Assessing Phrasal Representation 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

Deep transformer models have pushed performance on NLP tasks to new limits, suggesting sophisticated treatment of complex linguistic inputs, such as phrases. However, we have limited understanding of how these models handle representation of phrases, and whether this reflects sophisticated composition of phrase meaning like that done by humans. In this paper, we present systematic analysis of phrasal representations in state-of-the-art pre-trained transformers. We use tests leveraging human judgments of phrase similarity and meaning shift, and compare results before and after control of word overlap, to tease apart lexical effects versus composition effects. We find that phrase representation in these models relies heavily on word content, with little evidence of nuanced composition. We also identify variations in phrase representation quality across models, layers, and representation types, and make corresponding recommendations for usage of representations from these models.

1 Introduction

A fundamental component of language understanding is the capacity to combine meaning units into larger units—a phenomenon known as *composition*—and to do so in a way that reflects the nuances of meaning as understood by humans. Transformers (Vaswani et al., 2017) have shown impressive performance in NLP, particularly transformers using pre-training, like BERT (Devlin et al., 2019) and GPT (Radford et al., 2018, 2019), suggesting that these models may be succeeding at composition of complex meanings. However, because transformers (like other contextual embedding models) typically maintain representations for every token, it is unclear how and at what points they might be combining word meanings into phrase meanings. This contrasts with models that incorporate explicit phrasal composition into their architecture, e.g. RNNG (Dyer et al., 2016; Kim et al., 2019), recursive models for semantic composition (Socher et al., 2013), or transformers with attention-based composition modules (Yin et al., 2020).

In this paper we take steps to clarify the nature of phrasal representation in transformers. We focus on representation of two-word phrases, and we prioritize identifying and teasing apart two important but distinct notions: how faithfully the models are representing information about the *words* that make up the phrase, and how faithfully the models are representing the nuances of the *composed phrase meaning* itself, over and above a simple account of the component words. To do this, we begin with existing methods for testing how well representations align with human judgments of meaning similarity: similarity correlations and paraphrase classification. We then introduce controlled variants of these datasets, removing cues of word overlap, in order to distinguish effects of word content from effects of more sophisticated composition. We complement these phrase similarity analyses with classic sense selection tests of phrasal composition (Kintsch, 2001).

We apply these tests for systematic analysis of several state-of-the-art transformers: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019b), DistilBERT (Sanh et al., 2019), XLNet (Yang et al., 2019b) and XLM-RoBERTa (Conneau et al., 2019). We run the tests in layerwise fashion, to establish the evolution of phrase information as layers progress, and we test various tokens and token combinations as phrase representations. We find that when word overlap is not controlled, models show strong correspondence with human judgments, with noteworthy patterns of variation across models, layers, and representation types. However, we find that correspondence drops substantially once word overlap is controlled, suggesting that although these transformers contain faithful repre-

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 4896–4907,
November 16–20, 2020.

sentations of the lexical content of phrases, there is little evidence that these representations capture sophisticated details of meaning composition beyond word content. Based on the observed representation patterns, we make recommendations for selection of representations from these models. All code and controlled datesets are made available for replication and application to additional models.[1]

2 Related work

This paper contributes to a growing body of work on analysis of neural network models. Much work has studied recurrent neural network language models (Linzen et al., 2016; Wilcox et al., 2018; Chowdhury and Zamparelli, 2018; Gulordava et al., 2018; Futrell et al., 2019) and sentence encoders (Adi et al., 2016; Conneau et al., 2018; Ettinger et al., 2016). Our work builds in particular on analysis of information encoded in contextualized token representations (Bacon and Regier, 2019; Tenney et al., 2019b; Peters et al., 2018; Hewitt and Manning, 2019; Klafka and Ettinger, 2020) and in different layers of transformers (Tenney et al., 2019a; Jawahar et al., 2019). The BERT model has been a particular focus of analysis work since its introduction. Previous work has focused on analyzing the attention mechanism (Vig and Belinkov, 2019; Clark et al., 2019), parameters (Roberts et al., 2020; Radford et al., 2019; Raffel et al., 2020) and embeddings (Shwartz and Dagan, 2019; Liu et al., 2019a). We build on this work with a particular, controlled focus on the evolution of phrasal representation in a variety of state-of-the-art transformers.

Composition has been a topic of frequent interest when examining neural networks and their representations. One common practice relies on analysis of internal representations via downstream tasks (Baan et al., 2019; Ettinger et al., 2018; Conneau et al., 2019; Nandakumar et al., 2019; McCoy et al., 2019). One line of work analyzes word interactions in neural networks' internal gates as the composition signal (Saphra and Lopez, 2020; Murdoch et al., 2018), extending the Contextual Decomposition algorithm proposed by Jumelet et al. (2019). Another notable branch of work constructs synthetic datasets of small size to investigate compositionality in neural networks (Liška et al., 2018; Hupkes et al., 2018; Baan et al., 2019). Some work controls for word content, as we do, to study composition at the sentence level (Ettinger et al., 2018; Dasgupta et al., 2018). We complement this work with a targeted and systematic study of phrase-level representations in transformers, with a focus on teasing apart lexical properties versus reflections of accurate compositional phrase meaning.

