BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (Devlin et al., 2019)

Critique by Jingying Xu

October, 2025

Devlin et al. (2019) introduced BERT, a bidirectional Transformer model that set new benchmarks across a wide range of NLP tasks by leveraging masked language modeling and next-sentence prediction. Despite its impact, the paper has several key limitations:

1. *NSP does not teach fine-grained semantic inference.*

   The Next Sentence Prediction (NSP) objective has important limitations. Although NSP is intended to help the model learn relationships between sentences, the binary IsNext/NotNext distinction does not provide the type of information needed to capture true **semantic relations** such as *entailment*, *presupposition*, or *implicature*. Because the negative examples are simply random sentences, the model can succeed by detecting topical continuity rather than reasoning about whether the meaning of one sentence *entails*, *presupposes*, or *implies* the other. In other words, NSP does not give the model any supervision that distinguishes logical consequence from mere co-occurrence, so there is no way for the model to learn fine-grained semantic relations between sentence meanings from NSP alone.

2. *NSP does not capture multi-sentence discourse structure.*

   NSP is defined only over pairs of sentences, so the model receives no training signal about how meaning develops across three or more sentences. As a result, it does not learn how to track discourse coherence or information flow over longer stretches of text.

3. *Forced-choice tasks do not show real linguistic abilities.*

   Most of the benchmarks in this paper, such as CoLA, SST-2, MRPC, QQP, MNLI, QNLI, RTE, and SWAG, test forced-choice classification, where the model only needs to pick a label from a small set. where the model selects the label with the highest predicted probability. However, choosing the label with the highest probability does not mean the model actually understood the meaning of the sentence. It may simply be that one option received a slightly higher score than the others, not that the model formed a correct interpretation. Thus, the output reflects probability ranking, **not true linguistic understanding**. And importantly, these forced-choice decisions **do not always reflect human linguistic competence either** (Xu & Schmitt, 2025). They measure decision-making on a specific task format, which are known to vary across tasks, contexts, and elicitation formats.

References

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186.

Xu, J., & Schmitt, C. (2025). Pragmatic accommodation in judging event culmination. *Proceedings of Semantics and Linguistic Theory (SALT) 34*, 502–523. https://doi.org/10.3765/sgccnn77