On the Interplay Between Fine-tuning and Composition in Transformers

Lang Yu
Deptartment of Computer Science
University of Chicago
langyu@uchicago.edu

Allyson Ettinger
Department of Linguistics
University of Chicago
aettinger@uchicago.edu

Abstract

Pre-trained transformer language models have shown remarkable performance on a variety of NLP tasks. However, recent research has suggested that phrase-level representations in these models reflect heavy influences of lexical content, but lack evidence of sophisticated, compositional phrase information (Yu and Ettinger, 2020). Here we investigate the impact of fine-tuning on the capacity of contextualized embeddings to capture phrase meaning information beyond lexical content. Specifically, we fine-tune models on an adversarial paraphrase classification task with high lexical overlap, and on a sentiment classification task. After fine-tuning, we analyze phrasal representations in controlled settings following prior work. We find that fine-tuning largely fails to benefit compositionality in these representations, though training on sentiment yields a small, localized benefit for certain models. In follow-up analyses, we identify confounding cues in the paraphrase dataset that may explain the lack of composition benefits from that task, and we discuss potential factors underlying the localized benefits from sentiment training.

1 Introduction

Transformer language models like BERT (Devlin et al., 2019), GPT (Radford et al., 2018, 2019) and XLNet (Yang et al., 2019b), have improved the state-of-art in many NLP tasks since their introduction. The versatility of these pre-trained models suggests that they may acquire fairly robust linguistic knowledge and capacity for natural language "understanding". However, an emerging body of analysis demonstrates a level of superficiality in these models' handling of language (Niven and Kao, 2019; Kim and Linzen, 2020; McCoy et al., 2019; Ettinger, 2020; Yu and Ettinger, 2020).

In particular, although *composition*—a model's capacity to combine meaning units into more complex units reflecting phrase meanings—is an indispensable component of language understanding, when testing for composition in pre-trained transformer representations, Yu and Ettinger (2020) report that these representations reflect word content of phrases, but don't show signs of more sophisticated humanlike composition beyond word content. In the present paper we perform a direct follow-up of that study, asking whether models will show better evidence of composition after fine-tuning on tasks that are good candidates for requiring composition: 1) the Quora Question Pairs dataset in Paraphrase Adversaries from Word Scrambling (PAWS-QQP) (Zhang et al., 2019a), an adversarial paraphrase dataset forcing models to classify paraphrases with high lexical overlap, and 2) the Stanford Sentiment Treebank (Socher et al., 2013), a sentiment dataset with fine-grained phrase labels to promote composition. We base our analysis on the tests proposed by Yu and Ettinger (2020), which rely on alignment with human judgments of phrase pair similarities, and which leverage control of lexical overlap to target compositionality. We fine-tune and evaluate the same models and representation types tested in that paper, for optimal comparison.

We find that across the board, fine-tuning on PAWS-QQP does not improve compositionality—if anything, performance on composition metrics tends to degrade. Composition performance also remains low after training on SST, but we do see some localized improvements for certain models. Analyzing the PAWS-QQP dataset, we find reliable superficial cues to paraphrase labels (distance of word swap), explaining in part why fine-tuning on that task might fail to improve composition—and reinforcing the need for caution in interpreting difficulty of NLP tasks. We also discuss the contribution of variation in size of labeled phrases in SST, with respect to the benefits that result from fine-tuning on that task. All experimental code and

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 2279–2293

data are made available for further testing.[1]

2 Related work

Extensive work has studied the nature of learned representations in NLP models (Adi et al., 2016; Conneau et al., 2018; Ettinger et al., 2016; Durrani et al., 2020). Our work builds in particular on analysis of contextualized representations (Bacon and Regier, 2019; Tenney et al., 2019; Peters et al., 2018; Hewitt and Manning, 2019; Klafka and Ettinger, 2020; Toshniwal et al., 2020). Other work that has focused on transformers, as we do, has often focused on analyzing the attention mechanism (Vig and Belinkov, 2019; Clark et al., 2019), learned parameters (Roberts et al., 2020; Radford et al., 2019; Raffel et al., 2020) and redundancy (Dalvi et al., 2020; Voita et al., 2019; Michel et al., 2019). The evaluation that we use here follows the paradigm of classification-based probing (Kim et al., 2019; Wang et al., 2018; Zhang et al., 2019b; Yang et al., 2019a) and correlation with similarity judgments (Finkelstein et al., 2001; Gerz et al., 2016; Hill et al., 2015; Conneau and Kiela, 2018).

The current paper also builds on work subjecting trained NLP models to adversarial inputs, to highlight model weaknesses. One body of work approaches the problem by applying heuristic rules of perturbation to input sequences (Wallace et al., 2019; Jia and Liang, 2017; Zhang et al., 2019a), while another uses neural models to construct adversarial examples (Li et al., 2020, 2018) or manipulate inputs in embedding space (Jin et al., 2020). Our work also contributes to efforts to understand impacts and outcomes of the fine-tuning process (Miaschi et al., 2020; Mosbach et al., 2020; Wang et al., 2020; Perez-Mayos et al., 2021).

Phrase and sentence composition has drawn frequent attention in analysis of neural models, often focusing on analysis of internal representations and downstream task behavior (Ettinger et al., 2018; Conneau et al., 2019; Nandakumar et al., 2019; McCoy et al., 2019; Yu and Ettinger, 2020; Bhathena et al., 2020; Mu and Andreas, 2020; Andreas, 2019). Some work investigates compositionality via constructing linguistic (Keysers et al., 2019) and non-linguistic (Liška et al., 2018; Hupkes et al., 2018; Baan et al., 2019) synthetic datasets.

Most related to our work here is the finding of Yu and Ettinger (2020). They test for composition in two-word phrase representations from transformers, via similarity correlations and paraphrase detection. They find that baseline performance on these tasks is high, but once they control for amount of word overlap, performance drops dramatically, suggesting that observed correspondences rely on word content rather than phrase composition. We build directly on this work, testing whether these patterns will still hold after fine-tuning on tasks intended to encourage composition.

3 Fine-tuning Pre-trained Transformers

In response to the weaknesses observed by Yu and Ettinger (2020), we select two different datasets with promising characteristics for addressing these weaknesses. We fine-tune on these tasks, then perform layer-wise testing on contextualized representations from the fine-tuned models, comparing against results on the pre-trained models. Here we describe the two fine-tuning datasets.

