Transfer Fine-Tuning: A BERT Case Study

Yuki Arase[1⋆] and **Junichi Tsujii**[⋆2]
[1]Osaka University, Japan
⋆Artificial Intelligence Research Center (AIRC), AIST, Japan
[2]NaCTeM, School of Computer Science, University of Manchester, UK
arase@ist.osaka-u.ac.jp, j-tsujii@aist.go.jp

Abstract

A semantic equivalence assessment is defined as a task that assesses semantic equivalence in a sentence pair by binary judgment (*i.e.*, paraphrase identification) or grading (*i.e.*, semantic textual similarity measurement). It constitutes a set of tasks crucial for research on natural language understanding. Recently, BERT realized a breakthrough in sentence representation learning (Devlin et al., 2019), which is broadly transferable to various NLP tasks. While BERT's performance improves by increasing its model size, the required computational power is an obstacle preventing practical applications from adopting the technology. Herein, we propose to inject phrasal paraphrase relations into BERT in order to generate suitable representations for semantic equivalence assessment instead of increasing the model size. Experiments on standard natural language understanding tasks confirm that our method effectively improves a smaller BERT model while maintaining the model size. The generated model exhibits superior performance compared to a larger BERT model on semantic equivalence assessment tasks. Furthermore, it achieves larger performance gains on tasks with limited training datasets for fine-tuning, which is a property desirable for transfer learning.

1 Introduction

Paraphrase identification and semantic textual similarity (STS) measurements aim to assess semantic equivalence in sentence pairs. These tasks are central problems in natural language understanding research and its applications. In this paper, these tasks are defined as semantic equivalence assessments.

Sentence representation learning is the basis of assessing semantic equivalence. Unsupervised learning is becoming the preferred approach because it only requires plain corpora, which are now abundantly available. In this approach, a model is pre-trained to generate generic sentence representations that are broadly transferable to various natural language processing (NLP) tasks. Subsequently, it is fine-tuned to generate specific representations for solving a target task using an annotated corpus. Considering the high costs of annotation, a pre-trained model that efficiently fits the target task with a smaller amount of annotated corpus is desired.

Recently, Bidirectional Encoder Representations from Transformers (BERT) realized a breakthrough, which dramatically improved sentence representation learning (Devlin et al., 2019). BERT pre-trains its encoder using language modeling and by discriminating surrounding sentences in a document from random ones. Pre-training in this manner allows distributional relations between sentences to be learned. Intensive efforts are currently being made to pre-train larger models by feeding them enormous corpora for improvement (Radford et al., 2019; Yang et al., 2019). For example, a large model of BERT has 340M parameters, which is 3.1 times larger than its smaller alternative. Although such a large model achieves performance gains, the required computational power hinders its application to downstream tasks.

Given the importance of natural language understanding research, we focus on sentence representation learning for semantic equivalence assessment. Instead of increasing the model size, we propose the injection of semantic relations into a pre-trained model, namely BERT, to improve performance. Phang et al. (2019) showed that BERT's performance on downstream tasks improves by simply inserting extra training on data-rich supervised tasks. Unlike them, we inject semantic relations of finer granularity using phrasal paraphrase alignments automatically iden-

tified by Arase and Tsujii (2017) to improve semantic equivalent assessment tasks. Specifically, our method learns to discriminate phrasal and sentential paraphrases on top of the representations generated by BERT. This approach explicitly introduces the concept of the phrase to BERT and supervises semantic relations between phrases. Due to studies on sentential paraphrase collection (Lan et al., 2017) and generation (Wieting and Gimpel, 2018), a million-scale paraphrase corpus is ready for use. We empirically show that further training of a pre-trained model on relevant tasks transfers well to downstream tasks of the same kind, which we name as *transfer fine-tuning*.

The contributions of our paper are:

- We empirically demonstrate that transfer fine-tuning using paraphrasal relations allows a smaller BERT to generate representations suitable for semantic equivalence assessment. The generated model exhibits superior performance to the larger BERT while maintaining the small model size.
- Our experiments indicate that phrasal paraphrase discrimination contributes to representation learning, which complements simpler sentence-level paraphrase discrimination.
- Our model exhibits a larger performance gain over the BERT model for a limited amount of fine-tuning data, which is an important property of transfer learning.

We hope that this study will open up one of the crucial research directions that will make the approach of pre-trained models more practically useful. Our codes, datasets, and the trained models will be made publicly available at our web site.

2 Related Work

Sentence representation learning is an active research area due to its importance in various downstream tasks. Early studies employed supervised learning where a sentence representation is learned in an end-to-end manner using an annotated corpus. Among these, the importance of phrase structures in representation learning has been discussed (Tai et al., 2015; Wu et al., 2018). In this paper, we use structural relations in sentence pairs for sentence representations. Specifically, we employ phrasal paraphrase relations that introduce the notion of a phrase to the model.

The research focus of sentence representation learning has moved toward unsupervised learning in order to exploit the gigantic corpus. Skip-Thought, which was an early learning attempt, learns to generate surrounding sentences given a sentence in a document (Kiros et al., 2015). This can be interpreted as an extension of the distributional hypothesis on sentences. Quick-Thoughts, a successor of Skip-Thought, conducts classification to discriminate surrounding sentences instead of generation (Logeswaran and Lee, 2018). GenSen combines these approaches in massive multi-task learning (Subramanian et al., 2018) based on the premise that learning dependent tasks enriches sentence representations.