Our work relates closely to classic work on two-word phrases, which have used methods like landmark tests (Kintsch, 2001; Mitchell and Lapata, 2008, 2010), or compared against distribution-based phrase representations (Baroni and Zamparelli, 2010; Fyshe et al., 2015). Our work also draws on work using correlation with similarity judgments (Finkelstein et al., 2001; Gerz et al., 2016; Hill et al., 2015; Conneau and Kiela, 2018) and paraphrase classification (Ganitkevitch et al., 2013; Wang et al., 2018; Zhang et al., 2019; Yang et al., 2019a) to assess quality of models and representations. We build on this work by combining these methods together, applying them to a systematic analysis of transformers and their components, and introducing controlled variants of existing tasks to isolate accurate composition of phrase meaning from capturing of lexical information.

3 Testing phrase meaning similarity

Our methods begin with familiar approaches for assessing representations via meaning similarity: correlation with human phrase similarity judgments, and ability to identify paraphrases. The goal is to gauge the extent to which models arrive at representations reflecting the nuances of composed phrase meaning understood by humans. We draw on existing datasets, and begin by testing models on the original versions of these datasets—then we tease apart effects of word content from effects of more sophisticated meaning composition by introducing controlled variants of the datasets. The reasoning is that strong correlations with human similarity judgments, or strong paraphrase classification performance, could be influenced by artifacts that are not reflective of accurate phrase meaning composition per se. In particular, we may see strong performance simply on the basis of the amount of overlap in word content between phrases. To address this possibility, we create controlled datasets in which word overlap is no longer a cue to similarity.

As a starting point we focus on two-word phrases, as these are the smallest phrasal unit and the most conducive to these types of lexical con-

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

Normal Examples	
Source Phrase	Target Phrase & Score
average person	ordinary citizen (0.724)
	person average (0.518)
	country (0.255)
AB-BA Examples	
Source Phrase	Target Phrase & Score
law school	school law (0.382)
adult female	female adult (0.812)
arms control	control arms (0.473)

Table 1: Examples of correlation items. Numbers in parentheses are similarity scores between target phrase and source phrase. Upper half shows normal examples, and lower half shows controlled items.

trols, and because this allows us to leverage larger amounts of annotated phrase similarity data.

3.1 Phrase similarity correlation

We first evaluate phrase representations by assessing their alignment with human judgments of phrase meaning similarity. For testing this correspondence, we use the **BiRD** (Asaadi et al., 2019) dataset. BiRD is a bigram relatedness dataset designed to evaluate composition, consisting of 3,345 bigram pairs (examples in Table 1), with source phrases paired with numerous target phrases, and human-rated similarity scores ranging from 0 to 1.

In addition to testing on the full dataset, we design a controlled experiment to remove effects of word overlap, by filtering the dataset to pairs in which the two phrases consist of the same words. We refer to these pairs as "AB-BA" pairs (following terminology of the authors of the BiRD dataset), and show examples in the lower half of Table 1.

We run similarity tests as follows: given a model M with layers L, for ith layer $l_i \in L$ and a source-target phrase pair, we compute representations of source phrase $p^i_{rep}(\text{src})$ and target phrase $p^i_{rep}(\text{trg})$, where rep is a representation type from Section 4, and we compute their cosine $\cos(p^i_{rep}(\text{src}), p^i_{rep}(\text{trg}))$. Pearson correlation $\boldsymbol{r_i}$ of layer l_i is then computed between cosine and human-rated score for all source-target pairs.

3.2 Paraphrase classification

We further investigate the nature of phrase representations by testing their capacity to support binary paraphrase classification. This test allows us to explore whether we will see better alignment with human judgments of meaning similarity if we use more complicated operations than cosine similarity comparison. For the classification tasks, we draw on **PPDB 2.0** (Pavlick et al., 2015), a widely-used database consisting of paraphrases with scores generated by a regression model. To formulate our binary classification task, after filtering out low-quality paraphrases (discussed in Section 5), we use phrase pairs (source phrase, target phrase) from PPDB as positive pairs, and randomly sample phrases from the complete PPDB dataset to form negative pairs (source phrase, random phrase).

Because word overlap is also a likely cue for paraphrase classification, we filter to a controlled version of this dataset as well, as illustrated in Table 2. We formulate the controlled experiment here as holding word overlap between source phrase and target phrase to be exactly 50% for both positive and negative samples. Our choice of 50% word overlap in this case is necessary for construction of a sufficiently large, balanced classification dataset (AB-BA pairs in PPDB are too few to support classifier training, and AB-BA pairs are more likely to be non-paraphrases). Note, however, that by controlling word overlap to be exactly 50% for all phrase pairs, we still hold constant the *amount* of word overlap between phrases, which is the cue that we wish to remove. As an additional control, each source phrase is paired with an equal number of paraphrases and non-paraphrases, to avoid the classifier inferring labels based on phrase identity.