3.1 PAWS: fine-tuning on high word overlap

The core of the Yu and Ettinger (2020) finding is that model performance on the selected composition tests degrades significantly when cues of lexical overlap are controlled. It stands to reason, then, that a model trained to discern meaning differences under conditions of high lexical overlap may improve on these overlap-controlled composition tests. This drives our selection of the Paraphrase Adversaries from Word Scrambling (PAWS) dataset (Zhang et al., 2019b), which consists of sentence pairs with high lexical overlap. The task is formulated as binary classification of whether two sentences are paraphrases or not. State-of-the-art models achieve only $< 40\%$ accuracy before training on the dataset (Zhang et al., 2019a). Table 1 shows examples from this dataset. Due to the high lexical overlap, we might expect that in order to achieve non-trivial accuracy on this task, models must attend to more sophisticated meaning information than simple word content.

3.2 SST: fine-tuning on hierarchical labels

Another dataset that has been associated with training and evaluation of phrasal composition is the Stanford Sentiment Treebank, which contains syntactic phrases of various lengths, together with fine-grained human-annotated sentiment labels for these phrases. Because this dataset contains annotations

[1] Datasets and code available at https://github.com/yulang/fine-tuning-and-composition-in-transformers

Sentence 1	Sentence 2	Label
There are also specific discussions , public profile debates and project discussions .	There are also public discussions , profile specific discussions , and project discussions .	0
She worked and lived in Stuttgart , Berlin (Germany) and in Vienna (Austria) .	She worked and lived in Germany (Stuttgart , Berlin) and in Vienna (Austria) .	1

Table 1: Example pairs from PAWS-QQP. Both positive and negative pairs have high bag-of-words overlap.

of composed phrases of various sizes, we can reasonably expect that training on this dataset may foster an increased sensitivity to compositional phrase meaning. We formulate the fine-tuning task as a 5-class classification task following the setup in Socher et al. (2013). The models are trained to predict sentiment labels given phrases as input.

4 Representation evaluation

For optimal comparison of the effects of fine-tuning on the above tasks, we replicate the tests, representation types, and models reported on by Yu and Ettinger. Here we briefly describe these methods. For more details on the evaluation dataset and task setup, please refer to Yu and Ettinger (2020).

4.1 Evaluation tasks

Yu and Ettinger propose two analyses for measuring composition: similarity correlations and paraphrase classification. They focus on two-word phrases, using the BiRD bigram relatedness dataset (Asaadi et al., 2019) for similarity correlations, and the PPDB 2.0 paraphrase database (Ganitkevitch et al., 2013; Pavlick et al., 2015) for paraphrase classification. BiRD contains 3,345 bigram pairs, with source phrases paired with numerous target phrases, and human-annotated similarity scores ranging from 0 to 1. For similarity correlation, Yu and Ettinger take layer-wise correlations between these human phrase similarity scores and the cosine similarities of model representations for the same phrases. For paraphrase classification, Yu and Ettinger train a multi-layer perceptron classifier to label whether two phrase representations are paraphrases, drawing their positive phrase pairs from PPDB 2.0—which contains paraphrases with scores generated by a regression model—and randomly sampling negative pairs from the rest of the dataset. We replicate all of these procedures.

For both task types, Yu and Ettinger compare between "uncontrolled" and "controlled" tests, with the latter filtering the data to control word overlap within phrase pairs, such that amount of word overlap between two phrases can no longer be used as a cue for how similar the meanings are. It is on these controlled settings that Yu and Ettinger observe the significant drop in performance, suggesting that model representations lack the compositional knowledge to discern phrase meaning beyond word content. Below we will report results for both settings, with particular focus on controlled settings.

4.2 Representation types

Following Yu and Ettinger, for each input phrase we test as a potential representation 1) CLS token, 2) average of tokens within the phrase (Avg-Phrase), 3) average of all input tokens (Avg-All), 4) embedding of the second word of the phrase, intended to approximate the semantic head (Head-Word), and 5) SEP token. We test each of these representations at every layer of each model.[2]

5 Experimental setup

We fine-tune and analyze the same models that Yu and Ettinger test in pre-trained form: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), DistilBERT (Sanh et al., 2019), XLNet (Yang et al., 2019b) and XLM-RoBERTa (Conneau et al., 2019). In each case, the pre-trained "base" version is used as the starting point for fine-tuning. We use the implementation of Wolf et al. (2019)[3] based on PyTorch (Paszke et al., 2019).

We fine-tune these models on the two datasets described in Section 3. The Quora Question Pairs dataset in Paraphrase Adversaries from Word Scrambling (PAWS-QQP)[4] consists of a training set with 11,988 sentence pairs, and a dev/test set with 677 sentence pairs. Tuning on PAWS-QQP is formulated as binary classification. Sentences are passed as input and models are trained to pre-

[2] Like Yu and Ettinger, we also test both phrase-only input (encoder input consists only of two-word phrase plus special CLS/SEP tokens), as well as inputs in which phrases are embedded in sentence contexts.

[3] https://github.com/huggingface/transformers

[4] https://github.com/google-research-datasets/paws

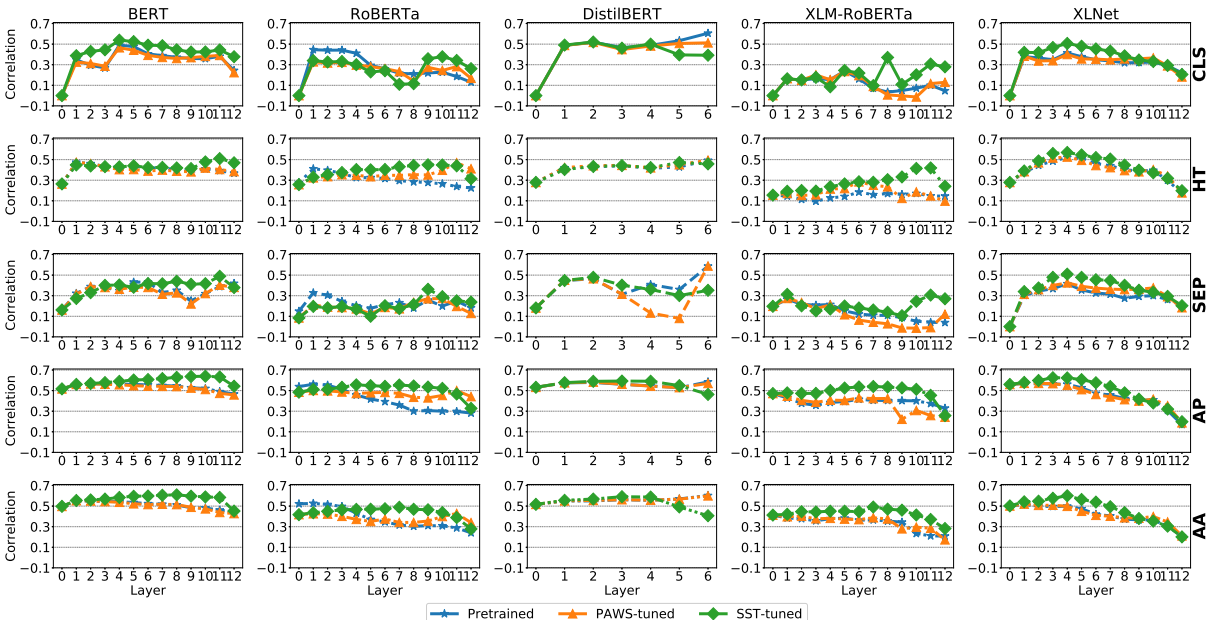

Figure 1: Similarity correlation on uncontrolled BiRD dataset, with phrase-only input. Columns correspond to models, and rows correspond to representation types ("HT" = Head-token, "AP" = Avg-Phrase and "AA" = Avg-All). For each model and representation type, the corresponding subplot shows correlations for pre-trained, PAWS-tuned and SST-tuned settings, respectively. For each subplot, X-axis corresponds to layer index, and Y-axis corresponds to correlation value. Layer 0 corresponds to input embeddings passed to the model.

