Specializing Unsupervised Pretraining Models for Word-Level Semantic Similarity

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 NLP applications. These models, however, retain some of the limitations of traditional static word embeddings. In particular, they encode only the distributional knowledge available in raw text corpora, incorporated through language modeling objectives. In this work, we complement such distributional knowledge with external lexical knowledge, that is, we integrate the discrete knowledge on word-level semantic similarity into pretraining. To this end, we generalize the standard BERT model to a multi-task learning setting where we couple BERT's masked language modeling and next sentence prediction objectives with an auxiliary task of binary word relation classification. Our experiments suggest that our "Lexically Informed" BERT (LIBERT), specialized for the word-level semantic similarity, yields better performance than the lexically blind "vanilla" BERT on several language understanding tasks. Concretely, LIBERT outperforms BERT in 9 out of 10 tasks of the GLUE benchmark and is on a par with BERT in the remaining one. Moreover, we show consistent gains on 3 benchmarks for lexical simplification, a task where knowledge about word-level semantic similarity is paramount.

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 (LM) objectives that exploit the knowledge encoded in large text corpora. BERT (Devlin et al., 2019), as one of the current state-of-the-art models, 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 encoded by word co-occurrences.

While several concurrent research threads are focused on making BERT optimization more robust (Liu et al., 2019b) or on imprinting external world knowledge on its representations (Zhang et al., 2019a,b; Sun et al., 2019; Liu et al., 2019a; Peters et al., 2019, *inter alia*), no study yet has been dedicated to mitigating a severe limitation that contextualized representations and unsupervised pretraining inherited from static embeddings: every model that relies on distributional patterns has a tendency to conflate together pure lexical semantic similarity with broad topic relatedness (Schwartz et al., 2015; Mrkšić et al., 2017).

In the past, a plethora of models have been proposed for injecting linguistic constraints (i.e., 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*) in order to emphasize a particular lexical relation in a *specialized* embedding space. For instance, lexically informed word vectors specialized for pure semantic similarity result in substantial gains in a number of downstream tasks where such similarity plays an important role, e.g., in dialog state tracking (Mrkšić et al., 2017; Ren et al., 2018) or for lexical simplification (Glavaš and Vulić, 2018; Ponti et al., 2019). Existing specialization methods are, however, not directly applicable to unsupervised pretraining models because they are either (1) tied to a particular training objective of a static word embedding model, or (2) predicated on the existence of an embedding space in which pairwise distances can be modified.

In this work, we hypothesize that supplementing unsupervised LM-based pretraining with clean lexical information from structured external resources

may also lead to improved performance in language understanding tasks. We propose a novel method to inject linguistic constraints, available from lexico-semantic resources like WordNet (Miller, 1995) and BabelNet (Navigli and Ponzetto, 2012), into unsupervised pretraining models, and steer them towards capturing word-level semantic similarity. To train Lexically Informed BERT (LIBERT), we (1) feed semantic similarity constraints to BERT as additional training instances and (2) predict lexico-semantic relations from the constraint embeddings produced by BERT's encoder (Vaswani et al., 2017). In other words, LIBERT adds lexical relation classification (LRC) as the third pretraining task to BERT's multi-task learning framework.

We compare LIBERT to a lexically blind "vanilla" BERT on the GLUE benchmark (Wang et al., 2018) and report their performance on corresponding development and test portions. LIBERT yields performance gains over BERT on 9/10 GLUE tasks (and is on a par with BERT on the remaining one), with especially wide margins on tasks involving complex or rare linguistic structures such as Diagnostic Natural Language Inference and Linguistic Acceptability. Moreover, we assess the robustness and effectiveness of LIBERT on 3 different datasets for lexical simplification (LS), a task proven to benefit from word-level similarity specialization (Ponti et al., 2019). We report LS improvements of up to 8.2% when using LIBERT in lieu of BERT. For direct comparability, we train both LIBERT and BERT from scratch, and monitor the gains from specialization across iterations. Interestingly, these do not vanish over time, which seems to suggest that our specialization approach is suitable also for models trained on massive amounts of raw text data.

