Subject–Verb Agreement Precision Analysis through BERT Prediction

1. Abstract

Agreement features like person and number are significant in Indo–European language family in natural language processing. The language framework Bidirectional Encoding Representation Transformer, or BERT, pre–trained based on texts from Wikipedia is mainly used to help computer understand ambiguation by using nearby texts to build contexts. In this paper, evaluation for BERT prediction performance on subject–verb agreement(SVA), one of English syntactic rules, is focused on. The pre–trained BERT is used to predict the singularity or plurality of verb or root of each sentence. Some factors to influence the decision by BERT are also analyzed.

2. Introduction

Grammatical Error Detection(GED, Leacock et al., 2014) is a common see, and is much important in real applications including writing assistant tools, self–assessment frameworks, language tutoring systems, machine translation, etc. GED mainly result from wrong syntactic parsing, irregular or exceptional grammatical rules, and incorrect labeling (Flachs et al., 2019).

Since Chomsky (1957), data like this has been taken as evidence that natural language grammars contain abstract syntactic rules that (i) are independent of the meaning of lexical items and (ii) obey hierarchical, rather than linear constraints. Number agreement (henceforth NA) between the subject (the cue) and the verb (the target) of the same clause in English is one of such rules.

1) The person who looks at those students is a kind man.
2) The person who looks at those students are a kind man.

The two examples above show SVA problem. The example 1) is a correct one while 2) is incorrect. “students” is not the real subject, but the the nearest noun to the root verb. In this case, the language models may easily predict the wrong root verb due to the distance between noun and root verbs.

In this case, many researches and studies have been done to explore the extent of language models to solve SVA problems and optimize them based on findings. Generally, there are basically three approaches on this issue: rule–based approaches, neural networks approaches, and the combination of these two approaches.

Rule–based approaches have been applied to solve this kind of problems, which mainly rely on the parsed syntactic trees. The recent neural networks approaches also have been used

including Recurrent Neural Networks (RNNs), Transformers–based models with large amount of data. Besides, the miscellaneous method is also adapted to SVA.

In this paper, BERT is used to predict root or main verb of each sentence on the number concerning singularity and plurality, so that its potential capacity on SVA can be evaluated. There are some findings to be provided to evaluate BERT performance considering several factors such as number of nouns, dependency distance between root verb and subject, number of clauses before root verb, etc.

3. Work Related

Rule–based approaches have been applied to solve this kind of problems, which mainly rely on the parsed syntactic trees. However, to parse the syntactic trees is easy to have errors in both manual pre–processing and automatic tools such as SpaCy.

The recent neural networks approaches also have been used. SVA error detection, one subtask of GED, is one example, and has been done with Recurrent Neural Networks (RNNs) trained on large amount of data. Also, the concept of attractors are introduced (Linzen et al., 2016). However, the public source data is small and the sentences given whose grammatical rules are regular, which may lead to overfitting. Gulordava et al. (2018) also consider subject–verb agreement, but in a “colorless green ideas” setting in which content words in naturally occurring sentences are replaced with random words with the same part–of–speech and inflection, thus ensuring a focus on syntax rather than on selectional–preferences based cues. Marvin and Linzen (2018) consider a wider range of syntactic phenomena (subject–verb agreement, reflexive anaphora, negative polarity items) using manually constructed stimuli, allowing for greater coverage and control than in the naturally occurring setting.

The BERT model is based on the “Transformer” architecture (Vaswani et al. , 2017), which—in contrast to RNNs—relies purely on attention mechanisms, and does not have an explicit notion of word order beyond marking each word with its absolute–position embedding. This reliance on attention may lead one1 to expect decreased performance on syntax–sensitive tasks compared to RNN (LSTM) models that do model word order directly, and explicitly track states across the sentence. Indeed, Tran et al. (2018) finds that transformer–based models perform worse than LSTM models on the Linzen et al. (2016) agreement prediction dataset. In contrast, (Tang et al., 2018) find that self–attention performs on par with LSTM for syntax sensitive dependencies in the context of machine–translation, and performance on syntactic tasks is correlated with the number of attention heads in multi–head attention. Goldberg (2019) used evaluation protocol and stimuli to the bidirectional setting required by BERT, and evaluate the pre–trained BERT models on SVA tasks. Although Transformer–based models show good performances, but their generalization cannot be understood. BERT has been examined the what extent it can perform lexically–independent subject–verb number agreement from the perspective of dependency distance.

3. Collected Data

From the previous studies, the data is generated from an automatic way with very some amount of data, i.e. the data is created with list of limited number of nouns, and verbs with Python. Or, the data is from previous studies. Some data is created originally, but the amount is quite small less than 100 sentences. Therefore, based on these three existing methods to have data and to ensure the originality of this research, I take the data totally from a new corpus name British National Corpus (BNC).

The data is all taken from BNC(British National Corpus), specifically, BNC XML Edition (2007). The XML Edition of the BNC contains 4049 texts and 96986707 words, covering different genres of source texts including written texts and spoken texts. The source materials in this corpus are selected from the time period from the year of 1960 to 1993. From this corpus, the .xml files can help recognize the part-of-speech tagging, but there is annotations for subject and root verb of each sentence. Therefore, once the sentences are selected according to the requirements, the .xml-format sentence is converted into .txt-format, so that subject and root verbs can roughly be detected with the package named SpaCy. With the specific regular expressions, around 10000 sentences are selected out from the big corpus, with 50% of sentences featuring singular subject matching singular verb, and the rest of 50% featuring plural subject matching plural verb. There are four kinds of regular expressions used for sentence selection. The regular expression examples are as follows:

(1) Regex1 = \ndet.+ nsubj.+ prep.+ pobj.+"
(2) Regex1 = "\nnsubj.+ prep det pobj.+"

(a) Sentence1 = "The rising threshold of competence needed in the job market and the relative decline in traditional semi-skilled or unskilled jobs means that the compulsory school can no longer hope to provide a marketable vocational education as it did for some usually lower-achieving children in the past"
(b) masked_Sentence1 = "The rising threshold of competence needed in the job market and the relative decline in traditional semi-skilled or unskilled jobs [MASK] that the compulsory school can no longer hope to provide a marketable vocational education as it did for some usually lower-achieving children in the past"
(c) Sentence2 = "the physical properties of natural gas require the compression ratio of the engine to be higher than in normal internal combustion engines , and the higher compression makes for greater efficiency "
(d) masked_Sentence2 = "the physical properties of natural gas [MASK] the compression ratio of the engine to be higher than in normal internal combustion engines , and the higher compression makes for greater efficiency"

Sentence1 is an example retrieved from the database with Regex1, while Sentence2 is selected out from the corpus with Regex2.