dict whether the input sentences are paraphrases or not. Models are trained on the training set, and validated on the dev/test set for convergence.

The Stanford Sentiment Treebank (SST)[5] (Socher et al., 2013) contains 215,154 phrases. 15% of the data is reserved for validation. The fine-tuning task is formulated as 5-class classification on sentiment labels, where models are given phrases as input, and asked to predict sentiment. In both tasks, we use the Adam optimizer (Kingma and Ba, 2014) with default weight decay. We train the models until convergence on the validation set.

The evaluation tasks consist of correlation analysis and paraphrase classification. For correlation in the uncontrolled setting, we use the complete BiRD dataset, containing 3,345 phrase pairs.[6] For the controlled test, we filter the complete dataset following the criteria in Yu and Ettinger (2020), resulting in 410 "AB-BA" mirror-image pairs with 100% word overlap (e.g., *law school* / *school law*). For the classification tasks, we use the preprocessed data released by Yu and Ettinger (2020).[7] We collect 12,036 source-target phrase pairs from the preprocessed dataset for our uncontrolled classification setting, and for the controlled classification setting, we collect 11,772 phrase pairs with exactly 50% word overlap in each pair, following the procedure from the original paper.

6 Results after fine-tuning

6.1 Full datasets

Figure 1 presents the original results from Yu and Ettinger (2020) on pre-trained models, alongside our new results after fine-tuning, on the full BiRD dataset. Since this is prior to the control of word overlap, these correlations can be expected to reflect effects of lexical content encoding, without yet having isolated effects of composition. We find that after fine-tuning on SST, most models and representation types show small improvements in peak correlations across layers, while fine-tuning on PAWS also yields improvements in peak correlations—albeit even smaller—in models other than BERT and XLM-RoBERTa. Overall, within a given representation type, improvements are generally stronger after fine-tuning on SST than on PAWS. Between representation types, Avg-Phrase and Avg-All remain consistently at the highest correlations after fine-tuning. Additionally, we see that the decline in correlation at later layers in pre-trained BERT, RoBERTa and XLM-RoBERTa is mitigated

[5] https://nlp.stanford.edu/sentiment/treebank.html

[6] http://saifmohammad.com/WebPages/BiRD.html

[7] https://github.com/yulang/phrasal-composition-in-transformers

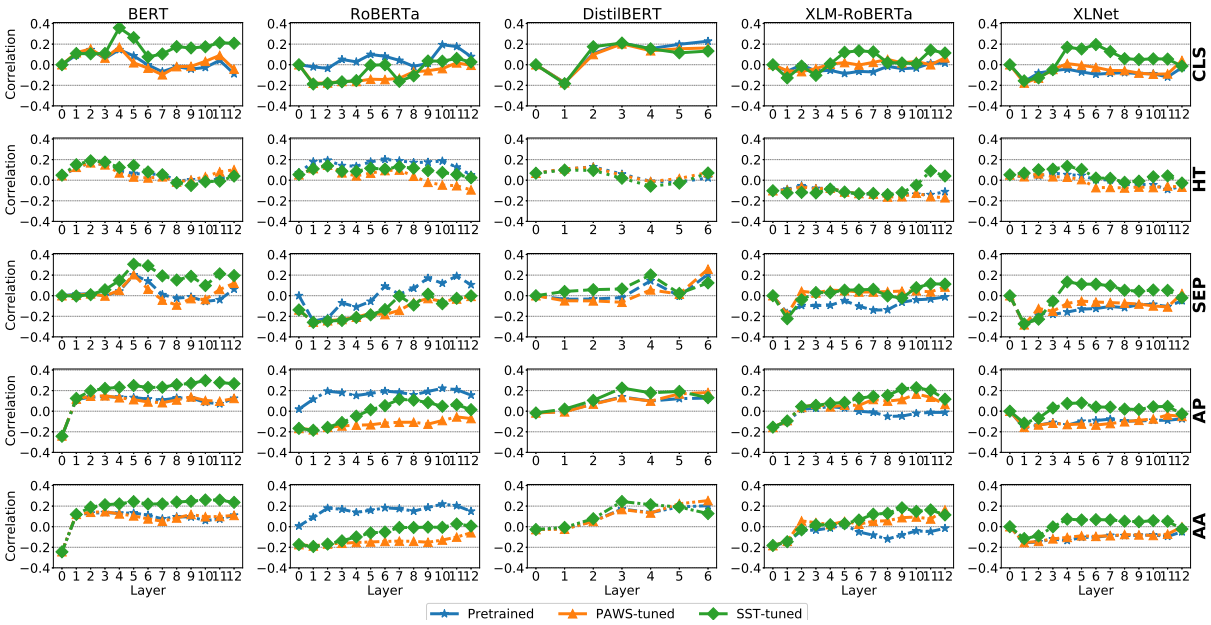

Figure 2: Similarity correlation on controlled BiRD dataset (AB-BA setting), with phrase-only input.

after fine-tuning. Model-wise, we see the most significant improvements in the RoBERTa model, for which the correlations become more consistent across layers for most representation types. As we discuss below, we take this as indication that the fine-tuning promotes more robust retention of word content information across layers, if not more robust phrasal composition.

For the sake of space, we present the plots of the uncontrolled paraphrase classification setting in Figure 7 of the Appendix. The overall improvements are even smaller than those seen in the correlations, but we do see comparable patterns in these paraphrase classification results, in particular with SST showing slightly stronger benefits than PAWS.

6.2 Controlled datasets

Above we see small benefits of fine-tuning for performance on the full, uncontrolled datasets. However, the critical question for our purposes is whether correlations also show improvements in word-overlap controlled settings, which better isolate effects of composition. Figure 2 shows correlations for all models on the controlled AB-BA (full word overlap) correlation test. Figure 3 shows the results for the controlled paraphrase classification setting, where both paraphrase and non-paraphrase pairs have exactly 50% word overlap.

