Which *BERT? A Survey Organizing Contextualized Encoders

Patrick Xia **Shijie Wu** **Benjamin Van Durme**
Johns Hopkins University
{paxia, vandurme}@cs.jhu.edu, shijie.wu@jhu.edu

Abstract

Pretrained contextualized text encoders are now a staple of the NLP community. We present a survey on language representation learning with the aim of consolidating a series of shared lessons learned across a variety of recent efforts. While significant advancements continue at a rapid pace, we find that enough has now been discovered, in different directions, that we can begin to organize advances according to common themes. Through this organization, we highlight important considerations when interpreting recent contributions and choosing which model to use.

1 Introduction

A couple years ago, Peters et al. (2018, ELMo) won the NAACL Best Paper Award for creating strong performing, task-agnostic sentence representations due to large scale unsupervised pretraining. Days later, its high level of performance was surpassed by Radford et al. (2018) which boasted representations beyond a single sentence and fine-tuning flexibility. This instability and competition between models has been a recurring theme for researchers and practitioners who have watched the rapidly narrowing gap between text representations and language understanding benchmarks. However, it has not discouraged research. Given the recent flurry of models, we often ask: "**What, besides state-of-the-art, does this newest paper contribute? Which encoder should *we* use?**"

The goals of this survey are to outline the areas of progress, relate contributions in text encoders to ideas from other fields, describe how each area is evaluated, and present considerations for practitioners and researchers when choosing an encoder. This survey does not intend to compare specific model metrics, as tables from other works provide comprehensive insight. For example, Table 16 in Raffel et al. (2019) compares the scores on a large suite of tasks of different model architectures, training objectives, and hyperparameters, and Table 1 in Rogers et al. (2020) details early efforts in model compression and distillation. We also recommend other closely related surveys on contextualized word representations (Smith, 2019; Rogers et al., 2020; Liu et al., 2020a), transfer learning in NLP (Ruder et al., 2019), and integrating encoders into NLP applications (Wolf et al., 2019). Complementing these existing bodies of work, we look at the ideas and progress in the scientific discourse for text representations from the perspective of discerning their differences.

We organize this paper as follows. §2 provides brief background on encoding, training, and evaluating text representations. §3 identifies and analyzes two classes of pretraining objectives. In §4, we explore faster and smaller models and architectures in both training and inference. §5 notes the impact of both quality and quantity of pretraining data. §6 briefly discusses efforts on probing encoders and representations with respect to linguistic knowledge. §7 describes the efforts into training and evaluating multilingual representations. Within each area, we conclude with high-level observations and discuss the evaluations that are used and their shortcomings.

We conclude in §8 by making recommendations to researchers: publicizing negative results in this area is especially important owing to the sheer cost of experimentation and to ensure evaluation reproducibility. In addition, probing studies need to focus not only on the models and tasks, but also on the pretraining data. We pose questions for users of contextualized encoders, like whether the compute requirement of a model is worth the benefits. We hope our survey serves as a guide for both NLP researchers and practitioners, orienting them to the current state of the field of contextualized encoders and differences between models.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 7516–7533,
November 16–20, 2020.

2 Background

Encoders Pretrained *text encoders* take as input a sequence of tokenized[1] text, which is encoded by a multi-layered neural model. The representation of each (sub)token, x_t, is either the set of hidden weights, $\{h_t^{(l)}\}$ for each layer l, or its weight on just the top layer, $h_t^{(-1)}$. Unlike fixed-sized word, sentence, or paragraph representations, the produced *contextualized representations* of the text depends on the length of the input text. Most encoders use the Transformer architecture (Vaswani et al., 2017).

Transfer: The Pretrain-Finetune Framework While text representations can be learned in any manner, ultimately, they are evaluated using specific *target tasks*. Historically, the learned representations (e.g. word vectors) were used as initialization for task-specific models. Dai and Le (2015) are credited with using *pretrained* language model outputs as initialization, McCann et al. (2017) use pretrained outputs from translation as frozen word embeddings, and Howard and Ruder (2018) and Radford et al. (2018) demonstrate the effectiveness of *finetuning* to different target tasks by updating the full (pretrained) model for each task. We refer to the embeddings produced by the pretrained models (or encoders) as contextualized text representations. As our goal is to discuss the encoders and their representations, we do not cover the innovations in finetuning (Liu et al., 2015; Ruder et al., 2019; Phang et al., 2018; Liu et al., 2019c; Zhu et al., 2020, *inter alia*).

Evaluation Widely adopted evaluations of text representations relate them to downstream natural language understanding (NLU) benchmarks. This full-stack process necessarily conflates representation power with finetuning strategies. Common language understanding benchmarks include (1) a diverse suite of sentence-level tasks covering paraphrasing, natural language inference, sentiment, and linguistic acceptability (GLUE) and its more challenging counterpart with additional commonsense and linguistic reasoning tasks (SuperGLUE) (Wang et al., 2019c,b; Clark et al., 2019a; De Marneffe et al., 2019; Roemmele et al., 2011; Khashabi et al., 2018; Zhang et al., 2018; Dagan et al., 2006; Bar Haim et al., 2006; Giampiccolo et al., 2007; Bentivogli et al., 2009; Pilehvar and Camacho-Collados, 2019; Rudinger et al., 2018; Poliak et al., 2018; Levesque et al., 2011); (2) crowdsourced questions derived from Wikipedia articles (Rajpurkar et al., 2016, 2018, SQuAD); and (3) multiple-choice reading comprehension (Lai et al., 2017, RACE).

[1]Unlike traditional word-level tokenization, most works decompose text into *subtokens* from a fixed vocabulary using some variation of byte pair encoding (Gage, 1994; Schuster and Nakajima, 2012; Sennrich et al., 2016)

3 Area I: Pretraining Tasks

To utilize data at scale, pretraining tasks are typically self-supervised. We categorize the contributions into two types: *token prediction* (over a large vocabulary space) and *nontoken prediction* (over a handful of labels). In this section, we discuss several empirical observations. While token prediction is clearly important, less clear is which variation of the token prediction task is the best (or whether it even matters). Nontoken prediction tasks appear to offer orthogonal contributions that marginally improve the language representations. We emphasize that in this section, we seek to outline the primary efforts in pretraining objectives and not to provide a comparison on a set of benchmarks.[2]

3.1 Token Prediction

Predicting (or generating) the next word has historically been equivalent to the task of language modeling. Large language models perform impressively on a variety of language understanding tasks while maintaining their generative capabilities (Radford et al., 2018, 2019; Keskar et al., 2019; Brown et al., 2020), often outperforming contemporaneous models that use additional training objectives.

