

Symmetric Regularization based BERT for Pair-wise Semantic Reasoning

Weidi Xu ^{1*}, Xingyi Cheng ^{1*}, Kunlong Chen ^{1*}, Wei Wang ², Bin Bi ²,
Ming Yan ², Chen Wu ², Luo Si ², Wei Chu ¹ and Taifeng Wang ¹

¹ Ant Financial Services Group

² DAMO, Alibaba Group

{weidi.xwd,fanyin.cxy,kunlong.ckl,hebian.ww,b.bi,yml19608,wuchen.wc,luo.si, weichu.cw,taifeng.wang}@alibaba-inc.com

Abstract

The ability of semantic reasoning over the sentence pair is essential for many natural language understanding tasks, e.g., natural language inference and machine reading comprehension. A recent significant improvement in these tasks comes from BERT. As reported, the next sentence prediction (NSP) in BERT, which learns the contextual relationship between two sentences, is of great significance for downstream problems with sentence-pair input. Despite the effectiveness of NSP, we suggest that NSP still lacks the essential signal to distinguish between entailment and shallow correlation. To remedy this, we propose to augment the NSP task to a 3-class categorization task, which includes a category for previous sentence prediction (PSP). The involvement of PSP encourages the model to focus on the informative semantics to determine the sentence order, thereby improves the ability of semantic understanding. This simple modification yields remarkable improvement against vanilla BERT. To further incorporate the document-level information, the scope of NSP and PSP is expanded into a broader range, i.e., NSP and PSP also include close but nonsuccessive sentences, the noise of which is mitigated by the label-smoothing technique. Both qualitative and quantitative experimental results demonstrate the effectiveness of the proposed method. Our method consistently improves the performance on the NLI and MRC benchmarks, including the challenging HANS dataset (McCoy, Pavlick, and Linzen 2019), suggesting that the document-level task is still promising for the pre-training.

1 Introduction

¹ The ability of semantic reasoning is essential for advanced natural language understanding (NLU) systems. Many NLU tasks that take sentence pairs as input, such as natural language inference (NLI) and machine reading comprehension (MRC), heavily rely on the ability of sophisticated semantic reasoning. For instance, the NLI task aims to determine whether the hypothesis sentence (e.g., a woman is sleeping) can be inferred from the premise sentence (e.g., a woman is talking on the phone). This requires the model to read and

understand sentence pairs to make the specific semantic inference.

Bidirectional Encoder Representations from Transformer (BERT) (Devlin et al. 2018) has shown strong ability in semantic reasoning. It was recently proposed and obtained impressive results on many tasks, ranging from text classification, natural language inference, and machine reading comprehension. BERT achieves this by employing two objectives in the pre-training, i.e., the masked language modeling (Masked LM) and the next sentence prediction (NSP). Intuitively, the Masked LM task concerns word-level knowledge, and the NSP task captures the global document-level information. The goal of NSP is to identify whether an input sentence is next to another input sentence. From the ablation study (Devlin et al. 2018), the NSP task is quite useful for the downstream NLI and MRC tasks (e.g., +3.5% absolute gain on the Question NLI (QNLI) (Wang et al. 2019a) task).

Despite its usefulness, we suggest that BERT has not made full use of the document-level knowledge. The sentences in the negative samples used in NSP are randomly drawn from other documents. Therefore, to discriminate against these sentences, BERT is prone to aggregating the shallow semantic, e.g., topic, neglecting context clues useful for detailed reasoning. In other words, the canonical NSP task would encourage the model to recognize the correlation between sentences, rather than obtaining the ability of semantic entailment. This setting weakens the BERT model from learning specific semantic for inference. Another issue that renders NSP less effective is that BERT is order-sensitive. Performance degradation was observed on typical NLI tasks when the order of two input sentences are reversed during the BERT fine-tuning phase. It is reasonable as the NSP task can be roughly analogy to the NLI task when the input comes as (premise, hypothesis), considering the causal order among sentences. However, this identity between NSP and NLI is compromised when the sentences are swapped.

Based on these considerations, we propose a simple yet effective method, i.e., introducing a *IsPrev* category to the classification task, which is a symmetric label of *IsNext* of NSP. The input of samples with *IsPrev* is the reverse of those with *IsNext* label. The advantages of using this previous sentence prediction (PSP) are three folds. (1) Learning the contrast between NSP and PSP forces the model to extract more detailed semantic, thereby the model is more

*Equal Contribution.

¹We withdrew the paper in ACL2020. The code was release at <https://github.com/ACL2020Anonymous/SymBERT> in Dec. 2019.

capable of discriminating the correlation and entailment. (2) NSP and PSP are symmetric. This symmetric regularization alleviates the influence of the order of the input pair. (3) Empirical results indicate that our method is beneficial for all the semantic reasoning tasks that take sentence pair as input.