The first comparison to note is between original and controlled settings, which allows us to establish the contributions of overlap information as opposed to composition. Comparing between Figure 1 and Figure 2, it is clear that fine-tuned models still show substantial reduction in correlation when overlap cues are removed. The same goes for Figure 3 (by comparison to Figure 7 of the Appendix)—we see that on the controlled dataset, accuracies hover just above chance-level performance both before and after fine-tuning, compared to over 90% accuracy on the uncontrolled dataset. This gap in performance between the original and controlled datasets mirrors the findings of Yu and Ettinger (2020), and suggests that even after fine-tuning, the majority of correspondence between model phrase representations and human meaning similarity judgments can be attributed to capturing of word content information rather than phrasal composition.

The second key comparison is between pre-trained and fine-tuned models within the overlap-controlled settings. While the prior comparison tells us that similarity correspondence is still dominated by word content effects, this second comparison can tell us whether fine-tuning shows at least some boost in meaning composition relative to pre-training. Comparing performance of pre-trained and fine-tuned models in Figure 2, we see that fine-tuning on PAWS-QQP actually slightly degrades correlations at many layers for a majority of models and representation types—with improvements largely restricted to XLM-RoBERTa and XLNet (perhaps notably, mostly in cases where pre-trained correlations are negative). This is despite the fact that models achieve strong validation performance on PAWS-QQP (as shown in Table 2), suggesting

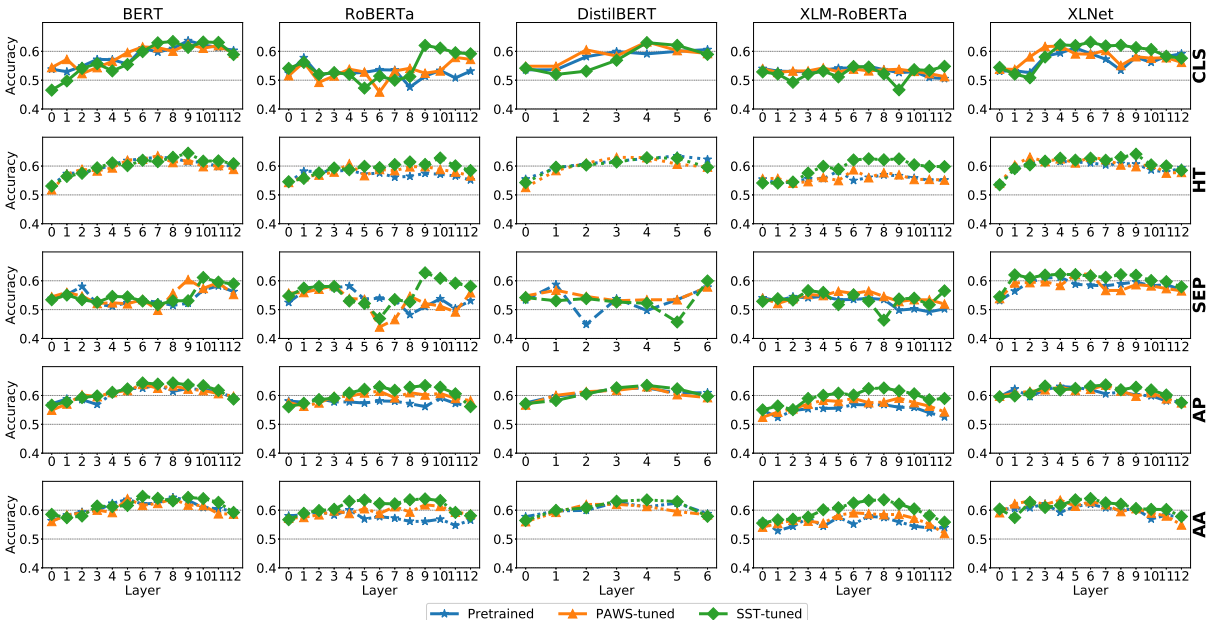

Figure 3: Paraphrase classification accuracy on controlled PPDB dataset (50% word overlap setting) with phrase-only input. Y-axis range is smaller relative to Figure 7, to make changes from pre-training more visible.

that learning this task does little to improve composition. We will explore the reasons for this below.

In Figure 3, we see that fine-tuning also does little to improve paraphrase classification accuracies in the controlled setting—though each model shows slight improvement in peak accuracy across layers and representation types (e.g., RoBERTa shows ∼3% increase in peak accuracy with SST tuning, and 2% with PAWS tuning). Even so, the best accuracies across models continue to be only marginally above chance. This, too, fails to provide evidence of any substantial composition improvement resulting from the fine-tuning process.

The story changes slightly when we turn to impacts of SST fine-tuning on correlations in Figure 2. While all correlations remain low after SST fine-tuning, we do see that correlations for BERT, XLM-RoBERTa and XLNet show some non-trivial benefits even in the controlled setting. In particular, SST tuning consistently improves correlation among all representation types in BERT (except for minor degradation in later layers for Head-token), boosting the highest correlation from ∼0.2 to ∼0.39. Between representation types, the greatest change is in the CLS token, with the most dramatic point of improvement being an abrupt correlation peak for CLS at BERT's fourth layer. We will discuss more below about this localized benefit.

A final important observation is that fine-tuning on either dataset produces clear degradation in correlations for all representation types in RoBERTa

Model	Accuracy(%)
BERT	80.13
RoBERTa	90.81
DistilBERT	81.98
XLM-RoBERTa	91.18
XLNet	88.24
Linear CLF	71.34

Table 2: Accuracy of fine-tuned models on PAWS-QQP dev/test set. Linear CLF is a baseline classifier with relative swapping distance as the only input feature.

under the controlled setting, by contrast to the general improvements seen for that and other models in the uncontrolled setting. This suggests that at least for that model, fine-tuning encourages retention or enhancement of lexical information, but may degrade compositional phrase information.[8]

7 Analyzing impact of fine-tuning

The presented results suggest that despite compelling reasons to think that fine-tuning on the selected tasks may improve composition of phrase meaning, these models mostly do not exhibit noteworthy benefits from fine-tuning. In particular, fine-

[8]Following Yu and Ettinger (2020), in addition to phrase-only inputs we also try embedding target phrases in sentence contexts. Consistent with the findings of Yu and Ettinger (2020), we see that presence of context words does boost overall correlation and accuracy, but does not alter the general trends. Moreover, models still show relatively weak performance on controlled tasks even with context available (see Figure 8 and Figure 9 in the Appendix for details).

