What noise reveals about RoBERTa's capacity to process language

Thomas Maher (tm3566@nyu.edu)

Sagar Patel (skp327@nyu.edu)

Aline-Priscillia Messi (alinepriscillia.messi@nyu.edu)

Abstract

We study the effect of introducing noise to a question-answering dataset using a Masked Language Model framework. Our aim is to inspect the degree to which the model shows human-like behaviour when comprehending common natural language understanding tasks. In particular, we're interested in the model's ability to utilize high level information, such as syntactical structure, to make predictions in the task. In order to investigate the model's sensitivity to the syntax and structure of language, we intentionally injected noise to both specific parts of speech and distinct positions within a sentence. Our results show that model performance was mostly a function of the position and quantity of noise. Noise introduced randomly at the end of the sentence and with a higher probability, had more of an effect on model performance than when it was targeted to a specific part of speech.

Keywords: RoBERTa; argument structure; noise

Introduction

Current approaches to language comprehension model it as the interplay between bottom-up (perception of sensory input) and top-down (prior linguistic knowledge) information. The introduction of noise is then thought to modulate this equilibrium as higher signal to noise ratios lead to a greater dependence on top-down information in order to properly parse the input: experiments with vocoded speech have shown that lexical information aides perceptual learning (Davis et al., 2005) as well as comprehension (Corps & Rabagliati, 2020).

Additionally, inspired by probabilistic theories of language comprehension, noisy channel models of speech comprehension (see Levy, 2008 for example) propose that the natural uncertainty of the input is offset by prior top-down information in order to predict the actual sequence. These models are rational as comprehenders will bias their comprehension of the input sequence in line with their prior beliefs influenced by factors such as plausibility (Levy, 2011).

Gibson et al. (2013) confirm the psychological reality of this approach in a series of experiments manipulating the insertion and deletion of material in sentence strings. They additionally find that literal syntactic interpretations are favored over rational interpretations based on plausibility for large edit distances.

NLP and language models

Machine language processing Computational models for natural language processing (NLP) have made huge advances in recent years. While machine learning approaches to NLP may be approaching human level accuracy in some tests (Nangia & Bowman, 2019), the types of information these models use for processing language is not clearly understood. Although there is some evidence to indicate that recent NLP models are able to capture some level of syntactic structure, the strength of these abilities seems to be limited (Linzen et al., 2016; Tran et al., 2018).

NLP models Linzen and Leonard (2018) found that while a simple recurrent neural network performs similarly to humans when determining the correct form of verbs in context, the error patterns differed from human patterns in important ways. Specifically, as sentence complexity increased, RNN error rates went up while human error rates decreased, potentially suggesting a lack in the model's ability to use high-level information in a sentence. More complex models such as LSTMs have increased the overall performance of recurrent models (Hochreiter & Schmidhuber, 1997). However, these models still lack the ability to capture complex sentences and syntactic structure.

Linzen et al. (2016) used subject-verb agreement tasks as a measure of an LSTM's ability to capture hierarchical syntactic structure and found that the model performed well only when provided explicit supervision for the task. This is likely due to the fact that even though LSTMs are able to track information in a way that is akin to a model of memory, they still rely on sequential structure. This limits their ability to capture hierarchical information in a sentence.

Transformer-based architectures such as BERT do not rely on sequential processing (Devlin et al., 2018; Vaswani et al., 2017). BERT utilizes a "masked language model" (MLM) pre-training objective. The MLM process uses the context of a word to determine its meaning, which removes the dependency on a unidirectional language model framework.

Transformer models are often used pre-trained and usually require fine-tuning in order to be primed for a given task. Devlin et al. (2018) argue that the choice of model architecture during pre-training is limited due to the fact that standard language models are unidirectional and introduce BERT as a solution. Despite the increase in the model's ability to approach human-level performance on some NLP tasks, the model still shows some pitfalls (Chernyavskiy et al., 2021).

