BURT: BERT-INSPIRED UNIVERSAL REPRESENTATION FROM LEARNING MEANINGFUL SEGMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Although pre-trained contextualized language models such as BERT achieve significant performance on various downstream tasks, current language representation still only focuses on linguistic objective at a specific granularity, which is not applicable when multiple levels of linguistic units are involved at the same time. Thus we present a universal representation model, **BURT** (**B**ERT-inspired **U**niversal **R**epresentation from learning meaningful segmen**T**), to encode different levels of linguistic unit into the same vector space. Specifically, we extract and mask meaningful segments based on point-wise mutual information (PMI) to incorporate different granular objectives into the pre-training stage. Our model surpasses BERT and BERT-wwm-ext on a wide range of downstream tasks in the ChineseGLUE (CLUE) benchmark. Especially, BURT-wwm-ext obtains 74.48% on the WSC test set, 3.45% point absolute improvement compared with its baseline model. We further verify the effectiveness of our unified pre-training strategy in two real-world text matching scenarios. As a result, our model significantly outperforms existing information retrieval (IR) methods and yields universal representations that can be directly applied to retrieval-based question-answering and natural language generation tasks.

1 INTRODUCTION

Representations learned by deep neural models have attracted a lot of attention in Natural Language Processing (NLP). However, previous language representation learning methods such as word2vec (Mikolov et al., 2013), LASER (Artetxe & Schwenk, 2019) and USE (Cer et al., 2018) focus on either words or sentences. Later proposed pre-trained contextualized language representations like ELMo (Peters et al., 2018), GPT(Radford et al., 2018), BERT (Devlin et al., 2018) and XLNet (Yang et al., 2019) may seemingly handle different sized input sentences, but all of them focus on sentence-level specific representation still for each word, which leads to unsatisfactory performance in real-world situations. Although the latest BERT-wwm-ext (Cui et al., 2019), StructBERT (Wang et al., 2020) and SpanBERT (Joshi et al., 2020) perform MLM on a higher linguistic level, the masked segments (whole words, trigrams, spans) either follow a pre-defined distribution or focus on a specific granularity. Such sampling strategy ignores important semantic and syntactic information of a sequence, resulting in a large number of meaningless segments.

In this paper, we introduce BURT, a pre-trained model that aims at learning universal representations for sequences of various lengths. Our model follows the architecture of BERT but differs from its original masking and training scheme. Specifically, we propose to efficiently extract and prune meaningful segments (n-grams) from unlabeled corpus with little human supervision, and then use them to modify the masking and training objective of BERT. The n-gram pruning algorithm is based on point-wise mutual information (PMI) and automatically captures different levels of language information, which is critical to improving the model capability of handling multiple levels of linguistic objects in a unified way, i.e., embedding sequences of different lengths in the same vector space.

Overall, our pre-trained models outperform BERT and BERT-wwm-ext on several downstream tasks, where BURT and BURT-wwm-ext reach 74.14% and 74.48% accuracy on WSC, respectively, surpassing BERT and BERT-wwm-ext by 3.45% absolute. Our models also exceed the baseline models by $0.2\% \sim 0.6\%$ point accuracy on three other CLUE tasks including TNEWS', IFLYTEK'

and CSL. Moreover, BURT can be easily applied to real-world applications such as Frequently Asked Questions (FAQ) and Natural Language Generation (NLG) tasks, where it encodes words, sentences and paragraphs into the same embedding space and directly retrieves sequences that are semantically similar to the given query based on cosine similarity. All of the above experimental results demonstrate that our well-trained model leads to universal representation that can adapt to various tasks and applications.

2 RELATED WORK

Representing words as real-valued dense vectors is a core technique of deep learning in NLP. Word embedding models (Mikolov et al., 2013; Pennington et al., 2014; Joulin et al., 2016) map words into a vector space where similar words have similar latent representations. ELMo (Peters et al., 2018) attempts to learn context-dependent word representations through a two-layer bi-directional LSTM network. In recent years, more and more researchers focus on learning sentence representations. The Skip-Thought model (Kiros et al., 2015) is designed to predict the surrounding sentences for an given sentence. Logeswaran & Lee (2018) improve the model structure by replacing the RNN decoder with a classifier. InferSent (Conneau et al., 2017) is trained on the Stanford Natural Language Inference (SNLI) dataset (Bowman et al., 2015) in a supervised manner. Subramanian et al. (2018); Cer et al. (2018) employ multi-task training and report considerable improvements on downstream tasks. LASER (Artetxe & Schwenk, 2019) is a BiLSTM encoder designed to learn multilingual sentence embeddings. Nevertheless, most of the previous work focused on a specific granularity. In this work we extend the training goal to a unified level and enables the model to leverage different granular information, including, but not limited to, word, phrase or sentence.