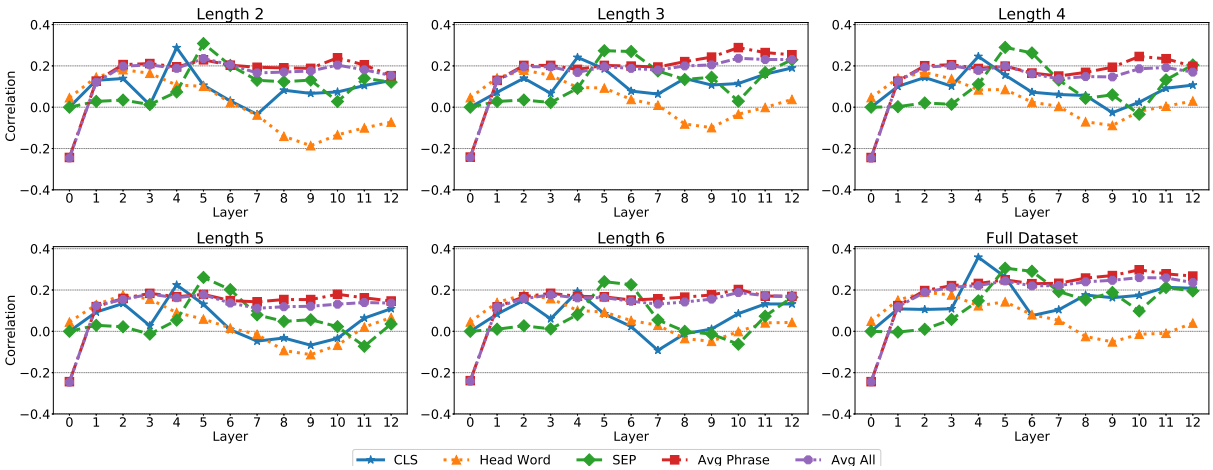

Figure 4: Layer-wise correlation of BERT fine-tuned on phrases of different lengths in SST.

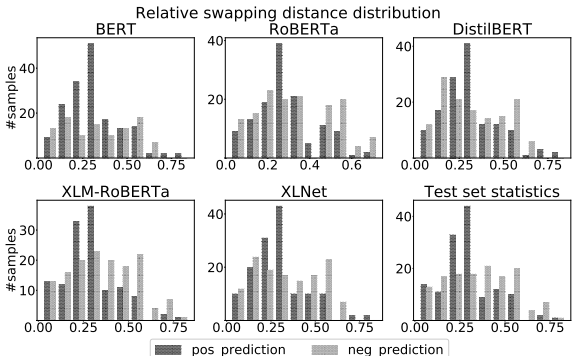

Figure 5: Distribution of positive and negative predictions made by tuned models. Last plot shows the statistics in the PAWS-QQP dev/test set. X-axis corresponds to relative swapping distance; Y-axis shows number of samples in the specific relative swapping distance bin.

tuning on the PAWS-QQP dataset often degrades performance on the controlled datasets taken to be most indicative of compositionality. As for SST, the benefits are minimal, but in localized cases like BERT's CLS token, we do see signs of improved compositionality. In this section, we conduct further analysis on the impacts of fine-tuning, and discuss why tuned models behave as they do.

7.1 Failure of PAWS-QQP

Table 2 shows accuracy of fine-tuned models on the dev/test set of PAWS-QQP.[9] It is clear that the models are learning to perform well on this dataset, but our results above indicate that this does not translate to improved composition sensitivity.

We explore the possibility that this discrepancy may be caused by trivial cues arising during the construction of the dataset, enabling models to infer paraphrase labels without needing to improve their understanding of the meaning of the sentence pair (c.f., Poliak et al., 2018; Gururangan et al., 2018). Sentence pairs in PAWS are generated via word swapping and back translation to ensure high bag-of-words overlap (Zhang et al., 2019a). We hypothesize that models may be able to achieve high performance in this task based on distance of the word swap alone, without requiring any sophisticated representation of sentence meaning.

To test this, given a sentence pair (s_1, s_2) with word counts l_1, l_2, respectively, we define "relative swapping distance" as

$$dist_{relative} = \frac{dist_{swap}}{max(l_1, l_2)}$$

where $dist_{swap}$ is defined as the index difference of the first swapping word in s_1 and s_2. For the example shown in the first row of Table 1, the first swapping word is "specific", with $dist_{swap} = 4$. Note that with this measure we focus on information from one word swap only, while some pairs in PAWS-QQP have multiple swapped words—so in reality, swapping distance information may be even stronger than our results below indicate.

In the last plot of Figure 5, we show an association between relative swapping distance and paraphrase labels in the PAWS dev/test set: sentence pairs with small swapping distance tend to be positive samples, while large swapping distance associates with negative labels. The other plots in Figure 5 show distribution of positive and negative predictions generated by each fine-tuned model with respect to relative swapping distance. We see a similar pattern, with models tending to generate negative labels when swapping distance is larger.

[9] The performance of BERT in the table is different from previous work mainly due to the fact that models in Zhang et al. (2019a) are tuned on concatenation of QQP and PAWS-QQP datasets rather than PAWS-QQP only.