In addition, to further incorporating the document-level knowledge, NSP and PSP are extended with non-successive sentences, where the label smoothing technique is adopted. The proposed method yields a considerable improvement in our experiments. We evaluate the ability of semantic reasoning on standard NLI and MRC benchmarks, including the challenging HANS dataset² (McCoy, Pavlick, and Linzen 2019). Analytical work on the HANS dataset provides a more comprehensible perspective towards the proposed method. Furthermore, the results on the Chinese benchmarks are provided to demonstrate its generality.

In summary, this work makes the following contributions:

- The supervision signal from the original NSP task is weak for semantic inference. Therefore, a novel method is proposed to remedy the asymmetric issue and enhance the reasoning ability.
- Both empirical and analytical evaluations are provided on the NLI and MRC datasets, which verifies the effectiveness of using more document-level knowledge.

2 Related Work

Pair-wise semantic reasoning Many NLU tasks seek to model the relationship between two sentences. Semantic reasoning is performed on the sentence pair for the task-specific inference. Pair-wise semantic reasoning tasks have drawn a lot of attention from the NLP community as they largely require the comprehension ability of the learning systems. Recently, the significant improvement on these benchmarks comes from the pre-training models, e.g., BERT, StructBERT (Wang et al. 2019b), ERNIE (Sun et al. 2019a; 2019b), RoBERTa (Liu et al. 2019) and XLNet (Yang et al. 2019). These models learn from unsupervised/self-supervised objectives and perform excellently in the downstream tasks. Among these models, BERT adopts NSP as one of the objectives in the pre-training and shows that the NSP task has a positive effect on the NLI and MRC tasks. Although the primary study of XLNet and RoBERTa suggests that NSP is ineffective when the model is trained with a large sequence length of 512, the effect of NSP on the NLI problems should still be emphasized. The inefficiency of NSP is likely because the expected context length will be halved for Masked LM when taking a sentence pair as the input. The models derived from BERT, e.g., StructBERT and ERNIE 1.0/2.0, aim to incorporating more knowledge by elaborating pre-training objectives. This work aims to enhance the NSP task and verifies whether document-level information is helpful for the pre-training. To probe whether our method achieves a better regularization ability, our approach is also evaluated on the HANS (McCoy, Pavlick, and Linzen 2019) dataset, which contains hard data samples constructed by three heuristics. Previous advanced models such

as BERT fail on the HANS dataset, and the test accuracy can barely exceed 0% in the subset of test examples.

Unsupervised learning from document In recent years, many unsupervised pre-training methods have been proposed in the NLP fields to extract knowledge among sentences (2015; 2017; 2018; 2019). The prediction of surrounding sentences endows the model with the ability to model the sentence-level coherence. Skip-Thought (Kiros et al. 2015) consists of an encoder and two decoders. When a sentence is given and encoded into a vector by the encoder, the decoders are trained to predict the next sentence and the previous sentence. The goal is to obtain a better sentence representation that is useful for reconstructing the surrounding context. Considering that the estimation of the likelihood of sequences is computationally expensive and time-consuming, the Quick-Thought method (Logeswaran and Lee 2018) simplifies this in a manner similar to sampled softmax (Jean et al. 2015), which classifies the input sentences between surrounding sentences and the other. Note that Quick-Thought does not distinguish between the previous and next sentence as it is functionally rotation invariant. However, BERT is order-dependent, and the discrimination can provide more supervision signal for semantic learning. InferSent (Conneau et al. 2017) instead pre-trains the model in a manner of supervised learning. It uses a large-scale NLI dataset as the pre-training task to learn the sentence representation. In our work, we focus on designing a more effective document-level objective, extended from the NSP task. The proposed method will be described in the following section and validated by providing extensive experimental results in the experiment part.

3 Method

Our method follows the same input format and the model architecture with original BERT. The proposed method solely concerns the NSP task. The NSP task is a binary classification task, which takes two sentences (A and B) as input and determines whether B is the next sentence of A . Although it has been proven to be very effective for BERT, there are two major deficiencies. (1) Discrimination between `IsNext` and `DiffDoc` (the label of the sentences drawn from different documents via negative sampling) is semantically shallow as the signal of sentence order is absent. The correlation between two successive sentences could be obvious, due to, for example, lexical overlap or the conjunction used at the beginning of the second sentence. As reported (Devlin et al. 2018), the final pre-trained model is able to achieve 97%-98% accuracy on the NSP task. (2) BERT is order-sensitive, i.e., $f_{\text{BERT}}(A, B) \neq f_{\text{BERT}}(B, A)$, while NSP is unidirectional. When the order of the input NLI pair is reversed, the performance will degrade. For instance, the accuracy decreases by about 0.5% on MNLI (Williams, Nangia, and Bowman 2018) and 0.4% on QNLI after swapping the sentences in our experiments³.

²Heuristic Analysis for NLI Systems

³The comparison was conducted for 5 times, and the averaged gap is reported.

