Deeper dive into understanding the coherence of Conversation Entailment tasks

William Wang and Xiangyu Qin
Computer Science and Engineering Division
University of Michigan
Ann Arbor, MI 48109, USA
{wiljwang, qinx}@umich.edu

1 Introduction: Problem Statement

The problem we are tackling for this final project is: given a series of conversation segments (a multi-turn natural language dialogue), we want to be able to predict whether a given hypothesis can be inferred from the dialogue or not. This is considered a *binary* text classification problem as the output is expected to be either True or False, which indicates whether the hypothesis is supported or not.

The problem with the existing research and approaches is that although the state-of-the-art transformer-based language models like RoBERTa and DeBERTa were successful in greatly improving the accuracy of the prediction over baseline system performance, it is suspected that the understanding of the model is still incoherent. That is, the model may be focusing on spurious intermediate evidence rather than the entire input data.

Our goal for this paper is to experiment with ways to improve the coherence, and see what effects it may bring to the accuracy. Because the state-of-the-art approach have produced high results in accuracy, we will be basing our approaches on transformers as well. To improve coherence, we will be attempting to use other transformer models that perform well with the given data input. While our primary goal is to improve coherence, we hypothesize that improvement in coherence would likely result in an increase in accuracy as well, since it will better utilize the structure of the input.

Accuracy and coherence are often correlated because a system that is able to produce accurate and correct output is more likely to produce output that is coherent and makes sense to a human reader. Coherence in natural language can be seen as the ability of a text or language to be logical and easy to understand, and accuracy is an important factor in achieving this.

2 Proposed Approaches

We will attempt to tackle the problem by measuring and evaluating coherence and accuracy separately using two different variations of BERT that have their own unique strengths and weaknesses.

In an attempt to improve coherence, we will make use of ALBERT by Google. ALBERT is "A Lite" version of BERT, because it utilizes two parameter reduction techniques to overcome the scaling problem of pre-trained models. We believe that the parameter reduction technique may also be helpful in preventing over-fitting to spurious intermediate evidence, so we will be paying particular attention to coherence measurements compared to the original paper's results. We are not sure how the accuracy might turn out for ALBERT, because on one hand, we could expect a better accuracy

Additionally, to tackle the objective of improving the accuracy of the binary classification, we will use XLNet by Carnegie Mellon University. XLNet is good at language tasks involving long context and it also does better in natural language inference when compared to BERT. XLNet achieves this by being able to look at context in both directions by utilizing randomized tokens when training. By being able to consider context in both the forward and backwards direction, we hypothesize that this will help the model understand the overall structure of the conversation better. This fits our problem statement well since the state-of-the-art transformer models RoBERTa and DeBERTa struggled to incorporate the dialogues across a long context. We will be noting the accuracy yielded from using XLNet compared to RoBERTa, but also its coherence to see if accuracy and coherence are natural tradeoffs in this domain.

Between these two pre-trained model types, and potentially more that we encounter along the way during implementation, we believe the possibilities of improving on accuracy and coherence can be tested. Further, we will be able to evaluate the types of models that result in tradeoffs between accuracy and coherence.

3 Data Set

The data set has been made available by the SLED Lab at the University of Michigan on GitHub and on Hugging Face. The data set contains information about the sequence of speakers that the dialogue is spoken in, the conversation segments spoken by each speaker, the hypothesis, and the labeled boolean flag for whether the conversation entails the hypothesis or not.

We will be using this very dataset while experimenting with different pre-trained models. Shane and Chai provided with two datasets to train and test on. The first dataset is the one introduced in 2009 by Zhang and Chai ((Zhang and Chai, 2009), labelled as the CE dataset, and the second is the Abductive Reasoning in narrative Text (ART) dataset, introduced by Bhagavatula et al. when they examined a similar problem, but for a multiple choice text plausibility classification task (Bhagavatula et al., 2019). While the ART dataset has a lot more data entries to train on, we would be using the CE dataset for two reasons. 1. We have a limited time to work on this project, and we believe it will be a better use of our time to experiment with better methods than to spend the time running the model. 2. Previous work has results using the CE dataset as well, so even with a smaller dataset, we will still be able to compare our results with existing methods. If our methods show promising results, we can attempt it on the ART dataset and evaluate how well it extends to a multiple choice text plausibility classification task.

The CE dataset consists of 703 entries in the training set, 110 entries in the development set, and 172 entries in the testing set, contributing to a total of 985 data entries.

