Is Semantic-aware BERT more Linguistically Aware? A Case Study on Natural Language Inference

Ling Liu[*]
University of Colorado
ling.liu@colorado.edu

Ishan Jindal
IBM Research
ishan.jindal@ibm.com

Yunyao Li[†]
Apple
yunyaoli@apple.com

Abstract

Recent work has shown that predicate-argument label representations from semantic role labeling (SRL) can be concatenated with BERT representations to improve natural language understanding tasks such as natural language inference (NLI) and reading comprehension. Two natural questions that arise are whether infusing SRL representations with BERT 1) improves model performance and 2) increases the model's linguistic awareness. This paper aims at answering both questions with a case study on the NLI task. We start by analyzing whether and how infusing SRL information helps BERT learn linguistic knowledge. We compare model performance on two benchmark datasets, SNLI and MNLI. We also conduct in-depth analysis on two probing datasets, Breaking NLI and HANS, which contain abundant examples where SRL information is expected to be helpful. We found that combining SRL representations with BERT representations does outperform BERT-only representations in general, with better awareness in lexical meaning and world knowledge but not in logic knowledge. We also found that infusing SRL information via predicate-wise concatenation with BERT word representations followed by an interaction layer is more effective than sentence-wise concatenation.[1]

1 Introduction

Recent advances in Transformer-based language models like BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019b), have surpassed the performance of non-expert humans when fine-tuned and evaluated on various benchmark datasets such as GLUE (Wang et al., 2018), SuperGLUE (Wang et al., 2019a). However, performing well on benchmark datasets does not mean the model has achieved human-like natural language understanding (NLU) competence (Naik et al., 2018; Talman and Chatzikyriakidis, 2019). It has been found that the transformer models when fine-tuned for a particular tasks are prone to adopt shallow heuristics (McCoy et al., 2019). For instance, in a natural language inference (NLI) system — a task of determining whether a premise sentence entails a hypothesis sentence (Dagan et al., 2005; MacCartney and Manning, 2009), a model is more likely to assign a label "contradiction" if the word *not* is present in the hypothesis and assign a label "entailment" if all words in a hypotheses are present in the premise (Naik et al., 2018).

Consider the following sentence pair where hypothesis contradicts premise:

> P: The lawyers were recommended by the doctor
> H: The lawyers recommended the doctor

A RoBERTa model fine-tuned on MNLI dataset, benchmark NLI dataset (Williams et al., 2018), completely ignores the semantic meaning of these sentences and predicts *entailment*[2]. This wrong prediction indeed happened because the model learns to generalize over the simple heuristic present in the training datasets that is predicting entailment when there is 100% lexical overlap between premise and hypothesis. At this point, an important question arises: Does providing semantic information to the language models in the form of external knowledge enhance models' linguistic knowledge, if yes, how?

Semantic role labeling (SRL) is a shallow semantic parsing task that identifies "*who did what to whom when*, *where* etc" for each predicate in a sentence, called *predicate-argument structure*. SRL represents the semantic meaning of a sentence and

[*] Work done during an internship at IBM research

[†] Work done while at IBM Research

[1] Our data and code can be found at https://github.com/System-T/LingBert.

[2] https://demo.allennlp.org/textual-entailment/roberta-mnli/s/the-lawyers-were-recommended-doctor/P8E4F0W1U9

this information is essential to natural language understanding (Palmer et al., 2005; He et al., 2017; Kasai et al., 2019; Jindal et al., 2020; Marcheggiani and Titov, 2020). For the sentence pair above, The semantic roles associated with the verb *recommend* are as follows:

> P: [ARG1 : The lawyers] were [V : recommended] by [ARG0 : the doctor]
> H: [ARG0 : The lawyers] [V : recommended] [ARG1 : the doctor]

This semantic information clearly distinguishes the *Recommender ARG0* in the premise and the hypothesis: the *doctor* is the *Recommender* in the premise and *lawyer* is the *Recommender* in the hypothesis. In this example, both the premise and the hypothesis have identical predicate-argument structures with the same predicate but swapped arguments. We expect that infusing SRL information would be helpful in such cases as it highlights the difference in the arguments, and thus the model is directed to the lexical meaning difference encoded by the tokens to correctly classify the example.

Recently, Zhang et al. (2019b) propose SemBERT to incorporate representations of semantic role labels (SRLs) into the BERT representations. In SemBERT, both representations were processed independently and get concatenated for fine-tuning. This concatenation alone shows slight improvement over BERT on NLU benchmark datasets but fails to improve the model's linguistic awareness as shown by its poor performance on probing datasets. In this work, we propose a simple yet more effective improvement over SemBERT to improve the model's awareness of various language phenomena such as lexical knowledge, world knowledge. Specifically, our main contributions include:

- We present detailed analysis of SRL-infused models on two probing datasets, Breaking NLI (Glockner et al., 2018) and HANS (McCoy et al., 2019) to understand what cases benefit or suffer from the infused SRL information.
- We propose ***Ling***uistic ***BERT*** (LingBERT) that concatenates predicate-wise SRL representations with BERT representations followed by an interaction layer. We demonstrate its effectiveness over existing baselines on two commonly used NLI benchmark datasets, SNLI and MNLI.
- We found that LingBERT does learn lexical and world knowledge better. However, how to leverage SRL information to improve the model's awareness of other linguistic phenomena remains an open challenge.

