Informing Unsupervised Pretraining with External Linguistic Knowledge

Anne Lauscher[1], Ivan Vulić[2], Edoardo Maria Ponti[2], Anna Korhonen[2], and Goran Glavaš[1]

[1]Data and Web Science Group, University of Mannheim, Germany
[2]Language Technology Lab, University of Cambridge, UK
[1]{anne,goran}@informatik.uni-mannheim.de,
[2]{iv250,ep490,alk23}@cam.ac.uk

Abstract

Unsupervised pretraining models have been shown to facilitate a wide range of downstream applications. These models, however, still encode only the distributional knowledge, incorporated through language modeling objectives. In this work, we complement the encoded distributional knowledge with external lexical knowledge. We generalize the recently proposed (state-of-the-art) unsupervised pretraining model BERT to a multi-task learning setting: we couple BERT's masked language modeling and next sentence prediction objectives with the auxiliary binary word relation classification, through which we inject clean linguistic knowledge into the model. Our initial experiments suggest that our "linguistically-informed" BERT (LIBERT) yields performance gains over the linguistically-blind "vanilla" BERT on several language understanding tasks.

1 Introduction

Unsupervised pretraining models, such as GPT and GPT-2 (Radford et al., 2018, 2019), ELMo (Peters et al., 2018), and BERT (Devlin et al., 2019) yield state-of-the-art performance on a wide range of natural language processing tasks. All these models rely on language modeling objectives that exploit the knowledge encoded in large corpora. BERT (Devlin et al., 2019), as the current state-of-the-art model, is pretrained on a joint objective consisting of two parts: (1) masked language modeling (MLM), and (2) next sentence prediction (NSP). Through both of these objectives, BERT still consumes only the distributional knowledge.

A plethora of models have been proposed for injecting linguistic constraints (e.g., lexical knowledge) from external resources to static word embeddings (Faruqui et al., 2015; Wieting et al., 2015; Mrkšić et al., 2017; Ponti et al., 2018, *inter alia*). Linguistically-informed word vectors produced by these models produce substantial gains in a number of downstream tasks, e.g., in dialog state tracking (Mrkšić et al., 2017; Ren et al., 2018), text simplification (Glavaš and Vulić, 2018; Saggion, 2017), and taxonomy induction (Nguyen et al., 2017; Nickel and Kiela, 2018). Like static embedding models, unsupervised pretraining models also operate only on large text corpora. We hypothesize that supplementing them with clean linguistic information from structured external resources may also lead to their improved downstream performance.

We aim to inject linguistic constraints, available from lexico-semantic resources like WordNet (Miller, 1995) and BabelNet (Navigli and Ponzetto, 2012), into unsupervised pretraining models. As a first step in this direction, we present linguistically-informed BERT (LIBERT), a simple and effective augmentation of BERT made aware of external linguistic knowledge. We (1) feed linguistic constraints (synonyms and direct hypernym-hyponym pairs) to BERT as additional training instances and (2) predict lexico-semantic relations from constraint embeddings produced by BERT's encoder (Vaswani et al., 2017): we add lexical relation classification (LRC) as the third pretraining task.

For direct comparability, we train the same model from scratch – with the augmentation (LIBERT) and without it (BERT). We then fine-tune both models on the training portions of datasets from the GLUE benchmark (Wang et al., 2018) and report their performance on corresponding development and test portions. LIBERT yields performance gains over BERT on 8/10 GLUE tasks.

2 Linguistically-Informed BERT

LIBERT is illustrated in Figure 1. LIBERT is also a pretraining model: it augments BERT's two pretraining tasks – masked language modeling (1.

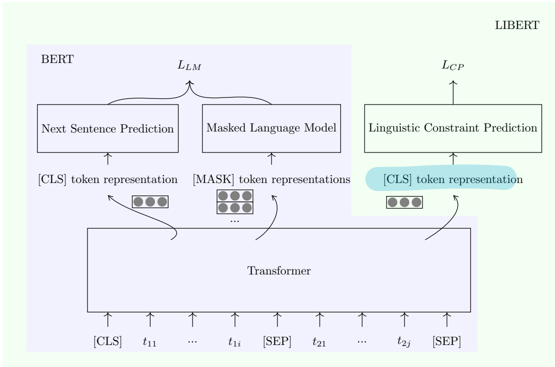


Figure 1: Architecture of LIBERT – linguistically-informed BERT.