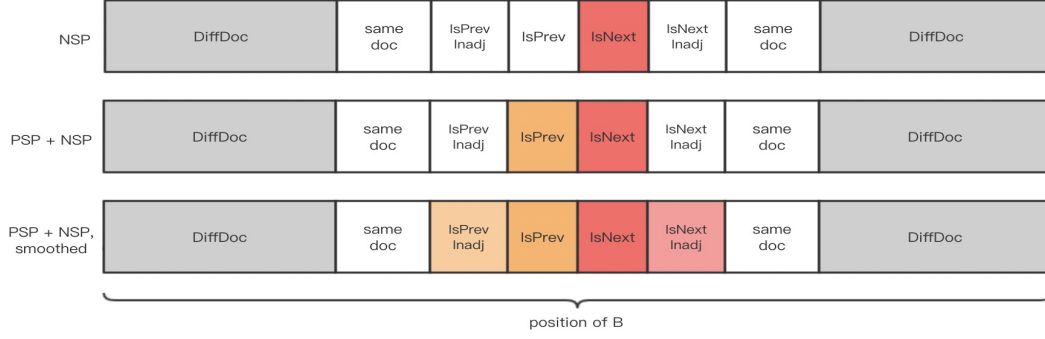


Figure 1: An illustration of the proposed method. B denotes the second input sentence. (1) Top: original NSP task. (2) Middle: 3-class categorization task with *DiffDoc*, *IsNext* and *IsPrev*. (3) Bottom: 3-class task, but with a wider scope of NSP and PSP. The in-adjacent sentences are assisted with a label smoothing technique to reduce the noise.

Motivated by these problems, we propose to extend the NSP task with previous sentence prediction (PSP). Despite its simplicity, empirical results show that this is beneficial for downstream tasks, including both NLI and MRC tasks. To further incorporate the document-level information, the scope is also expanded to include more surrounding sentences, not just the adjacent. The method is briefly illustrated in Fig. 1.

3.1 Previous Sentence Prediction

Learning to recognize the previous sentence enables the model to capture more compact context information. One would argue that *IsPrev* (the label of PSP) is redundant as it plays a similar role of *IsNext* (the label of NSP). In fact, Quick-Thought uses the sampled softmax to approximate the sentence likelihood estimation of Skip-Thought, and it actually does not differentiate between the previous and next sentences. However, we suggest the order discrimination is essential for BERT pre-training. Quick-Thought aims at extracting sentence embedding, and it uses a rotating symmetric function, which makes *IsPrev* redundant in Quick-Thought. In contrast, BERT is order-sensitive, and learning the symmetric regularization is rather necessary. Another advantage of PSP is to enhance document-level supervision. In order to tell the difference between NSP and PSP, the model has to extract the detailed semantic for inference.

3.2 Gathering More Document-level Information

Beyond NSP and PSP, which enable the model to learn the short-term dependency between sentences, we also propose to expand the scope of discrimination task to further incorporate the document-level information.

Specifically, we also include the in-adjacent sentences in the sentence-pair classification task. The in-adjacent sentences next to the *IsPrev* and *IsNext* sentences are sampled, labeled as *IsPrevInadj* and *IsNextInadj* (cf. the bottom of Fig. 1). Note that these in-adjacent sentences will introduce much more training noise to the model. Therefore, the label smoothing technique is adopted to reduce the noise of these additional samples. It achieves this

by relaxing our confidence on the labels, e.g., transforming the target probability from (1.0, 0.0) to (0.8, 0.2) in a binary classification problem.

In summary, when A is given, the pre-training example for each label is constructed as follows:

- *IsNext*: Choosing the adjacent following sentence as B.
- *IsPrev*: Choosing the adjacent previous sentence as B.
- *IsNextInadj*: Choosing the in-adjacent following sentence as B. There is a sentence between A and B.
- *IsPrevInadj*: Choosing the in-adjacent previous sentence as B. There is a sentence between A and B.
- *DiffDoc*: Drawing B randomly from a different document.

4 Experiment Settings

This section gives detailed experiment settings. The method is evaluated on the BERTbase model, which has 12 layers, 12 self-attention heads with a hidden size of 768.

To accelerate the training speed, two-phase training (Devlin et al. 2018) is adopted. The first phase uses a maximal sentence length of 128, and 512 for the second phase. The numbers of training steps of two phases are 50K and 40K for the BERTBase model. We used AdamW (Loshchilov and Hutter 2019) optimizer with a learning rate of $1e-4$, a β_1 of 0.9, a β_2 of 0.999 and a L2 weight decay rate of 0.01. The first 10% of the total steps are used for learning rate warming up, followed by the linear decay schema. We used a dropout probability of 0.1 on all layers. The data used for pre-training is the same as BERT, i.e., English Wikipedia (2500M words) and BookCorpus (800M words) (Zhu et al. 2015). For the Masked LM task, we followed the same masking rate and settings as in BERT.

We explore three method settings for comparison.

- **BERT-PN**: The NSP task in BERT is replaced by a 3-class task with *IsNext*, *IsPrev* and *DiffDoc*. The label distribution is 1:1:1.
- **BERT-PN5cls**: The NSP task in BERT is replaced by a 5-class task with two additional labels *IsNextInadj*, *IsPrevInadj*. The label distribution is 1:1:1:1:1.

- BERT-PNsmth: It uses the same data with BERT-PN5cls, except that the `IsPrevInadj` (`IsNextInadj`) label is mapped to `IsPrev` (`IsNext`) with a label smoothing factor of 0.8.