Most recently, the pre-trained language model BERT (Devlin et al., 2018) has shown its powerful performance on various downstream tasks. BERT is trained on a large amount of unlabeled data including two training targets: Masked Language Model (MLM) for modeling deep bidirectional representations, and Next Sentence Prediction (NSP) for understanding the relationship between two sentences. Lan et al. (2019) introduce Sentence-Order Prediction (SOP) as a substitution of NSP. Wang et al. (2020) develop a sentence structural objective by combining the random sampling strategy of NSP and continuous sampling as in SOP. However, Liu et al. (2019) and Joshi et al. (2020) use single contiguous sequences of at most 512 tokens for pre-training and show that removing the NSP objective improves the model performance. Besides, BERT-wwm (Cui et al., 2019), StructBERT (Joshi et al., 2020), SpanBERT (Wang et al., 2020) perform MLM on higher linguistic levels, augmenting the MLM objective by masking whole words, trigrams or spans, respectively. Nevertheless, we concentrate on enhancing the masking and training procedures from a broader and more general perspective.

3 METHODOLOGY

Our BURT follows the Transformer encoder (Vaswani et al., 2017) architecture where the input sequence is first split into subword tokens and a contextualized representation is learned for each token. We only perform MLM training on single sequences as suggested in Joshi et al. (2020). The basic idea is to mask some of the tokens from the input and force the model to recover them from the context. Here we propose a unified masking method and training objective considering different grained linguistic units.

Specifically, we apply an pruning mechanism to collect meaningful n-grams from the corpus and then perform n-gram masking and predicting. Our model differs from the original BERT and other BERT-like models in several ways. First, instead of the token-level MLM of BERT, we incorporate different levels of linguistic units into the training objective in a comprehensive manner. Second, unlike SpanBERT and StructBERT which sample random spans or trigrams, our n-gram sampling approach automatically discovers structures within any sequence and is not limited to any granularity.

3.1 N-GRAM PRUNING

In this subsection, we introduce our approach of extracting a large number of meaningful n-grams from the monolingual corpus, which is a critical step of data processing.

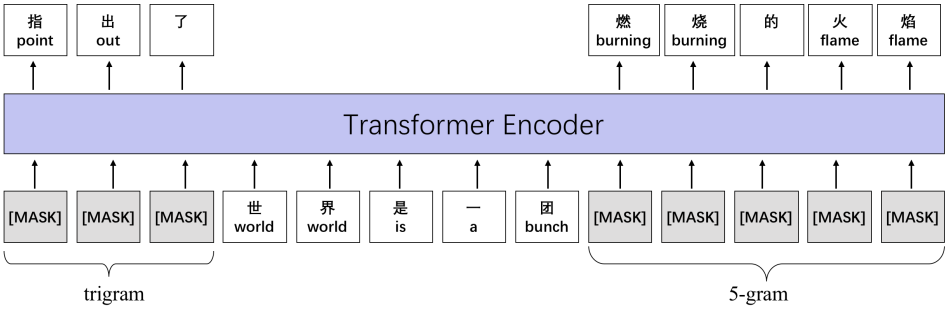

Figure 1: An illustration of *n*-gram pre-training.

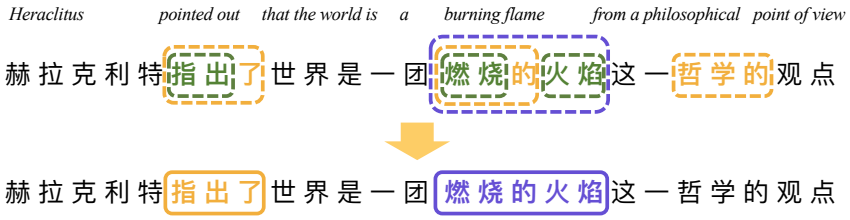

Figure 2: An example from the Chinese Wikipedia corpus. *n*-grams of different lengths are marked with dashed boxes in different colors in the upper part of the figure. During training, we randomly mask *n*-grams and only the longest *n*-gram is masked if there are multiple matches, as shown in the lower part of the figure.