Formally, for each model layer l_i and representation type rep, we train

$$\text{CLF}^i_{rep} = \text{MLP}([\boldsymbol{pair^i_{rep}}])$$

where $\boldsymbol{pair^i_{rep}}$ represents embedding concatenations of each source phrase and target phrase:

$$\boldsymbol{pair^i_{rep}} = [\boldsymbol{p^i_{rep}(src)}; \boldsymbol{p^i_{rep}(trg)}]$$

The classifier is trained on binary classification of whether concatenated inputs represent paraphrases.

4 Representation types

A variety of approaches have been taken for representing sentences and phrases when all tokens output contextualized representations, as in our tested transformers. To clarify the phrasal information present in different forms of phrase representation, we experiment with a number of different combinations of token embeddings as representation types.

Formally, let $[T_0, \cdots, T_k]$ be an input sequence of length $k+1$, with corresponding embeddings

Normal Examples	
Source Phrase	Target Phrase
are crucial	is absolutely vital (pos)
	was a matter of concern (neg)
	is an essential part (pos)
	are exacerbating (neg)
Controlled Examples	
Source Phrase	Target Phrase
communication infrastructure	telecommunications infrastructure (pos)
	data infrastructure (neg)

Table 2: Examples of classification items. Classification labels between target phrase and source phrase are in parentheses. Upper half shows normal examples, and lower half shows controlled items.

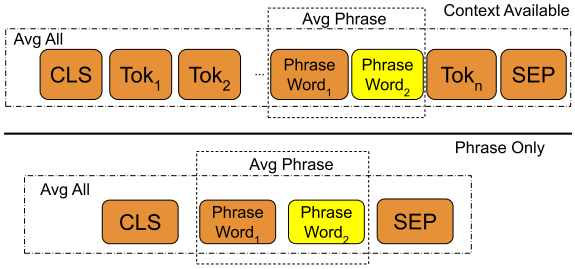

Figure 1: Example input sequences (BERT format). CLS is a special token at beginning of sequence. Tokens in yellow correspond to Head-Word. Avg-Phrase contains element-wise average of phrase word embeddings. Avg-All averages embeddings of all tokens.

at ith layer $[\boldsymbol{e_0^i}, \cdots, \boldsymbol{e_k^i}]$. Assume the phrase spans the sequence $[a, b]$, where $0 \leq a \leq b \leq k$. Because two-word phrases are atypical inputs for these models, we experiment both with inputs of the two-word phrases alone ("phrase-only"), as well as inputs with the phrases embedded in sentences ("context-available"). This is illustrated in Figure 1 along with phrase representation types.

We test the following forms of phrase representation, drawn from each model and layer separately:

CLS Depending on specific models, this special token can be the first or last token of the input sequence (i.e. $\boldsymbol{e_0^i}$ or $\boldsymbol{e_k^i}$). In many applications, this token is used to represent the full input sequence.

Head-Word In each phrase, the head word is the semantic center the phrase. For instance, in the phrase "public service", "service" is the head word, expressing the central meaning of the phrase, while "public" is a modifier. Because phrase heads are not annotated in our datasets, we approximate the head by taking the embedding of the final word of the phrase. This representation is proposed as a potential representation of the whole phrase, if information is being composed into a central word:

$$\boldsymbol{p_{hw}^i} = \boldsymbol{e_b^i}$$

Avg-Phrase For this representation type we average the embeddings of the tokens in the target phrase (dashed box in Figure 1). This type of averaging of token embeddings is a common means of aggregate representation (Wieting et al., 2015).

$$\boldsymbol{p_{ap}^i} = \frac{1}{b-a+1}\sum_{x=a}^{b}\boldsymbol{e_x^i}$$

Avg-All Expanding beyond the tokens in "Avg-Phrase", this representation averages embeddings from the full input sequence.

$$\boldsymbol{p_{aa}^i} = \frac{1}{k+1}\sum_{x=0}^{k}\boldsymbol{e_x^i}$$

SEP With some variation between models, the SEP token is typically a separator for distinguishing input sentences, and is often the last token ($\boldsymbol{e_k^i}$) or second to last token ($\boldsymbol{e_{k-1}^i}$) of a sequence.

5 Experimental setup

Embeddings of each token are obtained by feeding input sequences through pre-trained contextual encoders. We investigate the "base" version of five transformers: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019b), DistilBERT (Sanh et al., 2019), XLNet (Yang et al., 2019b) and XLM-RoBERTa (Conneau et al., 2019). For the models analyzed in this paper, we are using the implementation of Wolf et al. (2019),[2] which is based on PyTorch (Paszke et al., 2019).