BERT-PN is used to verify the feasibility of PSP. The comparison with BERT-PN5cls illustrates whether more document-level information helps. BERT-PNsmth, which is the label-smoothed version of BERT-PN5cls, is used to compare with BERT-PN5cls to see whether the noise reduction is necessary.

In the following, we first show that BERT is order-sensitive and the use of PSP remedies this problem. Then we provide experimental results on the NLI and MRC tasks to verify the effectiveness of the proposed method. At last, the proposed method is evaluated on several Chinese datasets.

5 Order-invariant with PSP

NSP in the pre-training is useful for NLI and MRC task (Devlin et al. 2018). However, we suggested that BERT trained with NSP is order-sensitive, i.e., the performance of BERT depends on the order of the input sentence pair. To verify our assumption, a primary experiment was conducted. The order of the input pair of NLI samples is reversed in the fine-tuning phase, and other hyper-parameters and settings keep the same with the BERT paper. Table 1 shows the accuracy on the validation set of the MNLI⁴ and QNLI datasets. For the BERTBase model, when the sentences are swapped, the accuracy decreases by 0.5% on the MNLI task and 0.4% on the QNLI task. These results confirm that BERT trained with NSP only is indeed affected by the input order. This phenomenon motivates us to make the NSP task symmetric. The results of BERT-PN verify that BERT-PN is order-invariant. When the input order is reversed, the performance of BERT-PN remains stable. These results indicate that our method is able to remedy the order-sensitivity problem.

Task	Model	P&H	H&P (reversed)
MNLI	BERTBase	91.5	91.0 (-0.5)
	BERTBase-PN	91.9	92.0 (+0.1)
QNLI	BERTBase	84.4	84.0 (-0.4)
	BERTBase-PN	85.0	84.9 (-0.1)

Table 1: The accuracy of BERT and BERT-PN on the validation set of the MNLI and QNLI dataset. P&H denotes that the input is (premise, hypothesis), which is the order used in BERT. The reported accuracy is the average after 5 runs.

6 Results of NLI Tasks

6.1 GLUE

A popular benchmark for evaluation of language understanding is GLUE (Wang et al. 2019a), which is a collection of three NLI tasks (MNLI, QNLI and RTE), three semantic textual similarity (STS) tasks (QQP, STS-B and MRPC),

⁴The matched set is used for evaluation.

two text classification (TC) tasks (SST-2 and CoLA). Although the method is motivated for pair-wise reasoning, the results of other problems are also listed.

Our implementation follows the same way that BERT performs in these tasks. The fine-tuning was conducted for 3 epochs for all the tasks, with a learning rate of $2e-5$. The predictions were obtained by evaluating the training checkpoint with the best validation performance.

Table 2 illustrates the experimental results, showing that our method is beneficial for all of NLI tasks. The improvement on the RTE dataset is significant, i.e., 4% absolute gain over the BERTBase. Besides NLI, our model also performs better than BERTBase in the STS task. The STS tasks are semantically similar to the NLI tasks, and hence able to take advantage of PSP as well. Actually, the proposed method has a positive effect whenever the input is a sentence pair. The improvements suggest that the PSP task encourages the model to learn more detailed semantics in the pre-training, which improves the model on the downstream learning tasks. Moreover, our method is surprisingly able to achieve slightly better results in the single-sentence problem. The improvement should be attributed to better semantic representation.

When comparing between PN and PN5cls, PN5cls achieves better results than PN. This indicates that including a broader range of the context is effective for improving inference ability. Considering that the representation of `IsNext` and `IsNextInadj` should be coherent, we propose BERTBase-PNsmth to mitigate this problem. PNsmth further improves the performance and obtains an averaged score of 81.0.

6.2 HANS

Although BERT has shown its effectiveness in the NLI tasks. McCoy, Pavlick, and Linzen pointed out that BERT is still vulnerable in the NLI task as it is prone to adopting fallible heuristics. Therefore, they released a dataset, called The Heuristic Analysis for NLI Systems (HANS), to probe whether the model learns inappropriate inductive bias from the training set. It is constructed by three heuristics, i.e., lexical overlap heuristic, sub-sequence heuristic, and constituent heuristic. The first heuristic assumes that a premise entails all hypotheses constructed from words in the premise, the second assumes that a premise entails all of its contiguous sub-sequences and the third assumes that a premise entails all complete sub-trees in its parse tree. BERT and other advanced models fail on this dataset and barely exceeds 0% accuracy in most cases (McCoy, Pavlick, and Linzen 2019).