MLM) and next sentence prediction (2. NSP) with an additional task of identifying (i.e., classifying) valid lexico-semantic relations from an external resource (3. LRC). After pretraining, LIBERT is, like BERT, fine-tuned on training sets of downstream tasks. For completeness, we first briefly describe the base BERT model and then provide the details of our linguistically-informed augmentation.

2.1 BERT: Transformer-Based Encoder

The core of the BERT model is a multi-layer bidirectional Transformer (Vaswani et al., 2017), pretrained using two objectives: (1) masked language modeling (MLM) and (2) next sentence prediction (NSP). MLM is a token-level prediction task, also referred to as *Cloze* task (Taylor, 1953): from input data, a certain percentage of tokens is masked out and needs to be predicted. NSP operates on the sentence-level and can, therefore, be seen as a higher-level language modeling task that captures information across sentences. NSP predicts if two given sentences are adjacent in text (negative examples are created by randomly pairing sentences).

2.2 Linguistically-Informed Pretraining

The base BERT model consumes only the distributional information. We aim to make the model more informed by exposing it to clean external knowledge presented as the set of linguistic constraints $C = \{(w_1, w_2)_i\}_{i=1}^{N}$, i.e., pairs of words that stand in a particular relation in some external lexico-semantic resource. Following the successful work on semantic specialization of static word embeddings (Wieting et al., 2015; Mrkšić et al., 2017; Vulić and Mrkšić, 2018), in this work we select pairs of synonyms (e.g., *car* and *automobile*) and direct hyponym-hypernym pairs (e.g., *car* and *vehicle*) as our constraints.[1]

We transform the constraints from C into a BERT-compatible input format and feed them as additional training examples for the model. The encoding of a constraint is then forwarded to the relation classifier, which predicts whether the input word pair represents a valid lexical relation.

From Linguistic Constraints to Training Instances. We start from a set of linguistic constraints $C = \{(w_1, w_2)_i\}_{i=1}^{N}$ and an auxiliary static word embedding space $\mathbf{X_{aux}} \in \mathbb{R}^d$. Each constraint $c = (w_1, w_2)$ corresponds to a true lexical relation, and thus represents a *positive* training example for the model. For each positive example c, we create corresponding negative examples as follows. We first group positive constraints from C in mini-batches B_p of size k. For each positive example $c = (w_1, w_2)$, we create two negative instances $\hat{c}_1 = (\hat{w}_1, w_2)$ and $\hat{c}_2 = (w_1, \hat{w}_2)$ such that $\hat{w}_1$ is the word from batch B_p (other than w_1) closest to w_2 and $\hat{w}_2$ the word (other than w_2) closest to w_1, respectively, in terms of the cosine similarity of their vectors in $\mathbf{X_{aux}}$. This way we create a batch B_n of $2k$ negative training instances from a batch B_p of k positive training instances.

Next, we transform each instance (i.e., a pair of words) into a "BERT-compatible" format, i.e., into a sequence of WordPiece (Wu et al., 2016) tokens.[2] We split both w_1 and w_2 into WordPiece tokens, insert the special separator token (with a randomly initialized embedding) before and after the tokens of w_2 and prepend the whole sequence with BERT's sequence start token, as shown in this example for the constraint (*mended*, *regenerated*):[3]

[CLS]	men	#ded	[SEP]	reg	#ener	#ated	[SEP]
0	0	0	0	1	1	1	1

As in the original work (Devlin et al., 2019), we sum the WordPiece embedding of each token with the embeddings of the segment and position of the token. We assign the segment ID of `0` to the `[CLS]` token, all w_1 tokens, and the first `[SEP]`

[1]The goal is to inform the BERT model on the relation of *true semantic similarity* between words (Hill et al., 2015): according to prior work on static word embeddings, the sets of both synonym pairs and direct hyponym-hypernym pairs are useful to boost the model's ability to capture true semantic similiarity, which in turn has a positive effect on downstream language understanding applications (Vulić, 2018).

[2]We use the same 30K WordPiece vocabulary as Devlin et al. (2019). Sharing WordPieces helps our word-level task as lexico-semantic relationships are similar for words composed of the same morphemes.

[3]The sign # denotes split WordPiece tokens.

token; segment ID `1` is assigned to all tokens of w_2 and the final `[SEP]` token.