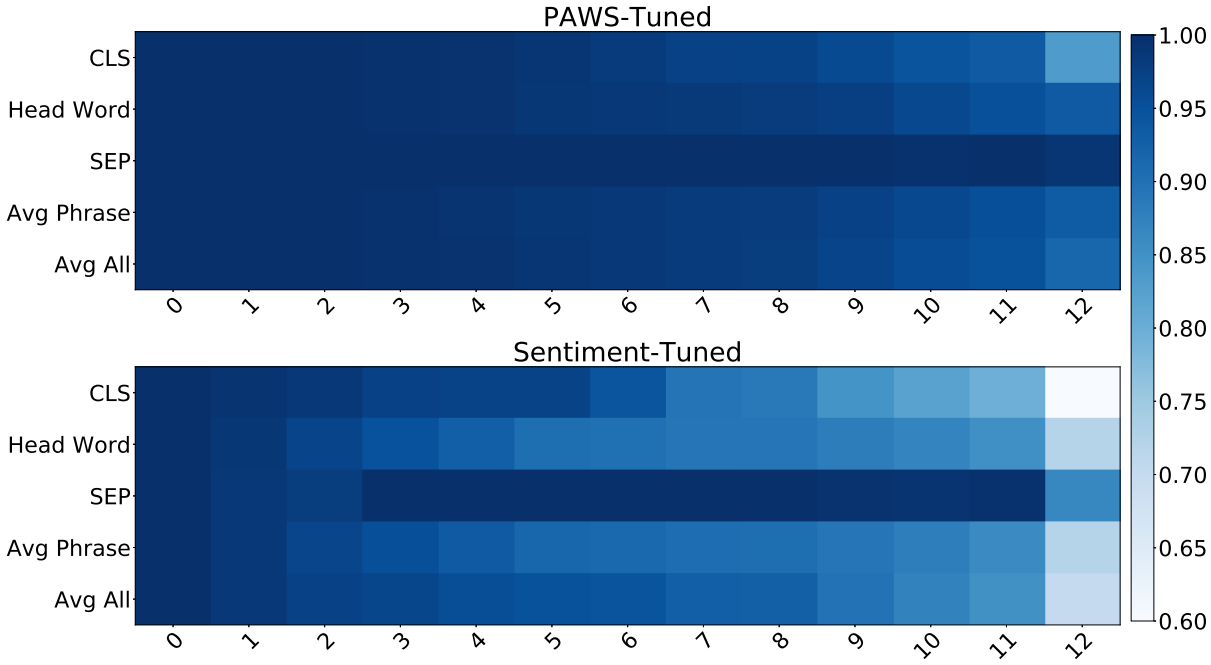

Figure 6: Average layer-wise embedding similarity between fine-tuned and pre-trained BERT. The upper half shows the comparison between PAWS-QQP tuned and pre-trained BERT. And the lower half presents Sentiment Treebank-tuned v.s. pre-trained. Embeddings are evaluated using full BiRD dataset for input.

To verify the viability of this cue, we train a simple linear classifier on PAWS-QQP, with relative swapping distance as the only input feature. The results are reported as "Linear CLF" in Table 2. Even without access to the content of the sentences, we see that this simple model is able to achieve non-trivial and comparably good classification accuracy on the dev/test set. The strong performance of the linear classifier and the distribution of predictions are consistent with the hypothesis that when we tune on PAWS-QQP, rather than forcing models to learn nuanced meaning in the absence of word overlap cues, we may instead encourage models to focus on lower-level information having little to do with the sentence meaning, further degrading their performance on the composition tasks.

7.2 Localized impacts of SST

Fine-tuning on sentiment shows a bit of a different pattern—while it mostly shows only minor changes from pre-training, and the correlations and classification accuracies remain at decidedly low levels on the controlled settings, we do see in certain models some distinctive changes in levels of similarity correlation as a result of tuning on SST. Notably, since these improvement patterns are seen in the similarity correlations but not in the classification accuracies, this suggests that these two tasks are picking up on slightly different aspects of phrasal compositionality. To investigate these effects further, we focus our attention on BERT, which shows the most distinctive improvement in correlations.

The obvious candidate for the source of the localized SST benefit is the dataset's inclusion of labeled syntactic phrases of various sizes. The benefits seen from SST tuning suggest that this may indeed encourage models to gain some finer-grained sensitivity to compositional impacts of phrase structure (at least those relevant for sentiment). To examine this further, we filter the SST dataset to subsets with phrases of the same length, from 2 to 6 words, and tune pre-trained BERT on each subset.

Figure 4 shows the correlations for BERT, fine-tuned on each phrase length, on the overlap-controlled BiRD dataset. We see that tuning on the full dataset (mixed phrase lengths) gives the strongest fourth-layer boost in CLS correlation performance—but among the size subsets, a semblance of the fourth-layer CLS peak is seen across phrase lengths, with length-2 training yielding the strongest peak, and length-6 training the smallest. This suggests an amount of size-based specialization—sentiment training on phrases of (or closer to) two words has more positive impact on similarity correlations for our two-word phrases.[10] However, we also see that phrases of

[10] Although subset size can potentially contribute to correlation performance, we find that subset size does not correlate

other sizes contribute non-trivially to the ultimate correlation improvement observed from training on the full dataset. This is consistent with the notion that training on diverse phrase sizes encourages fine-grained attention to compositionality, while training on phrases of similar size may have slightly more direct benefit.

Representation changes For further comparison of fine-tuning effects between tasks, we analyze changes in BERT representations at each layer before and after the fine-tuning process. Figure 6 shows the average layer-wise representation similarity between fine-tuned and pre-trained BERT given identical input. We see substantial differences between tasks in terms of representation changes: while SST fine-tuning produces significant changes across representations and layers, PAWS fine-tuning leaves representations largely unchanged (further supporting the notion that this task can be solved fairly trivially). We also see that after SST tuning, BERT's CLS token shows robust similarity to pre-trained representations until the fifth layer, followed by a rapid drop in similarity. This suggests that the fourth-layer correlation peak may be enabled in part by retention of key information from pre-training, combined with heightened phrase sensitivity from fine-tuning. We leave in-depth exploration of this dynamic for future work.

8 Discussion

The results of our experiments indicate that despite the promise of PAWS-QQP and SST tasks for improving models' phrasal composition, fine-tuning on these tasks falls far short of resolving the composition weaknesses observed by Yu and Ettinger (2020). The majority of correspondence with human judgments can still be attributed to word overlap effects—disappearing once overlap is controlled—and improvements on the controlled settings are absent, very small, or highly localized to particular models, layers and representations. This outcome aligns with the increasing body of evidence that NLP datasets often do not require of models the level of linguistic sophistication that we might hope for—and in particular, our identification of a strong spurious cue in the PAWS-QQP dataset adds to the growing number of findings emphasizing that NLP datasets often have artifacts that can inflate performance (Poliak et al., 2018; Gururangan et al., 2018; Kaushik and Lipton, 2018).

We do see a ray of promise in the small, localized benefits for certain models from tuning on SST. These improvements do not extend to all models, and are fairly small in the models that do see benefits—but as we discuss above, it appears that training on fine-grained syntactic phrase distinctions may indeed confer some enhancement of compositional meaning in phrase representations—at least when model conditions are amenable. Since sentiment information constitutes only a very limited aspect of phrase meaning, we anticipate that training on fine-grained phrase labels that reflect richer and more diverse meaning information could be a promising direction for promoting composition more robustly in these models.

9 Conclusions and future directions

We have tested effects of fine-tuning on phrase meaning composition in transformer representations. Although we select tasks with promise to address composition weaknesses and reliance on word overlap, we find that representations in the fine-tuned models show little improvement on controlled composition tests, or show only very localized improvements. Follow-up analyses suggest that the PAWS-QQP dataset contains spurious cues that undermine learning of sophisticated meaning properties when training on that task. However, results from SST tuning suggest that training on labeled phrases of various sizes could prove effective for learning composition. Future work should investigate how model properties interact with fine-tuning to produce improvements in particular models and layers—and should move toward phrase-level training with meaning-rich annotations, which we predict will be a promising direction for improving models' phrase meaning composition.

Acknowledgments

We would like to thank three anonymous reviewers for valuable feedback for improving this paper. We also thank members of the University of Chicago CompLing Lab for helpful comments and suggestions on this work. This material is based upon work supported by the National Science Foundation under Award No. 1941160.

with the performance patterns we observe here. Phrase count of each subset: length 2 - 11,499; length 3 - 11,779; length 4 - 15,050; length 5 - 11,816; length 6 - 9,935.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *arXiv preprint arXiv:1608.04207*.

