Bootstrapped Unsupervised Sentence Representation Learning

Yan Zhang[*1] **Ruidan He**[*2] **Zuozhu Liu**[†3] **Lidong Bing**[2] **Haizhou Li**[1,4]
[1]National University of Singapore, Singapore [2]DAMO Academy, Alibaba Group
[3]ZJU-UIUC Institute [4]Kriston AI Lab, China
eleyanz@nus.edu.sg, ruidan.he@alibaba-inc.com
zuozhuliu@intl.zju.edu.cn, l.bing@alibaba-inc.com
haizhou.li@nus.edu.sg

Abstract

As high-quality labeled data is scarce, unsupervised sentence representation learning has attracted much attention. In this paper, we propose a new framework with a two-branch Siamese Network which maximizes the similarity between two augmented views of each sentence. Specifically, given one augmented view of the input sentence, the online network branch is trained by predicting the representation yielded by the target network of the same sentence under another augmented view. Meanwhile, the target network branch is bootstrapped with a moving average of the online network. The proposed method significantly outperforms other state-of-the-art unsupervised methods on semantic textual similarity (STS) and classification tasks. It can be adopted as a post-training procedure to boost the performance of the supervised methods. We further extend our method for learning multilingual sentence representations and demonstrate its effectiveness on cross-lingual STS tasks. Our code is available at https://github.com/yanzhangnlp/BSL.

1 Introduction

Sentence representation learning aims to map sentences into vectors that capture rich semantic information. Among previous approaches, supervised methods achieve state-of-the-art performance by leveraging quality sentence labels. For example, the recently proposed model Sentence-BERT (SBERT) (Reimers and Gurevych, 2019) fine-tunes a Siamese BERT network on natural language inference (NLI) tasks with labeled sentence pairs. It achieves state-of-the-art results on multiple semantic textual similarity (STS) tasks. However, such performance is mostly induced by high-quality supervision, while labeled data are difficult and expensive to obtain in practice. Zhang et al. (2020) showed that SBERT generalizes poorly on target tasks that differ significantly from NLI on which SBERT is fine-tuned.

Many unsupervised methods learn sentence representations by optimizing over various self-supervised learning (SSL) objectives on a large-scale unlabeled corpus. Early works often use auto-encoders (Socher et al., 2011; Hill et al., 2016) or next-sentence prediction (Kiros et al., 2015) for sentence representation learning. Recently, more efforts have been devoted to representation learning with transformer-based networks using masked language modeling (MLM). However, transformer-based methods do not directly produce meaningful sentence representations. Instead, significant supervised fine-tuning steps with labeled data are commonly required to form good representations (Reimers and Gurevych, 2019). Recently, Giorgi et al. (2020) and Zhang et al. (2020) proposed novel transformer-based frameworks to directly learn sentence representations from an unlabeled corpus, which even exhibited competitive performance to the supervised counterparts on some tasks. However, Giorgi et al. (2020) required long text during training while the contrastive learning strategy employed by Zhang et al. (2020) need a careful treatment of negative pairs. More important, there is still great room for improvement in terms of the quality of learned sentence representations.

In this paper, we introduce **B**ootstrapped **S**entence Representation **L**earning (BSL), a simple and lightweight framework that directly learns sentence representations without supervised fine-tuning. Our work is inspired by the recent success of Siamese networks (Bromley et al., 1994) for unsupervised visual representation learning (Chen et al., 2020; Grill et al., 2020; Caron et al., 2020; Chen and He, 2020), especially the BYOL framework (Grill et al., 2020). These models employed

[*] Equally Contributed.
[†] Corresponding author.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 5168–5180
August 1–6, 2021. ©2021 Association for Computational Linguistics

various kinds of unsupervised learning objectives to maximize the similarity between two augmented views of each image, yielding performance on par with supervised methods. Unlike contrastive learning-based methods, which demand a carefully negative sampling process and large batch sizes, BYOL could achieve great performance without negative pairs.

The proposed BSL works as follows. Given an input sentence, we first construct two augmented views through back-translation. These two views are simultaneously fed into the two branches of the Siamese network, i.e., an online network and a target network following the terminology in (Grill et al., 2020). In particular, the online and target networks use two pre-trained transformer networks with the same structure, e.g., BERT, to encode the two views separately. During learning, the online network is trained to predict the representation of the other augmented view generated by the target network, and its parameters are updated by minimizing a predefined prediction loss. As for the target network, we apply a stop-gradient strategy (Chen and He, 2020) and update it with a weighted moving average of the online network. Hence, the outputs of the target network are iteratively bootstrapped to serve as targets, enabling enhanced representation learning of the online network while avoiding trivial solutions.

Our method is evaluated through extensive experiments. Empirical results show that BSL significantly outperforms strong unsupervised baselines on a standard suite of STS and classification tasks from the SentEval benchmark (Conneau and Kiela, 2018). We also demonstrate that BSL can serve as an effective post-training approach to boost the performance of the state-of-the-art supervised SBERT model. We further extend our method for learning multilingual sentence representations and demonstrate that it is able to outperform strong multilingual baselines on cross-lingual STS tasks under both unsupervised and supervised settings. Detailed analysis of a few factors that could affect the model performance is provided as well to motivate future research.

2 Related Work

2.1 Sentence Representation Learning

Prior approaches for sentence representation learning include two main categories – supervised and unsupervised methods, while a few works might leverage on both of them. Most of the supervised methods are trained on labeled natural language inference (NLI) datasets including Stanford NLI (SNLI) (Bowman et al., 2015) and MultiNLI (Williams et al., 2018). Early methods demonstrate good performance on a wide range of tasks (Conneau et al., 2017; Cer et al., 2018). Recently, SBERT (Reimers and Gurevych, 2019) fine-tuned a pre-trained Siamese BERT network on NLI and demonstrated the state-of-the-art performance. Though effective, those methods highly rely on labeled data and could be problematic to port to new domains. Zhang et al. (2020) showed that SBERT generalizes poorly on target tasks with a data distribution significantly different from the NLI data.