Because of their State Of The Art performance on NLU (natural language understanding) tasks, transformer-based models have been assumed to capture key aspects of language comprehension. Adopting a noisy channel model of language comprehension as presented above, our paper examines the breakdown of model comprehension in noisy settings and to what extent these models are humanlike in their behavior. We introduced noise in either a positional or syntactically-

targeted manner in order to evaluate if the model was picking up and relying on subjacent syntactic structure in question-answer pairs.

Related work

BERT and noise Previous work investigating the robustness of BERT under noisy conditions includes a study by Sun et al. (2020) who reveal a dependency on word-spelling in BERT. They tested BERT in several experiments after adding typos to words in the input sentences. In a question answering task, BERT showed a decrease in accuracy even after adding typos to a single word in some cases.

Zhu et al. (2022) compare the classification accuracy of BERT when faced with injected and weakly supervised noise at the label level. While the model is robust against injected noise (showing a decrease of less than 4% in accuracy across noise levels), performance drops as low as 35% for weak supervision noise.

Closer to the scope of our paper, Kumar et al. (2020) evaluate the performance of BERT by introducing noise to datasets on the character level modeled after typos. They both test and train on noisy datasets and find a proportional decrease in accuracy as a function of percentage of total noise.

BERT and syntax Ettinger (2020) revealed valuable insights into the types of information that BERT uses in psycholinguistic tasks. In particular, this paper showed that, among many strengths and weaknesses, BERT fails at registering negation in sentences. Its predictions on a "cloze" task, in which the model returns the k most likely words for a missing word in a sentence, are identical for pairs of sentences in which the only difference is that the missing word is negated. For example, its predictions were the same for the sentences "A robin is a ___" and "A robin is not a ___."

Goldberg (2019) evaluates BERT's syntactic abilities in the case of subject-verb agreement by replicating the stimuli used in Gulordava et al. (2018), Linzen et al. (2016), and Marvin and Linzen (2018) to evaluate the syntactic capabilities of LSTM models. He evaluates predictions of masked verbs when the model is trained on various agreement structures. The high accuracy (above 80%) for stimuli of varying syntactic complexity suggests that BERT is able to capture both syntactic hierarchies and dependencies. The performance of the model across these tasks was consistently near human-level accuracy. In most cases, it performed at least as well as LSTM models. While the results indicate that BERT is able to perform the subject-agreement task well, they do not indicate what type of information the model uses to make its predictions.

Additionally, Jawahar et al. (2019) find that different layers of BERT encode varying degrees of hierarchical linguistic information as well lower layers encoding phrasal information [1]. They evaluate layer performance on different probe tasks (taken from the Conneau et al., 2018 benchmark) such as predicting sentence length or word content for surface information, sensitivity to violations of word order and more specifically bigram shift for syntactic information, and predicting the number of the main clause subject or object for semantic information. Highest performance is obtained for bottom layers for surface information, middle layers for semantic information, and top layers for syntactic information.

[1] This was observed through extracting span representations from each layer

Our approach

While most papers focus on testing the robustness to noise in order to improve accuracy and performance, we have chosen to use noise as a means of seeing if the model exhibits human-like behavior and ultimately evaluate the structures facilitating the model's comprehension of language. Our paper aims to evaluate the degree to which BERT captures underlying syntactic information in noisy settings. If BERT does use such information, then we expect performance to be lesser in conditions where noise would prevent the elaboration of reliable top-down predictions in humans. More specifically, we would expect a difference between adding random noise throughout the sentence to adding positionally targeted noise as well as a larger breakdown in performance depending on the parts of speech that noise is added to. This decrease would be informed by the relative importance of certain syntactic categories, such as verbs or nouns, to overall argument structure.

Methods

Dataset

We used a sub-sample of the QNLI Dataset (Warstadt et al., 2018) which is composed of question-and-answer pairs from text that was scraped from Wikipedia. Each pair is associated with a binary output label indicating if the question-and-answer pair contained a natural language inference. Training was done with a sub-sample of the dataset consisting of 1051 positive and 1098 negative samples.

