Symmetric Regularization based BERT for Pair-wise Semantic Reasoning

Weidi Xu*
Xingyi Cheng*
weidi.xwd,fanyin.cxy@alibaba-inc.com
Ant Financial Services Group
Hanzhou, Zhejiang, China

Kunlong Chen*
Taifeng Wang
kunlong.ckl,taifeng.wang@alibaba-inc.com
Ant Financial Services Group
Hanzhou, Zhejiang, China

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 is of great significance for downstream problems with sentence-pair input. Despite its effectiveness, 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 multi-class categorization task, which includes previous sentence prediction (PSP). This task encourages the model to learn the subtle semantics, thereby improves the ability of semantic understanding. Furthermore, by using a smoothing technique, the scopes of NSP and PSP are expanded into a broader range which includes close but nonsuccessive sentences. This simple method yields remarkable improvement against vanilla BERT. Our method consistently improves the performance on the NLI and MRC benchmarks by a large margin, including the challenging HANS dataset [5].

CCS CONCEPTS

• **Theory of computation** → *Semantics and reasoning*.

KEYWORDS

Natural language inference, BERT

ACM Reference Format:
Weidi Xu, Xingyi Cheng, Kunlong Chen, and Taifeng Wang. 2020. Symmetric Regularization based BERT for Pair-wise Semantic Reasoning. In *43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3397271.3401309

1 INTRODUCTION

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) [1] 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 [1], the NSP task is useful for the downstream NLI and MRC tasks.

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 (`DiffDoc`). Therefore, to discriminate against these sentences, BERT is prone to aggregating the shallow semantic, e.g., topic, neglecting context clues useful for detailed reasoning. This setting weakens the BERT model from learning specific semantic for inference.

Based on these considerations, we propose to include a previous sentence prediction (PSP) task with a `IsPrev` category, which is the symmetric label of `IsNext` of NSP. The input of samples with `IsPrev` is the reverse of those with `IsNext`. In addition, to further incorporating the document-level knowledge, NSP and PSP are extended with non-successive sentences. The advantages of our methods are two folds. (1) Learning the contrast between NSP and PSP forces the model to extract more detailed semantic, thereby the model is more 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. [1]

The proposed method yields a remarkable improvement in our experiments. We evaluate the ability of semantic reasoning on standard NLI and MRC benchmarks, including the challenging HANS dataset [2] [5]. Analytical work on HANS provides a more comprehensible perspective towards the proposed method.

*Both authors contributed equally to this research.

[1] Our method was released prior to ALBERT [3] and we also include non-successive sentences to leverage more document-level information.
[2] Heuristic Analysis for NLI Systems

SIGIR '20, July 25–30, 2020, Virtual Event, China

ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/3397271.3401309

2 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 in BERT 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 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 [1], the final pre-trained model is able to achieve 97%-98% accuracy on the NSP task. (2) BERT is order-sensitive, i.e., $f_{\text{BERT}}(\mathtt{A}, \mathtt{B}) \neq f_{\text{BERT}}(\mathtt{B}, \mathtt{A})$, while NSP is uni-directional. 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 [9] and 0.4% on QNLI after swapping the sentences in our experiments.

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.

2.1 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.

2.2 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 as illustrated by the confusion matrix in Figure 2. In particular, the pairs with IsNextInadj (IsPrevInadj) are likely to be labeled as IsNext and IsPrev. Therefore, a smooth technique is adopted to reduce the noise of these additional samples. It maps IsPrevInadj (IsNextInadj) to IsPrev (IsNext) and relaxes our confidence on the labels by 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, B is chosen as in Table 1:

Category	Definition
DiffDoc	The sentence from a different document.
IsNext	The adjacent following sentence.
IsPrev	The adjacent previous sentence.
IsNextInadj	The following sentence next to IsNext.
IsPrevInadj	The previous sentence next to IsPrev.

Table 1: The definition of different categories.

3 EXPERIMENTS

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. We used Adam 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 data used for pre-training is the same as BERT, i.e., English Wikipedia (2500M words) and BookCorpus (800M words). For the Masked LM task, we followed the same masking rate and settings as in BERT.

We explore three settings for comparison. (1) 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. (2) 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. (3) BERT-PNsmth: It uses the same data with BERT-PN5cls, except that the IsPrevInadj (IsNextInadj) label is mapped to IsPrev (IsNext) with a 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 beneficial.