There are also fruitful outcomes for unsupervised methods. Some early studies attempt to learn from the internal structures within each sentence (Socher et al., 2011; Hill et al., 2016; Le and Mikolov, 2014) or utilize a distributional hypothesis to encode contextual information with generative (Kiros et al., 2015; Hill et al., 2016) or discriminative objectives (Jernite et al., 2017; Logeswaran and Lee, 2018). Recently, transformer-based networks attract more attentions (Devlin et al., 2019; Liu et al., 2019), however, they do not yield meaningful sentence representations directly without supervised fine-tuning. Reimers and Gurevych (2019) show that sentence embeddings obtained from BERT without fine-tuning even underperform the GloVe embeddings (Pennington et al., 2014) in terms of semantic textual similarity.

More recently, a few unsupervised methods were proposed to learn sentence representations from transformer-based networks without supervised fine-tuning. Li et al. (2020) proposes to transform the representation obtained by a pre-trained language model to an isotropic Gaussian distribution. Giorgi et al. (2020) minimizes the distance between different spans sampled from the same document. However, it requires an extremely long document of 2,048 tokens as input, which limits its applications to domains with only short documents. Zhang et al. (2020) proposed IS-BERT to maximize the mutual information between the global embedding and local n-gram embeddings of a given sentence. However, IS-BERT requires careful negative sampling and the n-gram embeddings may be suboptimal in capturing sentence-level semantics.

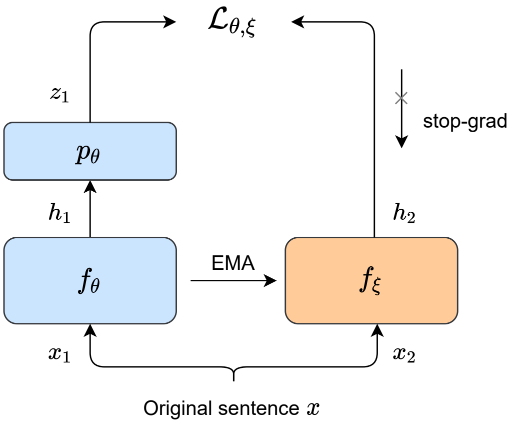

Figure 1: The proposed framework BSL. Two augmented views x_1 and x_2 of sentence x are encoded by the online network f_θ and the target network f_ξ, respectively. Both networks are initialised from the same pretrained language models but ξ are an exponential moving average (EMA) of θ during training. p denotes the predictor, which is a multi-layer perceptron and only applied on the online side. A stop-gradient operation is applied on the target side. The loss $\mathcal{L}_{\theta,\xi}$ maximise the similarity between online prediction z_1 and target representation h_2.

2.2 Unsupervised Representation Learning with Siamese Networks

Siamese networks have been increasingly used in various models (Chen and He, 2020; Grill et al., 2020; Caron et al., 2020) for unsupervised visual representation learning. These models typically maximize the similarity between two augmented views of an image encoded by the Siamese network. The main difference among these models is how they prevent undesired trivial solutions. Most works rely on contrastive learning with negative sampling (Chen et al., 2020; Tian et al., 2020) to avoid collapsing. Our method BSL is mainly inspired by BYOL (Grill et al., 2020), which shows that one can learn transferable visual representations via bootstrapping representations without negative sampling. We transfer this learning strategy from images to texts with different network architectures and augmenting methods.

3 BSL

3.1 Model Description

Given a sentence x sampled from the dataset $\mathcal{D}$ without label information, our goal is to learn a meaningful representation $h \triangleq f(x)$. In our framework, we adopt the idea from BYOL for unsupervised sentence representation learning with a Siamese network. The architecture of the proposed BSL is illustrated in Figure 1. Given a sentence x, we first obtain two augmented views $x_1 \triangleq \mathcal{T}(x)$ and $x_2 \triangleq \mathcal{T}'(x)$, where $\mathcal{T}$ and $\mathcal{T}'$ are augmentation transformations.

The two views are fed into the Siamese network separately. The online network contains an encoder network $f_\theta(\cdot)$ and a predictor network $p_\theta(\cdot)$. The target network contains an encoder network $f_\xi(\cdot)$ without a predictor, leading to an asymmetric framework. For the first augmented view x_1, the online network outputs a representation $z_1 \triangleq p_\theta(f_\theta(x_1))$. For the second augmented view, the target network outputs a representation $h_2 \triangleq f_\xi(x_2)$. Afterwards, we define a mean squared loss between the two normalized representations from the online and target networks, which can be simplified as minimizing their negative cosine similarity:

$$\mathcal{D}_{\theta,\xi}(z_1, h_2) = - < \frac{z_1}{\|z_1\|}, \frac{h_2}{\|h_2\|} >, \qquad (1)$$

where $\|\cdot\|$ denotes the l_2-norm and $<,>$ denotes the dot product between two vectors. As the loss is asymmetric over the two views, we also feed x_2 to the online network and x_1 to the target network to get $\tilde{z}_2 \triangleq p_\theta(f_\theta(x_2))$ and $\tilde{h}_1 \triangleq f_\xi(x_1)$, leading to the final objective:

$$\mathcal{L}_{\theta,\xi} = \frac{1}{2}\mathcal{D}_{\theta,\xi}(z_1, h_2) + \frac{1}{2}\mathcal{D}_{\theta,\xi}(\tilde{z}_2, \tilde{h}_1). \qquad (2)$$