For the models, we will be using the pre-trained model, and fine tuning it to our dataset. The pre-trained model for ALBERT is available on TensorFlow Hub (Abadi et al., 2015) and the pre-trainhed model for XLNet is being available by the original authors of XLNet (Yang et al., 2019)

4 Previous Work

4.1 Conversation Entailment

Textual entailment involves determining the relationship between two text segments. Specifically, given a pair of text segments, the task is to determine whether the meaning of one text segment (the "premise") entails the meaning of the other text segment (the "hypothesis").

For example, given the premise “The sky is blue” and the hypothesis “The sky is not blue,” the task would be to determine that the hypothesis contradicts the premise. On the other hand, given the premise “All dogs are mammals” and the hypothesis “My pet is a mammal,” the task would be to determine that the hypothesis is entailed by the premise.

Textual entailment is a important task because it can be used to identify relationships between text segments in a wide range of applications, such as information extraction, question answering, and summarization. It can also be used to evaluate the performance of natural language processing systems, as systems that are able to accurately identify relationships between text segments are likely to be more effective in other NLP tasks as well.

In the field of textual entailment, which our problem is a subset of, the approach has shifted over time from LSTM with attention (Liu et al., 2016) to the current state-of-the-art approach: transformers. The state-of-the-art pre-trained transformer RoBERTa has been successful in General Language Understanding Evaluation (GLUE) tasks, with an accuracy above 90% for 5 out of 9 of the GLUE tasks (Liu et al., 2019).

With regard to the field of conversation entailment, which was first examined in 2010, the baseline system performance was quite low at an accuracy of 60% (Zhang and Chai, 2010), which is not too much better than purely guessing at this binary task. Storks and Chai revisited this problem in 2021, applying the state-of-the-art pre-trained transformer models to this problem. A great increase in accuracy was seen as a result, where the highest test accuracy of 78.5% was obtained with RoBERTa + MNLI. However, despite the high obtained accuracy, the coherence score for each model suggests that while “the text classifiers can achieve high classification accuracy on CE and ART, they do not deeply understand the tasks” (Storks and Chai, 2021). Often, models and problems are evaluated by the accuracy score that they can achieve, but without strong coherence, there is little confidence that these results can be replicated in more diverse but structurally similar datasets.

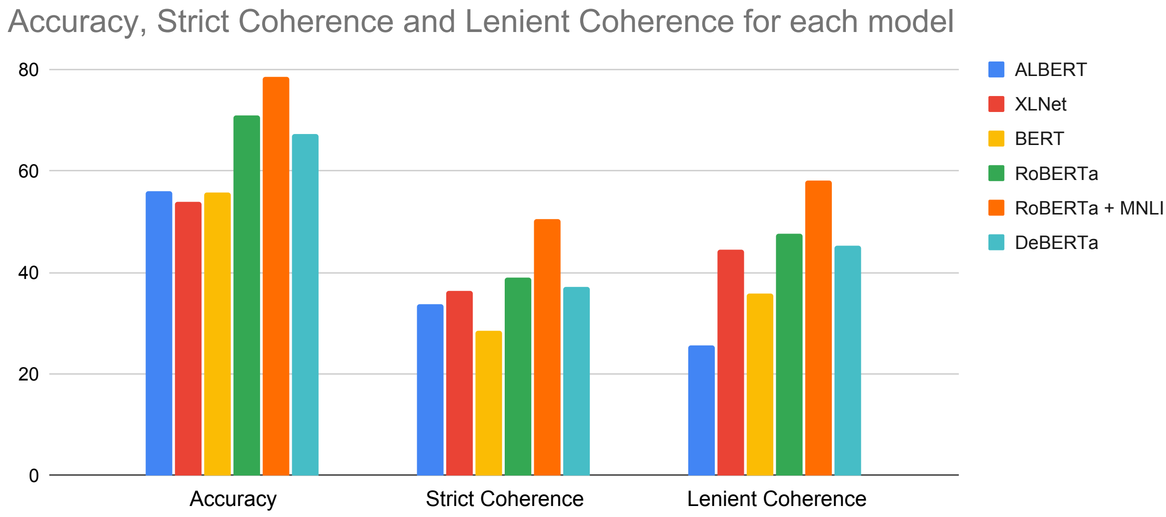

Figure 1: Accuracy, strict coherence, and lenient coherence on the CE dataset for the two proposed method and the methods covered in previous work (Liu et al., 2019). It can be observed that accuracy of both the ALBERT and XLNet performed significantly worse than the best performing model so far - RoBERTa + MNLI. However, both models seem to have a higher proportion of correctly classified hypothesis to be assessed as coherent as well.