Fig. 2 illustrates the accuracy of BERTBase and BERTBase-PNsmth on the HANS dataset. The evaluation is made upon the model trained on the MNLI dataset and the predicted neutral and contradiction labels are mapped into non-entailment. The BERTBase-PNsmth evidently outperforms the BERTBase with the non-entailment examples. For the non-entailment samples constructed using the lexical overlap heuristic, our model achieves 160% relative improvement over the BERTBase model. Some samples are constructed by swapping the entities in the sentence

	NLI			STS			TC		Average -
	MNLI 392k	QNLI 108k	RTE 2.5k	QQP 363k	STS-B 8.5k	MRPC 3.5k	SST-2 67k	CoLA 5.7k	
BiLSTM+ELMo+Attn	76.4/76.1	79.8	64.8	56.8	73.3	84.9	90.4	36.0	71.0
OpenAI GPT	82.1/81.4	87.4	56.0	70.3	80.0	82.3	91.3	45.4	75.1
BERTBase	84.6/83.4	90.5	66.4	71.2	85.8	88.9	93.5	52.1	79.6
BERTBase-PN	84.2/84.1	92.2	70.2	71.7	87.2	88.9	94.2	51.1	80.4
BERTBase-PN5cls	84.6/84.3	92.3	70.0	71.9	87.5	89.8	93.5	52.0	80.7
BERTBase-PNsmth	85.2/84.4	92.1	70.6	72.2	86.4	89.8	94.2	54.6	81.0

Table 2: Results on the test set of GLUE benchmark. The performance was obtained by the official evaluation server. The number below each task is the number of training examples. The "Average" column follows the setting in the BERT paper, which excludes the problematic WNLI task. F1 scores are reported for QQP and MRPC, Spearman correlations are reported for STS-B, and accuracy scores are reported for the other tasks. All the listed models are trained on the Wikipedia and the Book Corpus datasets. The results are the average of 5 runs.

(e.g., *The doctor saw the lawyer* \rightarrow *The lawyer saw the doctor*) and our method outperforms BERTBase by 20% in accuracy. We suggest that the Masked LM task can hardly model the relationship between two entities and NSP only is too semantically shallow to capture the precise meaning. However, the discrimination between NSP and PSP enhances the model to realize the role of entities in a given sentence. For example, to determine that A (*X is beautiful*) rather than \bar{A} (*Y is beautiful*) is the previous sentence of B (*Y loves X*), the model have to recognize the relationship between *X* and *Y*. In contrast, when PSP is absent, NSP can be probably inferred by learning the occurrence between *beautiful* and *loves*, regardless of the sentence structure. The detailed performance of the proposed method on the HANS dataset is illustrated in Fig. 3. The definition of each heuristic rules can be found in (McCoy, Pavlick, and Linzen 2019).

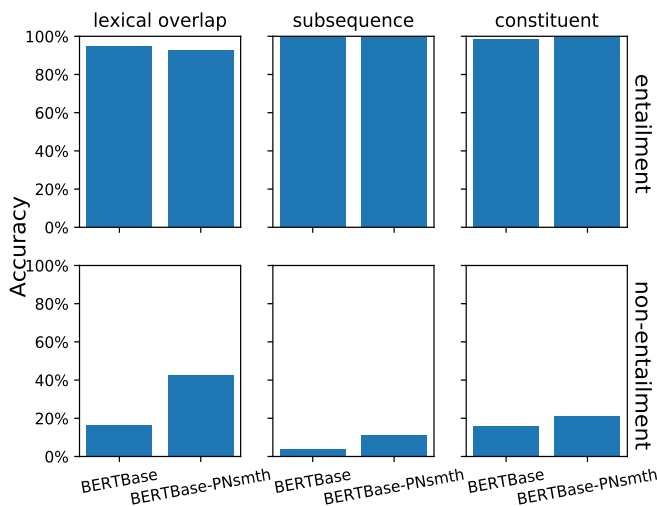


Figure 2: The accuracy on evaluation set of HANS. It has six sub-components, each defined by its correct label and the heuristic it addresses.

Model	Dev v1.1		Dev v2.0	
	EM	F1	EM	F1
RoBERTaBase	-	90.6	-	79.7
BERTBase	80.8	88.5	72.8	76.3
BERTBase-PN	83.2	90.5	76.5	79.6
BERTBase-PN5cls	83.3	90.6	77.0	80.3
BERTBase-PNsmth	83.6	90.6	77.4	80.6

Table 3: The performance of various BERT models fine-tuned on the SQuAD v1.1 and v2.0 dataset. EM means the percentage of exact match. The results of RoBERTa is the DOC-SENTENCES version retrieved from Table 2 in (Liu et al. 2019).

7 Results of MRC Tasks

7.1 SQuAD v1.1 and v2.0

We also evaluate our method on the MRC tasks. The Stanford Question Answering Dataset (SQuAD v1.1) is a question answering (QA) dataset, which consists of 100K samples (Rajpurkar et al. 2016). Each data sample has a question and a corresponding Wikipedia passage that contains the answer. The goal is to extract the answer from the passage for the given question.

In the fine-tuning procedure, we follow the exact way the BERT performed. The output vectors are used to compute the score of tokens being start and end of the answer span. The valid span that has the maximum score is selected as the prediction. And similarly, the fine-tuning training was performed for 3 epochs with a learning rate of $3e-5$.