2 Related Work

2.1 Specialization for Semantic Similarity

The conflation of disparate lexico-semantic relations in *static* word representations is an extensively researched problem. For instance, clearly discerning between true semantic similarity and broader conceptual relatedness in static embeddings benefits a range of natural language understanding tasks such as dialog state tracking (Mrkšić et al., 2017), text simplification (Glavaš and Vulić, 2018), and spoken language understanding (Kim et al., 2016). The most widespread solution relies on the use of specialization algorithms to enrich word embeddings with external lexical knowledge and steer them towards a desired lexical relation.

Joint specialization models (Yu and Dredze, 2014; Kiela et al., 2015; Liu et al., 2015; Osborne et al., 2016; Nguyen et al., 2017, *inter alia*) jointly train word embedding models from scratch and enforce the external constraints with an auxiliary objective. On the other hand, *retrofitting* models are post-processors that fine-tune pretrained word embeddings by gauging pairwise distances according to the external constraints (Faruqui et al., 2015; Wieting et al., 2015; Mrkšić et al., 2016; Mrkšić et al., 2017; Jo and Choi, 2018).

More recently, retrofitting models have been extended to specialize not only words found in the external constraints, but rather the entire embedding space. In *explicit retrofitting* models (Glavaš and Vulić, 2018), a (deep, non-linear) specialization function is directly learned from external constraints. *Post-specialization* models (Vulić et al., 2018; Ponti et al., 2018; Kamath et al., 2019), instead, propagate lexico-semantic information to unseen words by imitating the transformation undergone by seen words during the initial specialization. This family of models can also transfer specialization across languages (Glavaš and Vulić, 2018; Ponti et al., 2019).

The goal of this work is to move beyond similarity-based specialization of static word embeddings only. We present a novel methodology for enriching unsupervised pretraining models such as BERT (Devlin et al., 2019) with readily available discrete lexico-semantic knowledge, and measure the benefits of such semantic specialization on similarity-oriented downstream applications.

2.2 Injecting Knowledge into Unsupervised Pretraining Models

Unsupervised pretraining models do retain some of the limitations of static word embeddings. First, they still conflate separate lexico-semantic relations, as they learn from distributional patterns. Second, they fail to fully capture the world knowledge necessary for human reasoning: masked language models struggle to recover knowledge base triples from raw texts (Petroni et al., 2019). Recent work has, for the most part, focused on mitigating the latter limitation by injecting structured world knowledge into unsupervised pretraining and contextualized representations.

In particular, these techniques fall into the fol-

lowing broad categories: i) *masking* higher linguistic units of meanings, such as phrases or named entities, rather than individual WordPieces or BPE tokens (Zhang et al., 2019a); ii) including an *auxiliary task* in the objective, such as denoising auto-encoding of entities aligned with text (Zhang et al., 2019b), or continuous learning frameworks over a series of unsupervised or weakly supervised tasks (e.g., capitalization prediction or sentence reordering) (Sun et al., 2019); iii) *hybridizing* texts and graphs. Liu et al. (2019a) proposed a special attention mask and soft position embeddings to preserve their graph structure while preventing unwanted entity-word interactions. Peters et al. (2019) fuse language modeling with an end-to-end entity linker, updating contextual word representations with word-to-entity attention.

As the main contributions of our work, we incorporate external lexico-semantic knowledge, rather than world knowledge, in order to rectify the first limitation, namely the distortions originating from the distributional signal. In fact, Liu et al. (2019a) hybridized texts also with linguistic triples relating words to sememes (minimal semantic components); however, this incurs into the opposite effect of reinforcing the distributional signal based on co-occurrence. On the contrary, we propose a new technique to enable the model to distinguish between purely similar and broadly related words.