ELMo (Peters et al., 2018) is a BiLSTM model with a language modeling objective for the next (or previous) token given the forward (or backward) history. This idea of looking at the full context was further refined as a cloze[3] task (Baevski et al., 2019), or as a denoising Masked Language Modeling (MLM) objective (Devlin et al., 2019, BERT). MLM replaces some tokens with a `[mask]` symbol and provides both right and left contexts (bidirectional context) for predicting the masked tokens. The bidirectionality is key to outperforming a unidirectional language model on a large suite of natural language understanding benchmarks (Devlin et al., 2019; Raffel et al., 2019).

The MLM objective is far from perfect, as the use of `[mask]` introduces a pretrain/finetune vo-

[2]See Raffel et al. (2019) for comprehensive experiments.

[3]A cloze task is a fill-in-the-blank task.

cabulary discrepancy. Devlin et al. (2019) look to mitigate this issue by occasionally replacing `[mask]` with the original token or sampling from the vocabulary. Yang et al. (2019) convert the discriminative objective into an autoregressive one, which allows the `[mask]` token to be discarded entirely. Naively, this would result in unidirectional context. By sampling permutations of the factorization order of the joint probability of the sequence, they preserve bidirectional context. Similar ideas for permutation language modeling (PLM) have also been studied for sequence generation (Stern et al., 2019; Chan et al., 2019; Gu et al., 2019). The MLM and PLM objectives have since been unified architecturally (Song et al., 2020; Bao et al., 2020) and mathematically (Kong et al., 2020).

ELECTRA (Clark et al., 2020) replaces `[mask]` through the use of a small generator (trained with MLM) to sample a real token from the vocabulary. The main encoder, a discriminator, then determines whether each token was replaced.

A natural extension would mask units that are more linguistically meaningful, such as rarer words,[4] whole words, or named entities (Devlin et al., 2019; Sun et al., 2019b). This idea can be simplified to *random* spans of texts (Yang et al., 2019; Song et al., 2019). Specifically, Joshi et al. (2020) add a reconstruction objective which predicts the masked tokens using only the span boundaries. They find that masking random spans is more effective than masking linguistic units.

An alternative architecture uses an encoder-decoder framework (or denoising autoencoder) where the input is a corrupted (masked) sequence the output is the full original sequence (Wang et al., 2019d; Lewis et al., 2020; Raffel et al., 2019).

3.2 Nontoken Prediction

Bender and Koller (2020) argue that for the goal of natural language understanding, we cannot rely purely on a language modeling objective; there must be some grounding or external information that relates the text to each other or to the world. One solution is to introduce a secondary objective to directly learn these biases.

Self-supervised discourse structure objectives, such as text order, has garnered significant attention. To capture relationships between two *sentences*,[5] Devlin et al. (2019) introduce the next sentence prediction (NSP) objective. In this task, either sentence `B` follows sentence `A` or `B` is a random negative sample. Subsequent works showed that this was not effective, suggesting the model simply learned topic (Yang et al., 2019; Liu et al., 2019d). Jernite et al. (2017) propose a sentence order task of predicting whether `A` is before, after, or unrelated to `B`, and Wang et al. (2020b) and Lan et al. (2020) use it for pretraining encoders. They report that (1) understanding text order does contribute to improved language understanding; and (2) harder-to-learn pretraining objectives are more powerful, as both modified tasks have lower intrinsic performance than NSP. It is still unclear, however, if this is the best way to incorporate discourse structure, especially since these works do not use real sentences.

Additional work has focused on effectively incorporating multiple pretraining objectives. Sun et al. (2020a) use multi-task learning with *continual pretraining* (Hashimoto et al., 2017), which incrementally introduces newer tasks into the set of pretraining tasks from word to sentence to document level tasks. Encoders using visual features (and evaluated only on visual tasks) jointly optimize multiple different masking objectives over both token sequences and regions of interests in the image (Tan and Bansal, 2019).[6]

Prior to token prediction, discourse information has been used in training sentence representations. Conneau et al. (2017, 2018a) use natural language inference sentence pairs, Jernite et al. (2017) use discourse-based objectives of sentence order, conjunction classifier, and next sentence selection, and Nie et al. (2019) use discourse markers. While there is weak evidence suggesting that these types of objectives are less effective than language modeling (Wang et al., 2019a), we lack fair studies comparing the relative influence between the two categories of objectives.

3.3 Comments on Evaluation

We reviewed the progress on pretraining tasks, finding that token prediction is powerful but can be improved further by other objectives. Currently, successful techniques like span masking or arbitrarily sized "sentences" are linguistically unmotivated. We anticipate future work to further incorporate

[4] Clark et al. (2020) report negative results for rarer words.

[5] *Sentence* unfortunately refers to a text segment containing no more than a fixed number of subtokens. It may contain any (fractional) number of real sentences.

[6] Table 5 in Su et al. (2020) provides a recent summary of efforts in visual-linguistic representations.

more meaningful linguistic biases in pretraining.

Our observations are informed by evaluations that are compared across different works. **These benchmarks on downstream tasks do not account for ensembling or finetuning and can only serve as an approximation for the differences between the models.** For example, Jiang et al. (2020) develop a finetuning method over a supposedly weaker model which leads to gains in GLUE score over reportedly stronger models. Furthermore, these evaluations aggregate vastly different tasks. Those interested in the best performance should first carefully investigate metrics on their specific task. Even if models are finetuned on an older encoder,[7] it may be more cost-efficient and enable fairer future comparisons to reuse those over restarting the finetuning or reintegrating new encoders into existing models when doing so does not necessarily guarantee improved performance.