2 Motivation

Semantic role labeling in the PropBank format labels syntactic constituents for each predicate in the sentence with the corresponding generalized proto-thematic roles such as ARG0 , ARG1 etc. Therefore, though shallow, SRL provides explicit syntactic and semantic knowledge.

Existing studies (Naik et al., 2018; Talman and Chatzikyriakidis, 2019) point out that though large-size pre-trained language models can achieve good performance on classification tasks, such models are still far from a real understanding of language. Does incorporating SRL information explicitly at the fine-tuning stage increase the model's linguistic awareness and lead to performance improvement? This paper aims at answering this question by comparing the performance of BERT with models incorporating semantic role labels into BERT on two benchmark NLI datasets, SNLI and MNLI. We further analyze whether incorporating the additional SRL information increases the models' linguistic awareness by analyzing model performance on two probing datasets, Breaking NLI and HANS, which contain abundant cases where SRL information is expected to be helpful.

2.1 Issues with NLI Benchmark Datasets

Two commonly-used NLI benchmark datasets, SNLI[3] (Bowman et al., 2015) and MNLI[4] (Williams et al., 2018) are created via crowdsourcing. The models fine-tuned on these datasets perform well on their corresponding test sets. However, such models tend to adopt simple heuristics from the training data that are effective only for most frequent example types (McCoy et al., 2019). In other words, a model fine-tuned on SNLI training set may perform excellent (91.0% accuracy) on the corresponding test set but perform poorly (46.3% accuracy, in Table 5) when tested on a specific heuristics. **Therefore, it is imperative to test whether the fine-tuned models generalize over certain heuristics.**

[3] https://nlp.stanford.edu/projects/snli/

[4] https://cims.nyu.edu/~sbowman/multinli/

P	The	family	is	drinking	wine
	O	O	**B-V**	O	O
	B-ARG0	I-ARG0	O	**B-V**	B-ARG1
H_1	The	family	is	drinking	vodka
	O	O	**B-V**	O	O
	B-ARG0	I-ARG0	O	**B-V**	B-ARG1
H_2	The	family	is	drinking	gin
	O	O	**B-V**	O	O
	B-ARG0	I-ARG0	O	**B-V**	B-ARG1

Table 1: World knowledge heuristic examples from Breaking NLI dataset. SRLs (in blue) are expected to be helpful for such examples. The ground-truth NLI label for the premise and each hypothesis is *contradiction*. However, BERT predicts *neutral* for P-H_1, and *contradiction* P-H_2. Our proposed method classifies both cases correctly.

2.2 Probing Information

Various probing datasets have been developed (Glockner et al., 2018; McCoy et al., 2019) to analyze whether the trained model has learned the linguistic knowledge and reasoning generalization that humans resort to for the same task. Though the model has learned some linguistic features (Goldberg, 2019; Liu et al., 2019a; Hewitt and Manning, 2019; Clark et al., 2019; Lin et al., 2019; Tenney et al., 2019; Warstadt and Bowman, 2020; Manning et al., 2020; Ettinger, 2020; Michael et al., 2020), the poor performance of fined-tuned BERT models on the probing datasets indicates that it is not enough linguistic knowledge. We evaluate the fine-tuned models on two probing datasets: Breaking NLI[5] (Glockner et al., 2018) and HANS[6] (McCoy et al., 2019). Though not directly created to probe about SRL information, the two datasets contain abundant examples on different heuristics for which SRL information is expected to be helpful.

2.2.1 Lexical Meaning and World Knowledge

Breaking NLI provides examples to test whether the trained model has learned proper lexical meaning and world knowledge. It contains instances where the premise and hypothesis are identical except for one token as shown in Table 1. Infusing SRL information is expected to be helpful for this dataset because the extra predicate-argument information enforces the BERT model to focus more on distinctions encoded in the lexical meaning.

2.2.2 Lexical Overlap Heuristic

Lexical overlap is the first type of heuristic which the HANS dataset intends to probe about. It contains examples where hypotheses are constructed from all the words in premises. Both Breaking NLI and HANS provide examples to test this heuristics. Table 2 lists examples from the *lexical overlap* heuristics in Row 1. As can be seen, both the premise and the hypothesis have identical predicate-argument structures, where both have the same predicate with only the arguments swapped. We would expect infusing SRLs to be helpful in this case because it highlights the similarity in the predicate-argument structure of the text and thus the model is directed to the lexical meaning difference encoded by the tokens in order to correctly classify the example.

2.2.3 Subsequence Heuristic

Subsequence heuristic is the second type of heuristic HANS intends to probe about. In the *subsequence* heuristic example (Table 2 Row 2), the predicate-argument structures of the premise and the hypothesis are quite different. Here the premise has two predicates whereas the hypothesis has only one. This kind of tricky cases where the hypothesis is a sub-string of the premise is relatively hard for NLI models to learn. We also expect the predicate-argument structure to be helpful here because it highlights the syntactic difference of identical tokens: *the actors* is ARG1 for the predicate *believe* in the hypothesis while it is only a component of ARG1 for the same predicate in the premise.