Jacob Andreas. 2019. Measuring compositionality in representation learning. *arXiv preprint arXiv:1902.07181*.

Shima Asaadi, Saif Mohammad, and Svetlana Kiritchenko. 2019. Big bird: A large, fine-grained, bigram relatedness dataset for examining semantic composition. 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 505–516.

Joris Baan, Jana Leible, Mitja Nikolaus, David Rau, Dennis Ulmer, Tim Baumgärtner, Dieuwke Hupkes, and Elia Bruni. 2019. On the realization of compositionality in neural networks. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 127–137.

Geoff Bacon and Terry Regier. 2019. Does bert agree? evaluating knowledge of structure dependence through agreement relations. *arXiv preprint arXiv:1908.09892*.

Hanoz Bhathena, Angelica Willis, and Nathan Dass. 2020. Evaluating compositionality of sentence representation models. *ACL 2020*, page 185.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. 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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. 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.

Fahim Dalvi, Hassan Sajjad, Nadir Durrani, and Yonatan Belinkov. 2020. Analyzing redundancy in pretrained transformer models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4908–4926.

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.

Nadir Durrani, Hassan Sajjad, Fahim Dalvi, and Yonatan Belinkov. 2020. Analyzing individual neurons in pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4865–4880.

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*, 8:34–48.

Allyson Ettinger, Ahmed Elgohary, Colin Phillips, and Philip Resnik. 2018. Assessing composition in sentence vector representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1790–1801.

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.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. Ppdb: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764.

Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. Simverb-3500: A large-scale evaluation set of verb similarity. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2173–2182.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. 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 107–112.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word 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 4129–4138.

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

Dieuwke Hupkes, Anand Singh, Kris Korrel, German Kruszewski, and Elia Bruni. 2018. Learning compositionally through attentive guidance. *arXiv preprint arXiv:1805.09657*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8018–8025.

Divyansh Kaushik and Zachary C Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. *arXiv preprint arXiv:1808.04926*.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, et al. 2019. Measuring compositional generalization: A comprehensive method on realistic data. *arXiv preprint arXiv:1912.09713*.

Najoung Kim and Tal Linzen. 2020. Cogs: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105.

Najoung Kim, Roma Patelφ, Adam Poliak, Alex Wang, Patrick Xia, R Thomas McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, et al. 2019. Probing what different nlp tasks teach machines about function word comprehension. *NAACL HLT 2019*, page 235.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Josef Klafka and Allyson Ettinger. 2020. Spying on your neighbors: Fine-grained probing of contextual embeddings for information about surrounding words. *arXiv preprint arXiv:2005.01810*.

Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2018. Textbugger: Generating adversarial text against real-world applications. *arXiv preprint arXiv:1812.05271*.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. Bert-attack: Adversarial attack against bert using bert. *arXiv preprint arXiv:2004.09984*.

Adam Liška, Germán Kruszewski, and Marco Baroni. 2018. Memorize or generalize? searching for a compositional rnn in a haystack. *arXiv preprint arXiv:1802.06467*.

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

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448.

Alessio Miaschi, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2020. Linguistic profiling of a neural language model. *arXiv preprint arXiv:2010.01869*.

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

Marius Mosbach, Anna Khokhlova, Michael A Hedderich, and Dietrich Klakow. 2020. On the interplay between fine-tuning and sentence-level probing for linguistic knowledge in pre-trained transformers. *arXiv preprint arXiv:2010.02616*.

Jesse Mu and Jacob Andreas. 2020. Compositional explanations of neurons. *arXiv preprint arXiv:2006.14032*.

Navnita Nandakumar, Timothy Baldwin, and Bahar Salehi. 2019. How well do embedding models capture non-compositionality? a view from multiword expressions. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 27–34.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. *arXiv preprint arXiv:1907.07355*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca

Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. Ppdb 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 425–430.

Laura Perez-Mayos, Roberto Carlini, Miguel Ballesteros, and Leo Wanner. 2021. On the evolution of syntactic information encoded by bert's contextualized representations.

Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. *NAACL HLT 2018*, page 180.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel Bowman, Dipanjan Das, et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations, ICLR 2019*.

Shubham Toshniwal, Haoyue Shi, Bowen Shi, Lingyu Gao, Karen Livescu, and Kevin Gimpel. 2020. A cross-task analysis of text span representations. *arXiv preprint arXiv:2006.03866*.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing nlp. 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 2153–2162.

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 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Chengyu Wang, Minghui Qiu, Jun Huang, and Xiaofeng He. 2020. Meta fine-tuning neural language models for multi-domain text mining. *arXiv preprint arXiv:2003.13003*.

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.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019a. PAWS-X: A Cross-lingual Adversarial Dataset for Paraphrase Identification. In *Proc. of EMNLP*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019b. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in*

neural information processing systems, pages 5754–5764.

Lang Yu and Allyson Ettinger. 2020. Assessing phrasal representation and composition in transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4896–4907.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019a. Paws: Paraphrase adversaries from word scrambling. 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 1298–1308.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019b. PAWS: Paraphrase Adversaries from Word Scrambling. In *Proc. of NAACL*.