4 Area II: Efficiency

As models perform better but cost more to train, some have called for research into efficient models to improve deployability, accessibility, and reproducibility (Amodei and Hernandez, 2018; Strubell et al., 2019; Schwartz et al., 2019). Encoders tend to scale effectively (Lan et al., 2020; Raffel et al., 2019; Brown et al., 2020), so efficient models will also result in improvements over inefficient ones of the same size. In this section, we give an overview of several efforts aimed to decrease the computation budget (time and memory usage) during training and inference of text encoders. While these two axes are correlated, reductions in one axis do not always lead to reductions in the other.

4.1 Training

One area of research decreases wall-clock training time through more compute and larger batches. You et al. (2020) reduce the time of training BERT by introducing the LAMB optimizer, a large batch stochastic optimization method adjusted for attention models. Rajbhandari et al. (2020) analyze memory usage in the optimizer to enable parallelization of models resulting in higher throughput in training. By reducing the training time, models can be practically trained for longer, which has also been shown to lead to benefits in task performance (Liu et al., 2019d; Lan et al., 2020, *inter alia*).

[7] This the case with retrieval-based QA (Guu et al., 2020; Herzig et al., 2020), which builds on BERT.

Another line of research reduces the compute through attention sparsification (discussed in §4.2) or increasing the convergence rate (Clark et al., 2020). These works report hardware and estimate the reduction in floating point operations (FPOs).[8] These kinds of speedup are orthogonal to hardware parallelization and are most encouraging as they pave the path for future work in *efficient* training.

Note that these approaches do not necessarily affect the latency to process a single example nor the compute required during inference, which is a function of the size of the computation graph.

4.2 Inference

Reducing model size without impacting performance is motivated by lower inference latency, hardware memory constraints, and the promise that naively scaling up dimensions of the model will improve performance. Size reduction techniques produce smaller and faster models, while occasionally improving performance. Rogers et al. (2020) survey BERT-like models and present in Table 1 the differences in sizes and performance across several models focused on inference efficiency.

Architectural changes have been explored as one avenue for reducing either the model size or inference time. In Transformers, the self-attention pattern scales quadratically in sequence length. To reduce the asymptotic complexity, the self-attention can be sparsified: each token only attending to a small "local" set (Vaswani et al., 2017; Child et al., 2019; Sukhbaatar et al., 2019). This has further been applied to pretraining on longer sequences, resulting in sparse contextualized encoders (Qiu et al., 2019; Ye et al., 2019; Kitaev et al., 2020; Beltagy et al., 2020, *inter alia*). Efficient Transformers is an emerging subfield with applications beyond NLP; Tay et al. (2020) survey 17 Transformers that have implications on efficiency.

Another class of approaches carefully selects weights to reduce model size. Lan et al. (2020) use low-rank factorization to reduce the size of the embedding matrices, while Wang et al. (2019f) factorize other weight matrices. Additionally, parameters can be shared between layers (Dehghani et al., 2019; Lan et al., 2020) or between an encoder and decoder (Raffel et al., 2019). However, models that employ these methods do not always have *smaller* computation graphs. This greatly reduces the usefulness of parameter sharing compared to other

[8] We borrow this terminology from Schwartz et al. (2019).

methods that additionally offer greater speedups relative to the reduction in model size.

Closely related, model pruning (Denil et al., 2013; Han et al., 2015; Frankle and Carbin, 2018) during training or inference has exploited the overparameterization of neural networks by removing up to 90%-95% parameters. This approach has been successful in not only reducing the number of parameters, but also improving performance on downstream tasks. Related to efforts for pruning deep networks in computer vision (Huang et al., 2016), layer selection and dropout during both training and inference have been studied in both LSTM (Liu et al., 2018a) and Transformer (Fan et al., 2020) based encoders. These also have a regularization effect resulting in more stable training and improved performance. There are additional novel pruning methods that can be performed during training (Guo et al., 2019; Qiu et al., 2019). These successful results are corroborated by other efforts (Gordon et al., 2020) showing that low levels of pruning do not substantially affect pretrained representations. Additional successful efforts in model pruning directly target a downstream task (Sun et al., 2019a; Michel et al., 2019; McCarley, 2019; Cao et al., 2020a). Note that pruning does not always lead to speedups in practice as sparse operations may be hard to parallelize.

Knowledge distillation (KD) uses an overparameterized teacher model to rapidly train a smaller student model with minimal loss in performance (Hinton et al., 2015) and has been used for translation (Kim and Rush, 2016), computer vision (Howard et al., 2017), and adversarial examples (Carlini and Wagner, 2016). This has been applied to ELMo (Li et al., 2019) and BERT (Tang et al., 2019; Sanh et al., 2019; Sun et al., 2020b, *inter alia*). KD can also be combined with adaptive inference, which dynamically adjusts model size (Liu et al., 2020b), or performed on submodules which are later substituted back into the full model (Xu et al., 2020).

Quantization with custom low-precision hardware is also a promising method for both reducing the size of models and compute time, albeit it does not reduce the number of parameters or FPOs (Shen et al., 2020; Zafrir et al., 2019). This line of work is mostly orthogonal to other efforts specific to NLP.

4.3 Standardizing Comparison

There has yet to be a comprehensive and fair evaluation across all models. The closest, Table 1 in Rogers et al. (2020), compares 12 works in model compression. **However, almost no two papers are evaluated against the same BERT with the same set of tasks.** Many papers on attention sparsification do not evaluate on NLU benchmarks. We claim this is because finetuning is itself an expensive task, so it is not prioritized by authors: works on improving model efficiency have focused only on comparing to a BERT on a few tasks.

While it is easy for future research on pretraining to report model sizes and runtimes, it is harder for researchers in efficiency to report NLU benchmarks. We suggest extending versions of the leaderboards under different resource constraints so that researchers with access to less hardware could still contribute under the resource-constrained conditions. Some work has begun in this direction: the SustaiNLP 2020 Shared Task is focused on the energy footprint of inference for GLUE.[9]

5 Area III: (Pretraining) Data

Unsurprisingly for our field, increasing the size of training data for an encoder contributes to increases in language understanding capabilities (Yang et al., 2019; Raffel et al., 2019; Kaplan et al., 2020). At current data scales, some models converge before consuming the entire corpus. In this section, we identify a weakness when given *less* data, advocate for better data cleaning, and raise technical and ethical issues with using web-scraped data.