Most of sentences have attractors which serve as misleading nouns before root verbs to affect the decisions by BERT. For example, in Sentence1 whose agreement number is singular, "means" is the root verb of the sentence, and before it, there are several nouns which are singular and plural. In this list of nouns, "jobs" is plural before " means"m and it may take the influence to BERT's judgements on whether "means" is given higher prediction probability or "mean". Likewise in Sentence2 which belongs to the second category of

plurality, “properties” is the subject or nsubj, and “gas” is one attractor in this sentence; the language model may produce the result “requires” due to the attractor “gas” instead of predicting the correct root verb “require”. Therefore, it is clear to see that BERT performance is evaluated with these attractors in every sentence. Amongst these datasets, there are zero to four attractors in each sentence, and the following data analyses is made in the next section.

When the experiment data is prepared, the root verbs of sentences are masked, i.e. the root verb is replaced by “[MASK]”, and examples (b) and (d) can be referred. The part of “[MASK]” is what is to be predicted by BERT. The existing pre-trained model BERT is trained based on Wikipedia corpus. In this study, two versions of the model are used: “bert-large-uncase” and “bert-base-uncase”. The differences of these two versions lie in some parameters. “uncase” means the model does not distinguish the upper and lower cases. The models finally give out the probabilities of correct verbs and wrong verbs through the softmax function, and the higher probability of the predicted words are the results appended to the processed data sheets. The use of two versions of the model is to compare which can There are also some other factors which may affect BERT’s prediction decisions to be analyzed.

1. The length of sentences, or the number of tokens in one sentence (SL): from the example (1) and (2), there are 46 tokens in Sentence1, and 32 tokens in Sentence2.
2. The distance between subject and root verb (SVD), i.e. how many words between subject and root verb (root verb’s index minus subject’s index): SVD is 18 in Sentence1 and 4 in Sentence2. The hypothesis is that the longer the distance is, the lower the prediction accuracy will be. Therefore, it can be one of most important factors for BERT’s performance.
3. The number of non-subject nouns before root verb (All_NN_BEF): the number of nouns before root verb is 6 in Sentence1, including “competence”, “job”, “market”, “decline”, and “jobs”, and is 2 in Sentence2, including “gas”.
4. The number of verbs which are lemmatized tokens before root verb (NV_BEF): the verb number is 2 in Sentence1, including “rising” and “needed”, and is 0 in Sentence2. This way roughly help monitor how many sub-sentences there is in one sentence. More sub-sentences can lead to more complex syntactic structures.
5. The number of tokens before root verb (SL_BEF): the length of part of sentence before root verb is 20 in Sentence1, and is 6 in Sentence2.
6. The number of non-subject nouns whose number is opposite to that of subject before root verb (Attractor): in Sentence1, the non-subject plural noun is “jobs”, and the number of Attractor is 1; in Sentence2, the non-subject singular noun is “gas”, and the number of Attractor is 1.
7. The distance between root verb and the non-subject noun which is nearest among All_NN_BEF to root verb (NEAR_NVD): in Sentence1, the non-subject noun nearest to root verb is “jobs”, and the distance is 1; in Sentence2, the nearest noun is “gas”, and the distance is 1.

The whole datasets prepared and codes are available at https://github.com/louisefz/SVA/tree/data

4. Experiment Result Analyses

The effect of factors mentioned in the last section will be analyzed.

4.1 “bert_large_uncase” vs “bert_base_uncase”

“bert_large_uncase” and “bert_base_uncase” are two versions of BERT, with “bert–large–uncase” having 24–layer, 1024–hidden, 16–heads, 340M parameters, and “bert–base–uncase” having 12–layer, 768–hidden, 12–heads, 110M parameters. These two versions are applied into this study to compare which version can deliver better performances.

	Correct_root	Wrong_root	Accuracy
“bert_large_uncase”	8175	1825	81.75%
“bert_base_uncase”	8277	1723	82.77%

Figure 1. “bert_large_uncase” vs “bert_base_uncase”

Overall, “bert_base_uncase” performs slightly better than “bert_large_uncase” in this study.

4.2 Singularity vs Plurality

As there are singular and plural sentences with 5000 of each, the BERT’s performance on predicting whether singular verbs and plural verbs based on the number of subject can be evaluated.

	pl_correct	pl_wrong	sg_correct	sg_wrong
“bert_large_uncase”	4393	607	3782	1218
“bert_base_uncase”	4472	528	3805	1195

Figure 2. Singularity vs Plurality

Data trained by two versions of BERT shows that the accuracy of predicting the sentences’ root verb in plural files is higher than that in singular files. There are two hypotheses: 1) the model is better at predicting plural verbs that singular verbs; 2) plural data and singular data is not identically even due to the data selection and manual annotation and modification which is caused by the classification errors by SpaCy.

4.3 The length of sentences, or the number of tokens in one sentence (SL)

The number of tokens of each sentence can be seen as a potential factor to influence BERT's prediction.