Though we define the loss with parameters $\{\theta, \xi\}$, we only update θ during training, as shown in the stop-gradient operation Fig 1. This stop-gradient operation is empirically demonstrated effective for Siamese network (Grill et al., 2020; Chen and He, 2020). f_ξ is detached from the optimization graph of $\mathcal{L}_{\theta,\xi}$ and will be updated with a weighted moving average of f_θ. The updating dynamics becomes:

$$\theta_t \leftarrow \theta_{t-1} + \bigtriangledown_\theta \mathcal{L}_{\theta,\xi}, \qquad (3)$$
$$\xi_t \leftarrow \delta\xi_{t-1} + (1-\delta)\theta_t. \qquad (4)$$

Here δ is the momentum. When it is set to 1, the target network is never updated. When it is set to 0, the target network is instantaneously synchronized to the online network at each training step. At the inference stage, we obtain the representation of a sentence with the online encoder f_θ.

3.2 Architecture Details

Augmentation We use back-translation to obtain two augmented views x_1 and x_2. In this work, we only consider input sentence x in English. We use an English-to-German machine translation (MT) system to translate x to y_1, and subsequently use a German-to-English MT system to translate y_1 back to x_1 to obtain one augmented view. Similarly, we use English-to-French and French-to-English MT systems to obtain another augmented view x_2.[1] Besides back-translation, we also discuss other text augmentation approaches in § 4.4.

Architecture The online network f_θ and the target network f_ξ take x_1 and x_2 as inputs and output h_1 and h_2. We use pre-trained language models to initialize the weights in f_θ and f_ξ such that they benefit from the knowledge obtained at the pre-training stage. We apply average-pooling over outputs from the pre-trained language models to obtain h_1 and h_2. A multi-layer perceptron (MLP) p_θ is stacked on top of f_θ as the predictor to transform h_1 to predictions z_1 such as z_1 matches the target representation h_2.

4 Experiment

Design We conduct various experiments to evaluate the effectiveness of the proposed method. Following prior works (Reimers and Gurevych, 2019; Zhang et al., 2020), our major evaluations are conducted on the Semantic Textual Similarity (STS) tasks and the classification tasks with the SentEval toolkit (Conneau and Kiela, 2018). To demonstrate the flexibility of the proposed method, we further extend it for learning multilingual sentence representations and evaluate it on cross-lingual STS tasks.

Implementation The MLP contains three linear layers. Given an input vector of dimension d, the output dimensions of the three layers are $kd \rightarrow kd \rightarrow d$, where k is a hyperparameter controlling the hidden size. Batch normalization and rectified linear units (ReLU) are applied to the intermediate linear layers. We use BERT-base or RoBERTa-base to initialize the online and target networks in monolingual settings.

Hyperparameter We tune learning rate, batch size, momentum δ, and the hyperparameter k on the development set of STS-B (Cer et al., 2017). For all unsupervised experiments, we set learning rate to 5e-4, momentum to 0.999, and k to 8. Adam (Kingma and Ba, 2015) is used as the optimizer. [2]

Baselines Under a unsupervised learning setting, we compare to the **unigram-TFIDF** model, the Sequential Denoising Auto-Encoder (**SDAE**) (Hill et al., 2016), the **Skipthought** (Kiros et al., 2015) and **FastSent** (Hill et al., 2016). Those models are all trained on the Toronto book corpus with 70M sentences (Zhu et al., 2015). We also compare with sentence representations obtained with the average of GloVe embeddings (**GloVe avg.**), the average of BERT embeddings (**BERT avg.**), and the [CLS] representation of BERT (**BERT [CLS]**), as those are common ways to get sentence-level representations. We compare with **BERT-flow** (Li et al., 2020), a recent method that transforms the representation obtained by BERT to an isotropic Gaussian distribution. In addition, we compare with two unsupervised BERT fine-tuning methods. The first is to finetune BERT with masked language modeling (MLM) objective (**BERT-mlm**) (Gururangan et al., 2020). The second is **IS-BERT** (Zhang et al., 2020) which employs a mutual information maximization objective for fine-tuning BERT. We denote our model initialized by BERT-base (RoBERTa-base) as **BSL-BERT** (**BSL-RoBERTa**).

Under a supervised learning setting, we compared to **InferSent** (Conneau et al., 2017), Universal Sentence Encoder (**USE**) (Cer et al., 2018), and sentence BERT/RoBERTa (**SBERT/SRoBERTa**) (Reimers and Gurevych, 2019), which are all trained on the SNLI and MultiNLI datasets. To adapt BSL to a supervised learning setting, we first train a SBERT (SRoBERTa) model and then use the learned weights to initialize the online and target networks of BSL and perform BSL training. We denote this model variant as **BSL-SBERT** (**BSL-SRoBERTa**).

4.1 Semantic Textual Similarity (STS)

SentEval contains a suite of STS datasets including the STS tasks 2012-2016 (Agirre et al., 2012, 2013, 2014, 2015, 2016), the STS benchmark (STS-B) (Cer et al., 2017), and the SICK-Relatedness dataset (Marelli et al., 2014). These datasets con-

[1] We use Google translation engine. The datasets are released.