5.1 Data Quantity

There has not yet been observed a ceiling to the amount of data that can still be effectively used in training (Baevski et al., 2019; Liu et al., 2019d; Yang et al., 2019; Brown et al., 2020). Raffel et al. (2019) curate a 745GB subset of Common Crawl (CC),[10] which starkly contrasts with the 13GB used in BERT. For multilingual text encoding, Wenzek et al. (2020) curate 2.5TB of language-tagged CC. As CC continues to grow, there will be even larger datasets (Brown et al., 2020).

Sun et al. (2017) explore a similar question for computer vision, as years of progress iterated over 1M labeled images. By using 300M images, they improved performance on several tasks with a basic model. We echo their remarks that we should be cognizant of data sizes when drawing conclusions.

[9] https://sites.google.com/view/sustainlp2020/shared-task

[10] https://commoncrawl.org/ scrapes publicly accessible webpages each month.

Is there a floor to the amount of data needed to achieve current levels of success on language understanding benchmarks? As we decrease the data size, LSTM-based models start to dominate in perplexity (Yang et al., 2019; Melis et al., 2020), suggesting there are challenges with either scaling up LSTMs or scaling down Transformers. While probing contextualized models and representations is an important area of study (see §6), prior work focuses on pretrained models or models further pretrained on domain-specific data (Gururangan et al., 2020). We are not aware of any work which probes identical models trained with decreasingly less data. How much (and which) data is necessary for high performance on probing tasks?[11]

5.2 Data Quality

While text encoders should be trained on language, large-scale datasets may contain web-scraped and uncurated content (like code). Raffel et al. (2019) ablate different types of data for text representations and find that naively increasing dataset size does not always improve performance, partially due to data quality. This realization is not new. Parallel data and alignment in machine translation (Moore and Lewis, 2010; Duh et al., 2013; Xu and Koehn, 2017; Koehn et al., 2018, *inter alia*) and speech (Peddinti et al., 2016) often use language models to filter out misaligned or poor data. Sun et al. (2017) use automatic data filtering in vision. These successes on other tasks suggest that improved automated methods of data cleaning would let future models consume more *high-quality* data.

In addition to high quality, data uniqueness appears to be advantageous. Raffel et al. (2019) show that increasing the repetitions (number of epochs) of the pretraining corpus hurts performance. This is corroborated by Liu et al. (2019d), who find that random, unique masks for MLM improve over repeated masks across epochs. These findings together suggest a preference to seeing more *new* text. We suspect that representations of text spans appearing multiple times across the corpus are better shaped by observing them in unique contexts.

Raffel et al. (2019) find that differences in domain mismatch in pretraining data (web crawled vs. news or encyclopedic) result in strikingly different performance on certain challenge sets, and Gururangan et al. (2020) find that continuing pretraining on both domain and task specific data lead to gains in performance.

5.3 Datasets and Evaluations

With these larger and cleaner datasets, future research can better explore tradeoffs between size and quality, as well as strategies for scheduling data during training.

As we continue to scrape data off the web and publish challenge sets relying on other web data, we need to cautiously construct our training and evaluation sets. For example, the domains of many benchmarks (Wang et al. (2019c, GLUE), Rajpurkar et al. (2016, 2018, SQuAD), Wang et al. (2019b, SuperGLUE), Paperno et al. (2016, LAMBADA), Nallapati et al. (2016, CNN/DM)) now overlap with the data used to train language representations. Section 4 in Brown et al. (2020) more thoroughly discuss the effects of overlapping test data with pretraining data. Gehman et al. (2020) highlight the prevalance of toxic language in the common pretraining corpora and stress the important of pretraining data selection, especially for deployed models. We are not aware of a comprehensive study that explores the effect of leaving out targeted subsets of the pretraining data. We hope future models note the domains of pretraining and evaluation benchmarks, and for future language understanding benchmarks to focus on more diverse *genres* in addition to diverse *tasks*.

As we improve models by training on increasing sizes of crawled data, these models are also being picked up by NLP practitioners who deploy them in real-world software. These models learn biases found in their pretraining data (Gonen and Goldberg, 2019; May et al., 2019, *inter alia*). **It is critical to clearly state the source[12] of the pretraining data and clarify appropriate uses of the released models.** For example, crawled data can contain incorrect facts about living people; while webpages can be edited or retracted, publicly released "language" model are frozen, which can raise privacy concerns (Feyisetan et al., 2020).

6 Area IV: Interpretability

While it is clear that the performance of text encoders surpass human baselines, it is less clear what knowledge is stored in these models; how do they make decisions? In their survey, Rogers et al. (2020) find answers to the first question and also

[11] Conneau et al. (2020a) claim we need a few hundred MiB of text data for BERT.

[12] How was the data generated, curated, and processed?

raise the second. Inspired by prior work (Lipton, 2018; Belinkov and Glass, 2019; Alishahi et al., 2019), we organize here the major probing *methods* that are applicable to all encoders in hopes that future work will use comparable techniques.

6.1 Probing with Tasks

One technique uses the learned model as initialization for a model trained on a *probing task* consisting of a set of targeted natural language examples. The probing task's format is flexible as additional, (simple) diagnostic classifiers are trained on top of a typically frozen model (Ettinger et al., 2016; Hupkes et al., 2018; Poliak et al., 2018; Tenney et al., 2019b). Task probing can also be applied to the embeddings at various layers to explore the knowledge captured at each layer (Tenney et al., 2019a; Lin et al., 2019; Liu et al., 2019a). Hewitt and Liang (2019) warn that expressive (nonlinear) diagnostic classifiers can learn more arbitrary information than constrained (linear) ones. This revelation, combined with the differences in probing task format and the need to train, leads us to be cautious in drawing conclusions from these methods.