An example training sentence from this dataset is:

Text Input: The show starred Ted Danson as Dr. John Becker, a doctor who operated a small practice and was constantly annoyed by his patients, co-workers, friends, and practically everything and everybody else in his world. Who starred in 'True Love'?

Label Output: 1

Introduction of Noise

Word-level In order to create a noisy dataset, we injected noise into our test set. Instead of randomly introducing noise to text, we created noise in line with Kumar et al. (2020) by replacing a random character in a given word with one of its adjacent keys on a Qwerty keyboard. For example, an 'f' could be replaced with a 'd', 'e', 'r', 't', 'g', 'b', 'v', or 'c' to mimic common typos.

Dataset-level The injection of noise was defined using four parameters: part of speech, direction, probability and amount of target words.

Part of speech We defined five groupings of parts of speech. Each group represented varying degrees of importance with regards to argument structure. The scale is as follows: function words < mid < nouns < verbs < content words [2]. In our random condition which served as a control, noise was added to all words types, regardless of part of speech, using the same probability and amount parameters as for the part of speech groups.

Details of the composition of each group can be seen below in Table 1.

Group label	Parts of Speech
Verbs	Verbs (all tenses), Modal Auxiliaries
Mid	Adjectives, Prepositions
Nouns	Common and proper Nouns
Function	Determiners, Pronouns, Pre-Determiners
Content	Common and proper Nouns, Verbs (all tenses), Modal Auxiliaries
Random	All possible parts of speech

Table 1: Parts of Speech.

Probability Noise on the word-level was probabilistically-determined. Each selected word had a conditional probability of being noisy. We chose to evaluate performance with four probabilities: 25%, 50%, 75% and 90%.

An example of adding noise to all verbs with a 50% conditional probability is given below.

Clean Text Input: In most provinces a second Bachelor's Degree such as a Bachelor of Education is required to become a qualified teacher. What is the minimum required if you want to teach in Canada?

Noisy Text Input: In most provinces a second Bachelor's Degree such as a Bachelor of Education ks dequired to gecome a qualified tsacher. What ks the minimum dequired if you want to tsach in Canada?

Direction Noise was added either from the beginning of the text input to the end or from the end to the beginning.

[2]The content word group is simply composed of both verbs and nouns which makes it equivalent to verbs in terms of importance in argument structure.

Amount Our first approach to implementing noise was in terms of proportion. We selected words for a given part of speech group based on the percentage of the sentence that we wanted to inject noise into. The selected words than had a given chance of being noisy. Our initial values were: 25%, 50%, 75% and 90%. However, this approach meant that for a given part of speech, in each sentence only a subset of a subset of the total words would be noisy. This amount could not be controlled across parts of speech making comparison between them impossible.

In order to better control the amount of noise present in the dataset and ensure that any effects would be due to quality (part of speech and/or location) rather than sheer amount, we modified our approach. For each sentence, for a given group, we counted the applicable amount of parts of speech in that sentence. All selected words then had a conditional probability of being noisy. More importantly, this allowed us to directly compare applying noise to words within our groups of interest vs applying noise to all possible parts of speech in our random condition. Average count of words per sentence for each part of speech group have been indicated in Table 2. Generally sentences had more content words than function words.

Group	**Average** (Standard Deviation)
Verbs	5.24 (2.6)
Mid	9.69 (5.71)
Nouns	7.8 (4.79)
Function	7.04 (3.47)
Content	13.15 (6.27)

Table 2: Average word count per sentence of each target group

Model Development

Our model was built using Hugging Face's implementation of RoBERTa (Liu et al., 2019) which we then fine-tuned on a question-answer dataset for sentiment classification. Each question-and-answer pair from the dataset was concatenated to form a single input text for the model. Words were tokenized using the original base configuration of the RoBERTa tokenizer. The model we use contains 12 layers, a hidden size of 768, and 12 attention heads. Training was conducted over 3 epochs, a learning rate of 5e-05, and a batch size of 8. The resulting test accuracy on a clean (noiseless) test set was 81.81% using the optimized model. We used this accuracy as baseline for evaluating performance of the model with regards to noise.