[2] Hyperparameters and implementation details are attached in Appendix A

Model	STS-12	STS-13	STS-14	STS-15	STS-16	STS-B	SICK-R	Avg.
Unsupervised methods								
Unigram-TFIDF†	-	-	58.00	-	-	-	52.00	-
SDAE†	-	-	12.00	-	-	-	46.00	-
SkipThought†	-	-	27.00	-	-	-	57.00	-
FastSent†	-	-	63.00	-	-	-	61.00	-
GloVe avg.‡	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
BERT avg.‡	38.78	57.98	57.98	63.15	61.06	46.35	58.40	54.81
BERT [CLS]‡	20.16	30.01	20.09	36.88	38.08	16.50	42.63	29.19
BERT-mlm	48.86	64.76	56.97	70.86	64.65	64.33	67.76	62.60
IS-BERT*	56.77	69.24	61.21	75.23	70.16	69.21	64.25	66.58
BERT-flow$^{\circ}$	59.54	64.69	64.66	72.92	71.84	58.56	65.44	65.38
Ours: BSL-BERT	**67.83**	**71.40**	**66.88**	**79.97**	**73.97**	**73.74**	**70.40**	**72.03**
Ours: BSL-RoBERTa	**68.47**	**72.41**	**68.48**	**78.50**	**72.77**	**78.77**	**69.97**	**72.76**
Supervised methods								
InferSent‡	52.86	66.75	62.15	72.77	66.87	68.03	65.65	65.01
USE‡	64.49	67.80	64.61	76.83	73.18	74.92	76.69	71.22
SBERT‡	70.97	76.53	73.19	79.09	74.30	77.03	72.91	74.89
SROBERTA‡	71.54	72.49	70.80	78.74	73.69	77.77	74.46	74.21
BERT-flow$^{\circ}$	67.75	76.73	75.53	80.63	77.58	79.10	78.03	76.48
Ours: BSL-SBERT	**71.48**	**81.20**	73.78	79.08	**79.23**	**80.67**	76.95	**77.49**
Ours: BSL-SRoBERTa	**75.44**	**80.25**	**76.14**	**81.62**	**80.00**	**81.90**	77.02	**78.91**

Table 1: Spearman rank correlation ρ between the cosine similarity of sentence representations and the gold labels. $\rho * 100$ is reported. All BERT/RoBERTa-based models use BERT/RoBERTa-base as the transformer encoder. Results of baselines marked with † are obtained from (Hill et al., 2016) (with a different number of decimal places). Results of baselines marked with ‡, * and $^{\circ}$ are obtained from (Reimers and Gurevych, 2019), (Zhang et al., 2020) and (Li et al., 2020), respectively.

sist of sentence pairs with scores from 0 to 5, where a larger score indicates higher semantic relatedness of the two sentences. We use Spearman's rank correlation between the cosine-similarities of the sentence pairs and the gold scores as an evaluation metric, following prior works (Reimers and Gurevych, 2019; Zhang et al., 2020).

Most of the prior unsupervised methods were trained on the Toronto book corpus (Zhu et al., 2015), while the most recent and the best performed unsupervised method IS-BERT was trained on unlabeled texts from SNLI and Multi-Genre NLI (MultiNLI) datasets. To have a fair comparison with IS-BERT, we follow its setting to train BSL on unlabeled texts from the SNLI and MultiNLI datasets. The BERT-mlm baseline is also trained with the same setting for a fair comparison. We illustrate the effect of corpus choice in § 4.4. SNLI contains 570k sentence pairs and MultiNLI contains 430k sentence pairs from a wider range of genres of spoken and written texts. In both datasets, each sentence pair is labeled with *contradiction*, *entailment*, and *neutral*. Note that the labels are excluded when training BSL in unsupervised settings.

Table 1 presents the comparison results. Models are divided into two sets: trained on unlabeled data, or trained on labeled data. For unsupervised models, Unigram-TFIDF, SDAE, SkipThought and FastSent are trained on the Toronto book corpus while BERT-mlm, IS-BERT, BERT-flow and our proposed method are trained on NLI. In the supervised setting, BSL-SBERT and BSL-SRoBERTa only take labeled *entailment* pairs as the inputs to the online and target networks.

We make the following observations. First, BSL outperforms all prior unsupervised methods by large margins. On average, it outperforms IS-BERT and BERT-flow trained with the same encoder and training corpus by 5.45%, and 6.65%, respectively. It even outperforms supervised baselines InferSent and USE. Second, unsupervised BSL still underperforms SBERT since the latter was fine-tuned on labeled NLI data. We show that by using BSL as a post-training approach, BSL-SBERT (BSL-SRoBERTa) can further increase the average result

Model	MR	CR	SUBJ	MPQA	SST	TREC	MRPC	Avg.
				Unsupervised methods				
Unigram-TFIDF†	73.7	79.2	90.3	82.4	-	85.0	73.6	-
SDAE†	74.6	78.0	90.8	86.9	-	78.4	73.7	-
SkipThought†	76.5	80.1	93.6	87.1	82.0	92.2	73.0	83.50
FastSent†	70.8	78.4	88.7	80.6	-	76.8	72.2	-
GloVe avg.‡	77.25	78.30	91.17	87.85	80.18	83.0	72.87	81.52
BERT avg.‡	78.66	86.25	94.37	88.66	84.40	92.8	69.54	84.94
BERT [CLS]‡	78.68	84.85	94.21	88.23	84.13	91.40	71.13	84.66
BERT-mlm	79.92	85.78	94.82	85.97	86.00	92.40	74.14	85.57
IS-BERT*	81.09	87.18	94.96	88.75	85.96	88.64	74.24	85.91
Ours: BSL-BERT	**81.42**	86.89	**95.20**	**89.60**	**87.70**	**93.00**	74.09	**86.84**
Ours: BSL-RoBERTa	80.92	**90.41**	**93.80**	**89.96**	**91.10**	88.40	**75.07**	**87.09**
				Supervised methods				
InferSent‡	81.57	86.54	92.50	**90.38**	84.18	88.2	75.77	85.59
USE‡	80.09	85.19	93.98	86.70	86.38	93.2	70.14	85.10
SBERT‡	**83.64**	89.43	94.39	89.86	88.96	89.6	76.00	87.41
Ours: BSL-SBERT	83.34	**89.67**	**95.65**	89.97	88.58	88.60	**76.93**	**87.53**
Ours: BSL-SRoBERTa	83.50	**89.17**	**94.57**	89.31	**91.60**	**92.40**	**77.1**	**88.24**

Table 2: Evaluation accuracies (%) on SentEval classification tasks. Scores are based on a 10-fold cross-validation. Results of baselines marked with † are obtained from (Hill et al., 2016) (with a different number of decimal places). Results of baselines marked with ‡ and * are obtained from (Reimers and Gurevych, 2019) and (Zhang et al., 2020), respectively.

by 2.6% (4.7%) from SBERT. This suggests that BSL can also be used as an effective post-training approach after supervised fine-tuning.