6.2 Model Inspection

Model inspection directly opens the metaphorical black box and studies the model weights without additional training. For examples, the embeddings themselves can be analyzed as points in a vector space (Ethayarajh, 2019). Through visualization, attention heads have been matched to linguistic functions (Vig, 2019; Clark et al., 2019b). These works suggest inspection is a viable path to debugging specific examples. In the future, methods for analyzing and manipulating attention in machine translation (Lee et al., 2017; Liu et al., 2018b; Bau et al., 2019; Voita et al., 2019) can also be applied to text encoders.

Recently, interpreting attention as explanation has been questioned (Serrano and Smith, 2019; Jain and Wallace, 2019; Wiegreffe and Pinter, 2019; Clark et al., 2019b). The ongoing discussion suggests that this method may still be insufficient for uncovering the rationale for predictions, which is critical for real-world applications.

6.3 Input Manipulation[13]

Input manipulation draws conclusions by recasting the probing task format into the form of the pre-training task and observing the model's predictions. As discussed in §3, word prediction (cloze task) is a popular objective. This method has been used to investigate syntactic and semantic knowledge (Goldberg, 2019; Ettinger, 2020; Kassner and Schütze, 2019). For a specific probing task, Warstadt et al. (2019) show that cloze and diagnostic classifiers draw similar conclusions. As input manipulation is not affected by variables introduced by probing tasks and is as interpretable than inspection, we suggest more focus on this method: either by creating new datasets (Warstadt et al., 2020) or recasting existing ones (Brown et al., 2020) into this format. A disadvantage of this method (especially for smaller models) is the dependence on both the pattern used to elicit an answer from the model and, in the few-shot case where a couple examples are provided first, highly dependent on the examples (Schick and Schütze, 2020).

6.4 Future Directions in Model Analysis

Most probing efforts have relied on diagnostic classifiers, yet these results are being questioned. Inspection of model weights has discovered what the models learn, but cannot explain their causal structure. We suggest researchers shift to the paradigm of input manipulation. By creating cloze tasks that assess linguistic knowledge, we can both observe decisions made by the model, which would imply (lack of) knowledge of a phenomenon. Furthermore, it will also enable us to directly interact with these models (by changing the input) without additional training, which currently introduces additional sources of uncertainty.

Bender and Koller (2020) also recommend a top-down view for model analysis that focuses on the end-goals for our field over hill-climbing individual datasets. While language models continue to outperform each other on these tasks, they argue these models do not learn *meaning*.[14] If not meaning, what are these models learning?

We are overinvesting in BERT. While it is fruitful to understand the boundaries of its knowledge, we should look more across (simpler) models to see *how* and *why* specific knowledge is picked up as our models both become increasingly complex and perform better on a wide set of tasks. For example, how many parameters does a Transformer-based model need to outperform ELMo or even rule-based baselines?

[13] This is analogous to the "few-shot" and "zero-shot" analysis in Brown et al. (2020).

[14] A definition is given in §3 of Bender and Koller (2020).

7 Area V: Multilinguality

The majority of research on text encoders has been in English.[15] Cross-lingual shared representations have been proposed as an efficient way to target multiple languages by using multilingual text for pretraining (Mulcaire et al., 2019; Devlin et al., 2019; Lample and Conneau, 2019; Liu et al., 2020c, *inter alia*). For evaluation, researchers have devised multilingual benchmarks mirroring those for NLU in English (Conneau et al., 2018b; Liang et al., 2020; Hu et al., 2020). Surprisingly, without any explicit cross-lingual signal, these models achieve strong zero-shot cross-lingual performance, outperforming prior cross-lingual word embedding-based methods (Wu and Dredze, 2019; Pires et al., 2019).

A natural follow-up question to ask is why these models learn cross-lingual representations. Some answers include the shared subword vocabulary (Pires et al., 2019; Wu and Dredze, 2019), shared Transformer layers (Conneau et al., 2020b; Artetxe et al., 2020) across languages, and depth of the network (K et al., 2020). Studies have also found the geometry of representations of different languages in the multilingual encoders can be aligned with linear transformations (Schuster et al., 2019; Wang et al., 2019e, 2020c; Liu et al., 2019b), which has also been observed in independent monolingual encoders (Conneau et al., 2020b). These alignments can be further improved (Cao et al., 2020b).

7.1 Evaluating Multilinguality

All of the areas discussed in this paper are applicable to multilingual encoders. However, progress in training, architecture, datasets, and evaluations are occurring concurrently, making it difficult to draw conclusions. We need more comparisons between competitive multilingual and monolingual systems or datasets. To this end, Wu and Dredze (2020) find that monolingual BERTs in low-resource languages are outperformed by multilingual BERT. Additionally, as zero-shot (or few-shot) cross-lingual transfer has inherently high variance (Keung et al., 2020), **the variance of models should also be reported**.

We anticipate cross-lingual performance being a new dimension to consider when evaluating text representations. For example, it will be exciting to discover how a small, highly-performant monolingual encoder contrasts against a multilingual variant; e.g., what is the minimum number of parameters needed to support a new language? Or, how does model size relate to the phylogenetic diversity of languages supported?

8 Discussion

8.1 Limitations and Recommendations

This survey, like others, is limited to only what has been shared publicly so far. The papers of many models described here highlight their best parts, where potential flaws are perhaps obscured within tables of numbers. Leaderboard submissions that do not achieve first place may never be published. Meanwhile, encoders are expensive to work with, yet they are a ubiquitous component in most modern NLP models. We strongly encourage more **publication and *publicizing* of negative results** and limitations. In addition to their scientific benefits,[16] publishing negative results in contextualized encoders can avoid significant externalities of rediscovering what doesn't work: time, money, and electricity. Furthermore, we ask leaderboard owners to **periodically publish surveys** of their received submissions.

The flourishing research in improving encoders is rivaled by research in interpreting them, mainly focused on discovering the boundary of what knowledge is captured by the models. For investigations that aim to sharpen the boundary, it is logical to build off of these prior results. However, we raise a concern that these encoders are all trained on similar data and have similar sizes. Future work in **probing should also look across different sizes and domains of training data**, as well as study the effect of model size. This can be further facilitated by model creators who release (data) ablated versions of their models.