3 Specializing for Word-Level Similarity

LIBERT, illustrated in Figure 1, is a *joint* specialization model. It augments BERT's two pretraining tasks – masked language modeling (1. 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). LIBERT is first pretrained jointly on all three tasks. Similarly to BERT, after pretraining, LIBERT is fine-tuned on training datasets of downstream tasks. For completeness, we first briefly outline the base BERT model and then provide the details of our lexically informed augmentation.

3.1 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): among the input data, a certain percentage of tokens is masked out and needs to be recovered. NSP operates on the sentence-level and can, therefore, be seen as a higher-level sequence 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).

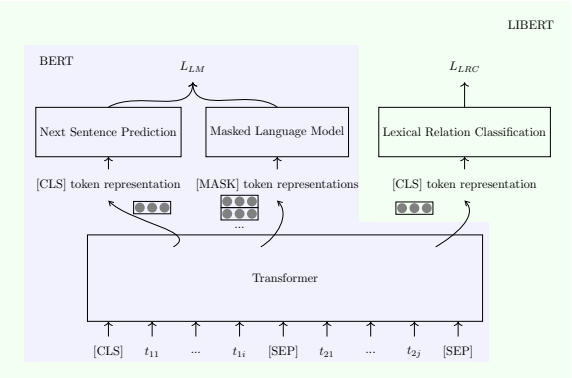


Figure 1: Architecture of LIBERT – lexically-informed BERT specialized with semantic similarity constraints.

3.2 LIBERT: Lexically-Informed (Specialized) Pretraining

The base BERT model consumes only the distributional information. We aim to steer the model towards capturing true semantic similarity (as opposed to conceptual relatedness) 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 the desired relation (i.e., true semantic similarity) in some external lexico-semantic resource. Following the successful work on semantic specialization of static word embeddings (see §2.1), 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 semantic similarity 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 In-

[1] As 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 (Vulić, 2018), the sets of both synonym pairs and direct hyponym-hypernym pairs are useful to boost the model's ability to capture true semantic similarity, which in turn has a positive effect on downstream language understanding applications.

stances. We start from a set of linguistic constraints $C = \{(w_1, w_2)_i\}_{i=1}^{N}$ and an auxiliary static word embedding space $\mathbf{X_{aux}} \in \mathbb{R}^d$. The space $\mathbf{X_{aux}}$ can be obtained via any standard static word embedding model such as Skip-Gram (Mikolov et al., 2013) or fastText (Bojanowski et al., 2017) (used in this work). Each constraint $c = (w_1, w_2)$ corresponds to a true/positive relation of semantic similarity, and thus represents a *positive* training example for the model. For each positive example c, we create corresponding negative examples following prior work on specialization of static embeddings (Wieting et al., 2015; Glavaš and Vulić, 2018; Ponti et al., 2019). We first group positive constraints from C into mini-batches B_p of size k. For each positive example $c = (w_1, w_2)$, we create two negatives $\hat{c}_1 = (\hat{w}_1, w_2)$ 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]` 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 desired lexico-semantic 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 formulated as follows:

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

with $\mathbf{W}_{LRC} \in \mathbb{R}^{H \times 2}$ and $\mathbf{b}_{LRC} \in \mathbb{R}^2$ 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:

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

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

4 Language Understanding Evaluation

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). In the first set of experiments, we probe the usefulness of injecting semantic similarity knowledge on the well-known suite of GLUE tasks (Wang et al., 2018), while we also present the results on lexical simplification, another task that has been shown to benefit from semantic similarity specialization (Glavaš and Vulić, 2018), later in §5.

4.1 Experimental Setup

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,

[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.

[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 lexically informed counterpart.

1995) and Roget's Thesaurus (Kipfer, 2009) and 326,187 direct hyponym-hypernym pairs (Vulić and Mrkšić, 2018) from WordNet.[5]

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 provided in Table 1:

CoLA (Warstadt et al., 2019): 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, recognizing question paraphrases;

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] the input sequence length is 128. The learning rate for both models is $2 \cdot 10^{-5}$ with a warm-up over the first $1,000$ 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 sentence 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 development sets. Finally, we evaluate fine-tuned BERT and LIBERT on all 10 test sets.