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

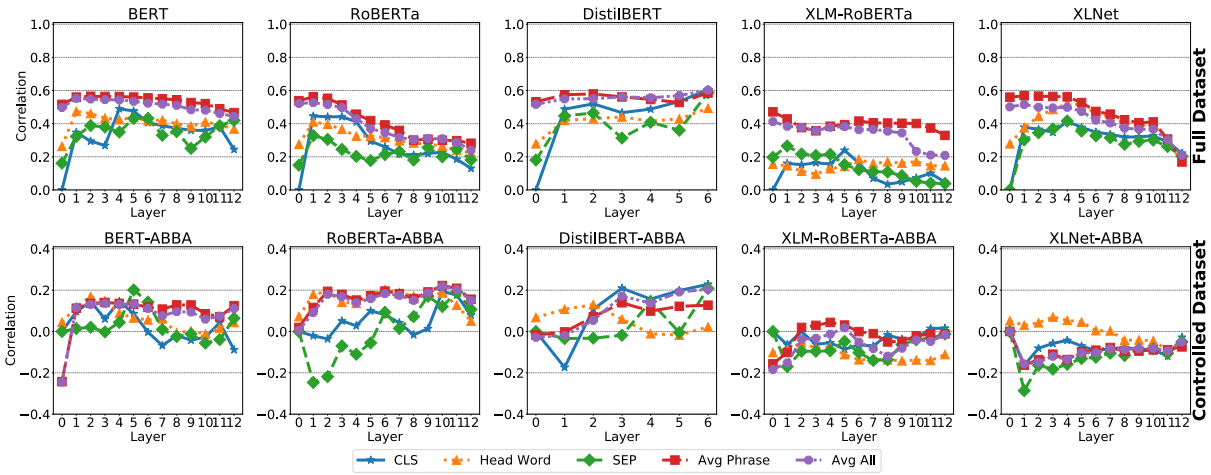

Figure 2: Correlation on BiRD dataset, phrase-only input setting. First row shows results on full dataset, and second row on controlled AB-BA pairs. Layer 0 corresponds to input embeddings passing to the model.

For correlation analysis, we first use the complete BiRD dataset, consisting of 3,345 phrase pairs.[3] We then test with our controlled subset of the data, consisting of 410 AB-BA pairs. For classification tasks, we first do preprocessing on PPDB 2.0,[4] filtering out pairs containing hyperlinks, non-alphabetical symbols, and trivial paraphrases based on abbreviation or tense change. For our initial classification test, we use 13,050 source-target phrase pairs (of varying word overlap) from this preprocessed dataset. We then test with our controlled dataset, consisting of 11,770 source-target phrase pairs (each with precisely 50% word overlap). For each paraphrase classification task, 25% of selected data is reserved for testing. We use a multi-layer perceptron classifier with a single hidden layer of size 256 with ReLU activation, and a softmax layer to generate binary labels. We use a relatively simple classifier following the reasoning of Adi et al. (2016), that this allows examination of how easily extractable information is in these representations.

For both correlation and classification tasks, we experiment with phrase-only inputs and context-available (full-sentence) inputs. To obtain sentence contexts, we search for instances of source phrases in a Wikipedia dump, and extract sentences containing them. For a given phrase pair, target phrases are embedded in the same sentence context as the source phrase, to avoid effects of varying sentence position between phrases of a given pair. [5]

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

[4] http://paraphrase.org

[5] Because context sentences are extracted based on source phrases, our use of the same context for source and target phrases can give rise to unnatural contextual fit for target phrases. We consider this acceptable for the sake of controlling sentence position—and if anything, differences in contextual fit may aid models in distinguishing more and less similar phrases. The slight boost observed on the full datasets (for Avg-Phrase) suggests that the sentence contexts do provide the intended benefit from using input of a more natural size.

6 Results

6.1 Similarity correlation

Full dataset The top row of Figure 2 shows correlation results on the full BiRD dataset for all models, layers, and representation types, with phrase-only inputs. Among representation types, Avg-Phrase and Avg-All consistently achieve the highest correlations across models and layers. In all models but DistilBERT, correlation of Avg-Phrase and Avg-All peaks at layer 1 and decreases in subsequent layers with minor fluctuations. Head-Word and SEP both show weaker, but non-trivial, correlations. The CLS token is of note with a consistent rapid rise as layers progress, suggesting that it quickly takes on properties of the words of the phrase. For all models but DistilBERT, CLS token correlations peak in middle layers and then decline.