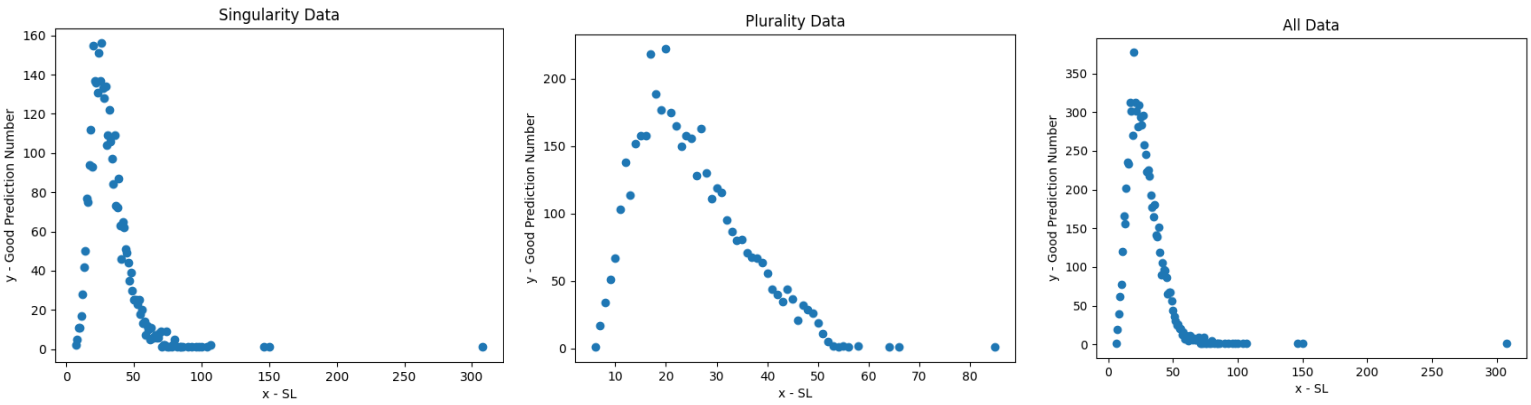

Figure 3. Correct prediction trends based on " bert_large_uncase

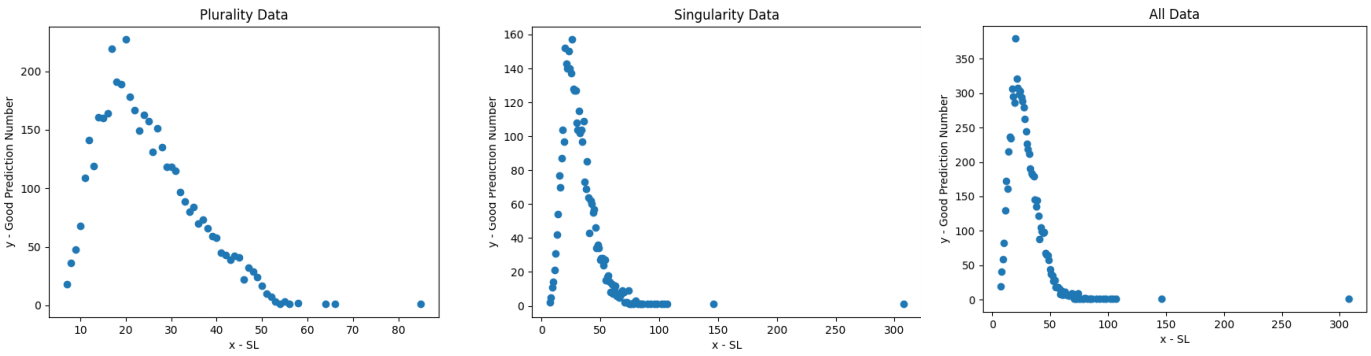

Figure 4. Correct prediction trends based on "bert_base_uncase"

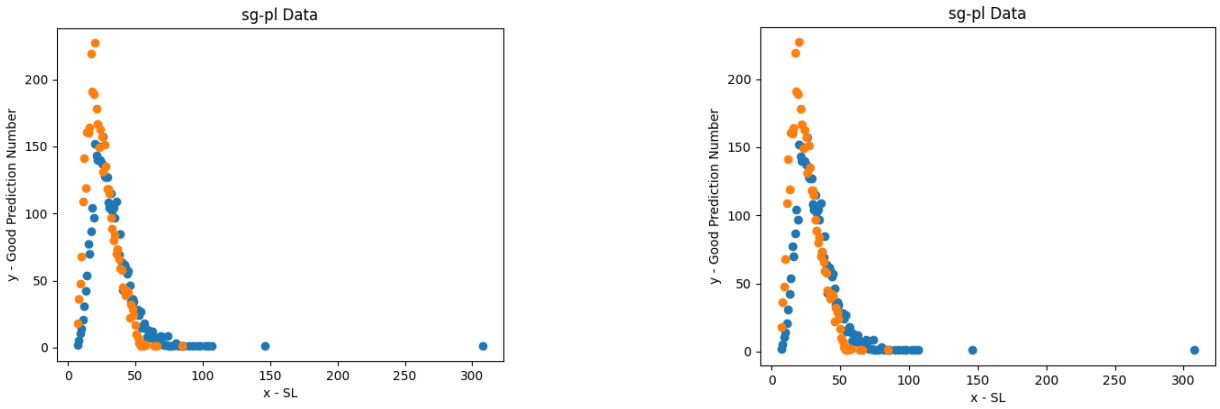

Figure 5. SL coverage between Singularity(blue points) and Plurality(yellow points) based on “bert_large_uncase” and “bert_base_uncase” respectively

From this six graphs which share the same tendency shows that the there is a strong connection between the length of each sentence. From the table below, it shows that when the length of the sentence is around 17 in plural sentences, the correct root verb can be correctly predicted by BERT with the best performance, and the best SL is around 24. Plurality’s number of correct prediction at the peak is twice of Singularity. Overall, BERT performance can be reached when SL is 20.

	pl_peak SL	sg_peak SL	all_peak SL
“bert_large_uncase”	(16,250)	(23.160)	(20,370)
“bert_base_uncase”	(18,250)	(24,160)	(20,370)

Figure 6. SL at the peak value

4.4 The distance between subject and root verb (SVD)

The hypothesis for this factor is that the shorter the SVD is, the more possible the model can predict the correct result.

