How effective is BERT without word ordering? Implications for language understanding and data privacy

Jack Hessel
Allen Institute for AI
jackh@allenai.org

Alexandra Schofield
Harvey Mudd College
xanda@cs.hmc.edu

Abstract

Ordered word sequences contain the rich structures that define language. However, it's often not clear if or how modern pretrained language models utilize these structures. We show that the token representations and self-attention activations within BERT are surprisingly resilient to shuffling the order of input tokens, and that for several GLUE language understanding tasks, shuffling only minimally degrades performance, e.g., by 4% for QNLI. While bleak from the perspective of language understanding, our results have positive implications for cases where copyright or ethics *necessitates* the consideration of bag-of-words data (vs. full documents). We simulate such a scenario for three sensitive classification tasks, demonstrating minimal performance degradation vs. releasing full language sequences.

1 Introduction

Masked language models (MLMs) like BERT (Devlin et al., 2019) use an ordered sequence of tokens as input. And rightfully so! Any model capable of "language understanding" undoubtedly *should* need access to the hierarchical, syntactic structures implicitly encoded in language. But are MLMs really doing better *because* they have access to full word sequences?

To assess this question, we first compare the internal representations of BERT and RoBERTa (Liu et al., 2019) when the sequence of unigrams is not available.[1] We do this by using the bag-of-words counts of an input to generate a random ordering of the unigrams, i.e., "shuffling" the input. For example, in a sentiment classification corpus, if an intact input was "The movie was great!", a possible shuffled ordering might be "movie the great was" (tokenization details are in §4). We find that, though BERT appears to become more sensitive to ordering in later layers, shuffled token representations and self-attention activations still closely resemble their unshuffled counterparts.

Following cues from prior work (Sugawara et al., 2020; Si et al., 2019; K et al., 2020), we next report the performance of pre-trained MLMs fine-tuned on GLUE, a suite of English-language understanding benchmarks, when given access only to unigram count information by handing models randomly ordered sequences of words (an approach we call BoW-BERT, for short). For most GLUE tasks, performance degradation when shuffling is minimal, e.g., MNLI, QQP, and QNLI accuracy degrade by less than 5 accuracy points.

The bad news: Despite BERT being trained on intact word sequences, BoW-BERT demonstrates that MLMs can readily ignore syntax (while maintaining strong performance) when fine-tuned for even carefully designed downstream language understanding tasks.[2] We thus advocate for reporting BoW-BERT's performance as a strong baseline.

The good news: BoW-BERT offers a practical modeling choice for researchers who *must* operate with only bag-of-words representations for legal or ethical reasons.[3] Bag-of-words data releases are sometimes the *only legal format* in which copyright-sensitive corpora may be distributed, e.g., HathiTrust[4] (16M historical volumes) (Christenson, 2011), Google N-grams (Michel et al., 2011), etc. And while ethical considerations sometimes preclude the full release of privacy-sensitive docu-

[1] We use the "base" models supplied by the authors

[2] Bowman and Dahl (2021) provide perspective on "fixing" NLU tasks.

[3] This is a surprisingly common case: our initial motivation for BoW-BERT was our experience in exploring such a corpus.

[4] In Authors Guild, Inc. v. HathiTrust (2014), the 2nd Circuit U.S. Court of Appeals ruled that showing only "*the number of times [a search term] appears on each page*" constitutes legal fair use, but "[displaying] to the user any text from the underlying copyrighted work" might not.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers), pages 204–211
August 1–6, 2021.

ments (e.g., medical transcriptions), bag-of-words data release offers the potential for compromise. While releasing unigram counts is one way of anonymizing documents (Gallé and Tealdi, 2015), recent work in differential privacy (Dwork, 2008; Fernandes et al., 2019; Schofield et al., 2019; Schein et al., 2019) has resulted in randomized algorithms capable of privatizing BoW count data (under varying definitions of privacy).[5]

We explore classification tasks on three sensitive corpora, simulating different input fidelity availability: full sequences, BoW counts, and differentially private (DP) BoW counts. We find that BoW-BERT often significantly outperforms prior BoW models, especially for shorter documents. And, for longer documents, BoW-BERT can even outperform full-sequence BERT. Finally, for the (naive) DP configuration we consider, BoW-BERT is a viable option for classifying shorter privatized documents, though linear BoW models remain competitive for longer documents.