2.2.4 Constituent Heuristic

Constituent heuristic is the third type of heuristic the HANS dataset intends to probe about. It contains examples where the premise entails all complete subtrees of its parse tree as shown in Table 2 Row 3. In comparison to other heuristics, the generalization over *constituent* heuristic examples requires more logical information rather than just predicate-argument structures. For the *PP_on_subject* sub-case of *constituent*, span-base SRL actually introduces more misleading information. Therefore, we do not expect the SRL information to be helpful for this case.

3 How to Incorporate Semantic Knowledge?

In last section we describe various heuristics where vanilla transformer model does not generalize well

[5] https://github.com/BIU-NLP/Breaking_NLI

[6] https://github.com/tommccoy1/hans

Heuristic		*non-entailment* Example								
Lexical overlap	P:	The	judge	encouraged	the	athlete	.			
		B-ARG0	I-ARG0	**B-V**	B-ARG1	I-ARG1	O			
	H:	The	athlete	encouraged	the	judge	.			
		B-ARG0	I-ARG0	**B-V**	B-ARG1	I-ARG1	O			
Subsequence	P:	The	scientists	believed	the	actors	saw	the	artists	.
		B-ARG0	I-ARG0	**B-V**	B-ARG1	I-ARG1	I-ARG1	I-ARG1	I-ARG1	O
		O	O	O	B-ARG0	I-ARG0	**B-V**	B-ARG1	I-ARG1	O
	H:	The	scientists	believed	the	actors	.			
		B-ARG0	I-ARG0	**B-V**	B-ARG1	I-ARG1	O			
Constituent	P:	Hopefully	the	president	introduced	the	doctors	.		
		B-ARGM-ADV	B-ARG0	I-ARG0	**B-V**	B-ARG1	I-ARG1	O		
	H:	The	president	introduced	the	doctors	.			
		B-ARG0	I-ARG0	**B-V**	B-ARG1	I-ARG1	O			

Table 2: *non-entailment* examples from HANS with SRLs (in blue) tags.

when fine-tuned on NLI datasets. In this section, we describe the limitations of existing semantic aware transformer models and describe the strategy to further improve their performance.

We follow Devlin et al. (2019) to formulate textual entailment as a classification problem and design the input to the BERT model as [CLS] P_i [SEP] H_i [SEP], where $P_i = \{w_{pi}^1, w_{pi}^2, \cdots, w_{pi}^{n_p}\}$ is the premise and $H_i = \{w_{hi}^1, w_{hi}^2, \cdots, w_{hi}^{n_h}\}$ is the hypothesis. The final hidden layer vector representation of the first symbol in the sequence [CLS] is used as input to the classification layer at the fine-tuning stage.

Zhang et al. (2019b) propose SemBERT to concatenate SRL embeddings with BERT representations. In addition to vanilla BERT embeddings, they apply an out-of-box span-based semantic role labeling model to label the premise and hypothesis text with SRLs in PropBank format. If the premise or the hypothesis is associated with multiple predicate-argument structures, a linear layer is used to convert the multiple semantic role labels associated with each token into one SRL embedding representation. Then they concatenate the BERT representation with the SRL representation tokenwise, and use the first token [CLS] representation for classification. We refer to the SemBERT approach as **sentence-wise concatenation** of the semantic role label representation and the BERT token representation.

We find the way SemBERT incorporates SRL information needs improvement for two reasons: (1) Semantic roles are predicate-specific. Therefore, fusing all the semantic role labels associated with a token without considering the predicate at all introduces too much noise. (2) SemBERT lacks interaction between lexical meaning (represented by BERT representation) and SRL, which makes the contribution of SRL minor. Therefore, we propose ***Ling***uistic ***BERT***(LingBERT) that performs predicate-wise concatenation and interaction of SRL representation with BERT word representation. We refer to our approach as **predicate-wise concatenation with interaction**. Specifically, we concatenate SRL representations with BERT word representations per predicate-argument structure, and add a linear layer to allow interaction between the word representation and the SRL embeddings. Figure 1 provides an illustration of our model architecture.

We apply the same out-of-box semantic role labeling model as SemBERT to label the premise and the hypothesis.[7] We obtain multiple predicate-argument structures for both the premise and the hypothesis, equal to the number of predicates in respective text. For a sentence $S = \{w^1, w^2, \cdots, w^n\}$ containing K predicates, we obtain K predicate-argument structures $[\{l_1^1, l_1^2, \cdots, l_1^n\}, \cdots, \{l_K^1, l_K^2, \cdots, l_K^n\}]$, where l_k^j represents a semantic role label for token j specific to the kth predicate. We represent the SRLs as embeddings and use a lookup table to map the SRLs to vectors, which are learned during training. Therefore, for sentence S we obtain $[\{v_1^1, v_1^2, \cdots, v_1^n\}, \cdots, \{v_K^1, v_K^2, \cdots, v_K^n\}]$ as SRL representations.

For each token w^j in sentence S we obtain its fused embedding based on its BERT representation and the corresponding SRL representations as $(\{e^j \circ v_1^j\}, \{e^j \circ v_2^j\}, \cdots, \{e^j \circ v_K^j\})$, where $\circ$ represents concatenation. For completeness and simplicity, we assume a "O " label for [CLS] and [SEP] tokens. This fused embedding is passed through a linear layer to increase interaction among

[7] Note that the SRL model is used only as a data preprocessing tool and is not trained, same for SemBERT.