Embeddings from Language Models (ELMo) made a significant step forward (Peters et al., 2018). ELMo uses language modeling with bidirectional recurrent neural networks (RNN) to improve word embeddings. ELMo's embedding contributes to the performance of various downstream tasks. OpenAI GPT (Radford et al., 2018) replaced ELMo's bidirectional RNN for language modeling with the Transformer (Vaswani et al., 2017) decoder. More recently, BERT combined the approaches of Quick-Thoughts (*i.e.*, a next-sentence prediction approach) and language modeling on top of the deep bidirectional Transformer. BERT broke the records of the previous state-of-the-art methods in eleven different NLP tasks. While BERT's pre-training generates generic representations that are broadly transferable to various NLP tasks, we aim to fit them for semantic equivalence assessment by injecting paraphrasal relations. Liu et al. (2019) showed that BERT's performance improves when fine-tuning with a multi-task learning setting, which is applicable to our trained model for further improvement.

3 Background

3.1 Phrase Alignment for Paraphrases

In order to obtain phrasal paraphrases, we used the phrase alignment method proposed in (Arase and Tsujii, 2017) and apply it to our paraphrase corpora. The alignment method aligns phrasal paraphrases on the parse forests of a sentential paraphrase pair as illustrated in Fig. 1.

According to the evaluation results reported in (Arase and Tsujii, 2017), the precision and recall of alignments are 83.6% and 78.9%, which are 89% and 92% of those of humans, respec-

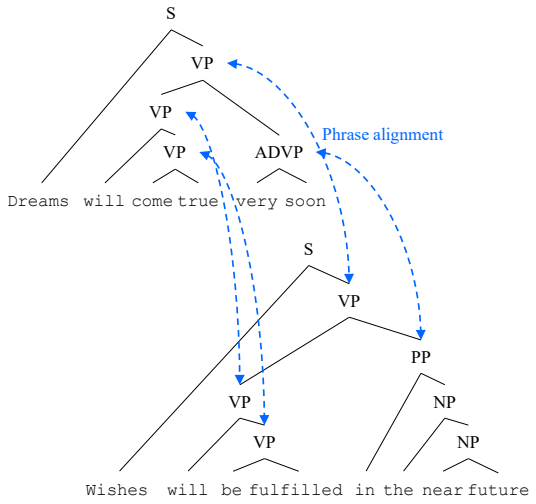

Figure 1: Phrasal paraphrases are obtained from (Arase and Tsujii, 2017); arrows indicate phrase alignments.

tively. Although alignment errors occur, previous studies show that neural networks are relatively robust against noise in a training corpus and still benefit from extra supervisions as demonstrated in (Edunov et al., 2018; Prabhumoye et al., 2018).

We collect all the spans of phrases in a sentential paraphrase pair and their alignments as pairs of phrase spans. Because the phrase alignment method allows unaligned phrases, not all of the phrases have aligned counterparts.

3.2 Pre-Training on BERT

BERT is a bidirectional Transformer that generates a sentence representation by conditioning both the left and right contexts of a sentence. A pre-trained BERT model can be easily fine-tuned for a wide range of tasks by just adding a fully-connected layer, without any task-specific architectural modifications. BERT achieved state-of-the-art performances for eleven NLP tasks, thereby outperforming the previous state-of-the-art methods by a large margin.

Pre-training in BERT accomplishes two tasks. The first task is masked language modeling, where some words in a sentence are randomly masked and the model then predicts them from the context. This task design allows the representation to fuse both the left and the right context. The second task predicts whether a pair of sentences are consecutive in a document to learn the relation between the sentences. Specifically, as illustrated in Fig. 2, BERT takes two sentences as input that are concatenated by a special token `[SEP]`.[1] The first token of every input is always the special token of `[CLS]`. The final hidden state corresponding to this `[CLS]` token is regarded as an aggregated representation of the input sentence pair. This is used to predict whether the sentence pair is composed of consecutive sentences in a document or not during pre-training.

BERT has a deep architecture. The BERT-base model has 12 layers of 768 hidden size and 12 self-attention heads. The BERT-large model has 24 layers of 1024 hidden size and 16 self-attention heads. Both BERT-bese and BERT-large models were pre-trained using BookCorpus (Zhu et al., 2015) and English Wikipedia (in total 3.3B words).

[1] Throughout the paper, `typewriter font` represents tokens and labels.

Algorithm 4.1 Paraphrasal Relation Injection

Input: Paraphrase sentence pairs $P = \{\langle s, t\rangle\}$, a pre-trained BERT model

Obtain a set of phrase alignments A as pairs of spans for each $\langle s, t\rangle \in P$
WordPiece tokenization of P
Accommodate phrase spans in A to BERT's token indexing: $A = \{\langle (j, k), (m, n)\rangle\}$
repeat
  for all mini-batch $b_t \in \{\langle P_i, A_i\rangle\}$ **do**
    Encode b_t by the BERT model
    Compute loss: L(Θ)
      For phrasal paraphrase task: $L_p(\Theta)$
      For sentential paraphrase task: $L_s(\Theta)$
      $L(\Theta) = L_p(\Theta) + L_s(\Theta)$
    Compute gradient: $\nabla(\Theta)$
    Update the model parameters
until convergence

4 Transfer Fine-Tuning with Paraphrasal Relation Injection

We inject semantic relations between a sentence pair into a pre-trained BERT model through classification of phrasal and sentential paraphrases. After the training, the model can be fine-tuned in exactly the same manner as with BERT models.

