2's perplexity and the sentence-level likelihood computed with BERT. This is a quite prominent result, especially considering that these two metrics are differently computed as a consequence of the two NLMs architectures.

Interestingly, we found the effectiveness of linguistic features modelling a wide set of morpho-syntactic and syntactic phenomena in predicting the perplexity of the two NLMs, especially for shorter sentences. Despite similar trends, we observed some differences between the two NLMs both at the level of regression accuracy and in the rankings of the features exploited in the prediction of perplexity. GPT-2's perplexity is better captured by the considered features and it resulted to be more affected by lexical parts-of-speech and features capturing the vocabulary richness of a sentence. On the contrary, BERT's perplexity seems to be best predicted by syntactic features highly sensitive to sentence length.

References

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.

Trevor Cohen and Serguei Pakhomov. 2020. A tale of two perplexities: Sensitivity of neural language models to lexical retrieval deficits in dementia of the Alzheimer's type. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1946–1957, Online. Association for Computational Linguistics.

V. Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109:193–210.

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.

Pablo Gamallo, Jose Ramom Pichel, and Iñaki Alegria. 2017. A perplexity-based method for similar languages discrimination. In *VarDial2017 workshop at EACL 2017. Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 109–114,Valencia, Spain, April 3, 2017. c©2017 Association for Computational Linguistics (http://web.science.mq.edu.au/ smalmasi/vardial4/index.html).*

M. González. 2015. An analysis of twitter corpora and the differences between formal and colloquial tweets. In *TweetMT@SEPLN*.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics, CMCL 2018, Salt Lake City, Utah, USA, January 7, 2018*, pages 10–18. 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.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

Adhiguna Kuncoro, Chris Dyer, Laura Rimell, Stephen Clark, and Phil Blunsom. 2019. Scalable syntax-aware language models using knowledge distillation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3472–3484, Florence, Italy. Association for Computational Linguistics.

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.

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

Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. 2020. Cross-linguistic syntactic evaluation of word prediction models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5523–5539, Online. Association for Computational Linguistics.

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

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

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

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

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

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

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

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*.

A Appendix A

fRaw Text Properties
[sent_length]: average length of sentences in a document, calculated in terms of the number of words per sentence
[char_per_tok]: average number of characters per word (excluded punctuation)
Vocabulary Richness
[ttr_lemma]: Type/Token Ratio (TTR) calculated with respect to the lemmata in a sentence. It ranges between 1 (high lexical variety) and 0 (low vocabulary richness)
[ttr_form]: Type/Token Ratio (TTR) calculated with respect to the word forms in a sentence. It ranges between 1 (high lexical variety) and 0 (low vocabulary richness)
Morphosyntactic information
[%_upos_*]: distribution of the part-of-speech categories defined in the Universal POS tags, as detailed at the following link: `https://universaldependencies.org/u/pos/index.html`
[%_xpos_*]: distribution of the part-of-speech categories defined in the Penn Treebank POS tags, as detailed at the following link: `https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html`
[lexical_density]: the value corresponds to the ratio between content words (nouns, proper nouns, verbs, adjectives, adverbs) over the total number of words in a sentence
Inflectional morphology
[%_aux_tense_*]: distribution of auxiliary verbs according to their tense: `https://universaldependencies.org/u/feat/Tense.html`
[%_aux_mood_dist_*]: distribution of auxiliary verbs according to their moods: `https://universaldependencies.org/u/feat/Mood.html`
[%_aux_form_*]: distribution of auxiliary verbs according to their forms: `https://universaldependencies.org/u/feat/VerbForm.html`
[verbs_gender_dist_*]: distribution of verbs according to the gender of participle forms, for the languages that have this features: `https://universaldependencies.org/u/feat/Gender.html`
[%_aux_num_pers_*]: distribution of auxiliary verbs according to their number and person: `https://universaldependencies.org/u/feat/Person.html`
Verbal Predicate Structure
[verbal_head]: average distribution of verbal heads in the document, out ot the total of heads.
[%_verbal_roo]: average distribution of roots headed by a lemma tagged as verb, out of the total of sentence roots;
[avg_verb_edges]: verbal arity, calculated as the average number of instantiated dependency links (covering both arguments and modifiers) sharing the same verbal head, excluding punctuation and auxiliaries bearing the syntactic role of copula according to the UD scheme
[verbal_arity*]: distribution of verbs for arity class (e.g. verbs with arity 1, 2, ...)
Global and Local Syntactic Tree Structures
[parse_depth]: mean of the maximum tree depths of the sentence. The maximum depth is calculated as the longest path (in terms of occurring dependency links) from the root of the dependency tree to some leaf
[clause_length]: average clause length, calculated in terms of the average number of tokens per clause, where a clause is defined as the ratio between the number of tokens in a sentence and the number of either verbal or copular head
[avg_links_len]: average number of words occurring linearly between each syntactic head and its dependent (excluding punctuation dependencies)
[max_links_len]: the value of the longest dependency link in the document, calculated in number of tokens
[prep_1]: distribution of prepositional chains 1-complement long. A prepositional chain is calculated as the number of embedded prepositional complements dependent on a noun
Relative order of elements
[%_obj_post]: distribution of objects following the verb
[%_subj_pre]: distribution of subjects preceding the verb
Syntactic Relations
[%_dep_*]: average distribution of the 37 universal syntactic relations used in UD (`https://universaldependencies.org/u/dep/index.html`)
Use of Subordination
[principal_prop_dist]: distribution of principal clauses
[%_subord_prop]: distribution of subordinate clauses, as defined in the UD scheme: `https://universaldependencies.org/u/overview/complex-syntax.html#subordination`
[subord_post]: distribution of subordinate clauses following the main clause
[avg_subord_chain]: average length of subordinate chains, where a subordinate 'chain' is calculated as the number of subordinate clauses embedded on a first subordinate clause
[subord_1]: distribution of subordinate chains 1-clause long