We also raise a concern about reproducibility and accessibility of evaluation. Already, several papers focused on model compression do not report full GLUE results, possibly due to the expensive finetuning process for each of the nine datasets. Finetuning currently requires additional compute and infrastructure,[17] and the specific methods used impact task performance. As long as finetuning is still an essential component of evaluating encoders, de-

[15] Of the monolingual encoders in other languages, core research in modeling has only been performed so far for a few non-English languages (Sun et al., 2019b, 2020a).

[16] An EMNLP 2020 workshop is motivated by better science (https://insights-workshop.github.io/).

[17] Pruksachatkun et al. (2020) is a library that reduces some infrastructural overhead of finetuning.

vising **cheap, accessible, and reproducible metrics for encoders is an open problem**.

Ribeiro et al. (2020) suggest a practical solution to both probing model errors and reproducible evaluations by creating tools that quickly generate test cases for linguistic capabilities and find bugs in models. This task-agnostic methodology may be extensible to both challenging tasks and probing specific linguistic phenomenon.

8.2 Which *BERT should we use?

Here, we discuss tradeoffs between metrics and synthesize the previous sections. We provide a series of questions to consider when working with encoders for research or application development.

Task performance vs. efficiency An increasingly popular line of recent work has investigated knowledge distillation, model compression, and sparsification of encoders (§4.2). These efforts have led to significantly smaller encoders that boast competitive performance, and under certain settings, non-contextual embeddings alone may be sufficient (Arora et al., 2020; Wang et al., 2020a). For downstream applications, ask: **Is the extra iota of performance worth the significant costs of compute?**

Leaderboards vs. real data As a community, we are hill-climbing on curated benchmarks that aggregate dozens of tasks. Performance on these benchmarks does not necessarily reflect that of specific real-world tasks, like understanding social media posts about a pandemic (Müller et al., 2020). Before picking the best encoder determined by average scores, ask: **Is this encoder the best for our specific task? Should we instead curate a large dataset and pretrain again?** Gururangan et al. (2020) suggest continued pretraining on in-domain data as a viable alternative to pretraining from scratch.

For real-world systems, practitioners should be especially conscious of the datasets on which these encoders are pretrained. **There is a tradeoff between task performance and possible harms contained within the pretraining data.**

Monolingual vs. Multilingual For some higher resource languages, there exist monolingual pretrained encoders. For tasks in those languages, those encoders are a good starting point. However, as we discussed in §7, multilingual encoders can, surprisingly, perform competitively, yet these metrics are averaged over multiple languages and tasks. Again, we encourage looking at the relative **performance for a specific task and language**, and whether **monolingual encoders (or embeddings) may be more suitable.**

Ease-of-use vs. novelty With a constant stream of new papers and models (without peer review) for innovating in each direction, we suggest using and building off **encoders that are well-documented with reproduced or reproducible results**. Given the pace of the field and large selection of models, unless aiming to reproduce prior work or improve underlying encoder technology, we recommend proceeding with caution when reimplementing ideas from scratch.

9 Conclusions

In this survey we categorize research in contextualized encoders and discuss some issues regarding its conclusions. We cover background on contextualized encoders, pretraining objectives, efficiency, data, approaches in model interpretability, and research in multilingual systems. As there is now a large selection of models to choose from, we discuss tradeoffs that emerge between models. We hope this work provides some assistance to both those entering the NLP community and those already using contextualized encoders in looking beyond SOTA (and Twitter) to make more educated choices.

Acknowledgments

We especially thank the (meta-)reviewers for their insightful feedback and criticisms. In addition, we thank Sabrina Mielke, Nathaniel Weir, Huda Khayrallah, Mitchell Gordon, and Shuoyang Ding for discussing several drafts of this work. This work was supported in part by DARPA AIDA (FA8750-18-2-0015) and IARPA BETTER (#2019-19051600005). The views and conclusions contained in this work are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, or endorsements of DARPA, ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Afra Alishahi, Grzegorz Chrupała, and Tal Linzen. 2019. Analyzing and interpreting neural networks for nlp: A report on the first blackboxnlp workshop. *Natural Language Engineering*, 25:543–557.

Dario Amodei and Danny Hernandez. 2018. AI and compute. https://openai.com/blog/ai-and-compute.

Simran Arora, Avner May, Jian Zhang, and Christopher Ré. 2020. Contextual embeddings: When are they worth it? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2650–2663, Online. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Alexei Baevski, Sergey Edunov, Yinhan Liu, Luke Zettlemoyer, and Michael Auli. 2019. Cloze-driven pretraining of self-attention networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5359–5368, Hong Kong, China. Association for Computational Linguistics.

Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Songhao Piao, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2020. Unilmv2: Pseudo-masked language models for unified language model pre-training.

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

Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2019. Identifying and controlling important neurons in neural machine translation. In *International Conference on Learning Representations*.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

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

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Qingqing Cao, Harsh Trivedi, Aruna Balasubramanian, and Niranjan Balasubramanian. 2020a. DeFormer: Decomposing pre-trained transformers for faster question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4497, Online. Association for Computational Linguistics.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020b. Multilingual alignment of contextual word representations. In *International Conference on Learning Representations*.

Nicholas Carlini and David A. Wagner. 2016. Towards evaluating the robustness of neural networks. *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57.

William Chan, Nikita Kitaev, Kelvin Guu, Mitchell Stern, and Jakob Uszkoreit. 2019. KERMIT: generative insertion-based modeling for sequences. *CoRR*, abs/1906.01604.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *URL https://openai.com/blog/sparse-transformers*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019a. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019b. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018a. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018b. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

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.

Andrew M. Dai and Quoc V. Le. 2015. Semi-supervised sequence learning. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3079–3087.

Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The commitmentbank: Investigating projection in naturally occurring discourse. *Proceedings of Sinn und Bedeutung*, 23(2):107–124.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. 2019. Universal transformers. In *International Conference on Learning Representations*.

