wang_2021_phrase_bert_improved_phrase_embeddings_from_bert_with_an_application_to_corpus_exploration

Year

2021

Author(s)

Wang, Shufan and Thompson, Laure and Iyyer, Mohit

Title

Phrase-BERT: Improved Phrase Embeddings from BERT with an Application to Corpus Exploration

Venue

EMNLP

Topic labeling

Fully automated

Focus

Secondary

Type of contribution

Novel

Underlying technique

Transformer-based

Topic labeling parameters

\

Label generation

PNTM is able to generate sentence-level topic interpretation, which offers an even more fine-grained understanding of learned topics.
Since the underlying BERT model of PNTM's embedding function is fine-tuned on both sentence and phrase-level data, its representations are semantically meaningful across multiple scales of text.
We also do not have to retrain the model to interpret topics with sentences; rather, we just have to encode the training sentences (or potentially sentences from an external corpus) with our embedding model (PNTM) and then add them to the vocabulary.

Interpreting with words / phrases
missourian, american history, county route, alabama, confederate, a state highway
Interpreting with sentences
1. At its 1864 convention , the Republican Party selected Johnson , a War Democrat from the Southern state of Tennessee , as his running mate .
2. Burnett also raised a Confederate regiment at Hopkinsville , Kentucky , and briefly served in the Confederate States Army .
3. Parker was nominated for Missouri 's 7th congressional district on September 13 , 1870 , backed by the Radical faction of the Republican party .

Table 11: Sentence-level interpretation makes it clear that this topic is about Civil War-era American history, while word and phrase interpretation offers a more high-level view.

The example shows a topic from a PNTM model trained with K = 50 on Wikipedia.
When interpreted with just words and phrases, the topic looks like it focuses on Southern and Midwestern U.S. states and their history. However, when interpreting the same topic with sentences from the training set, we observe that the most probable sentences for this topic all reference the Civil War / Reconstruction era of U.S. history.

Motivation

Offering (human readable) sentence-level interpretations of topics.
“This functionality has potential to help with automatic topic labeling”
“These kinds of observations might influence not only a practitioner’s labeling of a particular topic, but also how they use the topic model itself.”

Topic modeling

Transformer-based (BERT) with auto encoder (phrase-based neural topic model)

Topic modeling parameters

Nr of generated topics (K): 50
Embedding size (d): 768

Nr. of topics

50

Label

Complete sentences generated by the Phrase-based neural topic model

Label selection

\

Label quality evaluation

\

Assessors

\

Domain

Domain (paper): Transformer models
Domain (dataset):

Problem statement

Propose a contrastive fine-tuning objective that enables BERT to produce more powerful phrase embeddings.
Showing that Phrase-BERT embeddings can be easily integrated with a simple autoencoder to build a phrase-based neural topic model that interprets topics as mixtures of words and phrases by performing a nearest neighbor search in the embedding space.

Corpus

Dataset 1
Origin: Wikipedia
Nr. of documents: 35.255
Details: Wikipedia articles (Merity et al., 2017)

Dataset 2
Origin: Storium Evaluation Platform
Nr. of documents: 5743
Details: Fictional stories (Akoury et al., 2020)

Dataset 3
Origin: Amazon
Nr. of documents: 82.83 M
Details: Online user product reviews (He and McAuley, 2016)

Dataset	# Docs	# Words	# Phrases	Tok/doc
Wikipedia	304K	47.2K	75K	396
Storium	419K	44.0K	75K	190
Reviews	10K	32.4K	75K	101

Table 9: Corpus statistics for our three datasets, including the number of unique word and phrase types in our precomputed vocabulary. Note that we cap the number of unique phrases to the 75K most frequent.

Document

Dataset 1

A single wikipedia article

Dataset 2

A story written on STORIUM from January 2015 to August 2019

Dataset 3

A single amazon review

Pre-processing

```
@inproceedings{wang_2021_phrase_bert_improved_phrase_embeddings_from_bert_with_an_application_to_corpus_exploration,
    title = "Phrase-{BERT}: Improved Phrase Embeddings from {BERT} with an Application to Corpus Exploration",
    author = "Wang, Shufan  and
      Thompson, Laure  and
      Iyyer, Mohit",
    booktitle = "Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing",
    month = nov,
    year = "2021",
    address = "Online and Punta Cana, Dominican Republic",
    publisher = "Association for Computational Linguistics",
    url = "https://aclanthology.org/2021.emnlp-main.846",
    doi = "10.18653/v1/2021.emnlp-main.846",
    pages = "10837--10851",
    abstract = "Phrase representations derived from BERT often do not exhibit complex phrasal compositionality, as the model relies instead on lexical similarity to determine semantic relatedness. In this paper, we propose a contrastive fine-tuning objective that enables BERT to produce more powerful phrase embeddings. Our approach (Phrase-BERT) relies on a dataset of diverse
```

phrasal paraphrases, which is automatically generated using a paraphrase generation model, as well as a large-scale dataset of phrases in context mined from the Books3 corpus. Phrase-BERT outperforms baselines across a variety of phrase-level similarity tasks, while also demonstrating increased lexical diversity between nearest neighbors in the vector space. Finally, as a case study, we show that Phrase-BERT embeddings can be easily integrated with a simple autoencoder to build a phrase-based neural topic model that interprets topics as mixtures of words and phrases by performing a nearest neighbor search in the embedding space. Crowdsourced evaluations demonstrate that this phrase-based topic model produces more coherent and meaningful topics than baseline word and phrase-level topic models, further validating the utility of Phrase-BERT.",
}

#Thesis/Papers/Initial