First, we scan the corpus and extract all *n*-grams with lengths up to N using the SRILM toolkit[1] (Stolcke, 2002). In order to filter out meaningless *n*-grams and prevent the vocabulary from being too large, we apply pruning by means of point-wise mutual information (PMI) (Church & Hanks, 1990). To be specific, mutual information $I(x, y)$ describes the association between tokens x and y by comparing the probability of observing x and y together with the probabilities of observing x and y independently. Higher mutual information indicates stronger association between the two tokens.

$$I(x, y) = \log \frac{P(x, y)}{P(x)P(y)} \tag{1}$$

In practice, $P(x)$ and $P(y)$ denote the probabilities of x and y, respectively, and $P(x, y)$ represents the joint probability of observing x followed by y. This alleviates bias towards high-frequency words and allows tokens that are rarely used individually but often appear together such as "*San Francisco*" to have higher scores. In our application, an *n*-gram denoted as $w = (x_1, \ldots, x_{L_w})$, where L_w is the number of tokens in w, usually contains more than two words. Therefore, we present an extended PMI formula displayed as below:

$$PMI(w) = \frac{1}{L_w} \left(\log P(w) - \sum_{k=1}^{L_w} \log P(x_k) \right) \tag{2}$$

where the probabilities are estimated by counting the number of observations of each token and *n*-gram in the corpus, and normalizing by the size of the corpus. $\frac{1}{L_w}$ is an additional normalization factor which avoids extremely low scores for longer *n*-grams. Finally, *n*-grams with PMI scores below the chosen threshold are filtered out, resulting in a vocabulary of meaningful *n*-grams.

3.2 *N*-GRAM MASKING

For a given input $S = \{x_1, x_2, \ldots, x_L\}$, where L is the number of tokens in S, special tokens `[CLS]` and `[SEP]` are added at the beginning and end of the sequence, respectively. Before feeding the training data into the Transformer blocks, we identify all the *n*-grams in the sequence using

[1] http://www.speech.sri.com/projects/srilm/download.html

the aforementioned *n*-gram vocabulary. An example is shown in Figure 2, where there are overlap between *n*-grams, which indicates the multi-granular inner structure of the given sequence. In order to make better use of higher-level linguistic information, the longest *n*-gram is retained if multiple matches exist. Compared with other masking strategies, our method has two advantages. First, *n*-gram extracting and matching can be efficiently done in an unsupervised manner without introducing random noise. Second, by utilizing *n*-grams of different lengths, we generalize the masking and training objective of BERT to a unified level where different granular linguistic units are integrated.

Following BERT, we mask 15% of all tokens in each sequence. The data processing algorithm uniformly samples one *n*-gram at a time until the maximum number of masking tokens is reached. 80% of the time the we replace the entire *n*-gram with `[MASK]` tokens. 10% of the time it is replace with random tokens and 10% of the time we keep it unchanged. The original token-level masking is retained and considered as a special case of *n*-gram masking where $n = 1$. We employ dynamic masking as mentioned in Liu et al. (2019), which means masking patterns for the same sequence in different epochs are probably different.

3.3 PRE-TRAINING

3.3.1 TRANING OBJECTIVE

As depicted in Figure 1, the Transformer encoder generates a fixed-length contextualized representation at each input position and the model only predicts the masked tokens. Ideally, a universal representation model is able to capture features for multiple levels of linguistic units. Therefore, we extend the MLM training objective to a more general situation, where the model is trained to predict *n*-grams rather than subwords.

$$\max_{\theta} \sum_{w} \log P(w|\hat{\mathbf{x}};\theta) = \max_{\theta} \sum_{(i,j)} \log P(x_i,\ldots,x_j|\hat{\mathbf{x}};\theta) \tag{3}$$

where w is a masked *n*-gram and $\hat{\mathbf{x}}$ is a corrupted version of the input sequence. (i, j) represents the absolute start and end positions of w.