Results

In order to determine the quality and quantity of noise that leads to a breakdown in RoBERTa's performance, we trained

and tested the model on a clean dataset and evaluated its performance in different noisy conditions. We injected noise on the word level in a controlled manner to the validation set from a question and answering dataset and picked model accuracy as our evaluation metric for model performance.

We defined noise based on realistic typos and created different noise settings as a function of the conditional probability that a given word would be noisy and the amount of noise in the sentence. The amount of noise was itself defined by the number of words corresponding to the chosen part of speech so that for each sentence, the random condition and POS condition have the same number of noisy words.

Accuracy results across conditional probabilities for content, function, mid and random words have been reported in Tables 3 , 4 and 5. The point differences with regards to baseline are indicated in parentheses. The results for the remaining combinations of our parameters have been included in the Appendix in Tables 6 and 7.

[3]**Condition**	**Noise probability**			
	25%	50%	75%	90%
Content	81.81 (0)	74.55 (-7.26)	70.90 (-10.91)	67.27 (-14.54)
Random– start	81.81 (0)	74.55 (-7.26)	78.18 (-3.63)	70.90 (-10.91)
Random– end	74.55 (-7.26)	70.90 (-10.91)	**61.81** (-19.99)	**61.81** (-19.99)

Table 3: Accuracy (%) for the content condition

The model did not show a difference in performance between the content and random–start conditions at a conditional probability of 25 and 50%, though the point difference went from 0 to -7.26 for both conditions. For each noise probability, the model had the greatest drop in accuracy for the random–end condition. Lowest overall performance was for the random–end condition for noise probabilities of both 75 and 90 %. Interestingly, the model performed worse in the content condition with regards to the random–start condition for noise probabilities of 75 and 90%. While accuracy decreased relative to the amount of noise for both the content and random–end conditions, it increased at 75% probability for the random–start condition.

Condition	**Noise probability**			
	25%	50%	75%	90%
Function	80 (-1.81)	80 (-1.81)	80 (-1.81)	80 (-1.81)
Random– start	80 (-1.81)	76.34 (-5.47)	76.34 (-5.47)	**69.09** (-12.72)
Random– end	81.81 (0)	81.81 (0)	80 (-1.81)	76.36 (-5.45)

Table 4: Accuracy (%) for the function condition

The model had a slight drop in accuracy for the function condition, with a point difference of -1.81. However, it remained at 80% accuracy regardless of noise probability. As is the case for the content condition (exemplified in Table 3), model performance decreased as a function of the amount of noise determined by the noise probability of a given word. Unlike the content condition, lowest overall performance was for the random–start condition for a noise probability of 90%.

Condition	**Noise probability**			
	25%	50%	75%	90%
Mid	78.81 (-3.63)	76.36 (-5.45)	76.36 (-5.45)	76.36 (-5.45)
Random– start	80 (-1.81)	78.18 (-3.63)	76.34 (-5.45)	70.9 (-10.9)
Random– end	80 (-1.81)	74.55 (-7.26)	69.09 (-12.72)	**60** (-21.81)

Table 5: Accuracy (%) for the mid condition

Model performance in the mid condition was lower at 25% noise probability compared to the function and content conditions. However, it stayed constant from 50% onwards. Like in the previous conditions, model performance decreased as a function of noise probability in both the random–start and random–end conditions. For noise probabilities above 25%, the point difference relative to baseline for the random–end condition was twice as large as that of the random–start condition. Lowest overall accuracy across all groups and conditions was for the random–end with a difference of -21.81 points relative to baseline.