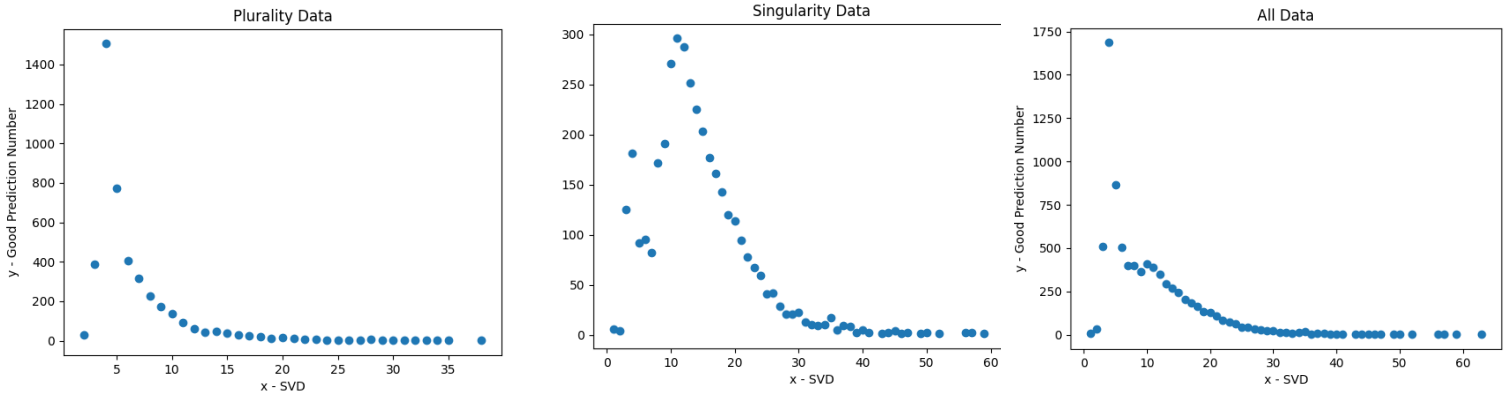

Figure 7. SVD scatter plot based on “bert_large_uncase”

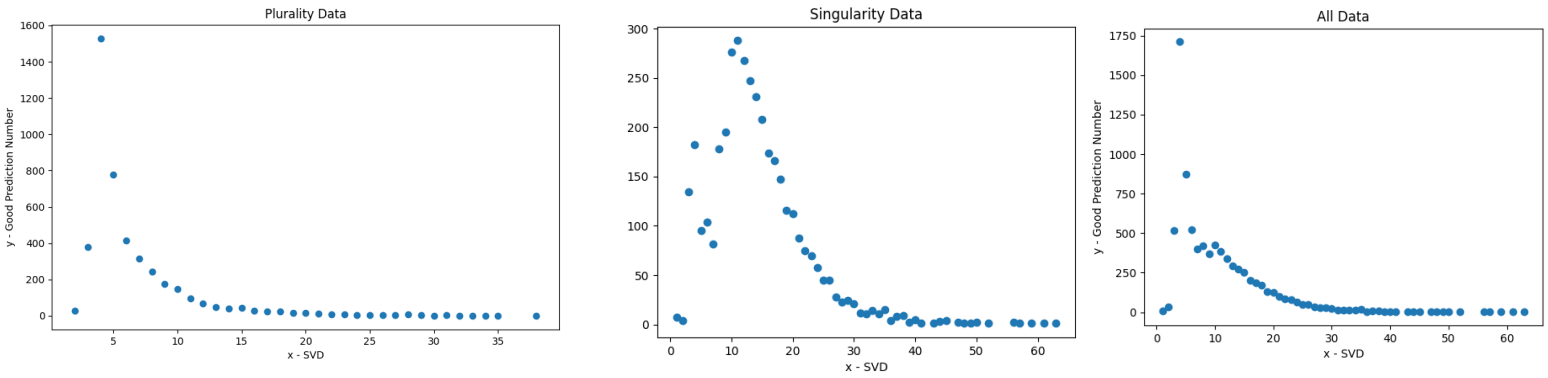

Figure 8. SVD scatter plot based on “bert_base_uncase”

The two versions of SVD graphical tendency is nearly identical. However, there is significant difference between the factor of singularity and plurality. The peak values are different with plural sentences at 5 and singular at 10. The best correct prediction for singularity when SVD ranging from 5 to 20, while for plurality when SVD ranging from 2 to 7.

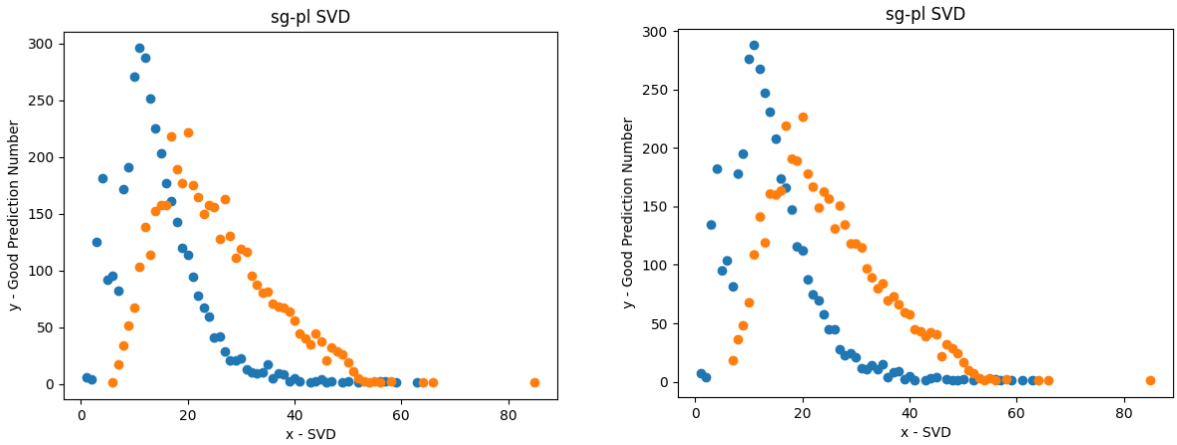

Figure 9. SVD coverage between Singularity(blue points) and Plurality(yellow points) based on “bert_large_uncase” and “bert_base_uncase” respectively

	pl_peak SVD	sg_peak SVD	all_peak SVD
“bert_large_uncase”	5,1560)	(10,290)	(5,1740)
“bert_base_uncase”	(5,1560)	(10,290)	(5,1740)

Figure 10. SVD at the peak value

4.5 The number of non-subject nouns before root verb (All_NN_BEF)

The number of non-subject nouns before the root verb in each sentence is seen as one factor by the author to affect the BERT's prediction. It is supposed that less All_NN_BEF can contribute better results.