3.3.2 TRAINING DETAILS

We download the Chinese Wikipedia Corpus[2] and pre-process with `process_wiki.py`[3], which extracts text from xml files. Then we convert the data into simplified characters using `OpenCC`. In order to extract high-quality *n*-grams, we remove non-Chinese characters and punctuation marks based on regular expressions, and finally get a corpus of 380M Chinese characters. In the *n*-gram pruning stage, we use maximum length of *n*-grams $N = 10$. After calculating the PMI scores of all *n*-grams, we try different thresholds from -15 to -10 at 0.5 intervals and manually evaluate the *n*-gram vocabulary. We find that when the threshold is high ($\geq$-11.5), nearly 50% of the *n*-grams contain $3 \sim 5$ characters, and only less than 0.5% *n*-grams are longer than 7. Although a lower threshold ($\leq$-12.5) can generate longer *n*-grams, it will cause too many meaningless *n*-grams. Therefore, we empirically set the threshold to -12, resulting in 58M *n*-grams with an average length of 6.4. As in BERT, sentence pairs are packed into a single sequence and the special `[CLS]` token is used for sentence-level predicting. While in accordance with Joshi et al. (2020), we find that single sentence training is better than the original sentence pair scenario. Therefore, in our experiments, the input is a continuous sequence with a maximum length of 512.

The model architecture is identical to BERT (Devlin et al., 2018). We use the pre-trained BERT and BERT-wwm-ext for initialization, each with 12 Transformer layers, 12 heads, 768 dimensional hidden states and 110M parameters in total. The former is an official release of Chinese BERT and the latter is trained from the Chinese BERT using whole word masking on extended data (Cui et al., 2019). We use Adam optimizer (Kingma & Ba, 2017) with initial learning rate of 5e-5 and linear warmup over the first 10% of the training steps. Batch size is set to 16 and dropout rate is 0.1. The model is trained for 56k steps on a single 1080Ti GPU.

[2]https://dumps.wikimedia.org

[3]https://github.com/panyang/Wikipedia_Word2vec/blob/master/v1/process_wiki.py

[4]https://www.cluebenchmarks.com/rc.html

Corpus	Task	Length	#Train	#Dev	#Test	#L
TNEWS'	Classificaiton	Short	53k	10k	10k	15
IFLYTEK'	Classification	Long	12k	2.6k	2.6k	119
AFQMC	Question Matching	Short-Short	34k	4.3k	3.9k	2
OCNLI	NLI	Short-Short	50k	3k	3k	3
WSC	Coreference Resolution	Long-Short	1.2k	304	290	2
CSL	Keyword Recognition	Long-Short	20k	3k	3k	2

Table 1: Statistics of six classification tasks in CLUE benchmark. #Train, #Dev, #Test are the size of training, development and test sets, respectively. #L is the number of labels. Sequences are simply divided into two categories according to their length: "Long" and "Short".

Batch size: 8, 16; Length: 128, 256; Epoch: 2, 3, 5, 50; lr: 1e-5, 2e-5, 3e-5

Models	AFQMC	TNEWS'	IFLYTEK'	OCNLI	WSC	CSL
BERT	73.70	56.58	60.29	72.20	62.00	80.36
+ MLM updates	73.32	56.46	60.23	70.63	70.69	79.27
SpanBERT	73.48	56.66	59.62	71.00	72.07	79.67
BURT	73.14	56.85	**60.50**	71.25	74.14	80.83
BERT-wwm-ext	**74.07**	56.84	59.43	**73.40**	61.10	80.63
+ MLM updates	73.97	56.70	59.42	72.83	71.03	80.40
SpanBERT-wwm-ext	73.22	56.87	58.50	71.60	73.79	80.23
BURT-wwm-ext	73.84	**57.29**	60.08	72.17	**74.48**	**80.97**

Table 2: CLUE test results scored by the evaluation server[4].

4 EXPERIMENTS

To evaluate the model ability of handling different linguistic units, we report the performance of our model on downstream tasks from CLUE benchmark. Moreover, we present an insurance FAQ task and a retrieval-based language generation task, where the key is to embed sequences of different lengths in the same vector space and retrieve the piece of text closest to the given query.

We compare our model with three variants: pre-trained models (BERT/BERT-wwm-ext), models trained with the same number of additional steps as BURT (+MLM updates), and models trained using random span masking with the same number of additional steps as BURT (SpanBERT/SpanBERT-wwm-ext). For SpanBERT and SpanBERT-wwm-ext, we simply replace our *n*-gram module with the masking strategy as proposed by Joshi et al. (2020), where the sampling probability of span length l is based on a geometric distribution $l \sim Geo(p)$. We follow the parameter setting that $p = 0.2$ and maximum span length $l_{max} = 10$.