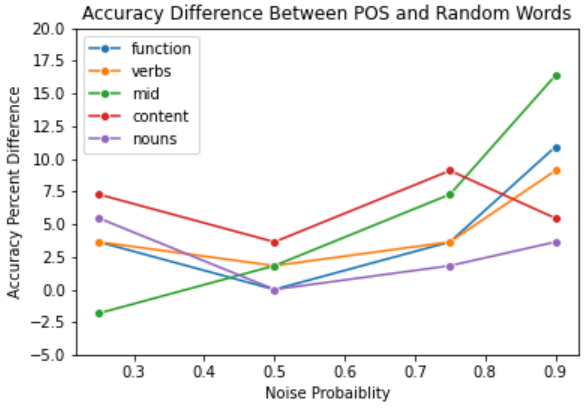

Figure 1: Difference in accuracy between specific parts of speech and random words

In Figure 1 we plotted the point difference in accuracy between all of the POS groups we evaluated the model on (listed in Table 1) and the random–end condition as a function of noise probability.

Point differences remain relatively uniform for all part of speech groups regardless of noise probability except for the mid condition. The mid condition shows the most variance across noise probabilities. Additionally, the point difference between the mid condition and random condition increases as a function of noise probability. The highest difference between random noise and the part of speech noise is in the mid condition for a noise probability of 0.9. In contrast to the

[3]Defined by the part of speech group was used to determine the amount of noise possible for each sentence.

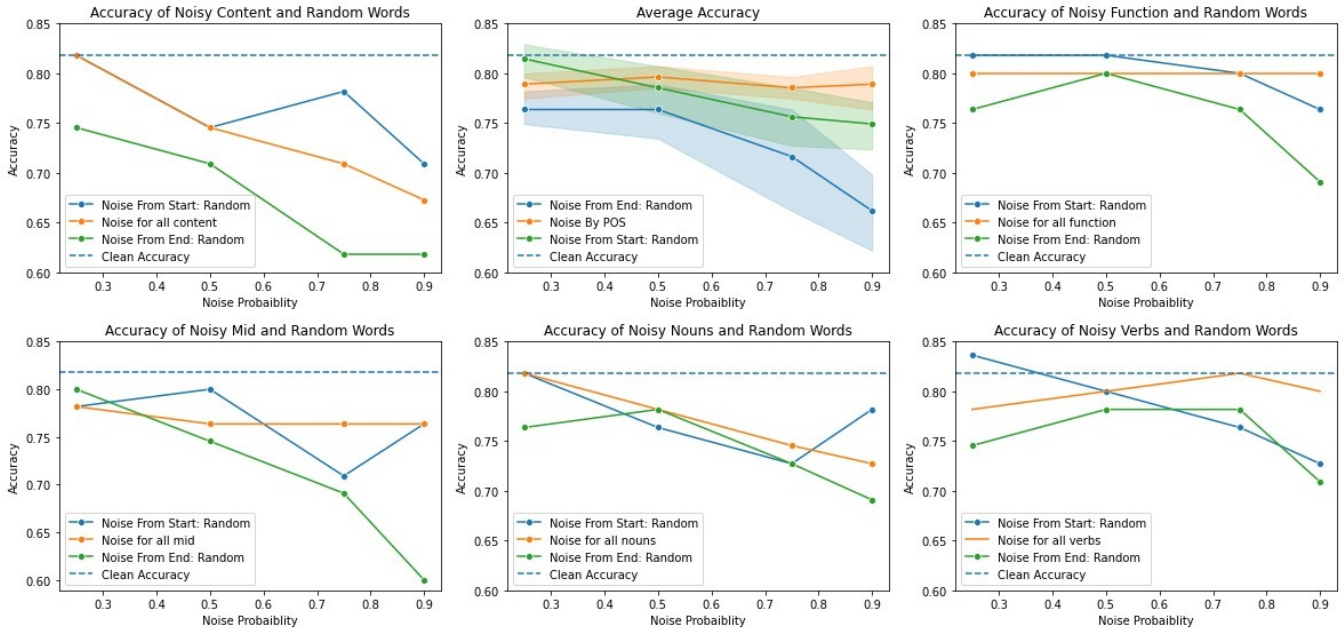

Figure 2: Plots of accuracy between specific parts of speech and random words with positional parameters

other part of speech groups, for a high noise probability of 0.9, adding noise to content words decreases the overall difference between noise-focused on content words and random noise.