A Appendix

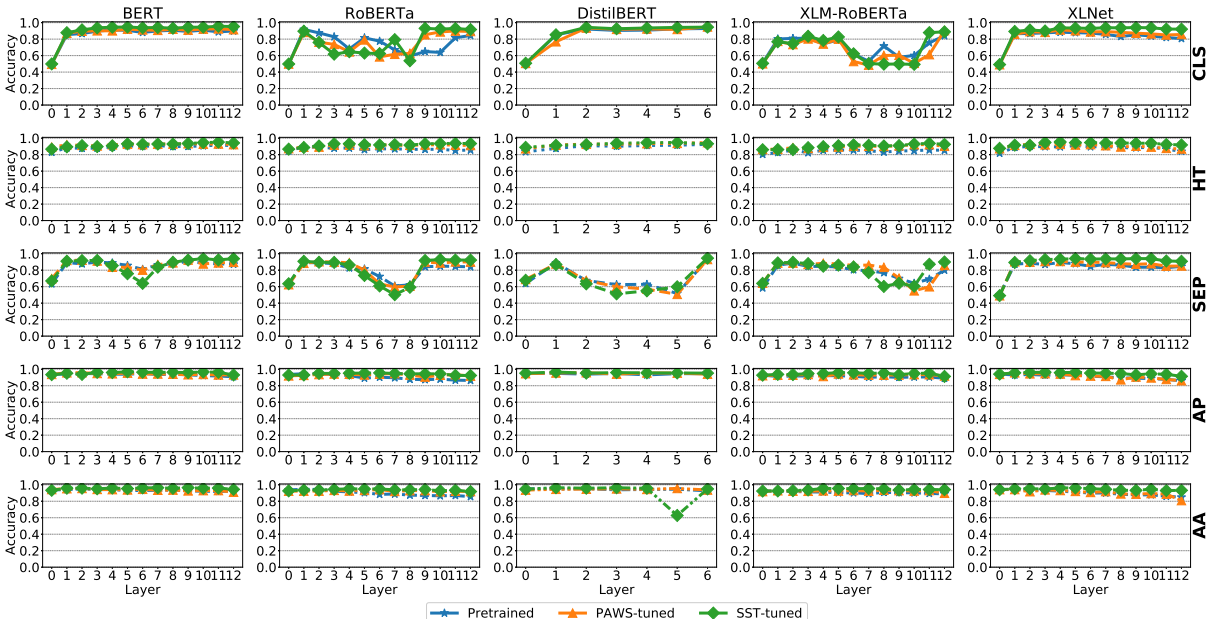

Figure 7: Paraphrase classification accuracy on uncontrolled PPDB dataset, with phrase-only input. Columns correspond to models, and rows correspond to representation types ("HT" = Head-Token, "AP" = Avg-Phrase and "AA" = Avg-All). For each model and representation type, the corresponding subplot shows accuracies for pre-trained, PAWS-tuned and SST-tuned settings, respectively. For each subplot, X-axis corresponds to layer index, and Y-axis corresponds to accuracy value. Layer 0 corresponds to input embeddings passed to the model.

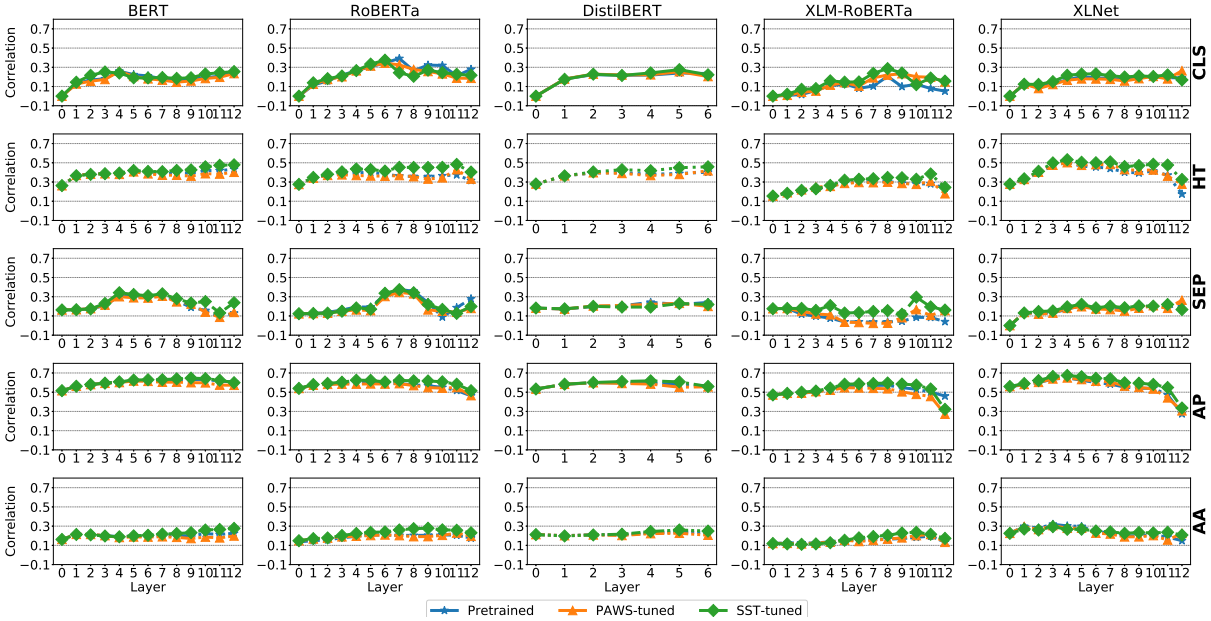

Figure 8: Similarity correlation on full BiRD dataset with phrases embedded in context sentence (context-available input).

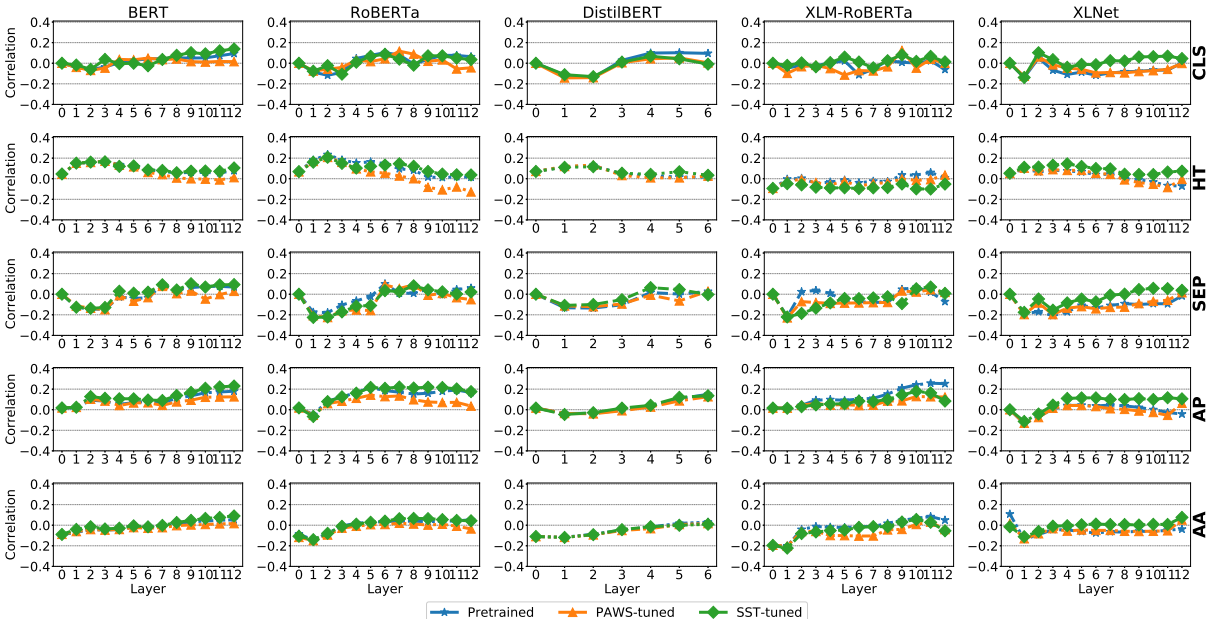

Figure 9: Similarity correlation on controlled BiRD dataset (AB-BA setting) with phrases embedded in context sentence (context-available input).