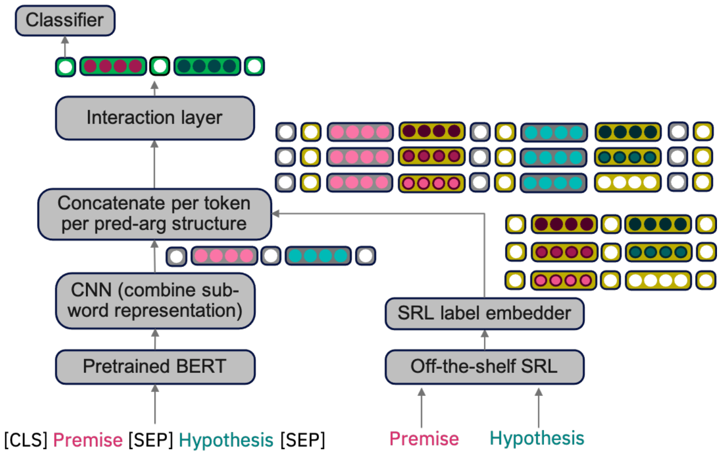
Classifier
Interaction layer
Concatenate per token per pred-arg structure
CNN (combine sub-word representation)
Pretrained BERT
SRL label embedder
Off-the-shelf SRL
[CLS] Premise [SEP] Hypothesis [SEP]
Premise
Hypothesis

Figure 1: LingBERT architecture

BERT and SRL representations, and the representation of the first input token `[CLS]` is then used as an input to the classifier.

Our model differs from SemBERT (Zhang et al., 2019b) in two aspects (see Figure 1for an illustration of our model). First, LingBERT concatenates the BERT representation for each token with the SRL representation for each predicate-argument structure respectively, rather than fuses SRLs with each token disregarding the lexical meaning of the predicate as SemBERT does. LingBERT is expected to be more effective than SemBERT because each predicate-argument structure is labeled with respect to the predicate (Palmer et al., 2005). Second, in SemBERT, the interaction between the BERT representation and the SRL embedding is only by concatenation. In contrast, LingBERT adds a linear layer after concatenating BERT representations and SRL representations to allow the interaction between them.

4 Experiments and Results

4.1 Model Setup

The pre-trained English *BERT-base-uncased* model is used in all experiments. For BERT fine-tuning, we used the Transformer (Vaswani et al., 2017) implementation by Wolf et al. (2019)[8]. For SemBERT, the implementation by Zhang et al. (2019b)[9] is used. LingBERT is implemented based on SemBERT. We adopt the same hyperparameters as SemBERT in all experiments. We trained each model multiple times and report the average accuracy. Following Zhang et al. (2019b), we obtain span-based [10] SRLs from AllenNLP (Gardner et al., 2018).

4.2 Overall Performance

As shown in Tables 3 and 4, both SRL-infused models, SemBERT and LingBERT, outperform BERT on SNLI test set and MNLI development sets. This result confirms that **infusing SRL representations with BERT representations does improve the model's performance on NLI.**

4.2.1 SNLI

LingBERT outperforms all semantic aware models including SemBERT on SNLI test set and on both the probing datasets, Breaking NLI and HANS (see Table 3). The evident advantage of LingBERT over SemBERT confirms that **predicate-wise concatenation with interaction is more effective than sentence-wise concatenation**.

4.2.2 MNLI

Similar to SNLI we observe that the semantic knowledge helps improve the model performance over simple BERT baselines when fine-tuned on MNLI dataset, in Table 4. We also observe a slight performance drop for LingBERT when compared with SemBERT on MNLI dev set, we explain this behaviour in Section 5.2.

4.3 Performance on Probing Datasets

4.3.1 Breaking NLI

Independent of training dataset, LingBERT outperforms BERT and SemBERT, achieving SoTA

[8] `https://github.com/huggingface/transformers`

[9] `https://github.com/cooelf/SemBERT`

[10] We obtain BIO tags for each token in the sentence. Throughout this work we use span-based SRLs.

Model	External knowledge	SNLI test	HANS	Breaking NLI
SNLI fine-tuned				
BERT$_{\text{Base}}$	-	90.30	58.83	93.84
(Pang et al., 2019)	SynParse	90.50	53.20	-
(Zhang et al., 2019a)	Semantic	89.60	-	-
(Kapanipathi et al., 2020)	KG	85.97	-	-
SemBERT$_{\text{Base}}$	Semantic	90.59*	57.89	93.16
LingBERT$_{\text{Base}}$	Semantic	**90.92**	**59.96**	**94.04**

Table 3: Performance summary and comparison with other models when **fine-tuned on *SNLI***. Best performance is highlighted in bold among SRL-infused models. Underline shows the best performance among all models. *Synparse* means syntactic parse, *KG* means knowledge graphs, and *Semantic* means Semantic Role Label information is used as external knowledge. *Though we use the same implementation and hyperparameters, we can not reproduce SemBERT reported performance.