4.2 Results and Discussion

Main Results. The main results are summarized in Table 2: we report both dev set and test set performance. After 1M MLM+NSP steps, LIBERT outperforms BERT on 8/9 tasks (dev) and 8/10 tasks (test). After 2M MLM+NSP steps, LIBERT is superior in 9/9 tasks (dev) and 9/10 tasks (test). For the test set of the tenth task (QNLI), LIBERT is on a par with BERT. While large gains are reported on CoLA, AX, and visible gains appear 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 lexical knowledge can be used to supplement unsupervised pretraining models with useful information which cannot be fully captured solely through large text data and their distributional signal. The results indicate that LIBERT, our lexically informed multi-task method, successfully blends such curated linguistic knowledge with distributional learning signals. It also further validates intuitions from relevant work on specializing 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 in general.

Fine-grained Analysis. To better understand how lexical information corroborates the model predic-

[5] Note again that similar to work of Vulić (2018), both WordNet synonyms and direct hyponym-hypernym pairs are treated exactly 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-documented issues.

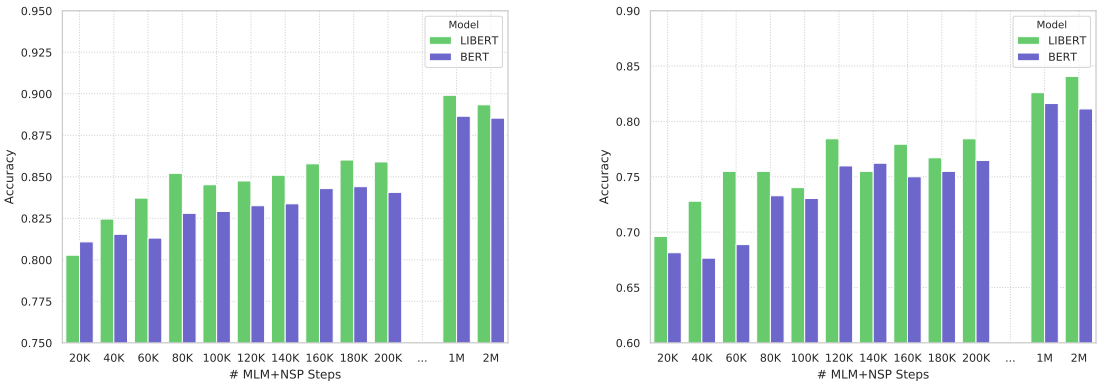
[7] Due to hardware restrictions, we train in smaller batches than in the the original work (Devlin et al., 2019) ($k = 256$). This means that for the same number of update steps, our models will have observed less training data than the original BERT model of Devlin et al. (2019).

	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: dataset sizes for tasks in the GLUE benchmark (Wang et al., 2018).

			CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	AX
			MCC	Acc	F1/Acc	Pears	F1/Acc	Acc	Acc	Acc	Acc	MCC
1M	Dev	BERT	29.4	88.7	87.1/81.6	86.4	85.9/89.5	78.2	**78.8**	86.2	63.9	–
		LIBERT	**35.3**	**89.9**	**87.9/82.6**	**87.2**	**86.3/89.8**	**78.5**	78.7	**86.5**	**65.3**	–
		Δ	+5.9	+1.2	+0.8/+1.0	+0.8	+0.4/+0.3	+0.3	-0.1	+0.3	+1.4	–
	Test	BERT	21.5	87.9	84.8/78.8	**80.8**	68.6/87.9	78.2	**77.6**	85.8	61.3	26.8
		LIBERT	**31.4**	**89.6**	**86.1/80.4**	80.5	**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	+0.4/+0.2	+0.2	-0.2	+0.4	+1.3	+6.0
2M	Dev	BERT	30.0	88.5	86.4/81.1	87.0	86.3/89.8	78.8	79.3	86.6	64.3	–
		LIBERT	**37.2**	**89.3**	**88.7/84.1**	**88.3**	**86.5/90.0**	**79.6**	**80.0**	**87.7**	**66.4**	–
		Δ	+7.2	+0.8	+2.3/+3.0	+1.3	+0.2/+0.2	+0.8	+0.7	+1.1	+2.1	–
	Test	BERT	28.8	89.7	84.9/79.1	81.1	69.0/88.0	78.6	78.1	**87.2**	63.4	30.8
		LIBERT	**35.3**	**90.8**	**86.6/81.7**	**82.6**	**69.3/88.2**	**79.8**	**78.8**	**87.2**	**63.6**	**33.3**
		Δ	+6.5	+1.1	+1.7/+2.6	+1.5	+0.3/+0.2	+1.2	+0.7	+0.0	+0.2	+2.5

Table 2: Results on 10 GLUE tasks after 1M and 2M 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.

tions, we perform a fine-grained analysis on the Diagnostic dataset (Wang et al., 2018), measuring the performance of LIBERT on specific subsets of sentences annotated for the linguistic phenomena they contain. We report the results in Table 3. As expected, *Lexical Semantics* is the category of phenomena that benefits the most (+43.7% for 1M iterations, +29.7% for 2M), but with significant gains also in phenomena related to *Logic* (+29.1% for 1M and +29.1% for 2M) and *Knowledge & Common Sense* (+51.7% for 1M). Interestingly, these results seem to suggest that knowledge about semantic similarity and lexical relations also partially encompasses factual knowledge about the world.

By inspecting even finer-grained phenomena related to *Lexical Semantics*, LIBERT outdistances its baseline by a large margin in: i) *Lexical Entailment* (+62.9% for 1M, +56.6% for 2M), as expected from the guidance of hypernym-hyponym pairs; ii) *Morphological Negation* (+75.8% for 1M, +40.4% for 2M). Crucially, the lower performance of BERT cannot be explained by the low frequency of morphologically derived words (prevented by the WordPiece tokenization), but exactly because of the distributional bias. iii) *Factivity* (+281.7% for 1M, +130.8% for 2M), which is a lexical entailment between a clause and the entire sentence it is embedded in. Since it depends on specific lexical triggers (usually verbs or adverbs), it is clear that lexico-semantic knowledge better characterizes the trigger meanings. The improvement margin for *Redundancy* and *Quantifiers* fluctuate across different amounts of iterations, hence no conclusions can be drawn from the current evidence.

	Model	All	Coarse-grained				Fine-grained					
			LS	PAS	Lo	KCS	LE	MN	Fa	Re	NE	Qu
	BERT	26.8	24.5	38.8	19.6	12.8	17.5	29.3	04.9	22.5	15.6	**57.2**
1M	LIBERT	**32.8**	**35.2**	**39.7**	**25.3**	**19.4**	**28.5**	**51.4**	**18.7**	**59.2**	**18.0**	56.9
	Δ	6.0	10.7	0.9	5.7	6.6	11.0	22.2	13.8	36.7	2.4	-0.3
	BERT	30.8	31.3	40.0	21.7	**19.7**	21.2	51.3	09.1	59.2	**21.0**	60.5
2M	LIBERT	**33.3**	**40.6**	39.9	**24.5**	18.3	**33.2**	**72.0**	**21.0**	59.2	18.3	**68.4**
	Δ	2.5	9.3	-0.1	2.8	-1.4	12.0	20.7	11.9	0.0	-2.7	7.9

Table 3: Linguistic analysis on the Diagnostic dataset. The scores are R_3 coefficients between gold and predicted labels, scaled by 100, for sentences containing linguistic phenomena of interest. We report all the coarse-grained categories: *Lexical Semantics* (**LS**), *Predicate-Argument Structure* (**PAS**), *Logic* (**Lo**), and *Knowledge and Common Sense* (**KCS**). Moreover, we report fine-grained categories for Lexical Semantics: *Lexical Entailment* (**LE**), *Morphological Negation* (**MN**), *Factivity* (**Fa**), *Redundancy* (**Re**), *Named Entities* (**NE**), and *Quantifiers* (**Qu**).