2 Related Work

Shuffling inputs to non-pretrained models. Word order shuffling has been tested as part of the full training process for non-pretrained models. Sankar et al. (2019) shuffle words in a dialog corpus, and find that LSTMs are more sensitive than Transformers to word order. Khandelwal et al. (2018) show that shuffling distant context words (e.g., beyond 50 tokens) has little effect in outcome for LM-LSTMs. Adi et al. (2017) show that LSTM autoencoders encode significant ordering information when fit to a corpus of Wikipedia sentences. Nie et al. (2019) report minimal performance decreases from word shuffling while training a number of model architectures, e.g., ESIM (Chen et al., 2017), for SNLI/MNLI tasks. In a multimodal setting, Cirik et al. (2018) show that shuffling doesn't affect performance for an LSTM in a referring expression task.

Shuffling inputs to pretrained MLMs. While at the time of submission of this work, shuffling results had not been fully reported on the popular GLUE taskset, prior results have used word-shuffling as a baseline with varying results.

Sugawara et al. (2020) operationalize ablations of reading comprehension skills from Kintsch (1988), and report that shuffling n-grams in 10 QA corpora results in 10-20% performance decreases for BERT. Si et al. (2019) report similar results when shuffling questions+answers in MCRC corpora, reporting absolute accuracy drops of between 5-20% when shuffling both passage/question words (e.g., BERT on DREAM drops from 63 → 41 accuracy relative to a 33% constant baseline). K et al. (2020) report that swapping tokens during pretraining of a multilingual BERT model results in moderate performance degradation for XNLI (e.g., 71 → 63 for en-es) but more significant performance degradation for NER (63 → 40 in the same setting). They find that a purely frequency-based corpus "is not enough for a reasonable cross-lingual performance."

Several works have examined shuffling inputs in multi-language scenarios (e.g., translation) when languages have variable syntax (Ahmad et al., 2019; Liu et al., 2020). Zhao et al. (2020) use a random token permutation to provide a baseline. Yang et al. (2019) find that self-attention networks are surprisingly bad at identifying two tokens that are swapped in the input. Ettinger (2020) show that shuffling BERT inputs decreases word cloze prediction performance on a corpus of 102 sentences without fine-tuning. Wang et al. (2020) incorporate a deshuffling objective into pre-training.

In some cases, shuffled inputs provide a stronger baseline than might be assumed, while in others, shuffling significantly degrades performance. At present, determining whether or not order is "needed" for a particular task is largely an experimental, empirical endeavor.

Syntax in MLMs. Prior works have investigated BERT's capacity to represent syntax: some researchers have designed prediction tasks that require syntactic knowledge (Linzen et al., 2016; Jawahar et al., 2019; Lin et al., 2019; Goldberg, 2019), while others have probed representations for linguistic information directly (Mareček and Rosa, 2018; Liu et al.; Hewitt and Manning, 2019; Reif et al., 2019). Tenney et al. (2019) find that contextual representations outperform lexical representations on many syntactic tasks, but not in a suite of semantic prediction tasks. Htut et al. (2019) and Clark et al. (2019) find that some attention heads encode information useful for dependency parsing. Glavaš and Vulić (2020) show that intermediate supervised training of a biaffine parser has little effect on downstream MLM performance.

[5] Releasing BoW counts is related to, but distinct from, the setting considered by Beigi et al. (2019), who produce private vector representations with uninterpretable dimensions.

A Bouquet of Contemporaneous Work. While this work was in submission, several related works were posted to arXiv. Gupta et al. (2021) examine NLI, paraphrase detection, and sentiment classification, and show that destructive interventions do not significantly affect either model predictions or model confidence. Sinha et al. (2020) find a similar result for NLI tasks, and, in follow-up work, Sinha et al. (2021) demonstrate pretraining is possible on unordered sequences. Pham et al. (2020) look specifically at GLUE classification for BERT-based models. Beyond contemporaneous confirmation of the GLUE results, our work contributes to this bouquet by: 1) examining internal activations/layers and 2) exploring classification settings where one might need to operate on (potentially differentially private) count-only data.