Misha Denil, Babak Shakibi, Laurent Dinh, Marc'Aurelio Ranzato, and Nando de Freitas. 2013. Predicting parameters in deep learning. In *NIPS*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kevin Duh, Graham Neubig, Katsuhito Sudoh, and Hajime Tsukada. 2013. Adaptation data selection using neural language models: Experiments in machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 678–683, Sofia, Bulgaria. Association for Computational Linguistics.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8(0):34–48.

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. 2016. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 134–139, Berlin, Germany. Association for Computational Linguistics.

Angela Fan, Edouard Grave, and Armand Joulin. 2020. Reducing transformer depth on demand with structured dropout. In *International Conference on Learning Representations*.

Oluwaseyi Feyisetan, Sepideh Ghanavati, and Patricia Thaine. 2020. Workshop on privacy in nlp (privatenlp 2020). In *Proceedings of the 13th International Conference on Web Search and Data Mining*,

WSDM '20, page 903–904, New York, NY, USA. Association for Computing Machinery.

Jonathan Frankle and Michael Carbin. 2018. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *ICLR*.

Philip Gage. 1994. A new algorithm for data compression. *C Users J.*, 12(2):23–38.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models.

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*, pages 1–9. Association for Computational Linguistics.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *ArXiv*, abs/1901.05287.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Mitchell Gordon, Kevin Duh, and Nicholas Andrews. 2020. Compressing BERT: Studying the effects of weight pruning on transfer learning. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 143–155, Online. Association for Computational Linguistics.

Jiatao Gu, Qi Liu, and Kyunghyun Cho. 2019. Insertion-based decoding with automatically inferred generation order. *Transactions of the Association for Computational Linguistics*, 7(0):661–676.

Fu-Ming Guo, Sijia Liu, Finlay S. Mungall, Xue Lin, and Yanzhi Wang. 2019. Reweighted proximal pruning for large-scale language representation.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. In *ICML*.

Song Han, Jeff Pool, John Tran, and William Dally. 2015. Learning both weights and connections for efficient neural network. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1135–1143. Curran Associates, Inc.

Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. 2017. A joint many-task model: Growing a neural network for multiple NLP tasks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1923–1933, Copenhagen, Denmark. Association for Computational Linguistics.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. 2020. TaPas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4320–4333, Online. Association for Computational Linguistics.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531.

Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization.

Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q. Weinberger. 2016. Deep networks with stochastic depth. In *ECCV*.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *J. Artif. Int. Res.*, 61(1):907–926.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Yacine Jernite, Samuel R. Bowman, and David Sontag. 2017. Discourse-based objectives for fast unsupervised sentence representation learning. *ArXiv*, abs/1705.00557.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2020. SMART: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2177–2190, Online. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8(0):64–77.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models.

Nora Kassner and Hinrich Schütze. 2019. Negated lama: Birds cannot fly.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A conditional transformer language model for controllable generation. *ArXiv*, abs/1909.05858.

Phillip Keung, Yichao Lu, Julian Salazar, and Vikas Bhardwaj. 2020. On the evaluation of contextual embeddings for zero-shot cross-lingual transfer learning. *arXiv preprint arXiv:2004.15001*.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. In *International Conference on Learning Representations*.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. 2018. Findings of the WMT 2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 726–739, Belgium, Brussels. Association for Computational Linguistics.

Lingpeng Kong, Cyprien de Masson d'Autume, Lei Yu, Wang Ling, Zihang Dai, and Dani Yogatama. 2020. A mutual information maximization perspective of language representation learning. In *International Conference on Learning Representations*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Jaesong Lee, Joong-Hwi Shin, and Jun-Seok Kim. 2017. Interactive visualization and manipulation of attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 121–126, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Liunian Li, Patrick Chen, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Efficient contextual representation learning with continuous outputs. *Transactions of the Association for Computational Linguistics*, 7(0):611–624.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. Xglue: A new benchmark dataset for cross-lingual pretraining, understanding and generation.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy. Association for Computational Linguistics.

Zachary C. Lipton. 2018. The mythos of model interpretability. *Queue*, 16(3):31–57.

Liyuan Liu, Xiang Ren, Jingbo Shang, Xiaotao Gu, Jian Peng, and Jiawei Han. 2018a. Efficient contextualized representation: Language model pruning for sequence labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1215–1225, Brussels, Belgium. Association for Computational Linguistics.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Qi Liu, Matt J. Kusner, and Phil Blunsom. 2020a. A survey on contextual embeddings.

Qianchu Liu, Diana McCarthy, Ivan Vulić, and Anna Korhonen. 2019b. Investigating cross-lingual alignment methods for contextualized embeddings with token-level evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 33–43, Hong Kong, China. Association for Computational Linguistics.

Shusen Liu, Tao Li, Zhimin Li, Vivek Srikumar, Valerio Pascucci, and Peer-Timo Bremer. 2018b. Visual interrogation of attention-based models for natural language inference and machine comprehension. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 36–41, Brussels, Belgium. Association for Computational Linguistics.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Haotang Deng, and Qi Ju. 2020b. Fastbert: a self-distilling bert with adaptive inference time.

Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-yi Wang. 2015. Representation learning using multi-task deep neural networks for semantic classification and information retrieval. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 912–921, Denver, Colorado. Association for Computational Linguistics.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019c. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020c. Multilingual denoising pre-training for neural machine translation.

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

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*, pages 6297–6308.

J. Scott McCarley. 2019. Pruning a bert-based question answering model. *ArXiv*, abs/1910.06360.

Gábor Melis, Tomás Kociský, and Phil Blunsom. 2020. Mogrifier lstm. In *International Conference on Learning Representations*.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems 32*, pages 14014–14024. Curran Associates, Inc.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden. Association for Computational Linguistics.

Phoebe Mulcaire, Jungo Kasai, and Noah A. Smith. 2019. Polyglot contextual representations improve crosslingual transfer. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3912–3918, Minneapolis, Minnesota. Association for Computational Linguistics.

Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar GuÌ‡lçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Allen Nie, Erin Bennett, and Noah Goodman. 2019. DisSent: Learning sentence representations from explicit discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4497–4510, Florence, Italy. Association for Computational Linguistics.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany. Association for Computational Linguistics.