4.1 CLUE BENCHMARK

The Chinese General Language Understanding Evaluation (ChineseGLUE or CLUE) benchmark aims at better serving Chinese language understanding and language model evaluation. We utilize six classification tasks from CLUE including single-sentence classification (TNEWS' and IFLYTEK'), Natural Language Inference (OCNLI), question matching (AFQMC), coreference resolution (WSC), and keyword recognition (CSL). Statistics of the datasets are listed in Table 1. Besides the diversity of task types, we also find that different datasets concentrates on sequences of different lengths, which satisfies our need to examine the model ability of representing multiple granular linguistic units.

Following BERT, in the fine-tuning procedure, pairs of sentences are concatenated into a single sequence with a special token `[SEP]` in between. For both single sentence and sentence pair tasks, the hidden state of the first token `[CLS]` is used for softmax classification. Table 2 shows the results on CLUE, where we find that training BERT with additional MLM steps can hardly bring any improvement except for the WSC task. SpanBERT is effective on WSC but is comparable to BERT

Query: 80岁还能买意外险吗? *Can 80-year-old people get accident insurance?*

Q: 受益人能否修改? *Can I change the beneficiary?*	A: 可以修改受益人。 *Yes, the beneficiary can be modified.*
Q: 高龄老人能否购买意外险? ***Can seniors buy accident insurance?***	**A: 你可以选择老年人的意外产品。** ***You can choose accident insurance for the elderly.***
Q: 如何进行线上理赔? *How to make an online claim?*	A: 首先给保险公司打电话报案申请理赔。第二… *First, call the insurance company to report the case and apply for a claim. Second, ...*

Figure 3: Examples of Question-Answer pairs from our insurance FAQ dataset. The correct match to the query is highlighted.

on other tasks. BERT-wwm-ext is better than our model on classification tasks involving pairs of short sentences such as AFQMC and OCNLI, which may be due to its relative powerful capability of modeling short sequences. Overall, both BURT and BURT-wwm-ext outperform the baseline models on 4 out of 6 tasks with considerable improvement, which sheds light on their effectiveness of modeling sequences of different lengths, and we find that our proposed PMI-based masking method is general and independent with model settings. The most significant improvement is observed on WSC (3.45% over the updated BERT and 2.07% over SpanBERT), where the model is trained to determine whether the given two spans refer to the same entity in the text. We conjecture that the model benefits from learning to predict meaningful spans in the pre-training stage, so it is better at capturing the meanings of spans in the text.

4.2 RETRIEVAL-BASED FAQ

Moving from word and sentence vectors towards representation for sequences of any lengths, a universal language model may have the ability of capturing semantics of free text and facilitating various applications that are highly dependent on the quality of language representation. Thus, we present an insurance FAQ task in this subsection and an NLG task in the next subsection to explore the effectiveness of BURT in real-world applications.

A Frequently Asked Question (FAQ) task involves a collection of Question-Answer (QA) pairs denoted as $\{(Q_1, A_1), (Q_2, A_2), ...(Q_N, A_N)\}$, where N is the number of QA pairs. The goal is to retrieve the most relevant QA pairs for a given query. We collect frequently asked questions and answers between users and customer service from our partners in a Chinese online financial education institution. It contains over 4 types of insurance questions, e.g., concept explanation ("*what*"), insurance consultation ("*why*", "*how*"), judgement ("*whether*") and recommendation. An example is shown in Figure 3. Our dataset is composed of 300 QA pairs that are carefully selected to avoid similar questions so that each query has only one exact match. Because queries are mainly paraphrases of the standard questions, we use query-Question similarity as the ranking score. The test set consists of 875 queries and the average lengths of questions and queries are 14 and 16, respectively.

Our baseline models include statistical methods such as TF-IDF and BM25, a sentence representation model LASER[5] (Artetxe & Schwenk, 2019), the pre-trained BERT/BERT-wwm-ext, and SpanBERT/SpanBERT-wwm-ext. The evaluation metric is Top-1 Accuracy and Mean Reciprocal Rank (MRR) because there is only one correct answer for each query. Results are reported in Table 4. As we can see, LASER and all pre-trained language models significantly outperform TF-IDF and BM25, indicating the superiority of embedding-based models over statistical methods. Besides, the continued BERT training is often beneficial. Among all the evaluated models, our BURT yields the highest accuracy (82.2%) and MRR (0.872). BURT-wwm-ext achieves a slightly lower accuracy (80.7%) compared with BURT but it still exceeds its baselines by 4.0% (+MLM updates) and 1.4% (SpanBert-wwm-ext) point, respectively.