3 Representation analysis

We might expect that shuffling the order of tokens in an input sentence would significantly corrupt the internal representations of BERT, but is that actually the case? We investigate with two new metrics. Consider applying a pre-trained, fixed BERT model to $x =$"the movie was great" and the shuffled $x' =$"movie the great was".

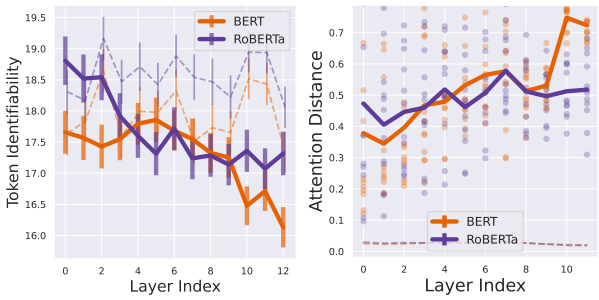
Token identifiability measures the similarity of BERT's vector representations of a word token (e.g., "movie") in x and x'. Identifiability is high if the model has similar representations for tokens after their order is shuffled.

Self-attention distance measures if BERT attends to similar tokens for each token in x and x' regardless of their order (e.g., is "the movie was great" $\approx$ "movie the great was" to BERT?). Self-attention distance is low if the model attends to the same tokens after input shuffling.

Token Identifiability. Let $\texttt{MLM}_l(x)$ be a $\mathbb{R}^{t\times d}$ matrix, where t is the number of tokens in sentence x, d is the MLM's dimension, and l is the layer index. In this setting, row i of $\texttt{MLM}_l(x)$ is the MLM's representation of the ith token in sentence x. We compare $\texttt{MLM}_l(x)$ to $\mathbb{E}[\texttt{MLM}_l(X')]$, where X' is drawn uniformly from the permutations of x: $perm(x)$. For a specific sample $x' \sim perm(x)$, we first take the row-wise cosine similarity of $\texttt{MLM}_l(x)$ and $\texttt{MLM}_l(x')$, and treat the resulting $t \times t$ matrix as an instance of a bipartite linear assignment problem. The *assignment accuracy* (AA) score for (x, x') is the proportion of assigned token pairs that have the same underlying word type. To avoid biasing towards shorter sentences, we take the ratio of the accuracy relative to chance, i.e.,

$$\text{ID-MLM}(x, l) = \frac{\mathbb{E}_{X'}[\text{AA}(\texttt{MLM}_l(x), \texttt{MLM}_l(X'))]}{\mathbb{E}_{\texttt{RAND}}[\text{AA}(\texttt{MLM}_l(x), \texttt{RAND})]}, \quad (1)$$

where RAND is a random matrix of reals $\mathbb{R}^{t\times d}$.[6]



(a) Identifiability (b) Self-Attention Distance

Figure 1: Token identifiability and attention distance by layer for BERT and RoBERTa; dashed lines represent baseline values of metrics with unshuffled sequences, error bars are 95% CI for mean, scatterplot=per-attention head result. Identifiability decreases towards 1 (pure random token features) when shuffled inputs produce very different embeddings from the intact inputs, while self-attention distance increases towards 1 (pure random attention) in this case. While later layers in both models are more order-sensitive, information is retained for shuffled inputs.

Self-Attention Distance. Let $\texttt{AMLM}_{l,h}(x)$ be the row-l_1-normalized $\mathbb{R}^{t\times t}$ matrix representing the self-attention matrix at layer l for attention head h. We can compute the same matrix for a shuffled input $\texttt{AMLM}_{l,h}(x')$, and then perform a transformation to re-order the rows and columns of this matrix to match the original order of tokens in x, yielding $\texttt{AMLM}^x_{l,h}(x')$. We then define the *row-wise Jensen-Shannon divergence* DS-JSD$(\texttt{AMLM}_{l,h}(x), \texttt{AMLM}^x_{l,h}(x'))$ as the mean row-wise JSD between $\texttt{AMLM}_{l,h}(x)$ and the DeShuffled reordered attention matrix $\texttt{AMLM}^x_{l,h}(x')$. As before, to reduce the effect of sentence length, we normalize using RND-JSD$(\texttt{AMLM}_{l,h}(x), \texttt{AMLM}^x_{l,h}(x'))$, which chooses a random row/column permutation. [7] The

[6] In practice, we simply compute the assignment step of AA using a $\mathbb{R}^{t\times t}$ matrix drawn from $U[0, 1)$.