Table 3 demonstrates the results on the SQuAD v1.1 dataset. The comparison between BERTBase-PN and BERTBase indicates that the inclusion of the PSP sub-task is beneficial (2.4% absolute improvement). When using BERTBase-PNsmth, another 0.3% increase in EM can be obtained. The experimental results on the SQuAD v2.0 (Rajpurkar, Jia, and Liang 2018) are also shown in Table. 3. The SQuAD v2.0 differs from SQuAD v1.1 by allowing the question-paragraph pairs that have no answer. For SQuAD v2.0, our method also achieved about 4% absolute improvement in both EM and F1 against BERTBase.

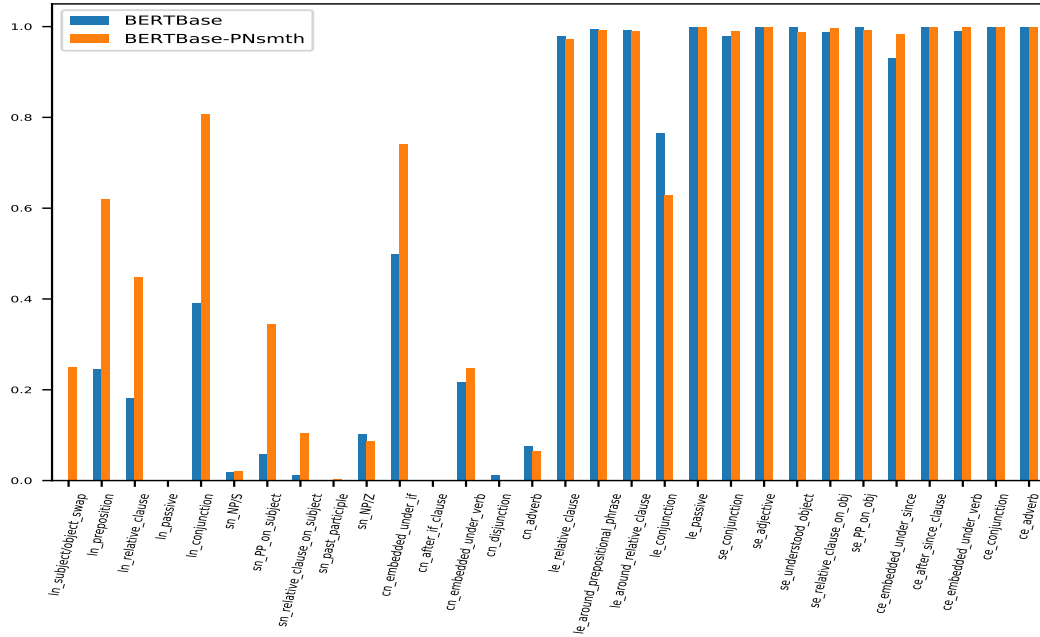


Figure 3: Performance on thirty detailed sub-components of the HANS evaluation set (30K instances). Each sub-component is defined by three heuristics, i.e., Lexical overlap, Sub-sequence and Constituent. For instance, in prefix “In”, “I” denotes lexical overlap heuristic, “n” denotes the non-entailment label. The suffix means a specific syntactic rule, e.g., subject/object_swap means in the hypothesis sentence, the subject and the object are swapped.

7.2 RACE

The ReAding Comprehension from Examinations (RACE) dataset (Lai et al. 2017) consists of 100K questions taken from English exams, and the answers are generated by human experts. This is one of the most challenging MRC datasets that require sophisticated reasoning.

In our implementation, the question, document, and option are concatenated as a single sequence, separated by [SEP] token. And each part is truncated by a maximal length of 40/432/40, respectively. The model computes for a concatenation a scalar as the score, which is then used in a softmax layer for the final prediction. The fine-tuning was conducted for 5 epochs, with a batch size of 32 and a learning rate of $5e-5$. As shown in Table 4, the proposed method significantly improve the performance on the RACE dataset. BERTBase-PN obtains 2.6% accuracy improvement, and BERTBase-PN5cls further brings 0.4% absolute gain.

The comparisons on the SQuAD v1.1, SQuAD v2.0, and RACE dataset demonstrate that the involvement of additional sentence and discourse information is not only beneficial for the NLI task but also the MRC task. This is reasonable as these tasks heavily rely on the global semantic understanding and sophisticated reasoning among sentences. And this ability can be effectively enhanced by our method.

8 Results of Chinese NLP Tasks

The experiments are also conducted on Chinese NLP tasks:

- XNLI (Conneau et al. 2018) a multi-lingual dataset. The data sample in XNLI is a sentence pair annotated with

Model	Middle	High	Accuracy
RoBERTaBase	-	-	65.6
BERTBase	71.8	63.6	66.0
BERTBase-PN	74.2	66.3	68.6
BERTBase-PN5cls	75.8	66.2	69.0
BERTBase-PNsmth	74.1	66.3	68.6

Table 4: The experimental results on test set of the RACE dataset. The results of RoBERTa is the DOC-SENTENCES version retrieved from Table 2 in (Liu et al. 2019). All the listed models are trained on the Wikipedia and the Book Corpus datasets.

textual entailment. The Chinese part is used.