Model	External knowledge	MNLI dev-m	MNLI dev-mm	HANS	Breaking NLI
MNLI fine-tuned					
BERT$_{\text{Base}}$	-	84.04	84.54	66.19	92.60
(Pang et al., 2019)	SynParse	84.70	84.7	-	-
(Cengiz and Yuret, 2020)	Semantic	-	-	**66.0***	-
SemBERT$_{\text{Base}}$	Semantic	**84.26**	**85.05**	53.44	92.97
LingBERT$_{\text{Base}}$	Semantic	84.21	84.70	58.58	**93.29**

Table 4: Performance summary and comparison with other models when **fine-tuned on *MNLI***. Best performance is highlighted in bold among SRL-infused models. Underline shows the best performance among all models. *MNLI dev-m* refers to the *matched* MNLI development set, i.e. data derived from the same sources as those in the MNLI training set; *MNLI dev-mm* refers to the *mismatched* MNLI development set, i.e. data derived from sources which are different from those in the MNLI training set. *Synparse* means syntactic parse, and *Semantic* means Semantic Role Label information is used as external knowledge. (* This paper uses HANS development set as validation set for MNLI training.)

accuracy 94.04% on Breaking NLI dataset. This indicates that LingBERT has learned lexical meaning and world knowledge better.

As mentioned in Section 2.2, the premise and hypothesis in this dataset have highly similar or identical predicate-argument structures. Table 1 provides examples where LingBERT consistently predicts correct labels while BERT performs inconsistently. Though in both examples only the objects of the verb differ, BERT model predicts *neutral* for the P-H$_1$ and predicts *contradiction* for the P-H$_2$. The key to the correct prediction is to know the lexical meaning difference between *wine* and *vodka*, and between *wine* and *gin*. LingBERT, with SRL infused, captures the similarity between the premise and hypothesis, and learns the lexical meaning and world knowledge distinction to make the correct prediction.

4.3.2 HANS

This dataset was created by manipulating the syntactic structure and constituents with prescribed rules. Better performance of a model on HANS means that the trained model is less vulnerable to *lexical overlap*, *subsequence*, or *constituent* heuristics present in the training data. As analyzed Section 2.2, we expect SRL information to help with cases of the *lexical overlap* and *subsequence* heuristics.

From Tables 3 and 4, we made the following observations for HANS dataset:

1. While LingBERT consistently outperforms SemBERT on HANS, it only outperforms BERT when fine-tuned on SNLI.

2. The BERT model fine-tuned on MNLI outperforms the BERT model fine-tuned on SNLI on HANS. However, SemBERT and LingBERT show the opposite tendency.

The detailed analysis of these observations are presented in the subsequent sections.

5 Detailed Analysis

In Section 2, we observe that there are some heuristics where infusing SRL information is useful. In this section, we analyze the performance of the SRL infused models on different heuristics in detail and analyze the effect of different training datasets on model performance.

5.1 Heuristic Level Analysis

When fine-tuned on SNLI (see Table 3), LingBERT achieves the best overall performance on HANS, followed by BERT and SemBERT. Considering that every model classifies most examples as *entailment* and achieves an accuracy of $\sim$99% for the ground-truth *entailment* examples, we examine the models trained on SNLI training set and evaluated on the HANS *non-entailment* examples for each heuristic type and its sub-cases to better interpret their performance. Details of model performance for these examples are summarized in Table 5.

5.1.1 Lexical Overlap Heuristic

As expected, LingBERT outperforms other models by a big margin on every sub-case of the lexi-

HANS Heuristics	***non-entailment* Examples**	**BERT**	**SemBERT**	**LingBERT**
Lexical Overlap Heuristic		46.33	43.02	**54.40**
ln_conjunction	P_1: The authors recognized the president and the judges .			
	H_1: The judges recognized the president .	40.93	33.57	**50.63**
ln_passive	P_2: The lawyers were recommended by the doctor .			
	H_2: The lawyers recommended the doctor .	17.90	30.77	**33.00**
ln_preposition	P_3: The senators behind the lawyer contacted the student .			
	H_3: The student contacted the senators .	58.37	49.20	**60.37**
ln_relative_clause	P_4: The student who the senators thanked stopped the scientist .			
	H_4: The scientist stopped the student .	46.67	40.20	**52.63**
ln_subject/object_swap	P_5: The student saw the managers .			
	H_5: The managers saw the student .	67.77	61.37	**75.37**
Subsequence Heuristic		**4.92**	3.69	4.01
sn_NP/S	*P_1: The author heard the presidents recommended the secretary .*			
	H_1: The author heard the presidents .	**0.70**	0.03	0.53
sn_NP/Z	*P_2: Although the managers hid the actors saw the athlete .*			
	H_2: The managers hid the actors .	**9.67**	6.43	5.27
sn_PP_on_subject	*P_3: The student near the secretaries supported the judges .*			
	H_3: The secretaries supported the judges .	**9.03**	6.9	7.83
sn_past_participle	P_4: The artist avoided the author paid in the laboratory .			
	H_4: The author paid in the laboratory .	**0.80**	0.27	**0.80**
sn_relative_clause_on_subject	P_5: The scientists that introduced the senator avoided the actor .			
	H_5: The senator avoided the actor .	4.40	4.83	**5.63**
Constituent Heuristic		**5.2**	2.44	3.03
cn_adverb	*P_1: **Hopefully** the presidents introduced the doctors .*			
	H_1: The presidents introduced the doctors .	**0.20**	0.00	0.00
cn_after_if_clause	*P_2: **Unless** the professor slept , the tourist saw the scientist.*			
	H_2: The tourist saw the scientist .	0.00	0.00	0.00
cn_disjunction	*P_3: The actor recommended the lawyers , **or** the managers stopped the author .*			
	H_3: The actor recommended the lawyers .	**0.33**	0.03	0.00
cn_embedded_under_if	*P_4: **If** the doctors mentioned the judge , the president thanked the student .*			
	H_4: The doctors mentioned the judge .	**25.3**	12.2	15.1
cn_embedded_under_verb	*P_5: The lawyers **believed** that the tourists shouted .*			
	H_5: The tourists shouted .	**0.13**	0.00	0.00