[7] If there are multiple possible valid permutations of x' that match x (e.g., if there are repeated words), DS-JSD will choose the order that minimizes the JSD, and RND-JSD will search through a number of random orderings equal to the number of valid permutations. If the number of valid permutations is > 16, 16 random valid permutations are sampled.

	MNLI-(m/mm) Acc/Acc	QQP F1/Acc	QNLI Acc	SST-2 Acc	CoLA MCC	STS-B PCC-r/SCC-ρ	MRPC F1/Acc	RTE Acc
RoBERTa (full seq)	87.3/87.1	72.0/88.8	92.9	95.8	58.8	89.5/88.8	90.2/86.6	69.9
BoW-RoBERTa	81.1/82.8	68.8/87.5	86.8	85.5	10.4	85.0/83.8	82.1/76.6	58.8
BERT (full seq)	84.2/83.2	71.6/89.1	90.6	92.6	50.7	87.3/86.4	87.5/82.8	68.4
BoW-BERT	79.8/79.7	68.3/87.5	86.2	86.7	14.3	81.8/80.3	82.9/75.2	60.4
CBOW GloVe	56.0/56.4	51.4/79.1	72.1	80.0	0.0	61.2/58.7	81.5/73.4	54.1

Table 1: GLUE test set prediction results.

final *attention distance* metric is defined as

$$\text{AD-MLM}(x, l, h) = \frac{\mathbb{E}_{X'}[\text{DS-JSD}(\texttt{AMLM}_{l,h}(x), \texttt{AMLM}^{x}_{l,h}(X'))]}{\mathbb{E}_{X'}[\text{RND-JSD}(\texttt{AMLM}_{l,h}(x), \texttt{AMLM}^{x}_{l,h}(X'))]} \quad (2)$$

Results. We randomly sample 100 sentences from each training set of 8 GLUE tasks, for a total of 800 sentences. To approximate expectations from Equations 1 and 2, we sample 32 random permutations per sentence. Figure 1 gives the per-layer token identifiability/attention similarity scores for both MLMs. For both metrics, later layers are more order sensitive to order, i.e., ID-MLM ↓ and AD-MLM ↑. Attention heads vary significantly in their order sensitivity: each attention head is a single point in the scatterplot of Figure 1b. But, even at late layers, both metrics suggest significantly more than random correspondence: internal representations of BoW-(Ro)BERT(a) clearly resemble their unshuffled counterparts.

4 BoW-BERT for Classification

We compare BERT and RoBERTa to their BoW counterparts on nine tasks from GLUE (Wang et al., 2019).[8] We run single-task training for six epochs, use early stopping, and optimize batch size ($\{16, 32\}$) and learning rate ($\{5, 2, 1, .5\} \times 10^{-5}$) via grid search on the validation set. To shuffle documents: we lowercase, tokenize, remove all tokens that consist only of punctuation, shuffle, then concatenate with whitespaces. We re-shuffle the training tokens each epoch, but fix validation and test tokens to one shuffled permutation.

[8]These tasks span NLI (MNLI (Williams et al., 2018), QNLI (Rajpurkar et al., 2016), and RTE (Dagan et al., 2006; Bar Haim et al., 2006; Giampiccolo et al., 2007; Bentivogli et al., 2009)); semantic similarity estimation (QQP,[9] MRPC (Dolan and Brockett, 2005), STS-B (Cer et al., 2017)); sentiment analysis (SST-2 (Socher et al., 2013)), and grammaticality judgement (CoLA (Warstadt et al., 2019)). We omit WNLI (Levesque et al., 2011) as is common (all models achieve chance performance on that corpus).

Results. Table 1 gives the GLUE test set results of our algorithms vs. GloVe CBOW, the best BoW baseline on the GLUE leaderboard at the time of submission. In all cases BoW-BERT outperforms CBOW. The extent to which BoW-BERT underperforms relative to BERT varies for each dataset, but in terms of relative percent performance decrease, ranges from over ↓70% for CoLA to only ↓3% QQP. Outside of CoLA, performance degradation never exceeds 10 absolute points for any task's metric.