Lexical Relation Classifier. Original BERT feeds Transformer-encoded token representations to two classifiers: MLM classifier (predicting the masked tokens), and the NSP classifier (predicting whether two sentences are adjacent). LIBERT introduces the third pretraining classifier: it predicts whether an encoded word pair represents a valid lexical relation (i.e., a positive example where two words stand in the relation of true semantic similarity – synonyms or hypernym-hyponym pairs) or not. Let $\mathbf{x}_{CLS} \in \mathbb{R}^h$ be the transformed vector representation of the sequence start token [CLS] that encodes the whole constraint (w_1, w_2). Our lexical relation predictor (LRC) is a simple softmax classifier:

$$\hat{\mathbf{y}} = \mathrm{softmax}(\mathbf{x}_{CLS}\mathbf{W}_{LRC}^{\top} + \mathbf{b}_{LRC})\,, \qquad (1)$$

with $W_{LRC} \in \mathbb{R}^{H\times 2}$ and $b_{LRC} \in \mathbb{R}^2$ as the classifier's trainable parameters. Relation classification loss L_{LRC} is then simply the negative log-likelihood over k instances in the training batch:

$$L_{LRC} = -\sum_{k} \ln \hat{\mathbf{y}}_k \cdot \mathbf{y}_k. \qquad (2)$$

where $\mathbf{y} \in \{[0,1],[1,0]\}$ is the true relation label for a word-pair training instance.

3 Experimental Setup

To isolate the effects of injecting linguistic knowledge into BERT, we train base BERT and LIBERT in the same setting: the only difference is that we additionally update the parameters of LIBERT's Transformer encoder based on the gradients of the LRC loss L_{LRC} from Eq. (2).

Pretraining Data. We minimize BERT's original objective $L_{MLM}+L_{NSP}$ on training examples coming from English Wikipedia.[4] We obtain the set of constraints C for the L_{LRC} term from the body of previous work on semantic specialization of static word embeddings (Zhang et al., 2014; Vulić et al., 2018; Ponti et al., 2018). In particular, we collect 1,023,082 synonymy pairs from WordNet (Miller, 1995) and Roget's Thesaurus (Kipfer, 2009) and 326,187 direct hyponym-hypernym pairs (Vulić and Mrkšić, 2018) from WordNet.[5]

[4]We acknowledge that training the models on larger corpora would likely lead to better absolute downstream scores; however, the main goal of this work is not to achieve state-of-the-art downstream performance, but to compare the base model against its linguistically-informed augmentation.

Fine-Tuning (Downstream) Tasks. We evaluate BERT and LIBERT on the the following tasks from the GLUE benchmark (Wang et al., 2018), where sizes of training, development, and test datasets for each task are given in Table 1:

CoLA (Warstadt et al., 2018): Binary sentence classification predicting if sentences from linguistic publications are grammatically acceptable;
SST-2 (Socher et al., 2013): Binary sentence classification, predicting sentiment (positive or negative) for movie review sentences;
MRPC (Dolan and Brockett, 2005): Binary sentence-pair classification predicting whether two sentences are mutual paraphrases;
STS-B (Cer et al., 2017): Sentence-pair regression task – predicting the degree of semantic similarity for a pair of sentences;
QQP (Chen et al., 2018): Binary classification task of recognizing question parahrases;
MNLI (Williams et al., 2018): Ternary natural language inference (NLI) classification of sentence pairs. Two test sets are given: a matched version (MNLI-m) in which the test domains match with training data domains, and a mismatched version (MNLI-mm) with different test domains;
QNLI: A binary classification version of the Stanford Q&A dataset (Rajpurkar et al., 2016);
RTE (Bentivogli et al., 2009): Another NLI dataset, ternary entailment classification for sentence pairs;
AX (Wang et al., 2018): A small, manually curated NLI dataset (i.e., a ternary classification task), with examples encompassing different linguistic phenomena relevant for entailment.[6]

Training and Evaluation. We train both BERT and LIBERT from scratch, with the configuration of the BERT$_{BASE}$ model (Devlin et al., 2019): $L = 12$ transformer layers with the hidden state size of $H = 768$, and $A = 12$ self-attention heads. We train in batches of $k = 16$ instances;[7] input sequence length is 128. The learning rate for both models is $2e-5$ with a warm-up over the first

[5]Note again that similar to work of Vulić (2018), both WordNet synonyms and direct hyponym-hypernym pairs are treated the same: as positive examples for the relation of true semantic similarity.

[6]Following Devlin et al. (2019), we do not evaluate on the Winograd NLI (WNLI), given its well-known issues.