- LCQMC (Liu et al. 2018) is a dataset for sequence matching. A binary label is annotated for a sentence pair in the dataset to indicate whether these two sentences have the same intention.
- NLPCC-DBQA (Duan and Tang 2017) formulates the domain-based question answering as a binary classification task. Each data sample is a question-sentence pair. The goal is to identify whether the sentence contains the answer to the question.
- CMRC-2018⁵ is the Chinese Machine Reading Comprehension dataset. Similar to SQuAD, the system needs to extract fragments from the text as the answer.

⁵<https://hfl-rc.github.io/cmrc2018>

	metrics	BERT 393M	BERT-wwm 393M	ERNIE -	ERNIE2.0 14988M	BERT-PN 10879M	BERT-PNsmth 10879M
<i>Single-task single base models on dev</i>							
XNLI	Accuracy	77.8 (77.4)	79.0 (78.4)	- (79.9)	- (81.2)	80.5 (79.9)	81.4 (81.0)
LCQMC	Accuracy	89.4 (88.4)	89.4 (89.2)	- (89.7)	- (90.9)	90.3 (89.4)	90.6 (90.1)
NLPCC-DBQA	F1	- (80.7)	- (-)	- (82.3)	- (84.7)	85.0 (84.6)	85.9 (85.4)
<i>Single-task single base models on test</i>							
XNLI	Accuracy	77.8 (77.5)	78.2 (78.0)	- (78.4)	- (79.7)	79.8 (79.4)	80.3 (79.9)
LCQMC	Accuracy	86.9 (86.4)	87.0 (86.8)	- (87.4)	- (87.9)	88.7 (87.5)	88.7 (88.0)
NLPCC-DBQA	F1	- (80.8)	- (-)	- (82.7)	- (85.3)	85.2 (84.9)	86.2 (85.9)

Table 5: Comparison on the Chinese NLP tasks. All the models are of “base” size. The results of BERT, BERT-wwm are retrieved from literature (Cui et al. 2019), except the results of NLPCC-DBQA which is from ERNIE 2.0 (2019b). The results of ERNIE, ERNIE 2.0 are retrieved from literature (Sun et al. 2019a; 2019b). The best result and the average (in bracket) of 5 runs are reported. The number below the model denotes the number of tokens in the pre-training data.

	CMRC-2018 (Dev)		DRCD (Dev)		DRCD (Test)	
metrics	F1	EM	F1	EM	F1	EM
BERTBase (ours)	84.7 (84.3)	64.1 (63.8)	90.2 (90.0)	83.5 (83.4)	89.0 (88.9)	82.0 (81.8)
BERTBase (Cui et al. 2019)	84.5 (84.0)	65.5 (64.4)	89.9 (89.6)	83.1 (82.7)	89.2 (88.8)	82.2 (81.6)
BERTBase (Sun et al. 2019b)	- (85.9)	- (66.3)	- (91.6)	- (85.7)	- (90.9)	- (84.9)
BERTBase-wwm	85.6 (84.7)	66.3 (65.0)	90.5 (90.2)	83.7 (83.5)	89.8 (89.4)	82.7 (82.1)
BERTBase-PN	87.5 (86.8)	66.6 (65.8)	92.3 (92.0)	86.4 (86.0)	92.3 (92.2)	86.1 (86.0)
BERTBase-PNsmth	86.4 (86.2)	66.5 (66.3)	93.0 (92.7)	86.8 (86.8)	92.6 (92.5)	86.7 (86.6)

Table 6: Results on the CMRC-2018 and DRCD datasets. Three BERTBase models are reported from our reproduction, BERT-wwm paper (Cui et al. 2019) and ERNIE 2.0 paper (Sun et al. 2019b), respectively. The results of BERTBase-wwm are obtained from the paper (Cui et al. 2019). EM denotes the percentage of exact matching. The best result and the average (in bracket) of 5 runs are reported.

- DRCD (Shao et al. 2018) is also a Chinese MRC data set. The data follows the format of SQuAD.

For Chinese NLP tasks, we pre-train the model using Chinese corpus. We collected textual data (10879M tokens in total) from the website, consisting of Hudong Baike data (6084M tokens)⁶, Zhihu data(465M tokens)⁷, Sohu News(3937M tokens)⁸ and Wikipedia data (393M tokens).

For the first 3 Chinese tasks, we follow the settings as in ERNIE (Sun et al. 2019a). The experimental results are given in Table 5. The proposed method is compared with four models, i.e., BERTBase (Devlin et al. 2018), BERTBase with whole word masking (Cui et al. 2019), ERNIE (Sun et al. 2019a) and ERNIE 2.0 (Sun et al. 2019b). Our method achieves comparable or even better results against ERNIE 2.0 (Sun et al. 2019b). Note that the Chinese ERNIE 2.0 is equipped with 5 different objectives and it uses more training data (14988M tokens in total) than ours. The results indicate that the proposed method is quite effective for the pair-wise semantic reasoning as simply including PSP can achieve the results on par with multiple objectives.