Model-wise, XLM-RoBERTa shows the weakest overall correlations, potentially due to the fact that it is trained to infer input language and to handle multiple languages. BERT retains fairly consistent correlations across layers, while RoBERTa and XLNet show rapid declines as layers progress, suggesting that these models increasingly incorporate information that deviates from human intuitions about phrase smilarity. DistilBERT, despite being of smaller size, demonstrates competitive correlation. The CLS token in DistilBERT is notable for its continuing rise in correlation strength across

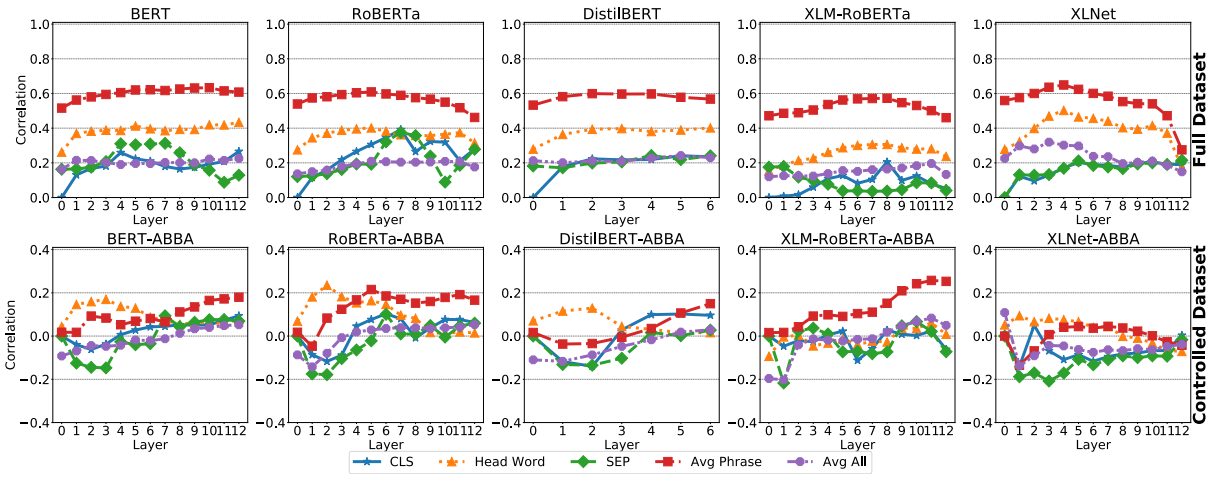

Figure 3: Correlation on BiRD dataset with phrases embedded in sentence context (context-available input setting).

layers. This suggests that DistilBERT in particular makes use of the CLS token to encode phrase information, and unlike other models, its representations retain the relevant properties to the final layer.

Controlled dataset Turning to our controlled AB-BA dataset, we examine the extent to which the above correlations indicate sophisticated phrasal composition versus effective encoding of information about phrases' component words. The bottom row of Figure 2 shows the correlations on this controlled subset. We see that performance of all models drops significantly, often with roughly zero correlation. Avg-All and Avg-Phrase no longer dominate the correlations, suggesting that these representations capture word information, but not higher-level compositional information. XLM-RoBERTa and XLNet show particularly low correlations, suggesting heavier reliance on word content. Notably, the CLS tokens in RoBERTa and DistilBERT stand out with comparatively strong correlations in later layers. This suggests that the rise that we see in CLS correlations for DistilBERT in particular may correspond to some real compositional signal in this token, and for this model the CLS token may in fact correspond to something more like a representation of the meaning of the full input sequence. The Avg-Phrase representation for RoBERTa also makes a comparatively strong showing.

Including sentence context Figure 3 shows the correlations when target phrases are embedded as part of a sentence context, rather than in isolation. As can be expected, Avg-Phrase is now consistently the highest in correlation on the full dataset—other tokens are presumably more impacted by the presence of additional words in the context. We also see that the Avg-Phrase correlations no longer drop so dramatically in later layers, suggesting that when given full sentence inputs, models retain more word properties in later layers than when given only phrases. This general trend holds also for Avg-All and Head-Word representations.

In the AB-BA setting, we see that presence of context does boost overall correlation with human judgment. Of note is XLM-RoBERTa's Avg-Phrase, which without sentence context has zero correlation in the AB-BA setting, but which with sentence context reaches our highest observed AB-BA correlations in its final layers. However, even with context, the strongest correlation across models is still less than 0.3. It is still the case, then, that correlation on the controlled data degrades significantly relative to the full dataset. This indicates that even when phrases are input within sentence contexts, phrase representations in transformers reflect heavy reliance on word content, largely missing additional nuances of compositional phrase meaning.