4.2 ALBERT Model

The motivation behind the authors to come up with a new variation of the BERT model was that models often have hundreds of millions or even billions of parameters, and with this many parameters, it is very easy to hit memory limitations as we try to scale the models. To overcome this, the authors have incorporated two parameter reduction techniques. The first technique is factorized embedding parametrization, where they decompose the large vocabulary embedding matrix that BERT uses into two smaller matrices. This separation makes it easy to increase the hidden layer size without significantly increasing the parameter sizes. The second technique, cross-layer parameter sharing, prevents the parameter from growing with the depth of the network. As a result of these techniques, the authors were able to reduce the size of the ALBERT model to have 18x fewer parameters than a BERT-large and also trained 1.7x faster (Lan et al., 2019).

Another benefit of using ALBERT is that they also introduce a self-supervised loss for sentence-order prediction (SOP). This allows ALBERT to focus on inter-sentence coherence and improve the performance of the model (Lan et al., 2019). We believe that this unique feature of ALBERT would not only improve the accuracy on the conversation entailment task, but also improve the coherence. This is because the existing models tested on conversation entailment task demonstrated a lack of the understanding of the structure, so the inter-sentence coherence of ALBERT may be successful in preventing that.

Lastly, ALBERT was designed to be smaller and more computationally efficient than RoBERTa, which means that it can naturally avoid overfitting that comes as a consequence of having a lot of training data to fit to. This could run counter to the "spurious intermediate evidence" being relied on.

4.3 XLNet Model

ALthough BERT is also copable of modeling bidirectional contexts, BERT neglects dependency between the masked positions and suffers from a pretrain-finetune discrepancy. XLNet does not rely on data corruption like BERT does, but instead introduces segment recurrence mechanism and relative enchoding scheme of Transformer-XL into pretraining, which empirically improves the performance especially for tasks involving a longer text sequence.

Further, XLNet may perform better than RoBERTa on specific NLP tasks, depending on the characteristics of the task and the training data. For example, XLNet has been shown to perform particularly well on tasks that require understanding long-term dependencies in language, such as language translation and language modeling. We hypothesize that this will apply to conversation entailment because the sequence of text is a reply, i. e. dependency, of previous text.

We believe that the text sequence in a conversation entailment task is considered a long text sequence, as not only does it have to learn through the span of an entire sentence, it has to do this for multiple sentences. Furthermore, conversation entailment complicates this further by alternating between two speakers, and the meaning of the speech would also be affected by who said it. Therefore,given these strength of XLNet, we belive that this= matched the problem that conversation entailement classificaiton task had, and could potentially be a solution to improving coherence in the classification.

5 Evaluation

A summary of the accuracy, strict coherence and lenient coherence metrics from our two proposed models compared with the other models introduced in previous work can be found in figure 1. The same coherence metrics as in Storks and Chai 2019 are used to measure both strict and lenient coherence in ALBERT and XLNet.

5.1 ALBERT Results

In order to confirm that the model is leaning something useful in the process and to observe the trend in how the accuracy changes with training, we first ran ALBERT on a smaller batch of inputs. We decided to run the model with 10% of the entire dataset. Since Storks and Chai combined the training dataset and the development dataset and performed cross-validation, we obtained 10% of the entries from each of the dataset before combining them into one dataset. With 703 training dataset and 110 development set, our initial smaller batch of inputs consisted of 81 entries. We ran 8 fold cross-validation on 10 epochs each, which is consistent with the hyper-parameter from the previous work in order to get comparable results. After training, we were able to obtain 52.7% accuracy with a strict coherence of 23.4 a and lenient coherence of 24.1.

Although the accuracy was only slightly higher than random guessing, this result was still very promising as we are able to see that specifically, that our strict coherence is already greater than half of that of BERT achieved based on the result from our previous work. This means that the model is learning the structure of the problem. Calculating this as a percentage, we can see that amongst all the hypthesis that were correctly identified, we can see that $\frac{23.4}{52.7ca} \times 100\% = 44.4\%$ of them were able to utilize the correct structure.