Performance over Time. Further, an analysis of performance over time (in terms of MLM+NSP 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 lexically-informed LIBERT persist at different time steps. This finding again indicates the complementarity of useful signals coded in large text data versus lexical resources (Faruqui, 2016; Mrkšić et al., 2017), which should be investigated more in future work.

5 Similarity-Oriented Downstream Evaluation: Lexical Simplification

Task Description. The goal of lexical simplification is to replace a target word w in a context sentence S with simpler alternatives of equivalent meaning. Generally, the task can be divided into two main parts: (1) generation of substitute candidates, and (2) candidate ranking, in which the simplest candidate is selected (Paetzold and Specia, 2017). Unsupervised approaches to candidate generation seem to be predominant lately (e.g., Glavaš and Štajner, 2015; Ponti et al., 2019). In this task, discerning between pure semantic similarity and broad topical relatedness (as well as from other lexical relations such as antonymy) is crucial. Consider the example: "*Einstein unlocked the door to the atomic age,*" where *unlocked* is the target word. In this context, the model should avoid confusion both with related words (e.g. *repaired*) and opposite words (e.g. *closed*) that fit in context but alter the original meaning.

Experimental Setup. In order to evaluate the simplification capabilities of LIBERT versus BERT, we adopt a standard BERT-based approach to lexical simplification (Qiang et al., 2019), dubbed BERT-LS. It exploits the BERT MLM pretraining task objective for candidate generation. Given the complex word w and a context sentence S, we mask w in a new sequence S'. Next, we concatenate S and S' as a sentence pair and create the BERT-style input by running WordPiece tokenization on the sentences, adding the `[CLS]` and `[SEP]` tokens before, in-between, and after the sequence, and setting segment IDs accordingly. We then feed the input either to BERT or LIBERT, and obtain the probability distribution over the vocabulary outputted by the MLM predictor based on the masked token $p(\cdot|S, S'\backslash\{w\})$. Based on this, we select the candidates as the top k words according to their probabilities, excluding morphological variations of the masked word.

For the substitution ranking component, we also follow Qiang et al. (2019). Given the set of candidate tokens C, we compute for each c_i in C a set of features: (1) BERT prediction probability, (2) loss of the likelihood of the whole sequence according to the MLM when choosing c_i instead of w, (3) semantic similarity between the fastText vectors (Bojanowski et al., 2017) of the original word w and the candidate c_i, and (4) word frequency of c_i in the top 12 million texts of Wikipedia and in the Children's Book Test corpus.[8] Based on the individual features, we next rank the candidates in C and consequently, obtain a set of ranks for each c_i. The best candidate is chosen according to its average rank across all features. In our experiments, we fix the number of candidates k to 6.

Evaluation Data. We run the evaluation on three standard datasets for lexical simplification:

(1) LexMTurk (Horn et al., 2014). The dataset

[8] A detailed description of these features can be found in the original work.