Figure 2 shows the measured accuracy for the random–start, random–end, and the specific part of speech for each grouping. Additionally, the average accuracy for all part of speech groupings is included along with the random noise. In general there is not a large difference between the three conditions in any group. In each plot, the random–end condition tends to have the most drastic drop in accuracy as the probability of noise increases. This is especially clear when looking at the plot for average accuracy among all groups. In comparison to the mid, function, and verb groups, content words and nouns seem to have a greater effect on accuracy.

Discussion

The fact that the lowest performance was quasi-consistently in the random conditions indicates that targeting specific parts of speech and thus violations of argument structure were not taken into account. This could be an inherent insensitivity of the model to implicit, high level linguistic structure. However, this could also be due to the fact that the task and training parameters did not allow the model to detect violations in the underlying syntactic structure.

Natural language inference tasks such as the one we chose leave syntax completely implicit. Other papers that evaluate the syntactic capabilities of BERT-based models rely on tasks where syntactic relations are more explicit such as predicting tense or subject-verb agreement (see Section BERT and syntax for more detail). It could be that with more training examples, more training epochs and a task where the underlying syntactic structure is more explicit, noising part of speech will be significantly different from random noise.

Overall, we found that the model was slightly more sensitive to noise when it was added from the end of the sentence. This sensitivity indicates that the model relies more on information towards the end of the input than it does on information at the beginning of the input or on any specific parts of speech.

The fact that we see a greater decrease in accuracy (for increasing noise probabilities) for content words and nouns compared to the other part of speech groups can probably be explained by the fact that these two groups contain the most words (see Table 2 for reference). Since the change in accuracy for these two categories was still smaller than it was when adding noise at random for the same number of words per-sentence, it does not appear to be an effect of the part of speech. Rather the decrease in model performance is a sole function of the quantity of noise present.

Sergio et al. (2020) improve the robustness of BERT in classification tasks with noisy data by adding a novel embracement layer to BERT. Their model does not rely on the *[cls]* token for classification but instead an embracement vector probabilistically composed of features from all tokens. The increase in accuracy of their model (trained on clean and tested on noise) as compared to BERT indicates a disproportionate end of sentence bias in classification tasks. The classification task and more specifically the information used in order to classify the input could account for the end of sentence bias exhibited by our model.

The fact that model's lowest overall performance was for the random–end condition for the mid group remains unaccounted for by our current explanation. Future work could

additionally study the positional distribution of each part of speech. It may be the case that the specific parts of speech of this group were skewed towards the end of the sentence.

Conclusion

In this paper, we studied the effects of adding noise to text on model performance during a question-answering task when introducing typos in a targeted manner to various parts of speech or randomly based on word-position in the sentence. We treated both the context and the question itself as a single text input when adding positional noise.

Our results indicate that the model does not take underlying syntactic structure into account as adding noise to parts of speech central to argument structure was less effective on inhibiting model performance than positional noise. We additionally found an end-of-sentence bias and a sensitivity to noise quantity.

Future work could expand upon our experiments by treating both the question and the answer as distinct inputs with separate noise parameters. Specifically, while we test the introduction of noise from the beginning and end of the concatenated text input, we do not explore the effects of adding noise to the beginning of the question portion itself.

References

Chernyavskiy, A., Ilvovsky, D., & Nakov, P. (2021). Transformers: "the end of history" for nlp? https://doi.org/10.48550/ARXIV.2105.00813

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., & Baroni, M. (2018). What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*.