However, we were surprised by the result when we ran this on the entire dataset. The final accuracy was 56.1%, showing almost no improvements at all from when we ran it on just 10% of the data. What we found more surprising was that the coherence on the other hand showed a massive improvement. The strict coherence has increased to 33.7, more than doubled from our test run, and the lenient coherence was 25.7

Interestingly, we observed that the strict coherence measure was reported to be higher than that of the lenient coherence measure. ALBERT was the only model out of the 6 models we have data on where the strict coherence was higher than the lenient coherence.

It is possible for strict coherence to be higher than lenient coherence in a text or speech if the text or speech meets strict criteria for logical connections and smooth flow, but does not meet the more lenient criteria. This could occur if the text or speech has a high degree of logical structure and clear transitions between ideas, but still has some disfluencies or ambiguities that do not meet the more lenient criteria for coherence.

For example, a text with strict coherence might be well-organized and have clear transitions between ideas, but still have some awkward phrasings or minor errors that do not meet the more lenient criteria for coherence. In this case, the text would have a high degree of strict coherence, but a lower degree of lenient coherence.

5.2 XLNet Results

Similar to the ALBERT model, we ran XLNet on the same smaller batch of inputs to ensure that the model is learning valuable features in the input and to observe any patterns and trends. The accuracy was almost identical to that of ALBERT, with an accuracy of 52.4%, strict coherence was 10.4% and the lenient coherence was 12.1%.

With the training on the smaller batch of inputs, we can see that although the accuracy of XLNet was similar to that of ALBERT, we see a pretty significant drop in coherence. This could suggests that potentially, the randomization of the input sequence may have actually guided XLNet in doing the opposite of what we wanted. Because XLNet is

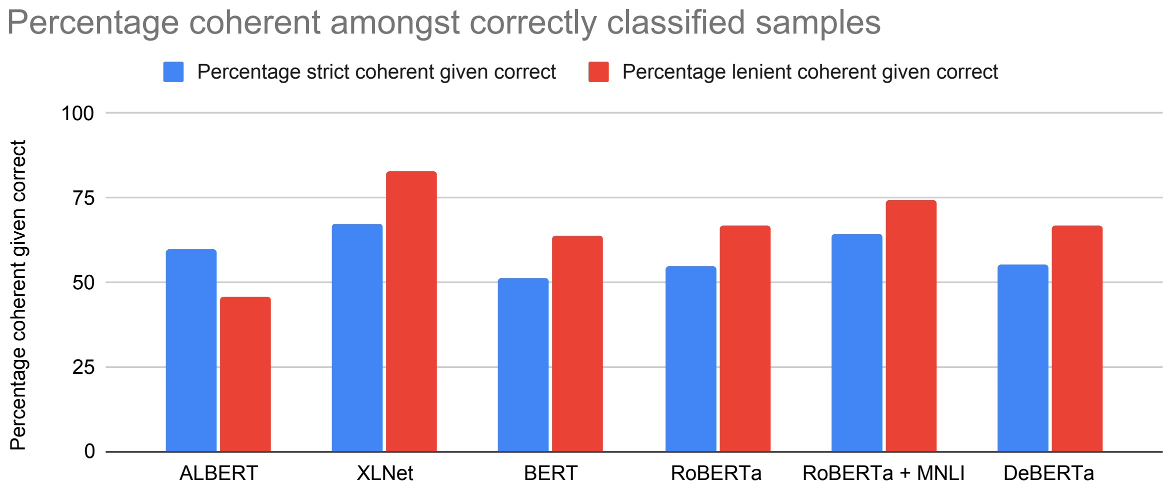

Figure 2: Percentage of tasks that are coherent given that it was classified correctly. While difference in measures between accuracy and coherence is valuable, the percentage of tasks that were classified correctly that are coherent is also important. This is because this allows us to know how likely it was that the model understood the structure well when making the correct decision. We can observe that XLNet performed significantly better under this metric. This indicates that XLNet when making the decision for whether the hypothesis is entailed or not, it effectively utilized the structure of the conversation as well, rather than basing it simply on spurious intermediate evidence.