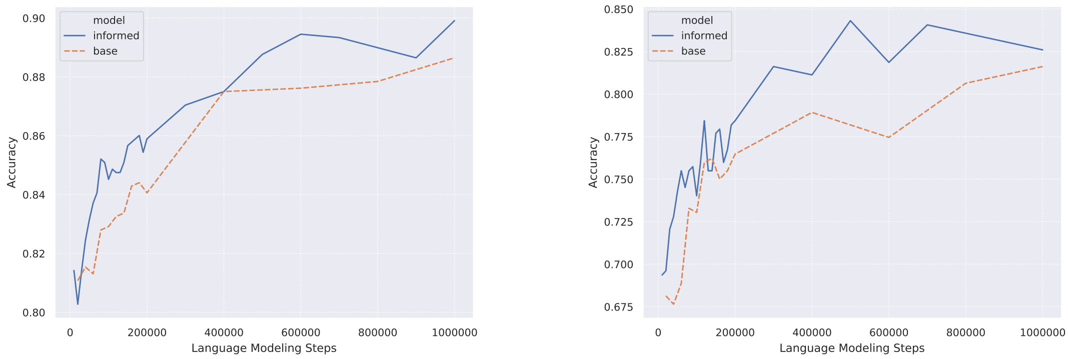
[7]Due to hardware restrictions, we train in batches that are half the size of the training batches from the original work (Devlin et al., 2019) ($k = 32$). This means that for the same number of update steps, our models will have seen half of the amount of the original BERT model of Devlin et al. (2019).

	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	AX
# Train	8,551	67,349	3,668	5,749	363,870	392,702	392,702	104,743	2,490	–
# Dev	1,042	872	408	1,501	40,431	9,815	9,832	5,463	278	–
# Test	1,063	1,821	1,725	1,379	390,964	9,796	9,847	5,463	3,000	1,104

Table 1: Data set sizes for tasks in the GLUE benchmark (Wang et al., 2018).

		CoLA MCC	SST-2 Acc	MRPC F1/Acc	STS-B Pears/Spearm	QQP F1/Acc	MNLI-m Acc	MNLI-mm Acc	QNLI Acc	RTE Acc	AX MCC
Dev	BERT	29.4	88.7	87.1/81.6	86.4/73.3	85.9/89.5	78.2	**78.8**	86.2	63.9	–
	LIBERT	**35.3**	**89.9**	**87.9/82.6**	**87.2/75.6**	**86.3/89.8**	**78.5**	78.7	**86.5**	**65.3**	–
	Δ	+5.9	+1,2	+0.8/+1.0	+0.8/2.3	+0.4/+0.3	+0.3	-0.1	+0.3	+1.4	–
Test	BERT	21.5	87.9	84.8/78.8	**80.8/79.3**	68.6/87.9	78.2	**77.6**	85.8	61.3	26.3
	LIBERT	**31.4**	**89.6**	**86.1/80.4**	80.5/78.8	**69.0/88.1**	**78.4**	77.4	**86.2**	**62.6**	**32.8**
	Δ	+9.9	+1.7	+1.3/+1.6	-0.3/-0.5	+0.4/+0.2	+0.2	-0.2	+0.4	+1.3	+6.5

Table 2: Results on the dev and test sets of 10 GLUE tasks after 1M MLM+NSP steps with BERT and LIBERT.



(a) SST-2

(b) MRPC

Figure 2: Accuracy over time for BERT and LIBERT on (a) SST-2 and (b) MRPC on the corresponding dev sets.

1, 000 training steps. Other hyperparameters are set to the values reported by Devlin et al. (2019).

LIBERT combines BERT's MLM and NSP objectives with our LRC objective in a multi-task learning setup. We update its parameters in a balanced alternating regime: (1) we first minimize BERT's $L_{MLM} + L_{NSP}$ objective on one batch of masked sentences pairs and then (2) minimize the LRC objective L_{LRC} on one batch of training instances created from linguistic constraints.

During fine-tuning, for each task we independently find the optimal hyperparameter configurations of the downstream classifiers for the pretrained BERT and LIBERT: this implies that it is valid to compare their performances on the downstream dev sets. Finally, we evaluate fine-tuned BERT and LIBERT on all 10 test sets.

4 Results and Discussion

The main results are summarized in Table 2. LIBERT outperforms BERT on 8/9 tasks (dev) and 8/10 tasks (test). While large gains are reported on CoLA, AX, and visible gains on SST-2 and MRPC, it is encouraging to see that slight and consistent gains are observed on almost all other tasks. These results suggest that available external linguistic knowledge can be used to supplement unsupervised pretraining models with useful information which cannot be fully captured solely using large text data. From another perspective, the results indicate that LIBERT, our linguistically informed multi-task method, successfully blends such curated linguistic knowledge with distributional learning signals. It also further validates intuitions from relevant work on specialising static word embeddings (Wieting et al., 2015; Mrkšić et al., 2017) that steering distributional models towards capturing true semantic similarity (as also done here) has a positive impact on language understanding applications.