Corps, R. E., & Rabagliati, H. (2020). How top-down processing enhances comprehension of noise-vocoded speech: Predictions about meaning are more important than predictions about form. *Journal of Memory and Language*, *113*, 104114.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General*, *134*(2), 222.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv*. https://doi.org/10.48550/ARXIV.1810.04805

Ettinger, A. (2020). What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models.

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110*(20), 8051–8056.

Goldberg, Y. (2019). Assessing bert's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. *arXiv preprint arXiv:1803.11138*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*, 1735–80. https://doi.org/10.1162/neco.1997.9.8.1735

Jawahar, G., Sagot, B., & Seddah, D. (2019). What does bert learn about the structure of language? *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Kumar, A., Makhija, P., & Gupta, A. (2020). Noisy text data: Achilles' heel of bert. *arXiv preprint arXiv:2003.12932*.

Levy, R. (2008). A noisy-channel model of human sentence comprehension under uncertain input. *Proceedings of the 2008 conference on empirical methods in natural language processing*, 234–243.

Levy, R. (2011). Integrating surprisal and uncertain-input models in online sentence comprehension: Formal techniques and empirical results. *Proceedings of the 49th annual meeting of the Association for Com-*

putational Linguistics: Human Language Technologies, 1055–1065.

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, *4*, 521–535.

Linzen, T., & Leonard, B. (2018). Distinct patterns of syntactic agreement errors in recurrent networks and humans. *arXiv preprint arXiv:1807.06882*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, *abs/1907.11692*. http://arxiv.org/abs/1907.11692

Marvin, R., & Linzen, T. (2018). Targeted syntactic evaluation of language models. *arXiv preprint arXiv:1808.09031*.

Nangia, N., & Bowman, S. R. (2019). Human vs. muppet: A conservative estimate of human performance on the glue benchmark. *arXiv preprint arXiv:1905.10425*.

Sergio, G. C., Moirangthem, D. S., & Lee, M. (2020). Attentively embracing noise for robust latent representation in bert. *Proceedings of the 28th International Conference on Computational Linguistics*, 3479–3491.

Sun, L., Hashimoto, K., Yin, W., Asai, A., Li, J., Yu, P., & Xiong, C. (2020). Adv-bert: Bert is not robust on misspellings! generating nature adversarial samples on bert. https://doi.org/10.48550/ARXIV.2003.04985

Tran, K., Bisazza, A., & Monz, C. (2018). The importance of being recurrent for modeling hierarchical structure. *arXiv preprint arXiv:1803.03585*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. https://doi.org/10.48550/ARXIV.1706.03762

Warstadt, A., Singh, A., & Bowman, S. R. (2018). Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.

Zhu, D., Hedderich, M. A., Zhai, F., Adelani, D. I., & Klakow, D. (2022). Is bert robust to label noise? a study on learning with noisy labels in text classification. *arXiv preprint arXiv:2204.09371*.

Additional tables and figures

Condition	**Noise probability**			
	25%	50%	75%	90%
Nouns	81.81 (0)	78.18 (-3.63)	74.54 (-7.27)	72.72 (-9.08)
Random– start	81.81 (0)	76.36 (-5.45)	72.72 (-9.08)	78.18 (-3.63)
Random– end	76.36 (-5.45)	78.18 (-3.63)	72.72 (-9.08)	**69.09** (-12.72)

Table 6: Accuracy (%) for the noun condition

Condition	**Noise probability**			
	25%	50%	75%	90%
Verbs	78.18 (-3.63)	80 (-1.81)	78.18 (-3.63)	80 (-1.81)
Random– start	83.63 (-1.83)	80 (-1.81)	76.36 (-5.45)	72.72 (-9.08)
Random– end	74.54 (-7.27)	78.18 (-3.63)	78.18 (-3.63)	**70.90** (-10.90)

Table 7: Accuracy (%) for the verb condition