6.2 Paraphrase classification

Full dataset Results for our full paraphrase classification dataset, with phrase-only inputs, are shown in the top row of Figure 4. Accuracies are overall very high, and we see generally similar patterns to the correlation tasks. Best accuracy is achieved by using Avg-Phrase and Avg-All representations. RoBERTa, XLM-RoBERTa, and XLNet show decreasing correlations for top-performing representations in later layers, while BERT and DistilBERT remain more consistent across layers. Performance of CLS requires a few

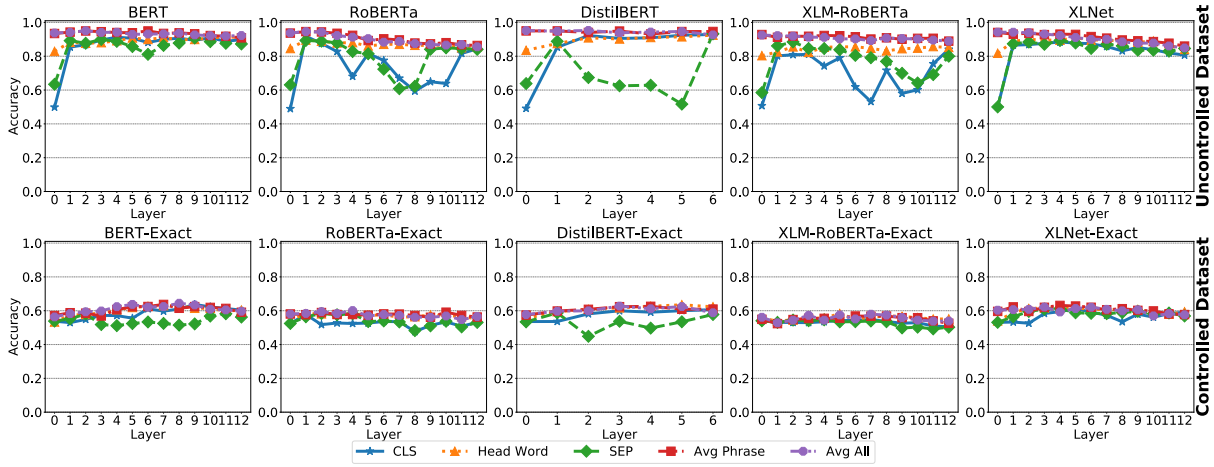

Figure 4: Classification accuracy on PPDB dataset (phrase-only input setting). First row shows classification accuracy on original dataset, and second row shows accuracy on controlled dataset.

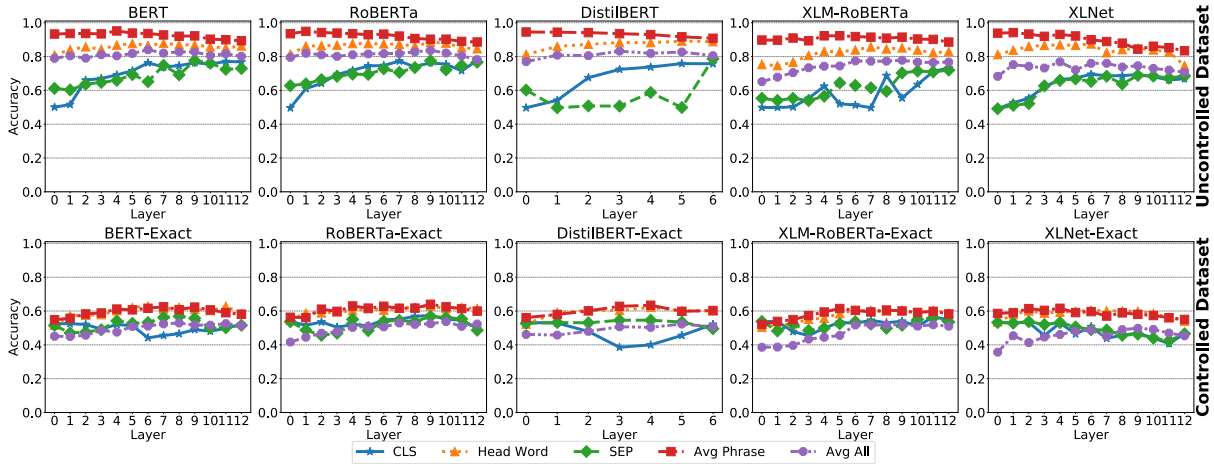

Figure 5: Classification accuracy on PPDB dataset with phrases embedded in sentence context. First row shows classification accuracy on original dataset, and second row shows accuracy on controlled dataset.

layers to peak, with top performance around middle layers, and in some models shows poor performance in later layers. SEP shows unstable performance compared to other representations, especially in DistilBERT and RoBERTa.

Controlled dataset The bottom row of Figure 4 shows classification accuracy when word overlap is held constant. Consistent with the drop in correlations on the controlled AB-BA experiments above, classification performance of all models drops down to only slightly above chance performance of 50%. This suggests that the high classification performance on the full dataset relies largely on word overlap information, and that there is little higher-level phrase meaning information to aid classification in the absence of the overlap cue. We see in some cases a very slight trend such that classification accuracy increases a bit toward middle layers—so to the extent that there is any compositional phrase information being captured, it may increase within representations in the middle layers. Overall, the consistency of these results with those of the correlation analysis suggests that the apparent lack of accurate compositional meaning information in our tested phrase representations is not simply a result of cosine correlations being inappropriate for picking up on correspondences.