# Steps		Candidate Generation BenchLS P	BenchLS R	BenchLS F1	LexMTurk P	LexMTurk R	LexMTurk F1	NNSeval P	NNSeval R	NNSeval F1	Full Simplification Pipeline BenchLS A	LexMTurk A	NNSeval A
1M	BERT	.2167	.1765	.1945	.3043	.1420	.1937	.1499	.1200	.1333	.3854	.5260	.2469
	LIBERT	**.2348**	**.1912**	**.2108**	**.3253**	**.1518**	**.2072**	**.1646**	**.1318**	**.1464**	**.4338**	**.6080**	**.2678**
	Δ	.0181	.0147	.0163	.0210	.0098	.0135	.0147	.0118	.0131	.0484	.0820	.0209
2M	BERT	.2408	.1960	.2161	.3267	.1524	.2079	.1583	.1267	.1408	.4241	.5920	.2594
	LIBERT	**.2766**	**.2252**	**.2483**	**.3700**	**.1727**	**.2354**	**.1925**	**.1541**	**.1712**	**.4887**	**.6540**	**.2803**
	Δ	.0358	.0292	.0322	.0433	.0203	.0275	.0342	.0274	.0304	.0646	.0620	.0209

Table 4: Results on the lexical simplification candidate generation task and for the full pipeline on three datasets: BenchLS, LexMTurk, and NNSeval. For each dataset we report the performance after 1M and 2M MLM+NSP steps (# Steps) with BERT and LIBERT in terms of Precision (P), Recall (R) and F1-measure (F1) for candidate generation and accuracy (A) for the full pipeline.

consists of 500 English instances, which are collected from Wikipedia. The complex word and the simpler substitutions were annotated by 50 crowd workers on Amazon Mechanical Turk.

(2) BenchLS (Paetzold and Specia, 2016) is a merge of LexMTurk and LSeval (De Belder and Moens, 2010) containing 929 sentences. The latter dataset focuses on text simplification for children. The authors of BenchLS applied additional corrections over the instances of the two datasets.

(3) NNSeval (Paetzold and Specia, 2017) is an English dataset focused on text simplification for non-native speakers and consists in total of 239 instances. Similar to BenchLS, the dataset is based on LexMTurk, but filtered for a) instances that contain a complex target word for non-native speakers, and b) simplification candidates that were found to be non-complex by non-native speakers.

We report the scores on all three datasets in terms of Precision, Recall and F1 for the candidate generation sub-task, and in terms of the standard lexical simplification metric of *accurracy* (A) (Horn et al., 2014; Glavaš and Štajner, 2015) for the full simplification pipeline. This metric computes the number of correct simplifications (i.e., when the replacement made by the system is found in the list of gold standard replacements) divided by the total number of target complex words.

Results and Discussion. The results for BERT and LIBERT for the simplification candidate generation task and for the full pipeline evaluation are provided in Table 4. We report the performance of both models after 1M and 2M MLM+NSP pretraining steps. We observe that LIBERT consistently outperforms BERT by at least 0.9 percentage points across all evaluation setups, measures, and for all three evaluation sets. Same as in GLUE evaluation, the gains do not vanish as we train both models for a longer period of time (i.e., compare the differences between the two models after 1M vs. 2M training steps). On the contrary, for the candidate generation task, the gains of LIBERT over BERT are even higher after 2M steps. The gains achieved by LIBERT are also visible in the full simplification pipeline: e.g., on LexMTurk, replacing BERT with LIBERT yields a gain of 8.2 percentage points. In sum, these results confirm the importance of similarity specialization for a similarity-oriented downstream task such as lexical simplification.

6 Conclusion

We have presented LIBERT, a lexically informed extension of the state-of-the-art unsupervised pretraining model BERT. Our model is based on a multi-task framework that allows us to steer (i.e., specialize) the purely distributional BERT model to accentuate a lexico-semantic relation of true semantic similarity (as opposed to broader semantic relatedness). The framework combines standard BERT objectives with a third objective formulated as a relation classification task. The gains stemming from such explicit injection of lexical knowledge into pretraining were observed for 9 out of 10 language understanding tasks from the GLUE benchmark, as well as for 3 lexical simplification benchmarks. These results suggest that complementing distributional information with lexical knowledge is beneficial for unsupervised pretraining models.

In the future, we will work on more sophisticated specialization methods, and we will investigate methods to encode the knowledge on asymmetric relations such as meronymy and lexical entailment. Finally, we will port this new framework to other languages and to resource-poor scenarios. We will release the code at: [URL]

References

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

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the ACL*, 5:135–146.