According to the GLUE diagnostic set (which tests 33 categories of linguistic phenomena) BoW-BERT has the most trouble with dealing with double negations (e.g., "I have never seen a hummingbird not flying.": MCC degrades 31.7 → -4.3 when switching BERT → BoW-BERT), quantifiers ("our sympathy to all [vs. some] of the victims": 61.8 → 46.1); and temporal logic ("Mary left before John entered": 8.0 → -8.6). Results for GLUE diagnostic meta-categories are: Knowledge (24.4 → 24.3); Pred-Arg Structure (39.2 → 39.1); Logic (24.7 → 22.1); Lexical Semantics (39.7 → 31.5).

Classification for Sensitive Texts

Privacy and legal concerns frequently necessitate BoW-only data releases. We ask: for potentially sensitive text classification tasks, *how does performance degrade if only bag of words counts are available (instead of full sequences)?* We consider three such tasks: Reddit controversy prediction on AskWomen/AskMen (CONT) (Hessel and Lee, 2019), offensiveness prediction in social media (SBF) (Sap et al., 2020), and sample medical transcript categorization (MTSAMP).[10] For each task, we compare models with access to sequences vs. models that can only access bag-of-words features. Our baselines are unigram/tfidf linear models, and CBOW models GloVe and fasttext (Mikolov et al., 2018). Table 2 contains corpus

[10]https://www.mtsamples.com/

statistics and prediction results. For CONT and SBF, BoW-BERT outperforms all BoW methods. For all tasks, performance drop-off from a full-sequence fine-tuned MLM to its BoW counterpart is less than 1%. CBOW/tfidf remain strong for MTSAMP, in which documents are longer.

Given that de-shuffling BoW representations is at least partially possible (Tao et al., 2021), we additionally consider a more robust *differentially private* (DP) unigram count data release (also known as the "local model" of DP) (Warner, 1965; Dwork et al., 2006; Schein et al., 2019). We follow a process similar to Schofield et al. (2019) by first compressing the original unigram count matrices via Gaussian random projection to 500D.[11] In the compressed space, we add noise per-entry with the Laplace mechanism (Dwork et al., 2006) with a per-feature privacy budget of ε. Then, we invert the random projection, normalize the vector to be a categorical word distribution, and sample (unordered) pseudodocuments from the resulting distribution with length $\sim Poisson(\ell)$.

We report results in an easier setting $\ell = 256, \varepsilon = 100$ and a harder setting $\ell = 128, \varepsilon = 50$ in the bottom half of Table 2. For these settings of DP, the linear baselines generally outperform BoW-(Ro)BERT(a). However, MLMs are again most competitive for the shortest document setting, SBF, where BoW-(Ro)BERT(a) exceeds the best linear model performance (60.4 vs. 62.0 F1).

Taken together, these results suggest 1) that releasing word counts instead of full document sequences is a viable data release strategy for some sensitive classification tasks; 2) BoW-BERT offers a means of accessing the representational power of modern MLMs in cases where only BoW information is available; and 3) for at least some local DP settings, linear models remain competitive particularly for long documents, while BoW-RoBERTa is viable when the underlying documents are shorter.

5 Conclusion and Future Work

We advocate for BoW-(Ro)BERT(a) as a surprisingly strong baseline for language understanding tasks, as well as a performant practical option for classifying (privatized) BoW texts when documents are short. Future work includes:

1. Evaluating BoW-BERT representations on BoW-only corpora in unsupervised text clustering scenarios (vs. classification) + designing self-supervised objectives for fine-tuning MLM weights from unlabelled domain-specific BoW corpora, e.g., HathiTrust.;
2. Extending (K et al., 2020) by further exploring BoW classification using non-English MLMs, where model dependence on syntactic information may differ;
3. Designing local private data release methods better adapted to MLM fine-tuning.

Acknowledgments. We thank David Mimno, Gregory Yauney, Chandra Bhagavatula, Keisuke Sakaguchi, and Max Chen for helpful discussions, comments, suggestions, and (in one case) physically rescuing files from an unplugged server. We also thank Gautam Kamath for feedback on an earlier version of the differential privacy section. Finally, we thank both our EACL and ACL reviewers for their thoughtful feedback.