4.2 SentEval Classification Tasks

Following prior works (Reimers and Gurevych, 2019; Zhang et al., 2020), we evaluate sentence representations on a set of classification tasks from SentEval. The evaluation is done by the SentEval toolkit. It takes sentence representations as fixed input features to a logistic regression classifier, which is trained in a 10-fold cross-validation setup and the prediction results is computed on the test-fold. The sentence encoder is not fine-tuned in the training process. This set of tasks is the common bechmark used to evaluate the transferability of sentence representations on downstream tasks.

Table 2 presents the comparison results. On average, BSL outperforms all prior unsupervised baselines. It also outperforms supervised baselines InferSent and USE, and only slightly underperforms SBERT. BSL-SBERT can marginally improve the results of SBERT. BSL-SRoBERTa achieves the best performance.

4.3 Multilingual STS

In this subsection, we show that BSL can be easily extended for learning multilingual sentence representations. Following (Reimers and Gurevych, 2020), we conduct evaluation on the multilingual STS 2017 dataset (Cer et al., 2017) which contains annotated pairs for EN-EN, AR-AR, ES-ES, EN-AR, EN-ES, EN-TR, EN-DE, and EN-FR.

To learn multilingual representations under the unsupervised setting, we process the NLI data as follows. We translate the English NLI sentences to AR, ES, TR, DE and FR using Google translation engine and pair the original English sentence to each of its translations. We obtain 5 pairs (EN-AR/ES/TR/DE/FR) from one sentence and treat the English sentence as one view and its translation as the other view. We concatenate all pairs as the training data. We use multilingual BERT (mBERT) to initialize f_θ and f_ξ, such that the token-level representations between the different languages are aligned. The remaining training procedure is the same as described in § 3. We denote our unsupervised model as **BSL-uns**. We compare with sentence representations obtained with mean pooling of **mBERT** and **XLM-R** (Conneau et al., 2020) embeddings under the unsupervised setting.

For supervised learning, we compare with meth-

Model	EN-EN	ES-ES	AR-AR	EN-AR	EN-DE	EN-TR	EN-ES	EN-FR
Unsupervised methods								
mBERT	54.4	56.7	50.9	16.7	33.9	16.0	21.5	33.0
XLM-R	50.7	51.8	25.7	17.4	21.3	9.2	10.9	16.6
Ours: BSL-uns	**76.9**	**81.2**	**68.3**	**71.6**	**71.5**	**72.7**	**69.5**	**75.6**
Supervised methods								
mBERT-nli-stsb	80.2	83.9	65.3	30.9	62.2	23.9	45.5	57.8
XLM-R-nli-stsb	78.2	83.1	64.4	44.0	59.5	42.4	54.7	63.4
mBERT ← SBERT-nli-stsb	82.5	83.0	78.8	77.2	78.9	73.2	79.2	78.8
XLM-R ← SBERT-nli-stsb	82.5	83.5	**79.9**	77.8	78.9	74.0	79.7	78.5
mUSE	**86.4**	**86.9**	76.4	79.3	**82.1**	75.5	79.6	82.6
LaBSE	79.4	80.8	69.1	74.5	73.8	72.0	65.5	77.0
Ours: BSL-sup	83.3	86.1	79.3	**80.6**	81.2	**78.9**	**82.0**	**83.5**

Table 3: Spearman's rank correlation ρ between the cosine similarity of sentence representations and the gold labels. ρ*100 is reported. Results of baselines are obtained from (Reimers and Gurevych, 2020).

ods from (Reimers and Gurevych, 2020): **mBERT- / XLM-R-nli-stsb** denotes the setting where we fine-tune XLM-R and mBERT on the English NLI and the English training set of the STS benchmark (STS-B); **mBERT- /XLM-R ← SBERT-nli-stsb** is the knowledge-distillation method proposed in their paper where we learn mBERT and XLM-R to imitate the output of the English SBERT trained on NLI and STS-B with multilingual parallel sentence pairs. We also compared to results of **mUSE** (Chidambaram et al., 2019) and **LaBSE** (Feng et al., 2020), which use dual encoder transformer architectures. mUSE was trained on question-answer pairs, SNLI, translated SNLI data, and parallel corpora over 16 languages. LaBSE was trained on 6 billion translation pairs for 109 languages. For BSL, we initialize our online and target networks with the learned weights from XLM-R ← SBERT-nli-stsb[3] and then perform BSL training in a same way as described above. We denote our model in this setting as **BSL-sup**.

Table 3 presents the results. Under the unsupervised setting, averaging the multilingual token representations yields poor results. BSL-uns achieves promising results with scores higher than 70. For the supervised methods, we observe that directly fine-tuning multilingual pre-trained models on English NLI and STS-B datasets does not generalize well in a cross-lingual setting. Knowledge distillation-based models are strong baselines. Applying BSL as a post-training approach can boost the results of the distilled models by large margins. These observations demonstrate that BSL has the flexibility to be applied to learning multilingual sentence representations.