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. Technical report, University of Waterloo.

Jan De Belder and Marie-Francine Moens. 2010. Text simplification for children. In *Proceedings of the SIGIR workshop on accessible search systems*, pages 19–26.

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 Sanja Štajner. 2015. Simplifying lexical simplification: Do we need simplified corpora? In *Proceedings of ACL-IJCNLP*, pages 63–68, Beijing, China.

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.

Colby Horn, Cathryn Manduca, and David Kauchak. 2014. Learning a lexical simplifier using Wikipedia. In *Proceedings of ACL*, pages 458–463.

Hwiyeol Jo and Stanley Jungkyu Choi. 2018. Extrofitting: Enriching word representation and its vector space with semantic lexicons. *CoRR*, abs/1804.07946.

Aishwarya Kamath, Jonas Pfeiffer, Edoardo Maria Ponti, Goran Glavaš, and Ivan Vulić. 2019. Specializing distributional vectors of all words for lexical entailment. In *Proceedings of the 4th Workshop on Representation Learning for NLP*, pages 72–83.

Douwe Kiela, Felix Hill, and Stephen Clark. 2015. Specializing word embeddings for similarity or relatedness. In *Proceedings of EMNLP*, pages 2044–2048.

Joo-Kyung Kim, Gokhan Tur, Asli Celikyilmaz, Bin Cao, and Ye-Yi Wang. 2016. Intent detection using semantically enriched word embeddings. In *Proceedings of SLT*.

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

Quan Liu, Hui Jiang, Si Wei, Zhen-Hua Ling, and Yu Hu. 2015. Learning semantic word embeddings based on ordinal knowledge constraints. In *Proceedings of ACL*, pages 1501–1511.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2019a. K-BERT: Enabling language representation with knowledge graph. *arXiv preprint arXiv:1909.07606*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NeurIPS*, pages 3111–3119.

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

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina Maria Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. In *Proceedings of NAACL-HLT*, pages 142–148.

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.

Dominique Osborne, Shashi Narayan, and Shay Cohen. 2016. Encoding prior knowledge with eigenword embeddings. *Transactions of the ACL*, 4:417–430.

Gustavo Paetzold and Lucia Specia. 2016. Benchmarking lexical simplification systems. In *Proceedings of LREC*, pages 3074–3080, Portorož, Slovenia.

Gustavo H Paetzold and Lucia Specia. 2017. A survey on lexical simplification. *Journal of Artificial Intelligence Research*, 60:549–593.

Matthew E. 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.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. 2019. Knowledge enhanced contextual word representations. In *Proceedings of the EMNLP-IJCNLP*, pages 43–54.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of EMNLP-IJCNLP*, pages 2463–2473.

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.

Edoardo Maria Ponti, Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Cross-lingual semantic specialization via lexical relation induction. In *Proceedings of the EMNLP-IJCNLP*, pages 2206–2217.

Jipeng Qiang, Yun Li, Zhu Yi, Yunhao Yuan, and Xindong Wu. 2019. A simple BERT-based approach for lexical simplification. *arXiv preprint arXiv:1907.06226*.

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.

Roy Schwartz, Roi Reichart, and Ari Rappoport. 2015. Symmetric pattern based word embeddings for improved word similarity prediction. In *Proceedings of CoNLL*, pages 258–267.

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.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2019. ERNIE 2.0: A continual pre-training framework for language understanding. *arXiv preprint arXiv:1907.12412*.

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. 2019. Neural network acceptability judgments. *Transactions of the ACL*, 7:625–641.

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*.

Mo Yu and Mark Dredze. 2014. Improving lexical embeddings with semantic knowledge. In *Proceedings of ACL*, pages 545–550.

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.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019a. ERNIE: Enhanced language representation with informative entities. In *Proceedings of ACL*, pages 1441–1451.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019b. ERNIE: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129*.