	CONT	SBF	MTSAMP
Mean len (toks)	111	23	578
# of docs	6.3K	45K	5.0K
# classes	2	2	40
	Acc	F1	Acc/W-F1
BERT (full seq)	65.2	84.1	30.1/27.4
BoW-BERT	**64.1**	**83.4**	34.3/29.6
RoBERTa (full seq)	*66.5*	*84.8*	31.5/29.1
BoW-RoBERTa	62.9	82.9	34.9/32.0
CBOW fasttext	61.7	77.7	***39.4/36.0***
CBOW GloVe	61.1	77.0	38.8/35.2
Unigram tfidf	57.3	78.9	36.2/25.0
Unigram Counts	58.0	79.5	33.5/20.6
Popular Class	50.0	0.0	20.7/7.1
Random Prediction	51.2	47.3	8.9/8.5
$DP^{\varepsilon=100}_{\ell=256}$ BoW-BERT	53.4	59.5	29.0/15.7
$DP^{\varepsilon=100}_{\ell=256}$ BoW-RoBERTa	53.0	62.0	28.9/14.9
$DP^{\varepsilon=100}_{\ell=256}$ Best Linear	57.7	60.4	31.3/21.5
$DP^{\varepsilon=50}_{\ell=128}$ BoW-BERT	50.5	57.0	22.4/10.8
$DP^{\varepsilon=50}_{\ell=128}$ BoW-RoBERTa	51.8	58.9	21.8/10.7
$DP^{\varepsilon=50}_{\ell=128}$ Best Linear	55.0	58.8	25.9/17.8

Table 2: Top: text classification prediction results on sensitive texts; **best BoW** bolded, *best overall* italicized. Bottom: *DP* = results on differentially private data; "Best Linear" is the most performant linear model, tfidf for $DP^{\varepsilon=100}_{\ell=256}$ and unigram counts for $DP^{\varepsilon=50}_{\ell=128}$.

[11] Our original submission used DP PCA instead. But it was brought to our attention that the paper proposing that algorithm was retracted for being non-private (+ discontinued in the library we used after we submitted). We have adjusted our code and recompiled our experiments using a comparable mechanism. Our intent isn't to advocate for this particular DP method, but rather, to fairly compare NLP algorithms on the same DP corpora.

References

Authors Guild, Inc. v. Hathitrust, 755 F.3d 87 (2d Cir. 2014).

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *ICLR*.

Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. In *NAACL*.

Roy Bar Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second PASCAL recognising textual entailment challenge.

Ghazaleh Beigi, Kai Shu, Ruocheng Guo, Suhang Wang, and Huan Liu. 2019. I am not what I write: Privacy preserving text representation learning. *arXiv preprint arXiv:1907.03189*.

Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, and Bernardo Magnini. 2009. The fifth PASCAL recognizing textual entailment challenge.

Samuel R Bowman and George E Dahl. 2021. What will it take to fix benchmarking in natural language understanding? In *NAACL*.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *ACL*.

Heather Christenson. 2011. Hathitrust. *Library Resources & Technical Services*, 55(2):93–102.

Volkan Cirik, Louis-Philippe Morency, and Taylor Berg-Kirkpatrick. 2018. Visual referring expression recognition: What do systems actually learn? In *NAACL*.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, pages 177–190. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

William B Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the International Workshop on Paraphrasing*.

Cynthia Dwork. 2008. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Theory of Cryptography Conference*, pages 265–284. Springer.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *TACL*.

Natasha Fernandes, Mark Dras, and Annabelle McIver. 2019. Generalised differential privacy for text document processing. In *ETAPS*. Springer-VDI-Verlag GmbH & Co. KG.

Matthias Gallé and Matías Tealdi. 2015. Reconstructing textual documents from n-grams. In *KDD*.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*.

Goran Glavaš and Ivan Vulić. 2020. Is supervised syntactic parsing beneficial for language understanding? an empirical investigation. *arXiv preprint arXiv:2008.06788*.