3.1 GLUE

A popular benchmark for evaluation of language understanding is GLUE [8], which is a collection of three NLI tasks (MNLI, QNLI and RTE), three semantic textual similarity (STS) tasks (QQP, STS-B and MRPC), two text classification (TC) tasks (SST-2 and CoLA). Although the method is motivated for pair-wise reasoning, the results of other GLUE tasks 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 in all of the NLI tasks. The improvement on the RTE dataset is significant, i.e., 4% absolute gain over the BERT-Base. Besides NLI, our model also performs better than BERTBase in the STS tasks, which is semantically similar to the NLI tasks. The improvements suggest that the PSP task encourages the model

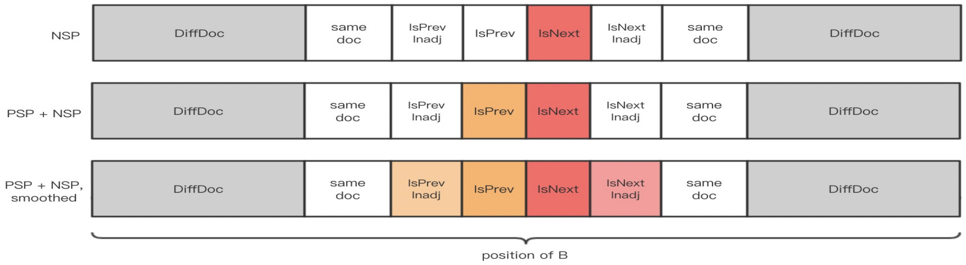

Figure 1: An illustration of the proposed method. B denotes the second input sentence. (1) Top: original NSP task. (2) Middle: 3-class categorization task with `DiffDoc`, `IsNext` and `IsPrev`. (3) Bottom: 3-class task, but with a wider scope of NSP and PSP.

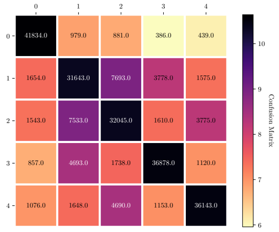

Figure 2: Confusion matrix of 226k sentence pair relationships from the trained BERT with extended 5 categories. The label distribution is 1:1:1:1:1.

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. The results of BERTBase (2:1), which has $\frac{2}{3}$ of `IsNext` and $\frac{1}{3}$ of `DiffDoc`, are also provided. Comparing BERTBase (2:1) with BERTBase, we can tell that the performance is insensitive to the data distribution. This means the improvement is not obtained from a different ratio of labels.

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.

3.2 HANS

McCoy et al. 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. BERT and other advanced models fail on this dataset and barely exceeds 0% accuracy in most cases [5].

Fig. 3 illustrates the accuracy of BERTBase and BERTBase-PNsmth on the HANS dataset. 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. 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.

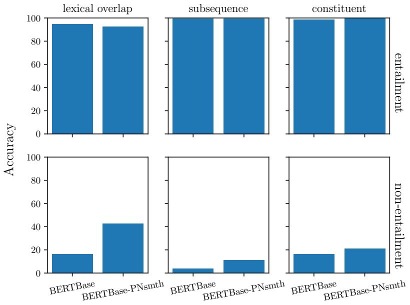

Figure 3: The accuracy on evaluation set of HANS. It has six sub-components, each defined by its correct label and the heuristic it addresses.

3.3 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 [7]. 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.

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 subtask 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 [6] 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, even surpassing the RoBERTa model.

Task Types	NLI			STS			TC		
Tasks	MNLI	QNLI	RTE	QQP	STS-B	MRPC	SST-2	CoLA	Average
Data Size	392k	108k	2.5k	363k	8.5k	3.5k	67k	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 (2:1)	84.4/83.4	90.4	65.0	71.1	86.9	87.1	92.7	52.4	79.3
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.

Dataset	Dev v1.1		Dev v2.0	
Metrics	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 [4].

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 [4]. All the listed models are trained on the Wikipedia and the Book Corpus datasets.

3.4 RACE

The ReAding Comprehension from Examinations (RACE) dataset [2] consists of 100K questions taken from English exams, and the answers are generated by human experts. As shown in Table 4, the proposed method significantly improves 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.

4 CONCLUSION

We generalized the standard NSP task to provide more document-level information in the pretraining. Specifically, we proposed a simple but important task, PSP, which effectively differentiates the dependency between different sentence orders, to achieve the symmetric property. Furthermore, we extended the scopes of NSP and PSP to capture more subtle semantic. The results of extensive experiments demonstrate that the newly pertained model acquires superior capacity in pair-wise semantic reasoning.

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[2] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard H. Hovy. 2017. RACE: Large-scale ReAding Comprehension Dataset From Examinations. In *EMNLP*. Association for Computational Linguistics, 785–794.

[3] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* (2019).

[4] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692 (2019). arXiv:1907.11692 http://arxiv.org/abs/1907.11692

[5] 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 Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. 3428–3448. https://www.aclweb.org/anthology/P19-1334/

[6] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*. 784–789. https://aclanthology.info/papers/P18-2124/p18-2124

[7] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100, 000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*. 2383–2392. http://aclweb.org/anthology/D/D16/D16-1264.pdf

[8] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. 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*. https://openreview.net/forum?id=rJ4km2R5t7

[9] 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)* (New Orleans, Louisiana). Association for Computational Linguistics, 1112–1122. http://aclweb.org/anthology/N18-1101