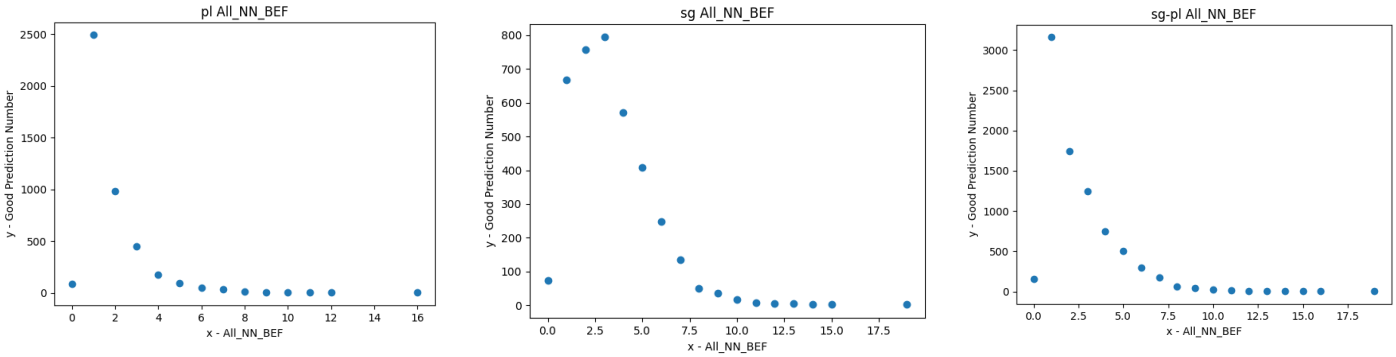

Figure 11. All_NN_BEF scatter plot based on "bert_large_uncase"

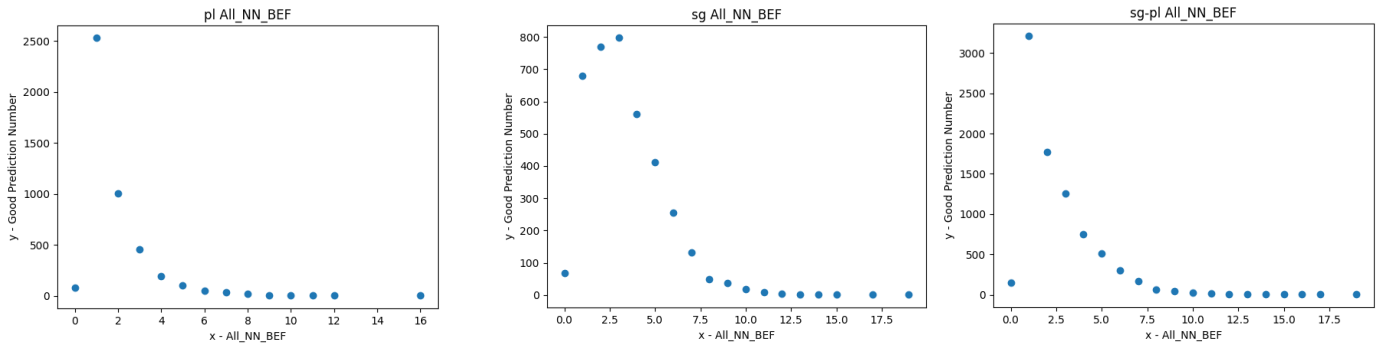

Figure 12. All_NN_BEF scatter plot based on "bert_base_uncase"

	pl_peak All_NN_BEF	sg_peak All_NN_BEF	all_peak All_NN_BEF
"bert_large_uncase"	(1,2500)	(3,800)	(1,3200)
"bert_base_uncase"	(1,2510)	(3,800)	(1,3200)

Figure 13. All_NN_BEF at the peak value

According to plots and table above, that the number of non–subject nouns before root verb is around 1 can result the best results.

4.6 The number of verbs which are lemmatized tokens before root verb (NV_BEF)

The selected sentences are designed to have several sub–sentences or clauses, and therefore, the emergence of verbs before the root verb can roughly be seen as the representation for one sub–sentence or clause. The hypothesis for this factor is that the less the sub–sentence there is, the more accurate the result can be.

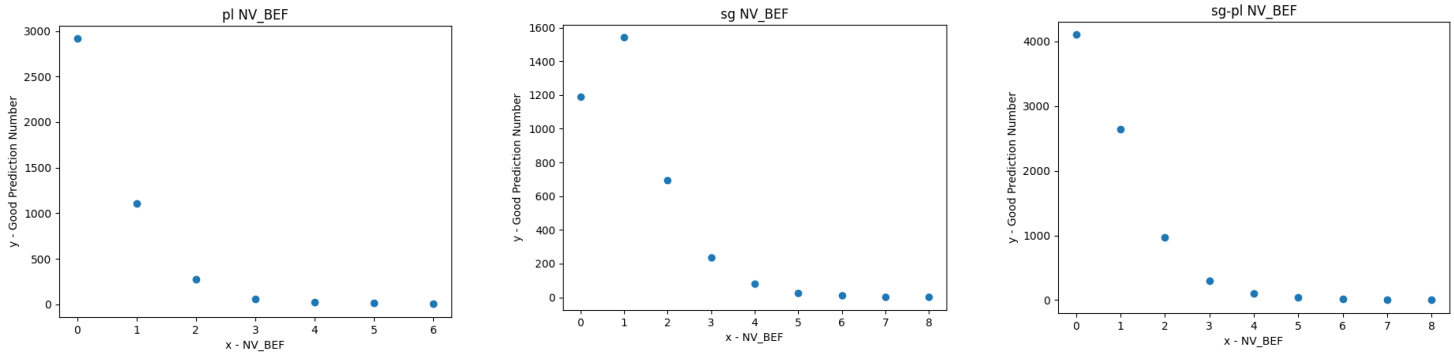

Figure 14. NV_BEF scatter plot based on “bert_large_uncase”