Including sentence context Figure 5 shows the classification results for representations of phrases embedded in sentence contexts. The patterns largely align with our observations from the correlation task. Performance on the full dataset is still high, with Avg-Phrase now showing consistently highest performance, being least influenced by the presence of new context words. In the controlled setting, we see the same substantial drop in per-

	horse ran	*color ran*
gallop	POS	NEG
dissolve	NEG	POS

Table 3: An example of landmark experiment of verb "run". Representations are expected to have higher cosine similarities between phrase and landmark word that are marked "POS".

formance relative to the full dataset—there is very slight improvement over the phrase-only representations, but the highest accuracy among all models is still around 0.6. Thus, the inclusion of sentence context again does not provide any additional evidence for sophisticated compositional meaning information in the tested phrase representations.

7 Qualitative analysis: sense disambiguation

The above analyses rely on testing models' sensitivity to meaning similarity between two phrases. In this section we complement these analyses with another test aimed at assessing phrasal composition: testing models' ability to select the correct senses of polysemous words in a composed phrase, as proposed by Kintsch (2001). Each test item consists of a) a central verb, b) two subject-verb phrases that pick out different senses of the verb, and c) two *landmark* words, each associating with one of the target senses of the verb. Table 3 shows an example with central verb "ran" and phrases "horse ran"/ "color ran". The corresponding landmark words are "gallop", which associates with "horse ran", and "dissolve", which associates with "color ran". The reasoning is that composition should select the correct verb meaning, shifting representations of the central verbs—and of the phrase as a whole—toward landmarks with closer meaning. For this example, models should produce phrase embeddings such that "horse ran" is closer to "gallop" and "color ran" is closer to "dissolve". We use the items introduced in Kintsch (2001), which consist of a total of 4 sets of landmark tests. We feed landmarks and phrases respectively through each transformer, without context, to generate corresponding representations p^i_{rep} for each layer l_i and representation type rep. Cosine similarity between each phrase-landmark pair is computed and compared against expected similarities.

Figure 6 shows the percentage of phrases that fall closer to the correct landmark word than to the incorrect one, averaged over 16 phrase-landmark word pairs. We see strong overall performance across models, suggesting that the information needed for this task is successfully captured by these models' representations. Additionally, we see that the patterns largely mirror the results above for correlation and classification on uncontrolled datasets. Particularly, Avg-Phrase and Avg-All show comparatively strong performance across models. RoBERTa and XLNet show stronger performance in early layers, dropping off in later layers, while BERT and DistilBERT show more consistency across layers. XLM-RoBERTa and XLNet show lower performance overall.

For this verb sense disambiguation analysis, the Head-Word token is of note because it corresponds to the central verb of interest, so its sense can only be distinguished by its combination with the other word of the phrase. XLM-RoBERTa has the weakest performance with Head-Word, while BERT and DistilBERT demonstrate strong disambiguation with this token. As for the CLS token, RoBERTa produces the highest quality representation at layer 1, and BERT outperforms other models starting from layer 6, with DistilBERT also showing strong performance across layers.

Notably, the observed parallels to our correlation and classification results are in alignment with the uncontrolled rather than the controlled versions of those tests. So while these parallels lend further credence to the general observations that we make about phrase representation patterns across models, layers, and representation types, it is worth noting that these landmark composition tests may be susceptible to lexical effects similar to those controlled for above. Since these test items are too few to filter with the above methods, we leave in-depth investigation of this question to future work.

8 Discussion

The analyses reported above yield two primary takeaways. First, they shed light on the nature of these models' phrase representations, and the extent to which they reflect word content versus phrasal composition. At many points in these models there is non-trivial alignment with human judgments of phrase similarity, paraphrase classification, and verb sense selection. However, when we control our correlation and classification tests to remove the cue of word overlap, we see little evidence that the representations reflect sophisticated

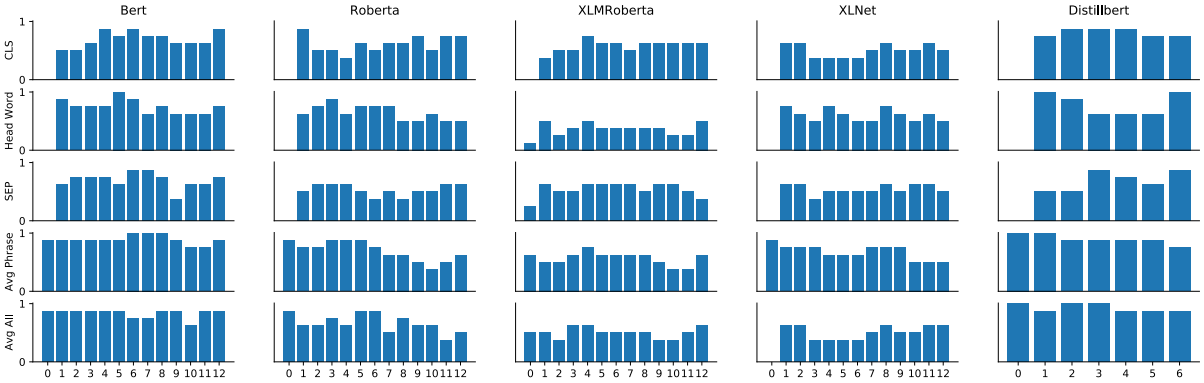

Figure 6: Landmark experiments. Y-axis denotes the percentage of samples that are shifted towards the correct landmark words in each layer. Missing bars occur when representations are independent of input at layer 0, such that cosine similarity between phrases and landmarks will always be 1.