4.4 Analysis

In this subsection, we discuss a few factors that could affect the model performance. We use BERT-base as the encoder for analysis.

Choice of Corpus Previous works (Hill et al., 2016; Cer et al., 2018) indicated that the dataset used for learning sentence representations in a supervised setting significantly impacts their performance on STS tasks. They found learning with NLI datasets is particularly useful and yields good results on common STS benchmarks. We have similar observations with the proposed unsupervised method. In Table 4, we show the results of training our model with a subset of 5 million sentences from the Toronto book corpus. This setting achieves an average result of 69.65 on STS tasks, still outperforming prior best unsupervised model IS-BERT by 3.07%, which again demonstrates the effectiveness of the proposed framework.

However, we observe that the average result obtained from training with the book corpus is 2.38% lower than the result of training with the NLI datasets even the number of training pairs of the latter is only 1 million. Training on both of them still underperforms training on NLI alone. This finding indicates that the choice of training corpus is a key factor that affects model performance. When evaluating the common STS benchmarks as used in our experiments, the NLI datasets are better choices as they are semantically related to the STS data. We also conduct an evaluation on an Argument Facet Similarity task, which is more domain-specific and

[3] Downloaded from https://www.sbert.net/docs/pretrained_models.html

Model	STS-12	STS-13	STS-14	STS-15	STS-16	STS-B	SICK-R	Avg.
Choice of training corpus								
Ours: NLI	**67.83**	**71.40**	**66.88**	**79.97**	**73.97**	73.74	**70.40**	**72.03**
5M Book	64.90	68.33	65.18	77.48	73.12	70.55	68.05	69.65
5M Book + NLI	68.93	68.23	66.13	79.72	72.54	**74.94**	69.76	71.46
Effect of data augmentation								
Ours: Back-translation	67.83	71.40	66.88	79.97	73.97	73.74	70.40	72.03
Synonym	62.31	69.73	63.37	77.78	67.94	70.04	66.36	68.21
MLM	61.47	69.58	66.91	78.86	69.62	70.55	68.78	69.39
NLI$_{entail}$	65.88	**72.62**	65.67	78.39	74.17	73.42	70.67	71.54
Back-translation + NLI$_{entail}$	**72.01**	**72.62**	**70.16**	**81.65**	**76.03**	**77.65**	**74.48**	**74.94**

Table 4: Results with 1) different training corpora; and 2) different augmentation techniques. Spearman rank correlation ρ between the cosine similarity of sentence representations and the gold labels. $\rho * 100$ is reported.

Original	The cats used to love plopping on the newspapers.
Synonym	The cats use to have sex flump on the newspapers.
MLM	The cats used to love plucking in the newspapers.
Back-translation	Cats loved to play in the newspapers.
Entailment	oh when i had the uh cats at my place as soon as i took out the newspaper to read it they would plop right down on top of it and just not move and just stay there forever.

Table 5: An example of augmentations generated by different approaches.

dissimilar to the NLI tasks. The results are provided in Appendix B. We find that in this scenario, training with NLI data yields poor generalization results on the target test set while training on the target raw text yields a much better performance. The results indicate that semantically related corpus to the target task should be adopted as the training set.

Augmentation Techniques It has been shown that data augmentation plays a crucial role in unsupervised visual representation learning (He et al., 2020; Chen et al., 2020; Grill et al., 2020). The images can be augmented easily by rotating, resizing, or cropping (Chen et al., 2020). However, less work has been done on augmentation techniques for texts (Fang et al., 2020; Giorgi et al., 2020). Here, we study how different augmentation techniques would affect the model performance. We present the results of another two augmentation approaches besides back-translation in Table 4. **Synonym** denotes the setting where we randomly replace a few words with their synonyms. **MLM** denotes the setting where we first randomly mask a few tokens and then use a pre-trained masked language model to generate the masked tokens. Specifically, for both methods, given a sentence x, we make $x_1 = x$ and obtain x_2 with the respective augmentation technique. We found that using one augmented view performs slightly better than using two augmented views for synonym- and MLM-based methods. One possible reason is that these methods may generate augmented sentences with semantics totally different from the original sentences as we will show in this subsection. Such kind of augmentation may bring in too much randomness and noise. Therefore using two augmented views might instead harm the model performance.

For *Synonym*, we select 30% of words and substitute them with similar words according to WordNet (Miller, 1995). For *MLM*, we mask 20% of tokens and use RoBERTa-base for token generation. In addition, we show results of a setting where we treat the sentence pairs labeled with *entailment* from the NLI datasets as the two views (**NLI**$_{entail}$) for our model, as well as a setting using the combination of NLI unlabeled text with back-translations and the entailment pairs as the training corpus(**Back-translation+NLI**$_{entail}$). The purpose is to illustrate how our model would perform with high quality augmented data.

The results in Table 4 show that our proposed framework can work with both *Synonym* and *MLM*, as they still outperform IS-BERT on the average result by 1.63% and 2.81%, respectively. However, they are less effective compared to *Backt-*

Momen.	0.5	0.9	0.99	0.999	1
	33.18	69.58	72.51	73.74	68.19

Table 6: Performance w.r.t. momentum on STS-B. Spearman rank correlation ρ ($*100$) is reported.

Methods	16	32	64	128
BSL	69.21	71.08	72.02	72.01
Contrastive	68.18	70.06	71.04	71.81

Table 7: Performance under different batch sizes. The average Spearman rank correlation across STS12-16, STS-B, and SICK-R is reported.

translation. We observe that training with entailment pairs yields good results, with only 300k training pairs, NLI_{entail} is comparable to the model trained on all data from the NLI datasets augmented with back-translation (1 million training pairs). In addition, when training on both (*Back-translation* + NLI_{entail}), a 2.91% improvement on the average result over *Back-translation* is observed. The results indicate that the quality of the augmented pairs directly affects the performance of the proposed framework.