4.1 Overview

Algorithm 4.1 provides an overview of our method. It takes a sentential paraphrase pair $\langle s, t\rangle$ as an input, which are referred to as the source and target, respectively, for the sake of clarity. First, a set of phrase alignments A is obtained for $\langle s, t\rangle$

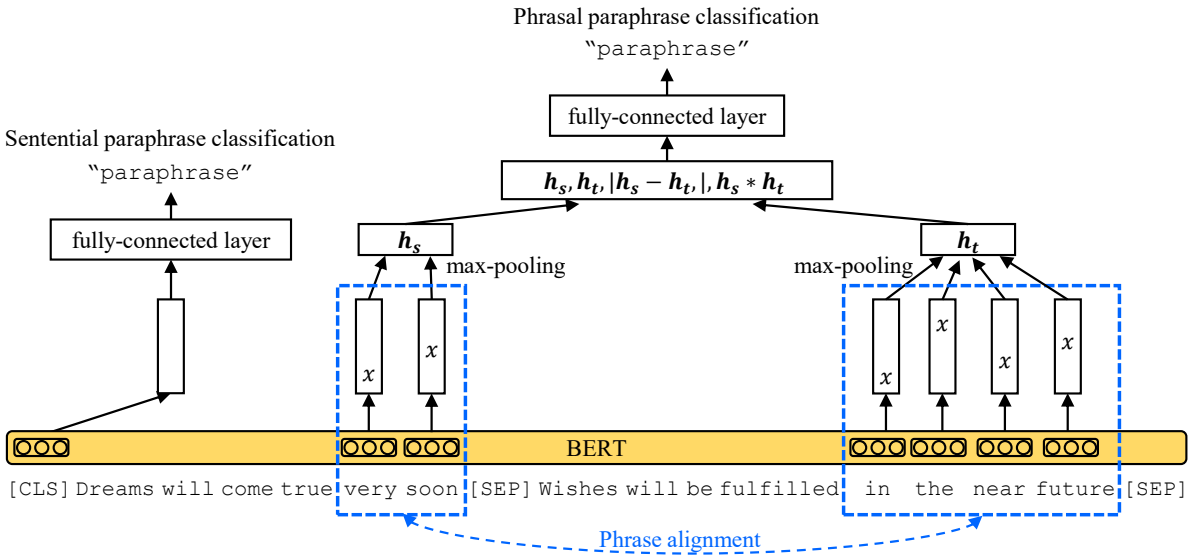

Figure 2: Our method injects semantic relations to sentence representations through paraphrase discrimination.

(line 1) as described in Sec. 3.1. Because BERT uses sub-words as a unit instead of words, all the input sentences are tokenized (line 2) by WordPiece (Wu et al., 2016). In addition, t is concatenated to s when being input to the BERT model, where the first token should always be `[CLS]` and the sentence pair is separated by `[SEP]` as described in Sec. 3.2. In order to accommodate to these factors, phrase spans in alignments A are adjusted accordingly (line 3).

Our method learns to discriminate phrasal and sentential paraphrases simultaneously as illustrated in Fig. 2. Cross-entropy is used as the loss functions for both tasks (line 8, 9).

Phrasal Paraphrase Classification The middle part of Fig. 2 illustrates phrasal paraphrase classification. We first generate phrase embedding for each aligned phrase as follows. The tokenized sentence pair is encoded by the BERT model. For the input sequence of N tokens $\{w_i\}_{i=1,\ldots,N}$, we obtain the final hidden states $\{\mathbf{h}_i\}_{i=1,\ldots,N}$ (*i.e.*, output of the bidirectional Transformer):

$$\mathbf{h}_i = \text{Transformer}(w_1, \ldots, w_N),$$

where $\mathbf{h}_i \in \mathbb{R}^\lambda$ and λ is the hidden size. We then combine $\{\mathbf{h}_i\}_i$ for a phrase pair with an alignment $\langle (j,k),(m,n)\rangle$ where $2 \leq j < k < m < n \leq N-1$ represent indexes of the beginning and ending of phrases (recall that the first and last tokens are always special tokens in BERT). As a combination function, we apply max-pooling that showed strong performance in (Conneau et al., 2017) to generate a representation of source and target phrases:

$$\mathbf{h}_s = \text{max-pooling}(\mathbf{h}_j, \ldots, \mathbf{h}_k), \quad (1)$$
$$\mathbf{h}_t = \text{max-pooling}(\mathbf{h}_m, \ldots, \mathbf{h}_n). \quad (2)$$

The max-pooling$(\cdot)$ function selects the maximum value over each dimension of the hidden units.