phrase composition beyond what can be gleaned from word content. While we see strong performance on classic sense selection items designed to test phrase composition, the observed results largely parallel those from the uncontrolled versions of the correlation and classification analyses, suggesting that success on this landmark test may reflect lexical properties more than sophisticated composition. Given the importance of systematic meaning composition for robust and flexible language understanding, based on these results we predict that we will see corresponding weaknesses as more tests emerge for these models' handling of subtle meaning differences in downstream tasks.

Our systematic examination of models, layers and representation types yields a second takeaway in the form of practical implications for selecting and extracting representations from these models. For faithful representations of word content, Avg-Phrase is generally the strongest candidate. If only the phrase is embedded, drawing from earlier layers is best in RoBERTa, XLM-RoBERTa, and XL-Net, while middle layers are better in BERT, and later layers in DistilBERT. If the phrase is input as part of a sentence, middle layers are generally best across models. Though the CLS token is often interpreted to represent a full input sequence, we find it to be a poor phrase representation even with phrase-only input, with the notable exception of the final layer of DistilBERT.

As for representations that reflect true phrase meaning composition, we have established that such representations may not currently be available in these models. However, to the extent that we do see weak evidence of potential compositional meaning sensitivity, this appears to be strongest in DistilBERT's CLS token in final layers, in RoBERTa's Avg-Phrase representation in later layers, and in XLM-RoBERTa's Avg-Phrase representation from later layers *only* when the phrase is contained within a sentence context.

9 Conclusions and future directions

We have systematically investigated the nature of phrase representations in state-of-the-art transformers. Teasing apart sensitivity to word content versus phrase meaning composition, we find strong sensitivity across models when it comes to word content encoding, but little evidence of sophisticated phrase composition. The observed sensitivity patterns across models, layers, and representation types shed light on practical considerations for extracting phrase representations from these models.

Future work can apply these tests to a broader range of models, and continue to develop controlled tests that target encoding of complex compositional meanings, both for two-word phrases and for larger meaning units. We hope that our findings will stimulate further work on leveraging the power of these generalized transformers while improving their capacity to capture compositional meaning.

Acknowledgments

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

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*.

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*.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193. Association for Computational Linguistics.

Shammur Absar Chowdhury and Roberto Zamparelli. 2018. RNN simulations of grammaticality judgments on long-distance dependencies. In *Proceedings of the 27th international conference on computational linguistics*, pages 133–144.

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.

Ishita Dasgupta, Demi Guo, Andreas Stuhlmüller, Samuel J Gershman, and Noah D Goodman. 2018. Evaluating compositionality in sentence embeddings. *arXiv preprint arXiv:1802.04302*.

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.

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A Smith. 2016. Recurrent neural network grammars. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 199–209.

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.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. 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 32–42.

Alona Fyshe, Leila Wehbe, Partha Talukdar, Brian Murphy, and Tom Mitchell. 2015. A compositional and interpretable semantic space. In *Proceedings of the 2015 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 32–41.

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.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

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*.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657.

Jaap Jumelet, Willem Zuidema, and Dieuwke Hupkes. 2019. Analysing neural language models: Contextual decomposition reveals default reasoning in number and gender assignment. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 1–11.

Yoon Kim, Alexander M Rush, Lei Yu, Adhiguna Kuncoro, Chris Dyer, and Gábor Melis. 2019. Unsupervised recurrent neural network grammars. 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 1105–1117.

Walter Kintsch. 2001. Predication. *Cognitive science*, 25(2):173–202.

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*.

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

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

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.

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

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.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *proceedings of ACL-08: HLT*, pages 236–244.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429.

W James Murdoch, Peter J Liu, and Bin Yu. 2018. Beyond word importance: Contextual decomposition to extract interactions from lstms. In *International Conference on Learning Representations*.

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.

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.

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.

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*.

Naomi Saphra and Adam Lopez. 2020. Word interdependence exposes how lstms compose representations. *arXiv preprint arXiv:2004.13195*.

Vered Shwartz and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

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, 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.

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

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 *Advances in neural information processing systems*, pages 5998–6008.

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.

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.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Towards universal paraphrastic sentence embeddings. *arXiv preprint arXiv:1511.08198*.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221.

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.

Da Yin, Tao Meng, and Kai-Wei Chang. 2020. Sentibert: A transferable transformer-based architecture for compositional sentiment semantics.

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