The results of CMRC-2018 and DRCD datasets are given in Table 6. Since the CMRC-2018 competition does not release the test set, the comparison on the test set is absent. Our

results are obtained using the open-sourced code of BERT-wwm⁹. We keep the hyper-parameters the same with that in ERNIE (Sun et al. 2019a), except that the batch size is 12 instead of 64 due to the memory limit. Under this setting, we achieved similar results of BERTBase in the BERT-wwm paper (Cui et al. 2019). However, this is worse than the results of BERTBase reported in the ERNIE 2.0 paper (Sun et al. 2019b) by about 1% in F1. This suggests that our results are currently incomparable with ERNIE 2.0. Overall, the results in Table 6 illustrate that our method is also effective for the Chinese QA tasks.

9 Conclusion

This paper aims to enrich the NSP task to provide more document-level information in the pre-training. Motivated by the in-symmetric property of NSP, we propose to differentiate between different sentence orders by including PSP. Despite the simplicity, extensive experiments demonstrate that the model obtains a better ability in pair-wise semantic reasoning. Our work suggests that the document-level objective is effective, at least for the BERTbase model. In the future, we will investigate the way to take advantages of both large-scale training and our method.

⁶<http://www.baik.com>

⁷<http://www.zhihu.com>

⁸<http://news.sohu.com>

⁹<https://github.com/ymcui/cmrc2018/tree/master/baseline>

References

- Conneau, A.; Kiela, D.; Schwenk, H.; Barrault, L.; and Bordes, A. 2017. Supervised learning of universal sentence representations from natural language inference data. In *EMNLP*, 670–680. Association for Computational Linguistics.
- Conneau, A.; Rinott, R.; Lample, G.; Williams, A.; Bowman, S. R.; Schwenk, H.; and Stoyanov, V. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Cui, Y.; Che, W.; Liu, T.; Qin, B.; Yang, Z.; Wang, S.; and Hu, G. 2019. Pre-training with whole word masking for chinese BERT. *CoRR* abs/1906.08101.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Duan, N., and Tang, D. 2017. Overview of the NLPCC 2017 shared task: Open domain chinese question answering. In *Natural Language Processing and Chinese Computing - 6th CCF International Conference, NLPCC 2017, Dalian, China, November 8-12, 2017, Proceedings*, 954–961.
- Jean, S.; Cho, K.; Memisevic, R.; and Bengio, Y. 2015. On using very large target vocabulary for neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, 1–10.
- Kiros, R.; Zhu, Y.; Salakhutdinov, R.; Zemel, R. S.; Urtasun, R.; Torralba, A.; and Fidler, S. 2015. Skip-thought vectors. In *NIPS*, 3294–3302.
- Lai, G.; Xie, Q.; Liu, H.; Yang, Y.; and Hovy, E. H. 2017. RACE: large-scale reading comprehension dataset from examinations. In *EMNLP*, 785–794. Association for Computational Linguistics.
- Liu, X.; Chen, Q.; Deng, C.; Zeng, H.; Chen, J.; Li, D.; and Tang, B. 2018. LCQMC: A large-scale chinese question matching corpus. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, 1952–1962.
- Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR* abs/1907.11692.
- Logeswaran, L., and Lee, H. 2018. An efficient framework for learning sentence representations. In *ICLR*. OpenReview.net.
- Loshchilov, I., and Hutter, F. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.
- McCoy, T.; Pavlick, E.; and Linzen, T. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, 3428–3448.
- Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, 2383–2392.
- Rajpurkar, P.; Jia, R.; and Liang, P. 2018. Know what you don’t know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, 784–789.
- Shao, C.; Liu, T.; Lai, Y.; Tseng, Y.; and Tsai, S. 2018. DRCD: a chinese machine reading comprehension dataset. *CoRR* abs/1806.00920.
- Sun, Y.; Wang, S.; Li, Y.; Feng, S.; Chen, X.; Zhang, H.; Tian, X.; Zhu, D.; Tian, H.; and Wu, H. 2019a. ERNIE: enhanced representation through knowledge integration. *CoRR* abs/1904.09223.
- Sun, Y.; Wang, S.; Li, Y.; Feng, S.; Tian, H.; Wu, H.; and Wang, H. 2019b. ERNIE 2.0: A continual pre-training framework for language understanding. *CoRR* abs/1907.12412.
- Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2019a. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.
- Wang, W.; Bi, B.; Yan, M.; Wu, C.; Bao, Z.; Peng, L.; and Si, L. 2019b. Structbert: Incorporating language structures into pre-training for deep language understanding. *arXiv preprint arXiv:1908.04577*.
- Williams, A.; Nangia, N.; and Bowman, S. 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)*, 1112–1122. Association for Computational Linguistics.
- Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J. G.; Salakhutdinov, R.; and Le, Q. V. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR* abs/1906.08237.
- Zhou, J.; Cheng, X.; and Zhang, J. 2019. An end-to-end neural network framework for text clustering. *CoRR* abs/1903.09424.
- Zhu, Y.; Kiros, R.; Zemel, R. S.; Salakhutdinov, R.; Urtasun, R.; Torralba, A.; and Fidler, S. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, 19–27.