Table 5 presents an example of augmentations generated to the same sentence.[4] We observe that *Synonym* substitutes words without considering the context while *MLM* generates words based on the context but losing the original word semantics. Back-translation yields a relatively better sentence, however, the drawback of which is that it relies on external machine translation systems. The *Entailment* refers to the sentence in the NLI datasets to which the original sentence has an entailment relation. It can be regarded as an ideal augmentation of the original sentence. How to automatically generate such augmentations remains an open question, and we leave it to future research.

Momentum The momentum δ in Equation (4) is an important hyperparameter. When it is set to 1, the target network is never updated and remains the same to its initialization. When it is set to 0, the target network is updated to the online network at each training step. Table 6 shows the results of our method with different values of momentum. We observe that our proposed method works better with larger momentum near but not equals to 1. A similar phenomenon has also been observed in BYOL (Grill et al., 2020). In addition, we find that although directly averaging the token embeddings from BERT yields poor sentence representations as shown in Table 1, initializing the target network using BERT and keeping it unchanged (set momentum to 1) during the learning procedure helps the online network learn much better representations, yielding a 21.84% improvement on STS-B.

Batch Size & Contrastive Learning Lastly, we analyze the effect of batch size. Table 7 shows how the proposed model performs with batch sizes in $\{16, 32, 64, 128\}$. We also compare to a setting where contrastive learning is used as the self-supervised learning objective since it is more commonly used in visual representation learning (Chen et al., 2020). Specifically, in this setting, given a batch of n augmented sentence pairs ($2n$ sentences), each of them is treated as a positive pair. For each positive pair, we treat the other $2(n-1)$ augmented examples within the minibatch as negative examples.

The results in Table 7 show that for BSL, setting the batch size to 64 yields the best result. Overall BSL is less sensitive to changes in batch size while contrastive learning tends to perform better with a larger batch size such that sufficient negative samples can be obtained. Contrastive learning may achieve better performance with a larger batch size while we leave it for future investigation due to its large memory consumption.

5 Conclusion

In this paper, we propose BSL for unsupervised sentence representation learning. The experimental results demonstrate that our method could significantly outperform the state-of-the-art unsupervised methods and it can be further extended for learning multilingual sentence representations. In future work, we expect both theoretically advance of Siamese networks for representation learning, e.g., why stop-gradient works so well and how to further improve the updating dynamics, as well as specifically designated ideas for NLP, e.g., augmentation or learning objectives.

Acknowledgements

This work is partly supported by Human-Robot Interaction Phase 1 (Grant No. 19225 00054), National Research Foundation (NRF) Singapore under the National Robotics Programme; Human Robot Collaborative AI for AME (Grant No. A18A2b0046), NRF Singapore.

[4] More examples are provided in Appendix C

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *Proc. of SemEval@ACL*.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proc. of SemEval@ACL*.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proc. of SemEval@ACL*.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proc. of SemEval@ACL*.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *sem 2013 shared task: Semantic textual similarity. In *The Second Joint Conference on Lexical and Computational Semantics*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proc. of EMNLP*.

J. Bromley, J. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, Eduard Säckinger, and R. Shah. 1994. Signature verification using a "siamese" time delay neural network. In *Proc. of NeurIPS*.

M. Caron, I. Misra, J. Mairal, Priya Goyal, P. Bojanowski, and Armand Joulin. 2020. Unsupervised learning of visual features by contrasting cluster assignments. In *Proc. of NeurIPS*.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and cross-lingual focused evaluation. In *Proc. of SemEval@ACL*.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proc. of EMNLP*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proc. of ICML*.

Xinlei Chen and Kaiming He. 2020. Exploring simple siamese representation learning.

Muthu Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yunhsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Learning cross-lingual sentence representations via a multi-task dual-encoder model. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proc. of ACL*.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proc. of LREC*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proc. of EMNLP*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL-HLT*.

Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding. *CoRR*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding. *CoRR*, abs/2007.01852.

John M. Giorgi, Osvald Nitski, Gary D. Bader, and Bo Wang. 2020. Declutr: Deep contrastive learning for unsupervised textual representations. *CoRR*.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. 2020. Bootstrap your own latent: A new approach to self-supervised learning. In *Proc. of NeurIPS*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proc. of ACL*.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. Momentum contrast for unsupervised visual representation learning.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. In *Proc. of NAACL-HLT*.

Yacine Jernite, Samuel R. Bowman, and David A. Sontag. 2017. Discourse-based objectives for fast unsupervised sentence representation learning. *arXiv preprint arXiv:1705.00557*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proc. of ICLR*.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Proc. of NeurIPS*.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proc. of ICML*.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proc. of EMNLP*.

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. *arXiv preprint arXiv:1907.11692*.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *Proc. of ICLR*.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proc. of LREC*.

G. Miller. 1995. Wordnet: a lexical database for english. *Commun. ACM*, 38:39–41.

Amita Misra, Brian Ecker, and Marilyn Walker. 2016. Measuring the similarity of sentential arguments in dialogue. In *Proc. of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proc. of EMNLP*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proc. of EMNLP-IJCNLP*.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proc. of EMNLP*.

Richard Socher, Eric H. Huang, Jeffrey Pennin, Christopher D Manning, and Andrew Y. Ng. 2011. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Proc. of NeurIPS*.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. 2020. What makes for good views for contrastive learning? In *Proc. of NeurIPS*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proc. of NAACL-HLT*.

Yan Zhang, Ruidan He, Zuozhu Liu, Kwan Hui Lim, and Lidong Bing. 2020. An unsupervised sentence embedding method by mutual information maximization. In *Proc. of EMNLP*.