Then $\mathbf{h}_s$ and $\mathbf{h}_t$ are converted to a single vector. To extract relations between $\mathbf{h}_s$ and $\mathbf{h}_t$, three matching methods are used (Conneau et al., 2017): (a) concatenating the representations $(\mathbf{h}_s, \mathbf{h}_t)$, (b) taking the element-wise product $\mathbf{h}_s * \mathbf{h}_t$, and (c) finding the absolute element-wise difference $|\mathbf{h}_s - \mathbf{h}_t|$. The final vector of $\mathbb{R}^{4\lambda}$ is fed into a classifier.[2]

Because our method aims to generate representations for semantic equivalence assessment, the classifier should be simple (Logeswaran and Lee, 2018). Otherwise, a sophisticated classifier would fit itself with the task instead of the representations. We use a single fully-connected layer culminating in a softmax layer as our classifier.

Previous studies have calculated interactions between words (He and Lin, 2016) and phrases (Chen et al., 2017) using the final hidden states of bidirectional RNN or recursive neural networks when composing a sentence representation. Our approach differs from these by giving explicit supervision of which phrase pairs have semantic interactions (*i.e.*, paraphrases).

[2]Our follow-up study confirms that a simpler feature generation improves the generality of our model to contribute not only to semantic equivalent assessment but also natural language inference. For details, please refer to the Appendix.

Negative Example Selection In paraphrase identification, non-paraphrases with large lexical differences are easy to discriminate. Discrimination becomes far more difficult when they contain a number of identical or related words. To effectively supervise the model by solving difficult discrimination problems, we designed a three-way classification task: discrimination of `paraphrase`, `random`, and `in-paraphrase` pairs.

The `random` examples are generated by pairing s to a random sentence t' from the training corpus, and then pairing all phrases in s to randomly chosen phrases in t'. The `in-paraphrase` examples aim to make the discrimination problem difficult, which requires distinguishing true paraphrases and phrases in the paraphrasal sentence pair t. These may provide sub-phrases or ancestor phrases of true paraphrases as difficult negative examples, which tend to retain the same topic and similar wordings. To prepare such examples, for each phrase pair $\langle (j,k),(m,n)\rangle \in A$, the target span (m,n) is replaced by a randomly chosen phrase span in t.

Phrasal paraphrase classification aims to give explicit supervision of semantic relations among phrases in representation learning. It also introduces structures in sentences, which is completely missed in BERT's pre-training. Swayamdipta et al. (2018) showed that supervision of phrase-based syntax improves the performance of a task relevant to semantics, *e.g.*, semantic role labeling.

Sentential Paraphrase Classification The left side of Fig. 2 illustrates the sentential paraphrase classification. The process is simple; the final hidden state of the `[CLS]` token, *i.e.*, $\mathbf{h}_1$, is fed into a classifier to discriminate whether a sentence pair is a paraphrase or a random sentence combination. Note that these random sentence pairs provide `random` phrases for the phrasal paraphrase classification described above.

4.2 Training Setting

We collected paraphrases from various sources as summarized in Table 1, which shows the numbers of sentential and phrasal paraphrase pairs after phrase alignment.[3] All the datasets were downloaded from the Linguistic Data Consortium (LDC) or authors' websites. The following bullets describe the sources.

Source	Sentence	Phrase
NIST OpenMT	47k	711k
Simple Wikipedia	97k	1.4M
Twitter URL corpus	50k	396k
Para-NMT	3.9M	26.7M
Total	4.1M	29.2M

Table 1: Numbers of sentential and phrasal paraphrases after the phrase alignment process.

- NIST OpenMT[4]: We randomly paired reference translations of the same source sentence as was done in (Arase and Tsujii, 2017).
- Twitter URL corpus (Lan et al., 2017): This corpus was collected from Twitter by linking tweets through shared URLs. We used a three-month collection of paraphrases.[5]
- Simple Wikipedia (Kauchak, 2013): This corpus aligned English Wikipedia and Simple English Wikipedia for text simplification. We used "sentence-aligned, version 2.0."[6]
- Para-NMT (Wieting and Gimpel, 2018): This corpus was created by translating the Czech side of a large Czech-English parallel corpus and pairing the translated English and originally target-side English as paraphrases. We used "Para-nmt-5m-processed."[7]

Note that these sentential and phrasal paraphrases are obtained by automatic methods. On the contrary, dataset creation for downstream tasks generally requires expensive human annotation.

We employed the pre-trained BERT-base model[8] and conducted paraphrase classification using the collected paraphrase corpora. Adam (Kingma and Ba, 2015) was applied as an optimizer with a learning rate of $5\mathrm{e}-5$. A dropout probability was 0.2 for the fully-connected layers in the classifiers. A development set and a test

[3] The numbers of sentential paraphrase pairs were reduced due to parsing and alignment failures.