Table 5: Performance on HANS *non-entailment* examples by models fine-tuned on *SNLI*. Examples **in black and normal font** are where BERT made wrong predictions and LingBERT made correct predictions. Examples **in *blue and italics*** are where none of the three models made the correct prediction. The last three columns are the accuracy in % on the *non-entailment* examples by BERT, SemBERT, and LingBERT respectively.

cal overlap heuristic (see Table 5 Row 1), indicating the contribution of effectively infusing SRL representations. In contrast, SemBERT, which also incorporates SRL representations, fails to outperform BERT for all sub-cases except for one (*ln_preposition*). This observation supports that **predicate-wise concatenation with interaction is more effective to infuse SRL than sentence-wise concatenation.**

Column 2 lists examples for each sub-case. The premise and hypothesis for all examples consist of one clause (except the premise for the *ln_relative_clase* example) with very similar predicate-argument structures. The similarity and simplicity of the predicate-argument structures of the premise and the hypothesis seem to be related to the contribution of SRL in LingBERT. When the premise and hypothesis have similar predicate-argument structures, **incorporating SRL representations emphasizes the similarity between them, thus enforces BERT to learn to distinguish lexical meaning and world knowledge**.

5.1.2 Subsequence Heuristic

All models perform poorly for the subsequence heuristic as shown in Table 5 Row 2. Overall, BERT outperforms SemBERT and LingBERT, though not a single model improves consistently over all the sub-cases. All examples (except *sn_relative_clause_on_subject* and *sn_PP_on_subject*) were created to confuse the model by leveraging the syntactic flexibility of the predicates. For instance, P_1-H_1 leverages the syntactic flexibility of *hear* that the ARG0 of this verb can be a clause as in P_1 or a noun phrase as in H_1. P_2-H_2 leverages the syntactic flexibility of *hide* that this verb can be intransitive as in P_2 or transitive as in H_2. Against our expectations discussed in the section on probing informa, SRL-infused models almost consistently perform worse than BERT. The poor performance of such models demands better methods of incorporating SRL embeddings with BERT representations.

Note that on *sn_PP_on_subject* sub-case, we observe the expected performance drop when SRL representations are infused with BERT representations. This drop is because the span-based predicate-argument structure introduces misleading information. For example, the predicate-argument structure of the premise of P_3-H_3 in Table 5 Row 2 is [$_{\text{ARG0}}$ The student near the secretaries] [$_{\text{V}}$ supported] [$_{\text{ARG1}}$ the judges] and that for the hypothesis is [$_{\text{ARG0}}$ The secretaries] [$_{\text{V}}$ supported] [$_{\text{ARG1}}$ the judges]. In both the premise and the hypothesis, *the secretaries* is within the ARG0 span, though in the premise it's only the modifier of the ARG0 which is not distinguished in span-based semantic role labeling. This information misleads the span-based SRL infused models to predict *entailment* opposed to *non-entailment*. We suspect head-based SRL information may improve the performance on the examples of this sub-case, which can be explored in future work.

5.1.3 Constituent Heuristic

As discussed earlier, we do not expect to observe LingBERT generalize over this heuristic, because SRL can not provide the critical logical knowledge to improve model performance on this heuristic. In Table 5 Row 3, all models perform poorly on this heuristic with basically no difference. For all the constituent heuristic examples, the truth condition of the premise is changed either by adding certain particular words or by using certain syntactic structures with particular connecting words. The words through which the premise's truth condition gets reversed are marked out in bold font in these examples. The poor performance of LingBERT and SemBERT for these examples indicates that **incorporating semantic role labels cannot help BERT learn logic knowledge as expected**.

5.2 Effect of Training Data

Tables 3 shows that there is performance boost on HANS in LingBERT when it is fine-tuned on SNLI. However, from the performance on HANS presented in Table 4, we observe the opposite for fine-tuning with MNLI. To understand why there is the opposite pattern, we examine how the following factors associated with training data may influence the effect of infusing SRL representations: (1) complexity of the training data, and (2) quality of semantic role labels.

5.2.1 Complexity of Training Dataset

From Tables 3 and 4, we observe that the performance of the BERT model on HANS varies largely with the training dataset. Therefore, we start by analyzing the effect of training data complexity on the fine-tuned BERT model. Since MNLI has premises of more diversified genres with longer and more complex sentences than SNLI, we hypothesize that the BERT model fine-tuned on longer and more complex sentences is less likely to overfit the three types of heuristics in HANS. Conversely, the BERT model fine-tuned on SNLI is more likely to overfit to these heuristics. Hence, the contribution of the SRL representations incorporated in the BERT representations may outweigh the noise it introduces when fine-tuned on SNLI in comparison to MNLI where BERT representations alone better generalize over such heuristics. To verify the hypothesis that longer and more complex NLI training data tend to have fewer supporting cases for HANS heuristics, we fine-tune BERT on the combined ANLI training set.[11] ANLI (Nie et al., 2020) is a newly created NLI dataset with much longer and more complex premises from more diversified domains than MNLI. **However, BERT model fine-tuned on this dataset achieves only** 56.44% **accuracy on HANS, even lower than BERT fine-tuned on SNLI (**58.83%**).** Therefore, we rule out the possibility that the complexity of training data is the only contributor towards the better performance of MNLI fine-tuned BERT model.