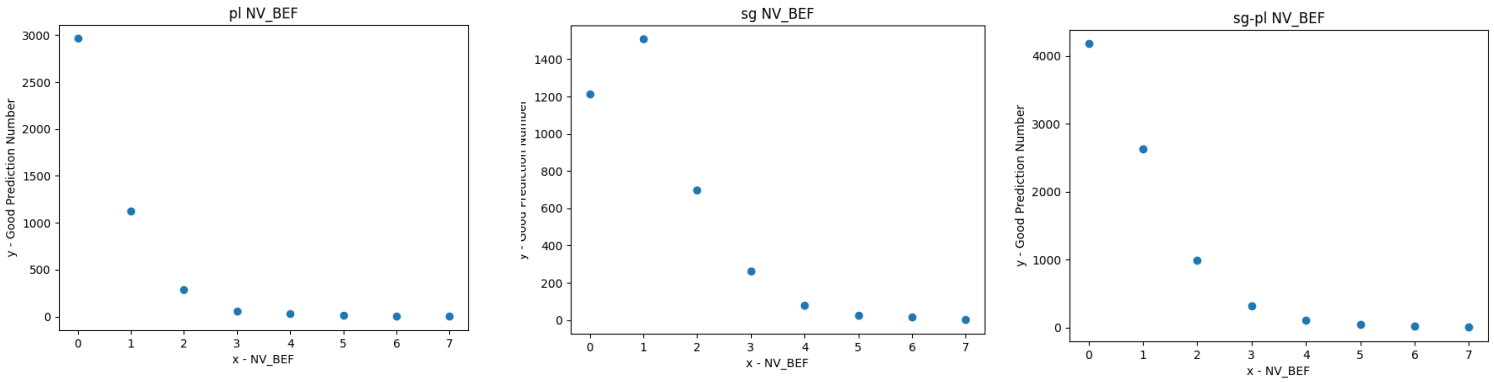

Figure 15. NV_BEF scatter plot based on “bert_base_uncase”

	pl_peak NV_BEF	sg_peak NV_BEF	all_peak NV_BEF
“bert_large_uncase”	(0, 3000)	(1,1580)	(0,4100)
“bert_base_uncase”	(0, 3000)	(1,1500)	(0,4200)

Figure 16. NV_BEF at the peak value

The overall tendency is identical to my hypothesis: less NV_BEF leads to higher accuracy.

4.7 The number of tokens before root verb (SL_BEF)

The logic of this factor is similar to the factor SL, but a more closer and detailed look is given to it, as only the number of tokens before root verb is to be looked into.

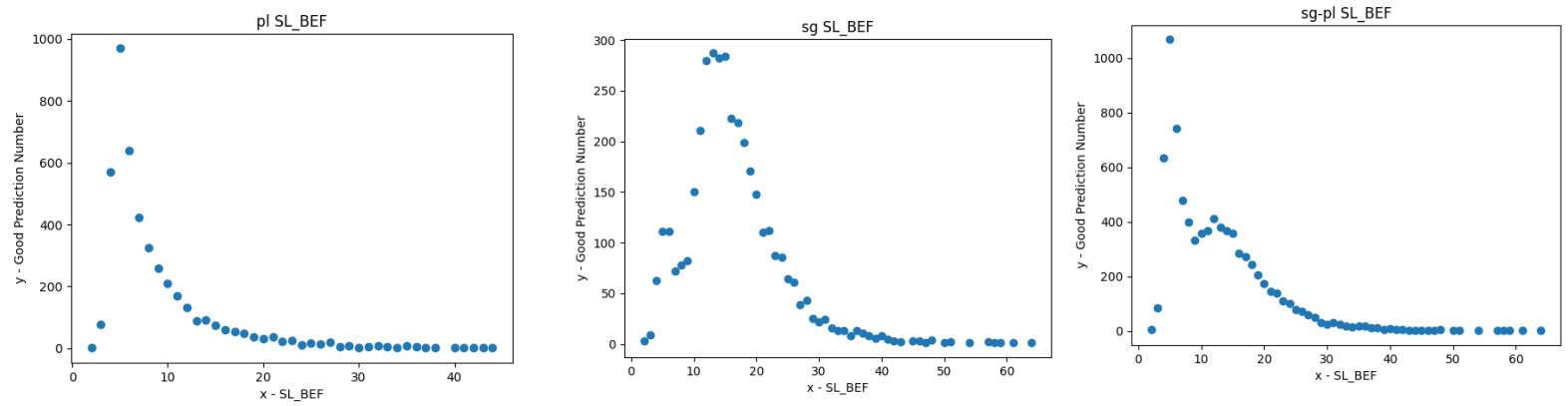

Figure 17. SL_BEF scatter plot based on “bert_large_uncase”

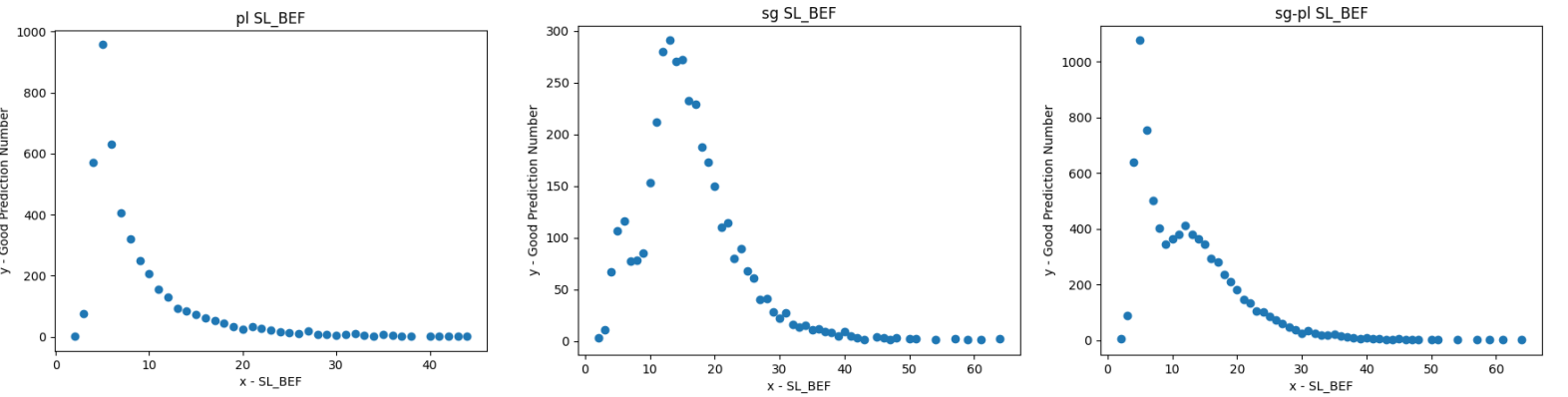

Figure 18. SL_BEF scatter plot based on “bert_base_uncase”

	pl_peak SL_BEF	sg_peak SL_BEF	all_peak SL_BEF
“bert_large_uncase”	(5,980)	(13,280)	(5,1100)
“bert_base_uncase”	(5,980)	(13,290)	(5,1100)

Figure 19. SL_BEF at the peak value

The conclusion from the analyzed data above is similar to that of SL. That the overall number of token is 5 can result in the best performances. When the number grows after the peak value, the precision for predictions by BERT falls sharply. However, the number of tokens of singular sentences which is 13 is more than that of plural ones.