Vijayaditya Peddinti, Vimal Manohar, Yiming Wang, Daniel Povey, and Sanjeev Khudanpur. 2016. Far-field ASR without parallel data. In *INTERSPEECH*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *ArXiv*, abs/1811.01088.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81, Brussels, Belgium. Association for Computational Linguistics.

Yada Pruksachatkun, Phil Yeres, Haokun Liu, Jason Phang, Phu Mon Htut, Alex Wang, Ian Tenney, and Samuel R. Bowman. 2020. jiant: A software toolkit for research on general-purpose text understanding models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 109–117, Online. Association for Computational Linguistics.

Jiezhong Qiu, Hao Ma, Omer Levy, Scott Wen tau Yih, Sinong Wang, and Jie Tang. 2019. Blockwise self-attention for long document understanding.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *The 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing*.

Timo Schick and Hinrich Schütze. 2020. It's not just size that matters: Small language models are also few-shot learners.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *International Conference on Acoustics, Speech and Signal Processing*, pages 5149–5152.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1599–1613, Minneapolis, Minnesota. Association for Computational Linguistics.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2019. Green AI. *ArXiv*, abs/1907.10597.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Sheng Shen, Zhen Dong, Jiayu Ye, Linjian Ma, Zhewei Yao, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. 2020. Q-BERT: Hessian based ultra low precision quantization of BERT. In *AAAI*.

Noah A. Smith. 2019. Contextual word representations: A contextual introduction. *ArXiv*, abs/1902.06006.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: Masked sequence to sequence pre-training for language generation. In *ICML*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and permuted pre-training for language understanding. *arXiv preprint arXiv:2004.09297*.

Mitchell Stern, William Chan, Jamie Ryan Kiros, and Jakob Uszkoreit. 2019. Insertion transformer: Flexible sequence generation via insertion operations. In *ICML*.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

Sainbayar Sukhbaatar, Edouard Grave, Piotr Bojanowski, and Armand Joulin. 2019. Adaptive attention span in transformers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 331–335, Florence, Italy. Association for Computational Linguistics.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting unreasonable effectiveness of data in deep learning era. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 843–852.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019a. Patient knowledge distillation for BERT model compression. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4322–4331, Hong Kong, China. Association for Computational Linguistics.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019b. Ernie: Enhanced representation through knowledge integration. *ArXiv*, abs/1904.09223.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020a. Ernie 2.0: A continual pre-training framework for language understanding. In *AAAI*.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020b. MobileBERT: a compact task-agnostic BERT for resource-limited devices. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2158–2170, Online. Association for Computational Linguistics.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5099–5110, Hong Kong, China. Association for Computational Linguistics.

Raphael Tang, Yao Lu, and Jimmy Lin. 2019. Natural language generation for effective knowledge distillation. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 202–208, Hong Kong, China. Association for Computational Linguistics.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2020. Efficient transformers: A survey.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

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. 2019b. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

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 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy. Association for Computational Linguistics.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4396–4406, Hong Kong, China. Association for Computational Linguistics.

Alex Wang, Jan Hula, Patrick Xia, Raghavendra Pappagari, R. Thomas McCoy, Roma Patel, Najoung Kim, Ian Tenney, Yinghui Huang, Katherin Yu, Shuning Jin, Berlin Chen, Benjamin Van Durme, Edouard Grave, Ellie Pavlick, and Samuel R. Bowman. 2019a. Can you tell me how to get past sesame street? sentence-level pretraining beyond language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4465–4476, Florence, Italy. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019b. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32*, pages 3266–3280. Curran Associates, Inc.

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

Liang Wang, Wei Zhao, Ruoyu Jia, Sujian Li, and Jingming Liu. 2019d. Denoising based sequence-to-sequence pre-training for text generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4003–4015, Hong Kong, China. Association for Computational Linguistics.

Sinong Wang, Madian Khabsa, and Hao Ma. 2020a. To pretrain or not to pretrain: Examining the benefits of pretrainng on resource rich tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2209–2213, Online. Association for Computational Linguistics.

Wei Wang, Bin Bi, Ming Yan, Chen Wu, Jiangnan Xia, Zuyi Bao, Liwei Peng, and Luo Si. 2020b. Structbert: Incorporating language structures into pre-training for deep language understanding. In *International Conference on Learning Representations*.

Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019e. Cross-lingual BERT transformation for zero-shot dependency parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5720–5726, Hong Kong, China. Association for Computational Linguistics.

Ziheng Wang, Jeremy Wohlwend, and Tao Lei. 2019f. Structured pruning of large language models. *ArXiv*, abs/1910.04732.

Zirui Wang, Jiateng Xie, Ruochen Xu, Yiming Yang, Graham Neubig, and Jaime G. Carbonell. 2020c. Cross-lingual alignment vs joint training: A comparative study and a simple unified framework. In *International Conference on Learning Representations*.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretic, and Samuel R. Bowman. 2019. Investigating BERT's knowledge of language: Five analysis methods with NPIs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2877–2887, Hong Kong, China. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Canwen Xu, Wangchunshu Zhou, Tao Ge, Furu Wei, and Ming Zhou. 2020. Bert-of-theseus: Compressing bert by progressive module replacing.

Hainan Xu and Philipp Koehn. 2017. Zipporah: a fast and scalable data cleaning system for noisy web-crawled parallel corpora. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2945–2950, Copenhagen, Denmark. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32*, pages 5753–5763. Curran Associates, Inc.

Zihao Ye, Qipeng Guo, Quan Gan, Xipeng Qiu, and Zheng Zhang. 2019. Bp-transformer: Modelling long-range context via binary partitioning.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. 2020. Large batch optimization for deep learning: Training bert in 76 minutes. In *International Conference on Learning Representations*.

Ofir Zafrir, Guy Boudoukh, Peter Izsak, and Moshe Wasserblat. 2019. Q8bert: Quantized 8bit bert. *The 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing*, abs/1910.06188.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. Record: Bridging the gap between human and machine commonsense reading comprehension.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2020. Freelb: Enhanced adversarial training for natural language understanding. In *International Conference on Learning Representations*.