Yukun Zhu, Ryan Kiros, S. Richard 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 *Proc. of ICCV*.

A Implementation Details

Our implementation is based on Python 3.6 and Pytorch 1.6.0. All experiments were conducted on a RTX 8000 GPU (CUDA version 10.2) configured on a standard workstation. The workstation is configured with 2 Intel Xeon Gold 6248R, 256GB RAM, and Ubuntu 18.04 operating system. We provide main hyperparameters of our model training on the NLI datasets in the Table 8. For cross-lingual experiments, we use bert-base-multilingual-cased and the other hyperparameters are the same. The NLI and related datasets can be downloaded from `https://huggingface.co/datasets`. The development results of BSL on the NLI dataset are shown in Table 9.

Hyperparameter	Size/Type
Batch Size	{16, 32, **64**, **128**}
Learning Rate	{1e-4, **2e-4, 5e-4**}
Weight Decay	{0.1, **0.01**, 0.001 }
Epsilon	1e-6
Optimizer	Adam
BERT Type	bert-base-uncased
BERT Embedding Size	768
K	{**8**, 4}
Pooling Strategy	Mean
Epoch Num	1

Table 8: Hyperparameters for training on the NLI dataset.

Method	STS-B-dev
BSL	79.42
BSL-SBERT	81.67

Table 9: Performance on STS-B development set. Spearman rank correlation ρ ($*100$) is reported.

B Argument Facet Similarity

We have demonstrated that the proposed method significantly outperforms other unsupervised baselines on a suite of STS and classification tasks that are commonly used in previous works. However, those tasks are less domain or task specific. Here, we further investigate the effectiveness of BSL in a domain-specific scenario. Following prior works (Reimers and Gurevych, 2019; Zhang et al., 2020), we conduct evaluations on an Argument Facet Similarity (AFS) (Misra et al., 2016) dataset. The dataset consists of 6k argument pairs on three controversial topics: *gun control*, *gay marriage*, and *death penalty*. Each pair was annotated on a scale from 0 (different) to 5 (equivalent). This dataset is more challenging compared to the STS benchmarks: the lexical gap between the sentences in AFS is larger and to be consider similar, a pair of arguments must not only make similar claims, but also provide a similar reasoning.

Model	r	ρ
Unigram-TFIDF†	46.77	42.95
GloVe avg.†	32.40	34.00
BERT avg.†	35.39	35.07
BERT-mlm	47.04	45.92
IS-BERT†	49.14	45.25
InferSent†	27.08	26.63
SBERT†	16.27	15.84
Ours: BSL	**51.56**	**50.47**

Table 10: Average Pearson correlation r and average Spearman's rank correlation ρ over three topics on the Argument Facet Similarity (AFS) corpus. Results marked with † are obtained from (Zhang et al., 2020).

We compare models in a setting where task- or domain-specific labeled data is not available. In this setting, supervised method such as SBERT and InferSent need to be trained on NLI data and perform cross-domain predictions on the AFS sentence pairs. Unsupervised methods such as BERT-mlm, IS-BERT and our proposed BSL can be directly trained on the task-specific raw texts.

Table 10 shows the comparison results. We present both Pearson correlation and Spearman's rank correlation. The results show that the proposed method still outperforms other methods. It is interesting to find that the two supervised methods InferSent and SBERT perform the worst in this setting. This is due to the fact that AFS data diferes significantly from NLI data. This suggests that the domain-relatedness between the training set and the target test set has a huge impact on the model performance, and the models learned with supervised methods are problematic to port to other distant domains.

C More Examples

More examples of augmentations generated by different approaches are provided in the Table 11.

Original	I realize she had written a new will .
Synonym	I realize she had drop a new will.
MLM	I realize she had bought a new will.
Back-translation	I realize that she had made a new will.
Entailment	I was now quite convinced that she had made a fresh will, and 67 had called the two gardeners in to witness her signature.
Original	There are people who believe that the interest on the national debt is a problem .
Synonym	There are the great unwashed who believe that the stake on the interior debt is a problem.
MLM	There are some who believe that compound interest on the national debt is a problem.
Back-translation	There are people who believe that national debt interest rates are a problem.
Entailment	But if Congress opts for debt over taxation, you can count on thoughtless commentators to denounce the interest payments on that debt as a second, and separate, outrage.
Original	According to numerous studies, music and suicide have little to no correlation .
Synonym	Harmonise to various survey, music and suicide have little to no correlation.
MLM	According to other studies, music and suicide have little to no correlation...
Back-translation	According to many studies, music and suicide have little or no correlation.
Entailment	Numerous studies show that there is no association between music and suicide.
Original	The earliest human remains found on Crete date back to the seventh millennium b.c.
Synonym	The earliest human remains found on Crete date backward to the 7th millenary b. degree celsius.
MLM	The earliest human remains found in the planet date back to the seventh millennium b.c.
Back-translation	The first human remains discovered in Crete date back to the seventh millennium BC.
Entailment	Crete has ancient human remains.
Original	It's a commitment to general education–a sequence of courses intended to develop critical thinking in a wide variety of disciplines–in opposition to early specialization.
Synonym	It ' s a commitment to general education – a sequence of course specify to educate critical thought in a wide variety of field – in opposition to other specialism.
MLM	It's a commitment to general education – a sequence of courses intended to develop critical thinking in a wide variety of disciplines – in opposition to early specialization.
Back-translation	It is a commitment to general education - a sequence of courses designed to develop critical thinking in a wide variety of disciplines - as opposed to early specialization.
Entailment	General education's focus is to develop students' critical thinking skills.

Table 11: More examples of augmentations generated by different approaches.