[5]https://github.com/facebookresearch/LASER

Category	Topics	
Daily Scenarios	Traveling, Recipe, Skin care, Beauty makeup, Pets	22
Sport & Health	Outdoor sports, Athletics, Weight loss, Medical treatment	15
Reviews	Movies, Music, Poetry, Books	16
Persons	Entrepreneurs, Historical/Public figures, Writers, Directors, Actors	17
General	Festivals, Hot topics, TV shows	6
Specialized	Management, Marketing, Commerce, Workplace skills	17
Others	Relationships, Technology, Education, Literature	14
All	-	107

Table 3: Details of the templates.

Method	Acc.	MRR
TF-IDF	73.7	0.813
BM25	72.1	0.802
LASER	79.9	0.856
BERT	76.8	0.831
+ MLM updates	78.3	0.843
SpanBERT	78.6	0.846
BURT	**82.2**	**0.872**
BERT-wwm-ext	76.7	0.834
+ MLM updates	76.7	0.834
SpanBERT-wwm-ext	79.3	0.856
BURT-wwm-ext	80.7	0.863

Table 4: Comparison of statistical methods, the sentence embedding model and pre-trained contextualized language models on the FAQ dataset. "Acc." represents Top-1 accuracy.

R	Judge1	Judge2	Avg.
BM25	60.3	61.8	61.1
LASER	63.9	61.6	62.8
BERT	65.9	67.3	66.6
+ MLM updates	65.0	67.5	66.3
SpanBERT	69.3	71.5	70.4
BURT	71.8	71.0	71.4
CM	Judge1	Judge2	Avg.
BM25	43.5	41.2	42.4
LASER	42.5	38.4	40.5
BERT	48.5	47.8	48.2
+ MLM updates	46.1	45.5	45.8
SpanBERT	51.6	53.9	52.8
BURT	54.2	56.5	55.4

Table 5: Results on NLG according to human judgment. "R" and "CM" represent the percentage of paragraphs that are "relevant" and "convey the meaning", respectively.

4.3 Natural Language Generation

In this subsection, we apply our model to a retrieval-based Natural Language Generation (NLG) task. The task is to generate articles based on manually created templates. Concretely, the goal is to retrieve one paragraph at a time from the corpus which best describes a certain sentence from the template and then combine the retrieved paragraphs into a complete passage. The main difficulty of this task lies in the need to compare semantics of sentence-level queries (usually contain only a few words) and paragraph-level documents (often consist of multiple sentences).

We use articles collected by our partners in a media company as our corpus. Each article is split into several paragraphs and each document contains one paragraph. The corpus has a total of 656k documents and cover a wide range of domains, including news, stories and daily scenarios. In addition, we have a collection of manually created templates in terms of 7 main categories, as shown in Table 3. Each template $T = \{s_1, s_2, \ldots, s_N\}$ provides an outline of an article and contains up to $N = 6$ sentences. Each sentence s_i describes a particular aspect of the topic.

The problem is solved in two steps. First, an index for all the documents is built using BM25. For each query, it will return a set of candidate documents that are related to the topic. Second, we use representation models to re-rank the top 100 candidates: each query-document pair $(\mathbf{q}, \mathbf{d})$ is mapped to a score $f(\mathbf{q}, \mathbf{d})$, where the scoring function f is based on cosine similarity. Quality of the generated passages was assessed by two native Chinese speakers, who were asked to examine whether the retrieved paragraphs were "relevant" to the topic and "conveyed the meaning" of the given sentence. Results are summarized in Table 5. Although nearly 62% of the paragraphs retrieved