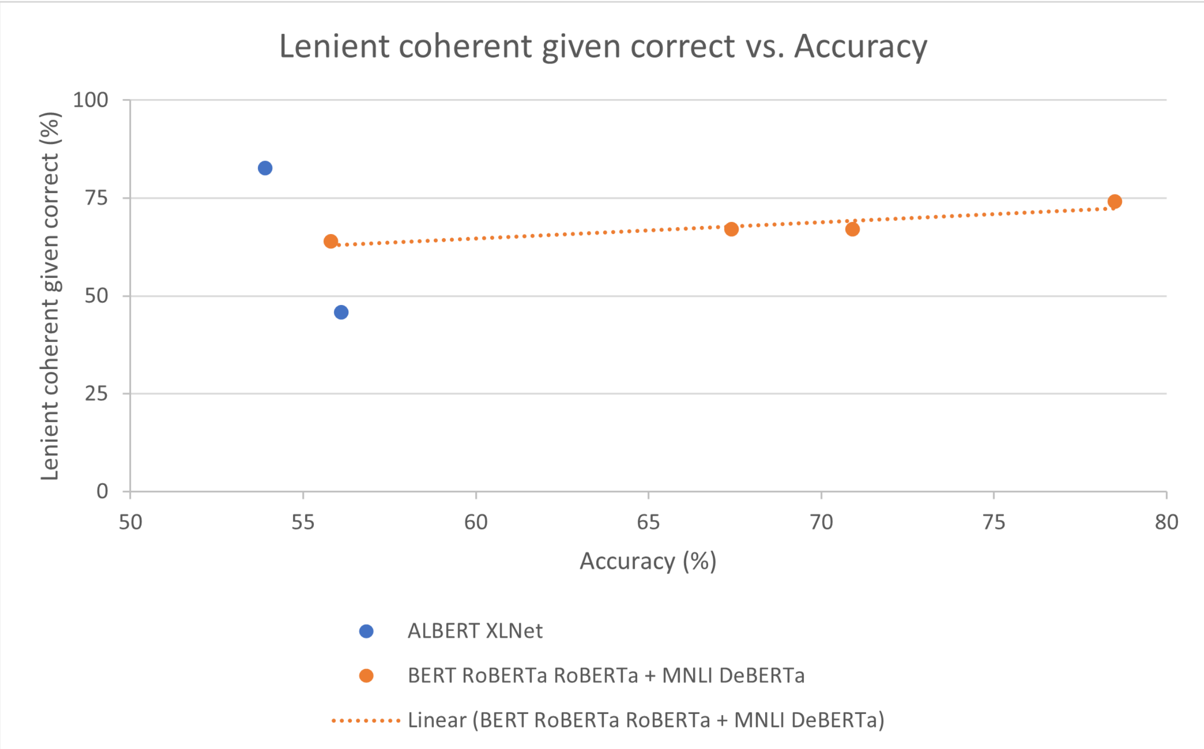

Figure 3: Scatter plot of accuracy vs the percentage of correctly classified samples that were also coherent. The blue points are those of our proposed models and the orange points are those investigated by previous work. We can see from the plot that for the models from previous work, there is a strong linear relationship between the accuracy and the percentage of lenient coherence. The gradient of the best fit is very small, indicating that while in general, as the accuracy goes up, we can expect more correctly classified samples to be more coherent as well, the amount this increases is almost trivial. This aligns with the findings from previous work where while they were able to achieve high accuracy, the transformer based models that were investigated struggled to incorporate the structure of the conversation.

trained with the input sequence randomized, this reduces the structure of the input when training. That is, the order in which the conversation happens, despite it being important when humans understand the meaning, would have been lacking when training the XLNet model.

When XLNet was ran on the entire dataset, we saw interesting results as well. Accuracy was reported to be 53.9% which was even worse than that of ALBERT. However on the other hand, the strict coherence measure was 36.3% and the lenient coherence was 44.6% performing significantly better than ALBERT. Taking a closer look at figure 1, we can see that this coherence performance for XLNet is actually not quite impressive when compared to the other better performing models. However if we shift our attention to figure 2, under the metric of percentage of correctly classified samples that are coherent, XLNet has outperformed all of the other models in both the strict and lenient coherence.

6 Discussion of results

Accuracy and coherence are two distinct aspects of language processing that can be evaluated separately. Accuracy refers to the degree to which a system's output (e.g., a machine translation or a text generation system) matches a reference or gold standard. Coherence, on the other hand, refers to the degree to which the information in a text or speech is logically connected and flows smoothly.

There is often a trade-off between accuracy and coherence in natural language processing systems. For example, a machine translation system that focuses on achieving high accuracy may produce translations that are more literal and faithful to the source text, but may be less fluent and coherent in the target language. On the other hand, a machine translation system that focuses on achieving high coherence may produce translations that are more fluent and coherent in the target language, but may be less accurate in terms of preserving the meaning of the source text.