[4] LDC catalogue number: LDC2010T14, LDC2010T17, LDC2010T21, LDC2010T23, LDC2013T03

[5] https://github.com/lanwuwei/Twitter-URL-Corpus

[6] http://www.cs.pomona.edu/~dkauchak/simplification/data.v2/sentence-aligned.v2.tar.gz

[7] https://drive.google.com/file/d/19NQ87gEFYu3zOIp_VNYQZgmnwRuSIyJd/view?usp=sharing

[8] https://github.com/google-research/bert

set, each with 50k sentence pairs, were subtracted from the paraphrase corpus. The rest of the corpus was used for training. The training was conducted on four NVIDIA Tesla V100 GPUs with a batch-size of 100. Early stopping was applied to stop training at the second time decrease in the accuracy of the phrasal paraphrase classification, which was measured on the development set. The final test-set accuracies were 98.1% and 99.9% for phrasal and sentential paraphrase classification, respectively.

5 Evaluation Setting

5.1 Hypotheses to Verify

BERT's pre-training learns to generate sentence representations broadly transferable to different NLP tasks. In contrast, our method gives more direct supervision to generate representations suitable for semantic equivalence assessment tasks. We set up the following hypotheses on features of our method, which will be empirically verified through evaluation:

- H1 Our method contributes to semantic equivalence assessment tasks.
- H2 Our method achieves improvement on downstream tasks that only have small amounts of training datasets for fine-tuning.
- H3 Our method moderately improves tasks if they are relevant to semantic equivalence assessment.
- H4 Our training does not transfer to distant downstream tasks that are independent to semantic equivalence assessment.
- H5 Phrasal and sentential paraphrase classification complementarily benefits sentence representation learning.

5.2 GLUE Datasets

We empirically verified the hypotheses H1 to H5 using the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2019)[9], which is the standard benchmark and provides collections of datasets for natural language understanding tasks. Table 2 summarizes the tasks and evaluation metrics at GLUE. All the scores reported in this paper are computed at the GLUE evaluation server unless stated otherwise. Accuracies on MRPC and QQP and Spearman correlation on STS-B are omitted due to space limitations. Note that they showed the same trends as F1 and Pearson correlation, respectively, in our experiment. WNLI was excluded because the GLUE web site reports its issues.[10]

Corpus	Task	Metrics
MRPC	paraphrase	F1
STS-B	STS	Pearson corr.
QQP	paraphrase	F1
MNLI-m	in-domain NLI	accuracy
MNLI-mm	cross-domain NLI	accuracy
RTE	NLI	accuracy
QNLI	QA/NLI	accuracy
SST	sentiment	accuracy
CoLA	acceptability	Matthews corr.

Table 2: GLUE tasks and evaluation metrics.

GLUE tasks can be categorized according to their aims as follows.

Semantic Equivalence Assessment Tasks (MRPC, STS-B, QQP) These are the primary targets of our method, which are used to verify hypothesis H1. Paraphrase identification assesses semantic equivalence in a sentence pair by binary judgments. Microsoft Paraphrase Corpus (MRPC) (Dolan et al., 2004) consists of sentence pairs drawn from news articles, while Quora Question Pairs (QQP)[11] consists of question pairs from the community QA website.

STS assesses semantic equivalence by grading. STS benchmark (STS-B) (Cer et al., 2017) provides sentence pairs drawn from heterogeneous sources, which are human-annotated with a level of equivalence from 1 to 5.

NLI Tasks (MNLI-m/mm, RTE, QNLI) We use natural language inference (NLI) tasks to verify hypothesis H3 because they constitute a class of problems relevant to semantic equivalence assessment. NLI tasks are different from semantic equivalence assessment in that they often require logical inference and understanding of commonsense knowledge. The Multi-Genre Natural Language Inference Corpus (MNLI) (Williams et al., 2018) is a crowd-sourced corpus and covers heterogeneous domains. MNLI-m is an in-domain

[9] https://gluebenchmark.com/

[10] https://gluebenchmark.com/faq

[11] https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs

Task / Model	Semantic Equivalence			NLI			Single-Sent.	
	MRPC	STS-B	QQP	MNLI (m/mm)	RTE	QNLI	SST	CoLA
BERT-base	88.3	84.7	71.2	84.3/83.0	59.8	89.1	93.3	52.7
BERT-large	88.6	86.0	**72.1**	**86.2/85.5**	**65.5**	**92.7**	**94.1**	**55.7**
Transfer Fine-Tuning	**89.2**	**87.4**	71.2	83.9/83.1	64.8	89.3	93.1	47.2

Table 3: GLUE test results scored by the GLUE evaluation server. The best scores are represented in **bold** and scores higher than those of BERT-base are underlined.

NLI task while MNLI-mm is a cross-domain NLI task. The Recognizing Textual Entailment (RTE) corpus[12] was created from news and Wikipedia. Question-answering NLI (QNLI) was created from The Stanford Question Answering Dataset (Rajpurkar et al., 2016) on which all the sentences were drawn from Wikipedia.

Single-Sentence Tasks (SST, CoLA) We use these tasks to verify hypothesis H4. They aim to estimate features in a single sentence, which has little interaction with semantic equivalence assessment in a sentence pair. The Stanford Sentiment Treebank (SST) (Socher et al., 2013) task is a binary sentiment classification, while The Corpus of Linguistic Acceptability (CoLA) (Warstadt et al., 2018) task is a binary classification of grammatical acceptability.