4.8 The number of non–subject nouns whose number is opposite to that of subject before root verb (Attractor)

The hypothesis for this factor is easily prone to be that more Attractors will make the bad results, since BERT is distracted by more opposite–number nouns.

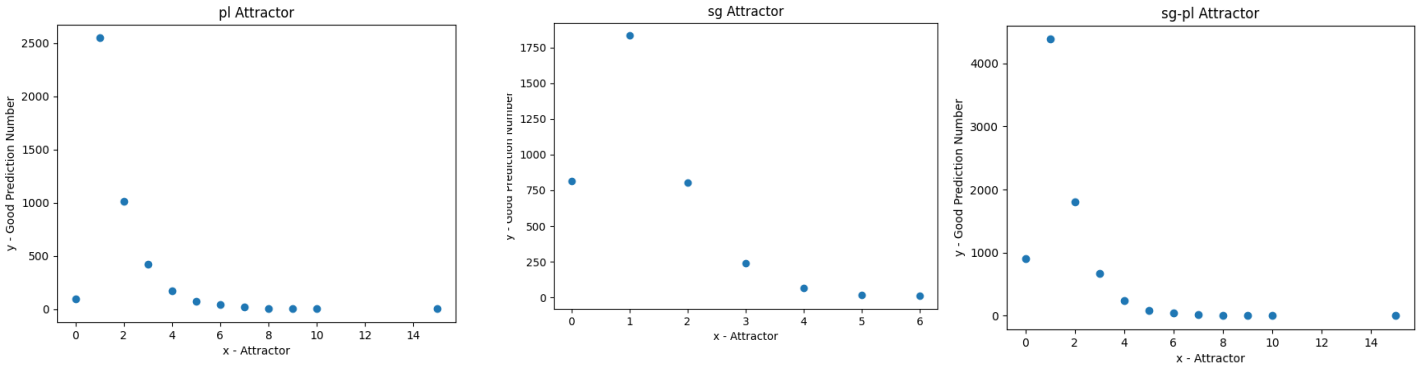

Figure 20. Attractor scatter plot based on “bert_large_uncase”

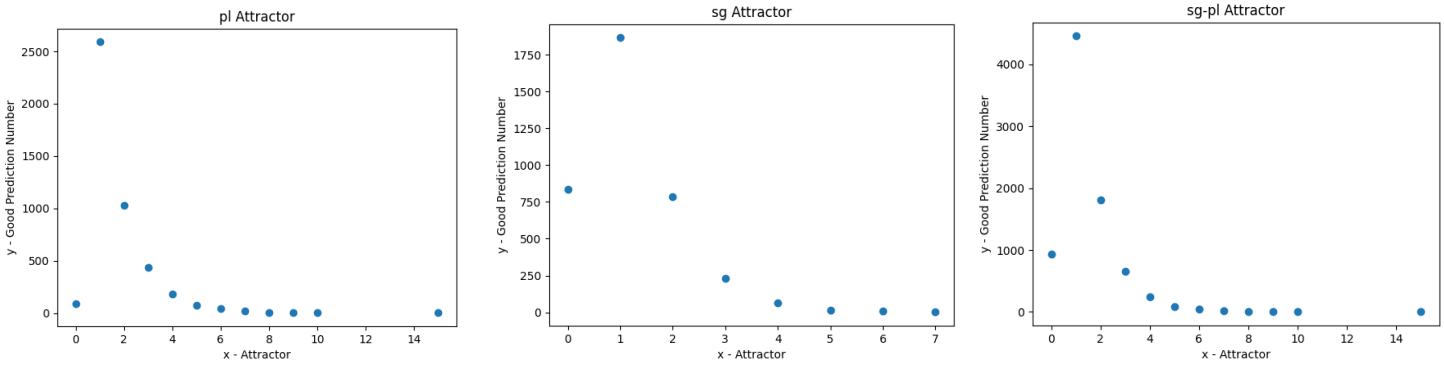

Figure 21. Attractor scatter plot based on “bert_base_uncase”

	pl_peak Attractor	sg_peak Attractor	all_peak Attractor
“bert_large_uncase”	(1,2600)	(1,1800)	(1,4400)
“bert_base_uncase”	(1,2600)	(1,1800)	(1,4400)

Figure 22. Attractor at the peak value

4.9 The distance between root verb and the non-subject noun which is nearest among All_NN_BEF to root verb (NEAR_NVD)

The hypothesis for this factor is that the nearer the distance between the nearest non-subject noun and root verb, the less accurate the prediction will be, as the nearest noun can be the most important attractor.

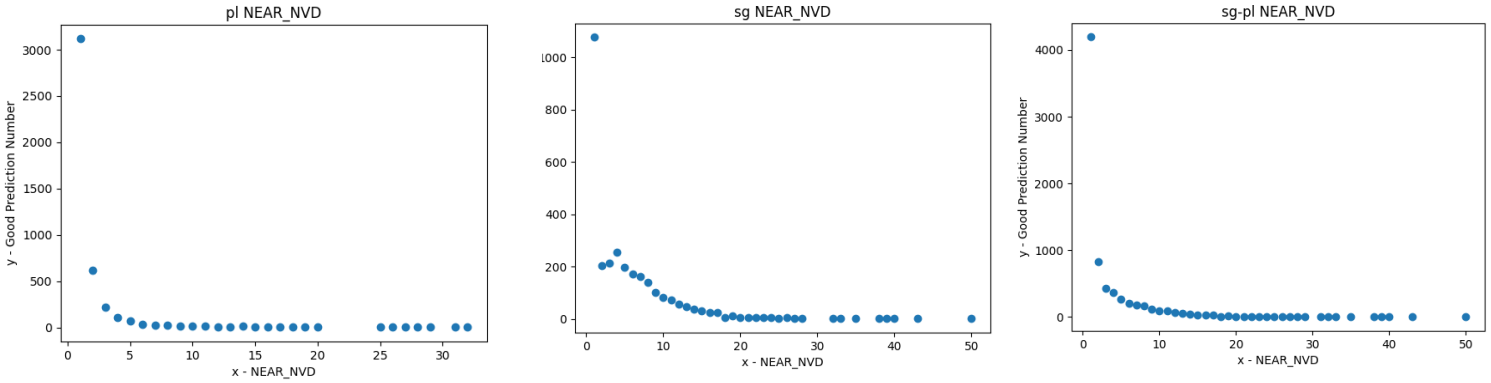

Figure 23. NEAR_NVD scatter plot based on “bert_large_uncase”