5.2.2 Quality of SRL

Similar to SNLI fine-tuned LingBERT we expect to see performance improvement on MNLI fine-tuned LingBERT over BERT. However, though LingBERT outperform SemBERT, we observe an obvious performance drop when compared to BERT. To understand it further we hypothesize that the noise in the SRL process may get carried over to the downstream NLI task with LingBERT and led to the performance drop with the infusion of SRL representations. To test this hypothesis, we combined the labeled SNLI training data and the la-

[11] https://dl.fbaipublicfiles.com/anli/anli_v1.0.zip

beled MNLI training data respectively with the CoNLL 2012 SRL training data (Pradhan et al., 2013)[12] to train an in-house SRL model. We also train the in-house SRL model on CoNLL 2012 training data only. When evaluated on the SRL test set, the F1-score for the SRL-data-only, SRL-data+labeled-SNLI, and SRL-data+labeled-MNLI training are 88%, 76.39% and 78.89% respectively. Given that SRL models perform similarly over SNLI and MNLI, **SRL quality is unlikely to be the reason for the difference.** Therefore, we refute both hypotheses and we plan to explore this direction further in the future.

6 Related Work

It has been acknowledged that though neural models can surpass human performance on large benchmark datasets, they are still far from actually understanding the language. This is true for NLI (Naik et al., 2018; Talman and Chatzikyriakidis, 2019; Gururangan et al., 2018; Poliak et al., 2018; Tsuchiya, 2018; Geva et al., 2019; Nie et al., 2020) and other tasks like question answering (Agrawal et al., 2016; Mudrakarta et al., 2018), reading comprehension (Kaushik and Lipton, 2018), commonsense reasoning (Branco et al., 2021) etc.

SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018) are two large benchmark datasets which motivate the application of neural models developed for the natural language inference task, but the performance of neural models rapidly reach the ceiling level on these two datasets. Adversarial NLI (Nie et al., 2020) was developed with a never-end learning scenario in response to the rapid model improvement. Various probing datasets have been developed to probe about whether the model has learned certain aspects of information humans usually resort to for conducting the same task. For example, Breaking NLI (Glockner et al., 2018) probes about whether the model can distinguish lexical meaning and world knowledge. HANS (McCoy et al., 2019) was created by manipulating the syntactic structure of the premise and the hypothesis to investigate whether the model relies on heuristics or real linguistic understanding. IMPPRES (Jeretic et al., 2020) was developed to investigate whether the model has learned pragmatic inferences like presupposition and implicature.

This paper is closely related to work on incorporating explicit linguistic or external knowledge to contextual word representations (Zhang et al., 2019b; Cengiz and Yuret, 2020; Peinelt et al., 2020; Wang et al., 2019b; Kapanipathi et al., 2020) and examining whether the model has learned human-like competence (Goldberg, 2019; Liu et al., 2019a; Hewitt and Manning, 2019; Clark et al., 2019; Lin et al., 2019; Tenney et al., 2019; Warstadt and Bowman, 2020; Manning et al., 2020; Ettinger, 2020; Michael et al., 2020). We propose to improve SemBERT (Zhang et al., 2019b) by predicate-wise concatenation of SRL representations with BERT representations followed with interaction, and analyze whether the infused SRL information increased the model's performance and linguistic awareness.

7 Conclusion

We conduct experiments to analyze whether incorporating SRL representations into BERT representations by concatenation can improve the models' performance on the NLI task, and help the models learn linguistic knowledge and human-like generalizations. In addition, we propose predicate-wise concatenation with interaction for combining SRL embeddings and BERT representations, which turns out to be more effective than the sentence-wise concatenation adopted in SemBERT. Through experiments, we observe that infusing SRL representations with BERT representations improves the model generalization on lexical and world knowledge, as evident from the consistently better performance of LingBERT on the Breaking NLI dataset. Performance improvement on HANS lexical overlap heuristic examples when the model is fine-tuned on SNLI further strengthens our claim. However, we do not observe any semantic-aware model fined-tuned on MNLI outperform BERT on the HANS dataset. This indicates that incorporating linguistic knowledge explicitly encoded by predicate-argument structures by just concatenating SRL representations with BERT representations does not increase the model's awareness of other linguistic knowledge and competence of reasoning generalization. Therefore, how to effectively infuse SRL information to improve the model's awareness of other linguistic phenomena remains an open challenge. With the current model, the extra explicit SRL information does not necessarily reduce the model's vulnerability to heuristics, and its performance is influenced by the training data.

[12] Since CoNLL 2012 SRL dataset contains multiple genres, we combined the data for training in order to alleviate the genre difference between SNLI and MNLI and make the comparison more reliable.