Yoav Goldberg. 2019. Assessing BERT's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

Ashim Gupta, Giorgi Kvernadze, and Vivek Srikumar. 2021. BERT & family eat word salad: Experiments with text understanding. *arXiv preprint arXiv:2101.03453*.

Jack Hessel and Lillian Lee. 2019. Something's brewing! early prediction of controversy-causing posts from discussion features. In *NAACL*, Minneapolis, Minnesota.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *NAACL*.

Phu Mon Htut, Jason Phang, Shikha Bordia, and Samuel R Bowman. 2019. Do attention heads in BERT track syntactic dependencies? *arXiv preprint arXiv:1911.12246*.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *ACL*.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual BERT: an empirical study. In *ICLR*.

Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. 2018. Sharp nearby, fuzzy far away: How neural language models use context. In *ACL*.

Walter Kintsch. 1988. The role of knowledge in discourse comprehension: A construction-integration model. *Psychological review*, 95(2):163.

Hector J Levesque, Ernest Davis, and Leora Morgenstern. 2011. The Winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, volume 46, page 47.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside BERT's linguistic knowledge. In *ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *TACL*, 4:521–535.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. Linguistic knowledge and transferability of contextual representations. In *NAACL*.

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. *arXiv preprint arXiv:1907.11692*.

Zihan Liu, Genta Indra Winata, Samuel Cahyawijaya, Andrea Madotto, Zhaojiang Lin, and Pascale Fung. 2020. On the importance of word order information in cross-lingual sequence labeling. *arXiv preprint arXiv:2001.11164*.

David Mareček and Rudolf Rosa. 2018. Extracting syntactic trees from transformer encoder self-attentions. In *Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A Nowak, and Erez Liberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Yixin Nie, Yicheng Wang, and Mohit Bansal. 2019. Analyzing compositionality-sensitivity of NLI models. In *AAAI*.

Thang M Pham, Trung Bui, Long Mai, and Anh Nguyen. 2020. Out of order: How important is the sequential order of words in a sentence in natural language understanding tasks? *arXiv preprint arXiv:2012.15180*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *EMNLP*.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of BERT. In *NeurIPS*.

Chinnadhurai Sankar, Sandeep Subramanian, Chris Pal, Sarath Chandar, and Yoshua Bengio. 2019. Do neural dialog systems use the conversation history effectively? An Empirical Study. In *ACL*.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *ACL*.

Aaron Schein, Zhiwei Steven Wu, Alexandra Schofield, Mingyuan Zhou, and Hanna Wallach. 2019. Locally private Bayesian inference for count models. In *ICML*.

Alexandra Schofield, Gregory Yauney, and David Mimno. 2019. Combatting the challenges of local privacy for distributional semantics with compression. In *PriML Workshop at NeurIPS*.

Chenglei Si, Shuohang Wang, Min-Yen Kan, and Jing Jiang. 2019. What does BERT learn from multiple-choice reading comprehension datasets? *arXiv preprint arXiv:1910.12391*.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. *arXiv preprint arXiv:2104.06644*.

Koustuv Sinha, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams. 2020. Unnatural language inference. *arXiv preprint arXiv:2101.00010*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*.

Saku Sugawara, Pontus Stenetorp, Kentaro Inui, and Akiko Aizawa. 2020. Assessing the benchmarking capacity of machine reading comprehension datasets. In *AAAI*.

Chongyang Tao, Shen Gao, Juntao Li, Yansong Feng, Dongyan Zhao, and Rui Yan. 2021. Learning to organize a bag of words into sentences with neural networks: An empirical study. In *NAACL*.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *ICLR*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *ICLR*.

Wei Wang, Bin Bi, Ming Yan, Chen Wu, Jiangnan Xia, Zuyi Bao, Liwei Peng, and Luo Si. 2020. StructBERT: Incorporating language structures into pretraining for deep language understanding. In *ICLR*.

Stanley L Warner. 1965. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *TACL*.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of NAACL-HLT*.

Baosong Yang, Longyue Wang, Derek F. Wong, Lidia S. Chao, and Zhaopeng Tu. 2019. Assessing the ability of self-attention networks to learn word order. In *ACL*.

Wei Zhao, Goran Glavaš, Maxime Peyrard, Yang Gao, Robert West, and Steffen Eger. 2020. On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation. In *ACL*.