Table 3: Linguistic features used in the experiments.

B Appendix B

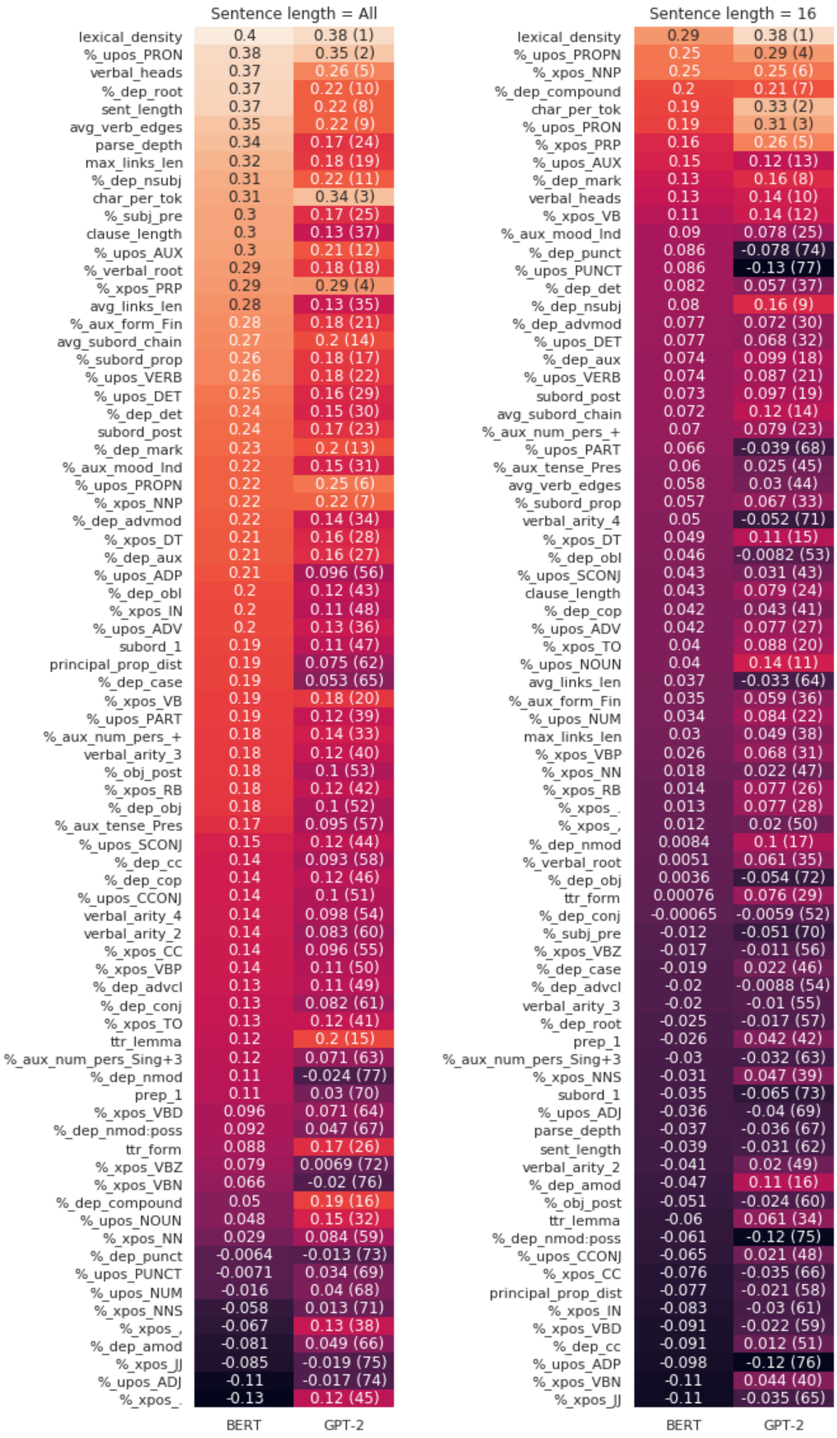

Figure 3: BERT and GPT-2 ρ scores obtained with the LinearSVR model using one feature at a time, for the whole UD dataset and sentences with lengths = 16.