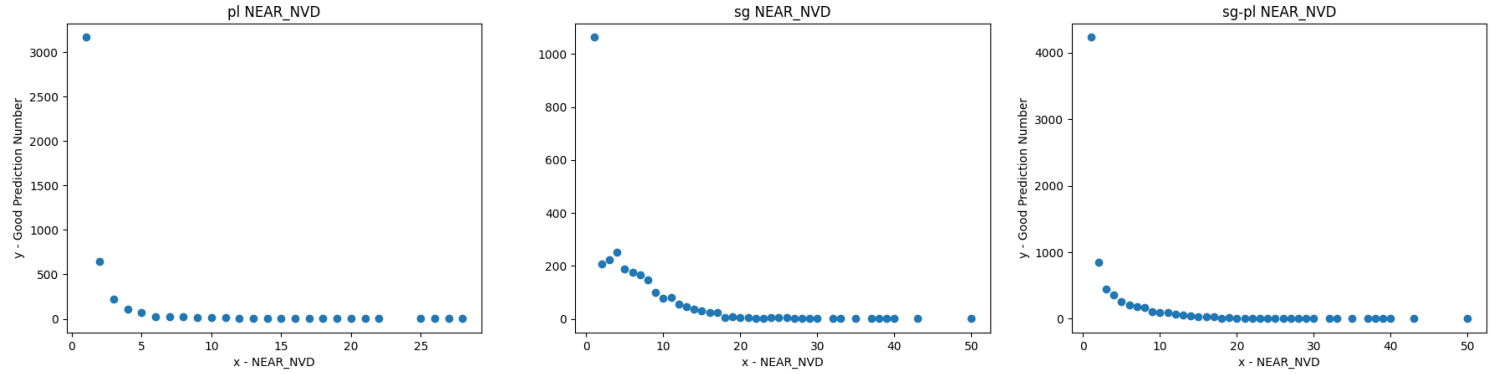

Figure 24. NEAR_NVD scatter plot based on “bert_base_uncase”

	pl_peak NEAR_NVD	sg_peak NEAR_NVD	all_peak NEAR_NVD
“bert_large_uncase”	(1, 3100)	(1, 1100)	(1, 4200)
“bert_base_uncase”	(1, 3100)	(1, 1100)	(1, 4200)

Figure 25. NEAR_NVD at the peak value

From the graphs above, the fact is just the opposite the hypothesis. When the non-subject noun is just before the root verb, the number of correct prediction reaches the highest number. Beside, when the distance between the nearest non-subject noun and root verb grows, the difference on good prediction rate does not decline sharply. The sentences of plurality and the overall data show that except for when the distance is 1, there is no difference among different good prediction numbers.

5. Conclusions

This study uses BERT, one kind of language model to predict the root verb of the sentence based on the subject-verb agreement in English. From the previous studies and data analyses, some conclusions can be reached.

In this study, two versions of BERT are used: "bert-large-uncase" and "bert-base-uncase". These two versions of model have different interior parameters, which may create in different results. In this research, "bert-base-uncase" has a slightly better performance than "bert-large-uncase".

Besides, the pre-processed data has been designed to have half-half singular and plural sentences. However, BERT's prediction performance on the sentences of plurality is better than those of singularity. There are two hypotheses: (1) the data is not evenly-scattered; (2) BERT's performance on plurality prediction is better than singularity.

As for the rest of factors, some of them show strong significance. The number of tokens of the whole sentence (SL), the number of tokens before root verb (SL_BEF) as well as SVD share similarities that when the number of token reaches 20, 5 and 5 respectively the performance can get the best result, and the tendencies are identical: rising to the peak and falling sharply. Beside, some factors show the best result at the first beginning when the x-axis number is near zero or one, such as The number of non-subject nouns before root verb (All_NN_BEF), the number of verbs which are lemmatized tokens before root verb (NV_BEF), the number of non-subject nouns whose number is opposite to that of subject before root verb (Attractor). All of these factors mentioned fits the author's hypotheses. However, the last factor, i.e. the distance between root verb and the non-subject noun which is nearest among All_NN_BEF to root verb (NEAR_NVD). The author's hypothesis is quite opposite the fact that the shorter distance can create better results.

Reference

el Tetreault. 2010. Automated Grammatical Error Correction for Language Learners. Synthesis lectures on human language technologies, 3(1):1— 134.

Marek Rei and Helen Yannakoudakis. 2016. Compositional Sequence Labeling Models for Error Detection in Learner Writing. In Proceedings of ACL 2016.

Marek Rei and Helen Yannakoudakis. 2017. Auxiliary Objectives for Neural Error Detection Models. In Proceedings of BEA 2017.

Øistein E. Andersen, Helen Yannakoudakis, Fiona Barker, and Tim Parish. 2013. Developing and testing a self–assessment and tutoring system. In Proceedings of BEA 2013.

Joel Tetreault, Claudia Leacock. 2014. Automated Grammatical Error Correction for Language Learners. Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Tutorial Abstracts, pages 8—10, Dublin, Ireland, August 23–29 2014.

Simon Flachs, Ophélie Lacroix, Marek Rei, Helen Yannakoudakis, Anders Søgaard. 2019. A Simple and Robust Approach to Detecting Subject–Verb Agreement Errors. Proceedings of NAACL-HLT 2019, pages 2418–2427 Minneapolis, Minnesota, June 2 - June 7, 2019. c 2019 Association for Computational Linguistics

Noam Chomsky. 1957. Syntactic Structures. Mouton, The Hague.

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 NAACL–HLT, pages 1195—1205.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencieTransactions of the Association for Computational Linguistics, 4:521–535.

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.

Gongbo Tang, Mathias Mller, Annette Rios, and Rico Sennrich. 2018. Why self-attention? a targeted evaluation of neural machine translatioIn Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4263–4272.

Ke Tran, Arianna Bisazza, and Christof Monz. 2018. The importance of being recurrent for modeling hierarchical structureIn Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4731–4736.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. CoRR, abs/1901.05287.