5.3 Fine-Tuning on Downstream Tasks

Once trained, our model can be used in exactly the same manner as the pre-trained BERT models. For fine-tuning our models and replicating BERT's results under the same setting, we set the hyper-parameter values to those recommended in (Devlin et al., 2019): a batch size of 32, a learning rate of $3e-5$, the number of training epochs to 4, and a dropout probability of 0.1. We fine-tuned all the models on downstream tasks using the script provided in the Pytorch version of BERT.[13] For STS-B, we modified the script slightly to conduct regression instead of classification. All other hyper-parameters were set to the default values defined in the BERT's fine-tuning script.

For fair comparison, we kept the same hyper-parameter settings described above across all tasks and models. Phang et al. (2019) discussed that BERT performances become unstable when a training dataset with fine-tuning is small. In our evaluation, performances were stable when setting the same hyper-parameters, but further investigation is our future work.

6 Results and Discussion

6.1 Effect on Semantic Equivalence Assessment Tasks

Table 3 shows fine-tuning results on GLUE; our model, denoted as Transfer Fine-Tuning, is compared against BERT-base and BERT-large. The first set of columns shows the results of semantic equivalence assessment tasks. Our model outperformed BERT-base on MRPC (+0.9 points) and STS-B (+2.7 points). Furthermore, it outperformed even BERT-large by 0.6 points on MRPC and by 1.4 points on STS-B, despite BERT-large having 3.1 times more parameters than our model. Devlin et al. (2019) described that the next-sentence prediction task in BERT's pre-training aims to train a model that understands sentence relations. Herein, we argue that such relations are effective at generating representations broadly transferable to various NLP tasks, but are too generic to generate representations for semantic equivalence assessment tasks. Our method allows semantic relations between sentences and phrases that are directly useful for this class of tasks to be learned.

These results support hypothesis H1, indicating that our approach is more effective than blindly enlarging the model size. A smaller model size is desirable for practical applications. We have also applied our method on the BERT-large model, but its performance was not much improved to warrant the larger model size. Further investigation regarding pre-trained model sizes is our future work.

6.2 Effect of the Amount of Fine-Tuning Datasets

Our method did not improve upon BERT-base for QQP. We consider this is because a large QQP training set (364k sentence pairs) allows the BERT model to converge to a certain optimum. This also

[12] https://aclweb.org/aclwiki/Recognizing_Textual_Entailment

[13] `run_classifier.py` in https://github.com/huggingface/pytorch-pretrained-BERT

Task	Train. size	BERT-base	Transfer Fine-Tuning
MRPC	1k	81.6	88.1 (+6.5)
	all (3.7k)	89.4	90.2 (+0.8)
STS-B	1k	83.4	86.2 (+2.8)
	all (5.7k)	88.1	90.1 (+2.0)
QQP	1k	69.9	71.4 (+1.5)
	5k	75.5	76.3 (+0.8)
	10k	77.0	77.6 (+0.6)
	20k	79.6	79.5 (−0.1)
	all (364k)	87.7	87.7 (±0.0)

Table 4: Development set scores of the BERT-base model and our model (and their differences) that were fine-tuned using subsamples and full-size training sets.

relates to hypothesis H2.

To investigate the effect of the sizes of training sets, we fine-tuned our model and BERT-base for semantic equivalence assessment tasks using randomly subsampled training sets. Table 4 shows scores on the development sets.[14] The result clearly indicates that our method is more beneficial when a training dataset is limited on a downstream task, which supports hypothesis H2. This property is preferable for a transfer learning scenario that unsupervised sentence representation learning assumes.

Another factor that may affect the performance is domain mismatch between our paraphrase corpora and QQP corpus. The former was mostly collected from news while the latter was extracted from a social QA forum. In the future, we will investigate the effects of domains by generating multi-domain paraphrase corpora using a method proposed by Wieting and Gimpel (2018).

6.3 Effect on NLI Tasks

The second set of columns in Table 3 shows the results on NLI tasks. Our model presents moderate improvements on most NLI tasks, which supports hypothesis H3. We consider this is because the majority of NLI tasks that require inferences in one-direction, contrary to bi-directional entailment relations of paraphrases, are uni-directional.

Another reason is that our elaborate feature generation for the phrasal paraphrase classifier tightly fits the model for paraphrase identification. This contributes to performance improvements on this task, but sacrifices the model's generality on relevant tasks. We tackle this issue in our follow-up study reported in the Appendix.

Among NLI tasks, our model largely outperformed BERT-base by 5.0 point on RTE. This may be again due to the property of our method that brings improvement on tasks with a limited training set as RTE has only 2.5k training sentence pairs.

6.4 Effect on Single-Sentence Tasks

The last two columns of Table 3 show results on single-sentence tasks; SST and CoLA, which are the most distant tasks from paraphrase classification. Our model presents a slightly lower score on SST compared to BERT-base and performed poorly on CoLA.

One potential reason for this degradation is that our training takes a sentence pair as input, which may weaken the ability to model a single sentence. Another cause is attributable to similarities between our training and fine-tuning tasks. For SST, sentiment analysis could be adversarial toward paraphrase discrimination tasks. Although paraphrasal sentences tend to have the same sentiments, sentences with the same sentiments do not generally hold paraphrastic relations. For CoLA, semantic relations unlikely contribute to determining grammatical acceptability, as required by CoLA task.