References

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the behavior of visual question answering models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1955–1960, Austin, Texas. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Ruben Branco, António Branco, João António Rodrigues, and João Ricardo Silva. 2021. Shortcutted commonsense: Data spuriousness in deep learning of commonsense reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1504–1521, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Cemil Cengiz and Deniz Yuret. 2020. Joint training with semantic role labeling for better generalization in natural language inference. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 78–88, Online. Association for Computational Linguistics.

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, Florence, Italy. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer.

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

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.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. 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 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 473–483.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.

Ishan Jindal, Ranit Aharonov, Siddhartha Brahma, Huaiyu Zhu, and Yunyao Li. 2020. Improved semantic role labeling using parameterized neighborhood memory adaptation. *arXiv preprint arXiv:2011.14459*.

Pavan Kapanipathi, Veronika Thost, Siva Sankalp Patel, Spencer Whitehead, Ibrahim Abdelaziz, Avinash Balakrishnan, Maria Chang, Kshitij P Fadnis, R Chulaka Gunasekara, Bassem Makni, et al. 2020. Infusing knowledge into the textual entailment task using graph convolutional networks. In *AAAI*, pages 8074–8081.

Jungo Kasai, Dan Friedman, Robert Frank, Dragomir Radev, and Owen Rambow. 2019. Syntax-aware neural semantic role labeling with supertags. 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 701–709.

Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015, Brussels, Belgium. Association for Computational Linguistics.

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

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

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*.

Bill MacCartney and Christopher D. Manning. 2009. An extended model of natural logic. In *Proceedings of the Eight International Conference on Computational Semantics*, pages 140–156, Tilburg, The Netherlands. Association for Computational Linguistics.

Christopher D Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*.

Diego Marcheggiani and Ivan Titov. 2020. Graph convolutions over constituent trees for syntax-aware semantic role labeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3915–3928, Online. Association for Computational Linguistics.

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, Florence, Italy. Association for Computational Linguistics.

Julian Michael, Jan A. Botha, and Ian Tenney. 2020. Asking without telling: Exploring latent ontologies in contextual representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6792–6812, Online. Association for Computational Linguistics.

Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. 2018. Did the model understand the question? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1896–1906, Melbourne, Australia. Association for Computational Linguistics.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Deric Pang, Lucy H Lin, and Noah A Smith. 2019. Improving natural language inference with a pretrained parser. *arXiv preprint arXiv:1909.08217*.

Nicole Peinelt, Marek Rei, and Maria Liakata. 2020. GiBERT: Introducing linguistic knowledge into bert through a lightweight gated injection method. *arXiv preprint arXiv:2010.12532*.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational*

Natural Language Learning, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.

Aarne Talman and Stergios Chatzikyriakidis. 2019. Testing the generalization power of neural network models across NLI benchmarks. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 85–94, Florence, Italy. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Xiaoyan Wang, Pavan Kapanipathi, Ryan Musa, Mo Yu, Kartik Talamadupula, Ibrahim Abdelaziz, Maria Chang, Achille Fokoue, Bassem Makni, Nicholas Mattei, et al. 2019b. Improving natural language inference using external knowledge in the science questions domain. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7208–7215.

Alex Warstadt and Samuel R Bowman. 2020. Can neural networks acquire a structural bias from raw linguistic data? *arXiv preprint arXiv:2007.06761*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910.

Zhuosheng Zhang, Yuwei Wu, Zuchao Li, and Hai Zhao. 2019a. Explicit contextual semantics for text comprehension. In *Proceedings of the 33rd Pacific Asia Conference on Language, Information and Computation*, pages 298–308. Waseda Institute for the Study of Language and Information.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2019b. Semantics-aware bert for language understanding. *arXiv preprint arXiv:1909.02209*.

A Detailed results on HANS dataset

	train on SNLI			**train on MNLI**		
	all heuristics			**all heuristics**		
	entail	non-entail	overall	entail	non-entail	overall
BERT	98.85	18.81	58.83	98.15	**34.24**	**66.19**
SemBERT	99.40	16.38	57.89	**99.30**	7.59	53.44
Ours	**99.45**	**20.48**	**59.96**	98.41	18.75	58.58
	lexical overlap heuristic			**lexical overlap heuristic**		
	entail	non-entail	overall	entail	non-entail	overall
BERT	97.35	46.33	71.84	95.20	**64.77**	**79.98**
SemBERT	**98.61**	43.02	70.82	**98.51**	14.15	56.33
Ours	98.42	**54.40**	**76.41**	96.34	41.01	68.68
	subsequence heuristic			**subsequence heuristic**		
	entail	non-entail	overall	entail	non-entail	overall
BERT	99.67	**4.92**	**52.29**	99.44	**14.27**	**56.85**
SemBERT	99.67	3.69	51.68	**99.98**	2.05	51.01
Ours	**99.96**	4.01	51.99	99.85	5.17	52.51
	constituent heuristic			**constituent heuristic**		
	entail	non-entail	overall	entail	non-entail	overall
BERT	99.53	**5.20**	**52.36**	**99.81**	**23.69**	**61.75**
SemBERT	99.90	2.44	51.17	99.43	6.56	52.99
Ours	**99.96**	3.03	51.49	99.03	10.07	54.55

Table 6: Performance on HANS. *entail* columns are where the ground-truth label is *entailment*, *non-entail* columns are where the ground-truth label is *non-entailment*, and *overall* columns are for both ground-truth labels together.