"B"-BM25, "L"-LASER, "S"-SpanBERT, "U"-BURT

Query: 端午节的由来 (*The **Origin** of the Dragon Boat Festival*)
B: 一个中学的高级教师陈老师生动地解读端午节的由来，深深打动了在场的观众... (*Mr. Chen, senior teacher at a middle School, vividly introduced the **origin** of the Dragon Boat Festival and deeply moved the audience...*)
L: 今天是端午小长假第一天...当天上午，在车厢内满目挂有与端午节相关的民俗故事及有关诗词的文字... (*Today is the first day of the Dragon Boat Festival holiday...There are folk stories and poems posted in the carriage...*)
S, U: ...端午节又称端阳节...是中华民族的传统节日。形成于先秦，发展于汉末魏晋，兴盛于唐... (*...Dragon Boat Festival, also known as Duanyang Festival...is a traditional festival of the Chinese nation. It is formed in the Pre-Qin Dynasty, developed in the late Han and Wei-Jin, and prospered in the Tang...*)
Comments: B and L is related to the topic but does not convey the meaning of the query.
Query: 狗的喂养知识 (*Dog **Feeding** Tips*)
B: ...创建一个“比特狗”账户，并支付99元领养一只“比特狗”。然后购买喂养套餐喂养“比特狗”，“比特狗”就能通过每天挖矿产生BTGS虚拟货币。 (*...First create a "Bitdog" account and pay 99 yuan to adopt a "Bitdog". Then buy a package to **feed** the "Bitdog", which can generate virtual currency BTGS through daily mining.*)
L: 要养成定时定量喂食的好习惯，帮助狗狗更好的消化和吸收，同时也要选择些低盐健康的狗粮... (*It is necessary to feed your dog regularly and quantitatively, which can help them digest and absorb better. Meanwhile, choose some low-salt and healthy dog food...*)
S: 泰迪犬容易褪色是受到基因和护理不当的影响，其次是饮食太咸...一定要注意正确护理，定期洗澡...还要给泰迪低盐营养的优质狗粮... (*Teddy bear dog's hair is easy to fade because of its genes and improper care. It is also caused by salty diet... So we must take good care of them, such as taking a bath regularly, and preparing dog food with low salt...*)
U: 还可以一周自制一次狗粮给狗狗喂食，就是买些肉类，蔬菜，自己动手做...日常的话，建议选择些适口性强的狗粮，有助磨牙，防止口腔疾病。 (*You can also make dog food once a week, such as meats and vegetables... In daily life, it is recommended to choose some palatable dog food to help their teeth grinding and prevent oral diseases.*)
Comments: B is not a relevant paragraph. S is relevant to the topic but is inaccurate.

Table 6: Examples of the retrieved paragraphs and corresponding comments from the judges.

by BM25 are relevant to the topic, only two-thirds of them actually convey the original meaning of the template. Despite LASER's comparable performance to BURT on FAQ, it is less effective when different granular linguistic units are involved at the same time. Re-ranking using BURT substantially improves the quality of the generated paragraphs. We show examples retrieved by BM25 , LASER, SpanBERT and BURT in Table 6, denoted by B, L, S and U, respectively. BM25 tends to favor paragraphs that contain the keywords even though the paragraph conveys a different meaning, while BURT selects accurate answers according to semantic meanings of queries and documents.

5 Conclusion

This paper formally introduces the task of universal representation learning and then presents a pre-trained language model for such a purpose to map different granular linguistic units into the same vector space where similar sequences have similar representations. Our method extends BERT's masking and training objective to a more general level, which leverage information from sequences of different lengths in a comprehensive way. Overall, our proposed BURT outperforms BERT and BERT-wwm on a wide range of downstream tasks with regard to sequences of different lengths, and generates high-quality vectors that can be applied to applications such as FAQ retrieval and natural language generation.

REFERENCES

Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Trans. Assoc. Comput. Linguistics*, 7:597–610, 2019. URL https://transacl.org/ojs/index.php/tacl/article/view/1742.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075.

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. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 169–174, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2029.

Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 670–680, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1070.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. Pre-training with whole word masking for chinese bert, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification, 2016.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in neural information processing systems*, pp. 3294–3302, 2015.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations, 2019.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

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

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pp. 2227–2237, 2018.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *URL https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*, 2018.

A. Stolcke. Srilm - an extensible language modeling toolkit. In *INTERSPEECH*, 2002.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Wei Wang, Bin Bi, Ming Yan, Chen Wu, Jiangnan Xia, Zuyi Bao, Liwei Peng, and Luo Si. Structbert: Incorporating language structures into pre-training for deep language understanding. In *International Conference on Learning Representations*, 2020.

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