In general, it is important for natural language processing systems to achieve both high accuracy and high coherence in order to produce output that is both faithful to the source material and easy for humans to understand. However, the relative importance of accuracy and coherence will depend on the specific task and the needs of the user.

As applied back to the problem of conversation entailment, figure 3 illustrates the relationship between the accuracy a model achieved and the percentage of the correctly classified samples that were coherent. It can be observed that both of our proposed approaches were outliers to the trend that was seen in previous work. ALBERT performed much worse in coherence than expected, and XLNet performed significantly better than what was expected.

For ALBERT looking at how the metrics improved from our smaller batch of training inputs, we can see that neither the accuracy nor the coherence has improved much when we ran it on the entire dataset. We hypothsize that this is because of the paramter reduction technique that was employed. By making the model simpler than the other models, we believe that it was able to obtain some meaningful understanding right away, with only a few parameters to train on. However, because of the lack of parameter, we believe that it also did not extend well when giving a larger dataset. That is, even with a larger dataset, it wasn't able to learn anything meaningful past what it did with just 10% of the total training samples. Furthermore, the coherence metric was the lowest for the ALBERT model, and this may be explained by because of the lack of parameters, ALBERT was not able to learn the complex structure of the conversation and depended more on the spurious intermediate evidences. Being able to learn the structure of the conversation is a difficult task, and the result from previous work where even with a high accuracy, the model still tended to base the classification on spurious intermediate evidence, demonstrating how difficult it is for models to learn the structure. This is the complete opposite of what we hypothesized, since our hypothesis was that ALBERT may perform better because spurious intermediate evidence is a lot more problem dependent than learning the structure. Thus, we believed that with fewer parameters, ALBERT would prioritize learning the structure to obtain meaningful understanding of the problem.

XLNet although performed the worst in accuracy out of all 6 models, it did perform exceptionally well in coherence. Based on our previous discussion on how learning the structure is a difficult task, we believe that is the exact reason why XLNet was able to perform better than the other models in terms of coherence metrics. XLNet is able to understand forward and backward relations between conversations, and this is enabled due to its unique way of training. It randomizes the order of the input, so that XLNet would start to recognize the relationships between different sentences. We believe that XLNet, contrary to ALBERT, focussed on learning the structure of the input rather than focussing too much on spurious intermediate evidence. As evidence, we can see that the coherence metrics of XLNet improved significantly from the smaller batch training data to when we used the entire training data. We believe that given the complexity of the problem, just 10% of a already small training data was not enough for XLNet to learn many meaningful features.

7 Conclusion

Although we primarily ran the two models on smaller batch of input data to ensure that the code is working and that the model is in fact learning something useful, we were able to make unexpected relation and analysis on how the relationship between coherence and accuracy for the two proposed methods.

While transformers may achieve high accuracy in terms of predicting the correct output for a given input, they may not always produce output that is coherent or easily understandable to humans. This is because transformers are trained to optimize for certain performance metrics, such as minimizing the cross-entropy loss or maximizing the likelihood of the output given the input, rather than for producing output that is grammatically correct or coherent.

In order to improve the coherence of the output produced by a transformer model, it may be necessary to fine-tune the model on a specific task or dataset, or to incorporate additional constraints or loss functions that encourage the model to produce more coherent output. In our case, we examined the unique qualities of various high-performing state-of-the-art transformer-based language models and attempted to improve coherence based on those qualities, to mixed success.

Another bigger picture conclusion that could be drawn is that transformers are more similar than they are different. Of course, their construction can be quite different, so the accuracy and coherence can vary significantly between them. These results are also generally indicative of better performance on other natural language processing tasks. But the high level results, such as the relationship between accuracy and coherence, are quite similar

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. *CoRR*, abs/1908.05739.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942.

Yang Liu, Chengjie Sun, Lei Lin, and Xiaolong Wang. 2016. Learning natural language inference using bidirectional LSTM model and inner-attention. *CoRR*, abs/1605.09090.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Shane Storks and Joyce Chai. 2021. Beyond the tip of the iceberg: Assessing coherence of text classifiers. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3169–3177, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

Chen Zhang and Joyce Chai. 2009. What do we know about conversation participants: Experiments on conversation entailment. In *Proceedings of the SIGDIAL 2009 Conference*, pages 206–215, London, UK. Association for Computational Linguistics.

Chen Zhang and Joyce Chai. 2010. Towards conversation entailment: An empirical investigation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 756–766, Cambridge, MA. Association for Computational Linguistics.