Together with the results in Sec. 6.3, hypothesis H4 is supported; the effectiveness of our method depends on relevance between paraphrase discrimination and downstream tasks. Our future work will be to examine what characteristics of NLP tasks make our method less effective.

6.5 Ablation Study

To verify hypothesis H5, we conducted an ablation study that investigates independent effects of sentential and phrasal paraphrase classification. Table 5 shows the results; the last three rows show performances when conducting only sentential paraphrase classification, phrasal paraphrase classification, and binary classification of `paraphrase` and `in-paraphrase` pairs, respectively. All the models were fine-tuned in the same manner as described in Sec. 5.3.

First, the results support the hypothesis; sentential and phrasal paraphrase classification complements each other on sentence representation learning. Our model achieved its best scores

[14] We used the development set because the GLUE server allows only two submissions per day. Note that the number of training epochs for fine-tuning is fixed in our experiments, hence, the development set was not used for other purposes.

Task / Model	Semantic Equivalence			NLI			Single-Sent.	
	MRPC	STS-B	QQP	MNLI (m/mm)	RTE	QNLI	SST	CoLA
Transfer Fine-Tuning	**89.2**	87.4	**71.2**	83.9/**83.1**	64.8	89.3	93.1	47.2
BERT-base	88.3	84.7	**71.2**	**84.3**/83.0	59.8	89.1	**93.3**	**52.7**
+sentence	88.2	**87.6**	71.1	83.2/82.8	**66.2**	**90.2**	92.4	39.8
+3way-PP	88.2	85.8	70.9	82.9/81.9	65.8	88.0	91.3	32.6
+binary-PP	87.7	82.8	70.7	83.7/82.2	61.2	87.6	92.5	42.1

Table 5: Results of the ablation study where the best scores are represented in **bold** and scores higher than those of BERT-base are underlined. The last three rows show performances when conducting only sentential paraphrase classification (+sentence), phrasal paraphrase classification (+3way-PP), and binary classification of phrasal paraphrase (+binary-PP), respectively.

on MRPC, MNLI-m/mm, SST, and CoLA tasks by conducting both sentential and phrasal paraphrase classification simultaneously. Interestingly, these scores are higher than those when sentential and phrasal paraphrase classification are conducted independently. This is reasonable considering the process of fine-tuning. Sentential paraphrase classification directly affects the representation of `[CLS]`, which is the primary tuning factor in fine-tuning for downstream tasks. Alternatively, phrasal paraphrase classification affects representations of phrases, which are the basis for generating the `[CLS]` representation. Simultaneously conducting both sentential and phrasal paraphrase classification thus creates synergy.

It is also obvious that the three-way classification of phrasal paraphrases, on which the model discriminates paraphrases, random combinations of phrases from a random pair of sentences, and random combinations of phrases in a paraphrasal sentence pair, is superior to binary classification. This shows that discriminating random combinations of phrases, which is a simpler and easier task, also contributes to representation learning.

7 Conclusion

We empirically demonstrate that sentential and phrasal paraphrase relations help sentence representation learning. While BERT's pre-training aims to generate generic representations transferable to a broad range of NLP tasks, our method generates representations suitable for the class of semantic equivalence assessment tasks. Our method achieves performance gains while maintaining the model size. Furthermore, it exhibits improvement on downstream tasks with limited amounts of training datasets for fine-tuning, which is a property crucial for transfer learning.

In the future, we plan to investigate the effects of our method on different sizes of BERT models. Additionally, we will apply our model to improve the alignment quality of the phrase alignment model.

Acknowledgments

We appreciate the anonymous reviewers for their insightful comments and suggestions to improve the paper. This work was supported by JST, ACT-I, Grant Number JPMJPR16U2, Japan.

References

Yuki Arase and Jun'ichi Tsujii. 2017. Monolingual phrase alignment on parse forests. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1–11, Copenhagen, Denmark.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*, pages 1–14, Vancouver, Canada.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced lstm for natural language inference. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1657–1668, Vancouver, Canada.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 670–680, Copenhagen, Denmark.

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 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 350–356, Geneva, Switzerland.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 489–500, Brussels, Belgium.

Hua He and Jimmy Lin. 2016. Pairwise word interaction modeling with deep neural networks for semantic similarity measurement. pages 937–948, San Diego, California.

David Kauchak. 2013. Improving text simplification language modeling using unsimplified text data. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1537–1546, Sofia, Bulgaria.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Proceedings of Conference on Neural Information Processing Systems (NeurIPS)*, pages 3294–3302.

Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. 2017. A continuously growing dataset of sentential paraphrases. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1224–1234, Copenhagen, Denmark.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. *arXiv*, 1901.11504.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 2227–2237, New Orleans, Louisiana.

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2019. Sentence encoders on STILTs: Supplementary training on intermediate labeled-data tasks. *arXiv*, 1811.01088.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 866–876, Melbourne, Australia.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Technical report, OpenAI.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2383–2392, Austin, Texas.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1631–1642, Seattle, Washington, USA.

Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J Pal. 2018. Learning general purpose distributed sentence representations via large scale multi-task learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Swabha Swayamdipta, Sam Thomson, Kenton Lee, Luke Zettlemoyer, Chris Dyer, and Noah A. Smith. 2018. Syntactic scaffolds for semantic structures. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3772–3782.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *ACL-IJCLNLP*, pages 1556–1566, Beijing, China.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of Conference on Neural Information Processing Systems (NeurIPS)*, pages 5998–6008.