Further, an analysis of performance over time (in terms of training steps for BERT and LIBERT) for one single-sentence task (SST-2) and one sentence-pair classification task (MRPC) is reported in Figures 2a-2b. The scores clearly suggest that the impact of external knowledge does not vanish over time: the gains with the linguistically informed

LIBERT persist at different time steps. This finding again hints on the complementarity of useful signals coded in large text data vs. lexical resources (Faruqui, 2016; Mrkšić et al., 2017) which should be investigated more in future work.

5 Conclusion

We presented LIBERT, a linguistically-informed extension of the state-of-the-art unsupervised pretraining model BERT. LIBERT (1) uses BERT's Transformer network to additionally encode clean external lexico-semantic constraints and (2) couples BERT's two pretraining tasks – masked language modeling and next sentence prediction – with a lexical relation classifier in a multi-task learning setup. LIBERT yields improvements over BERT on 8 out of 10 language understanding tasks from the GLUE benchmark, suggesting that the complementarity between distributional and clean linguistic information is beneficial for unsupervised pretraining and warrants further investigation.

Acknowledgments

IV, EMP, and AK are supported by the ERC Consolidator Grant LEXICAL: Lexical Acquisition Across Languages (no 648909) awarded to AK.

References

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The Fifth PASCAL recognizing textual entailment challenge. In *Proceedings of TAC*.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of SemEval*, pages 1–14.

Zihan Chen, Hongbo Zhang, Xiaoji Zhang, and Leqi Zhao. 2018. Quora question pairs.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Manaal Faruqui. 2016. *Diverse Context for Learning Word Representations*. Ph.D. thesis, Carnegie Mellon University.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of NAACL-HLT*, pages 1606–1615.

Goran Glavaš and Ivan Vulić. 2018. Explicit retrofitting of distributional word vectors. In *Proceedings of ACL*, pages 34–45.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Barbara Ann Kipfer. 2009. *Roget's 21st Century Thesaurus (3rd Edition)*. Philip Lief Group.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic specialisation of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the ACL*, 5:309–324.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Kim Anh Nguyen, Maximilian Köper, Sabine Schulte im Walde, and Ngoc Thang Vu. 2017. Hierarchical embeddings for hypernymy detection and directionality. In *Proceedings of EMNLP*, pages 233–243.

Maximilian Nickel and Douwe Kiela. 2018. Learning continuous hierarchies in the Lorentz model of hyperbolic geometry. In *Proceedings of ICML*, pages 3776–3785.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.

Edoardo Maria Ponti, Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. 2018. Adversarial propagation and zero-shot cross-lingual transfer of word vector specialization. In *Proceedings of EMNLP*, pages 282–293.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *OpenAI Technical Report*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1:8.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of EMNLP*, pages 2383–2392.

Liliang Ren, Kaige Xie, Lu Chen, and Kai Yu. 2018. Towards universal dialogue state tracking. In *Proceedings of EMNLP*, pages 2780–2786.

Horacio Saggion. 2017. Automatic text simplification. *Synthesis Lectures on Human Language Technologies*, 10(1):1–137.

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 *Proceedings of EMNLP*, pages 1631–1642.

Wilson L. Taylor. 1953. cloze procedure: A new tool for measuring readability. *Journalism Bulletin*, 30(4):415–433.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NeurIPS*, pages 5998–6008.

Ivan Vulić. 2018. Injecting lexical contrast into word vectors by guiding vector space specialisation. In *Proceedings of The 3rd Workshop on Representation Learning for NLP*, pages 137–143.

Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. 2018. Post-specialisation: Retrofitting vectors of words unseen in lexical resources. In *Proceedings of NAACL-HLT*, pages 516–527.

Ivan Vulić and Nikola Mrkšić. 2018. Specialising word vectors for lexical entailment. In *Proceedings of NAACL-HLT*, pages 1134–1145.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the Blacbox NLP Workshop*, pages 353–355.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2018. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. From paraphrase database to compositional paraphrase model and back. *Transactions of the ACL*, 3:345–358.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of NAACL-HLT*, pages 1112–1122.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Jingwei Zhang, Jeremy Salwen, Michael Glass, and Alfio Gliozzo. 2014. Word semantic representations using bayesian probabilistic tensor factorization. In *Proceedings of EMNLP*, pages 1522–1531.