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 *Proceedings of the International Conference on Learning Representations (ICLR)*.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2018. Neural network acceptability judgments. *arXiv*.

John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 451–462, Melbourne, Australia.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. pages 1112–1122, New Orleans, Louisiana.

Wei Wu, Houfeng Wang, Tianyu Liu, and Shuming Ma. 2018. Phrase-level self-attention networks for universal sentence encoding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3729–3738, Brussels, Belgium.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv*, 1810.04805.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. *arXiv*, 1906.08237.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of International Conference on Computer Vision (ICCV)*, pages 19–27.

Appendix: Transfer Fine-Tuning with Simple Features

To further investigate effects of transfer fine-tuning using paraphrase relations on BERT, we designed a model that generates a simplest feature to input into the classifier in Fig. 2. We assume that this method transmits learning signals to the underlying BERT in a more effective manner. Specifically, we use mean-pooling to generate representations of source and target phrases in Eq. (1) and Eq. (2), respectively. These representations are simply concatenated as a feature representation and then fed into the classifier.

Table 6 compares this new model (denoted as Simple Transfer Fine-Tuning) to BERT models as well as our model with the elaborate feature generation described in Sec. 4 (denoted as Transfer Fine-Tuning) on semantic equivalent assessment and NLI tasks of GLUE benchmark. Table 7 reports an ablation study. The results and findings are summarized as follows.

- Our model with simple feature generation (Simple Transfer Fine-Tuning) on BERT-base outperformed BERT on both semantic equivalent assessment and NLI tasks. Furthermore, it performed on-par against BERT-large on MRPC and outperformed it on STS-B and RTE, despite BERT-large having 3.1 times more parameters than our model.

- The same trend was confirmed on the model trained on BERT-large, where our model outperformed BERT-large on all the tasks except QNLI.

- Simple Transfer Fine-Tuning also outperformed our model with elaborate feature generation (Transfer Fine-Tuning) on all semantic equivalent assessment and NLI tasks except MRPC. This result implies that elaborate feature generation tightly fits the model to paraphrase identification while sacrifices its generality to relevant tasks. Further investigation will be our future work.

- Sentential and phrasal paraphrase classification complements each other on sentence representation learning when using simple feature generation, as also confirmed when using the elaborate feature generation in Table 5. Simple Transfer Fine-Tuning achieved higher scores on STS-B, RTE, and QNLI tasks than models trained either with only sentential (+sentence) or phrasal paraphrase (+3way-PP [Simple Feature]) classification.

- Simple feature generation improves the performance of the model trained with only phrasal paraphrase classification; +3way-PP

Task / Model	Semantic Equivalence			NLI		
	MRPC	STS-B	QQP	MNLI (m/mm)	RTE	QNLI
BERT-base	88.3	84.7	71.2	84.3/83.0	59.8	89.1
Transfer Fine-tuning	89.2	87.4	71.2	83.9/83.1	64.8	89.3
Simple Transfer Fine-Tuning	88.6	**87.7**	71.5	84.7/83.6	67.0	91.1
BERT-large	88.6	86.0	72.1	86.2/85.5	65.5	**92.7**
Simple Transfer Fine-Tuning	**89.9**	87.1	**72.5**	**86.5/85.6**	**68.2**	92.2

Table 6: Test results on semantic equivalence assessment and NLI tasks scored by the GLUE evaluation server. The best scores for each task are represented in **bold**. The scores higher than those of BERT counterparts (against BERT-base and BERT-large, respectively) are underlined. Our models with simple feature generation (Simple Transfer Fine-Tuning) consistently outperformed the BERT models and achieved the best scores for six out of seven tasks.

Task / Model	Semantic Equivalence			NLI		
	MRPC	STS-B	QQP	MNLI (m/mm)	RTE	QNLI
Simple Transfer Fine-Tuning	88.6	**87.7**	**71.5**	**84.7/83.6**	**67.0**	**91.1**
BERT-base	88.3	84.7	71.2	84.3/83.0	59.8	89.1
+sentence	88.2	87.6	71.1	83.2/82.8	66.2	90.2
+3way-PP [Elaborate Feature]	88.2	85.8	70.9	82.9/81.9	65.8	88.0
+3way-PP [Simple Feature]	**89.0**	86.6	**71.5**	**84.7/83.6**	65.6	90.6

Table 7: Results of the ablation study where the best scores are represented in **bold** and scores higher than those of BERT-base are underlined. The third row shows performances when conducting only sentential paraphrase classification (+sentence) and the fourth row shows those when conducting only phrasal paraphrase classification with elaborate feature generation (+3way-PP [Elaborate Feature]), as reported in Table 5. The last row shows performances when conducing phrasal paraphrase classification with simple feature generation (+3way-PP [Simple Feature]). Results indicate that sentential and phrasal paraphrase classification complementarily contributes to Simple Transfer Fine-Tuning modeling.

[Simple Feature] outperformed +3way-PP [Elaborate Feature] on all tasks except RTE.