he_2021_automatic_topic_labeling_model_with_paired_attention_based_on_pre_trained_deep_neural_network

Year

2021

Author(s)

He, Dongbin and Ren, Yanzhao and Khattak, Abdul Mateen and Liu, Xinliang and Tao, Sha and Gao, Wanlin

Title

Automatic Topic Labeling model with Paired- Attention based on Pre-trained Deep Neural Network

Venue

IJCNN

Topic labeling

Fully automated

Focus

Primary

Type of contribution

Novel approach

Underlying technique

Transformer-based (BERT with additional layers)

BASELINES: Lexrank [38], Textrank [32], Submodular [3], and TLRank [16], BertSum [24]

Topic labeling parameters

- length of the summary topic label was limited to 100 words
- Nr of training steps: 60,000
- batch size: 100
- learning rate schedule of optimizer Adam with warmingup on first 4,000 steps.
- Besides, we set P_sel =0.5 in Eq. (5).

Label generation

Bhatia et al. [12] proposed a simple [...] approach [that] was able to find a succinct phrase label for each topic, and outperformed other compared topic labeling systems.

We [...] propose an extractive summarisation topic labeling model based on a BERT
Inspired by Liu [24], we transform the topic labeling process into a single-document summarisation task.

To further improve the quality of topic labeling, we proposed a novel Topic Labeling model with a Paired-Attention (TLPA), consisting of three layers: BERT layer, extracting layer, and summarisation layer.

The BERT layer is located at the bottom, which is used to encode the input sentences.
On top of the model, there is an extractive summarization layer to extract the appropriate sentence to composite the summarization topic label.
The output of the BERT layer is the contextual embeddings of sentences, which can be directly fed to the top layer to learn how to generate topic labels.
To improve TLPA's performance, we trained an additional middle layer that, from the BERT layer, extracts the important features and filters the noise information.

Two-phase topic labeling
we introduced a two phases topic labeling model, TLPA

First phase

we applied Wan and Wang [3] approach to score sentences of the current corpus for each discovered topic. They reported that most sentences in the corpus were irrelevant to the given topic. Thus, the needs was to extract the top-500 sentences from AP/SIGMOD for a given topic, and then consider them as Candidate Sentences Set (CSSet).

KLD is used to measure the relevance between sentences and the discovered topics. The equation is defined as follows.

$$KLD(T,S) = \sum_{w \in S \cup T} P_T(w) * \log \frac{P_T(w)}{tf(w,S)/|S|}$$

Where $KLD(T,S)$ represents the KLD between topic T and sentence S, $tf(w,S)$ represents the frequency of word w in S, and $|S|$ denotes the word count of S. According to Wan and Wang [3], if a word w is not in sentence S, the $tf(w,S)/|S|$ would be replaced with 0.00001.

After the scoring process, we obtain the KLD between each sentence and each discovered topic in the current corpus. It can be used as the golden-standard data to train the extracting layer

Second phase

We used a novel summarization approach to label discovered topics if the candidate sentences are viewed as a single document.

We proposed a BERT-based summarization model to label the discovered topics. This transformed a topic labeling task into a textual summarization task.

Specifically, in the process of generating a topic label, we observe the issue of whether a sentence is included in the final summarization topic label as the optimization problem of binary classification.

Our model consists of three parts, as shown in Fig. 1.

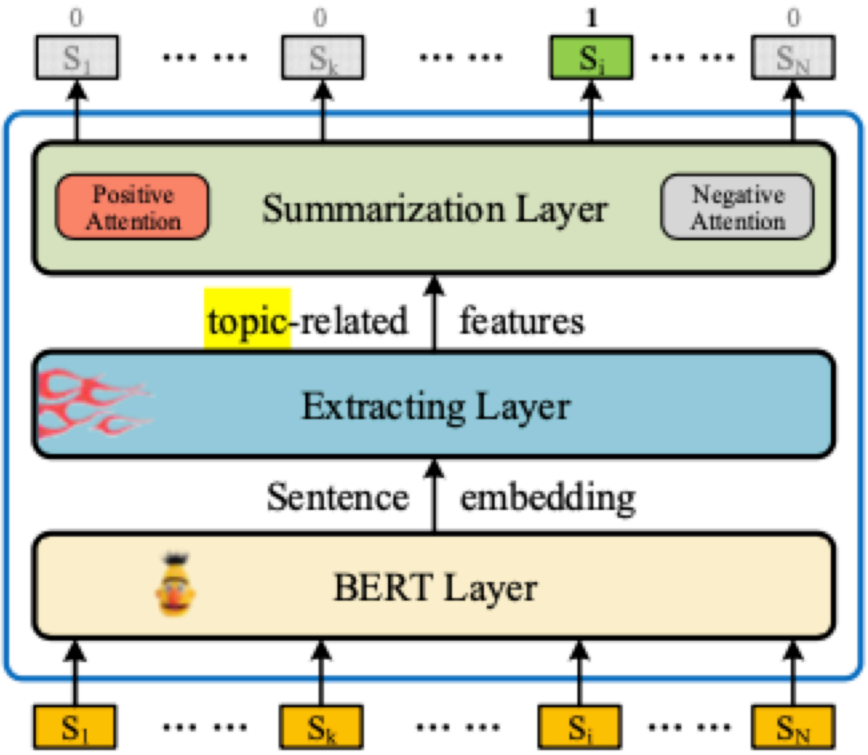

Fig. 1. The three layers of TLPA: the bottom layer is a pre-trained BERT that operates at the sentence level; the top layer is a classifier that runs over document level, which summarizes the discovered topics with paired-attention; the middle layer is a sentence features extractor that runs at the sentence level.

Training data

CURRENTLY SKIPPED

BERT layer

we used the pre-trained language model BERT to encode the input sentence to contextual embedding.

Document D consisted of N sentences [S1, S2, ..., SN], as a batch of data, is fed to the BERT layer.

The BERT encoded the D to a vector X=[X1, X2, ..., XN], which can be viewed as a batch of sentences embeddings, while the Xi represents i-th component of X corresponding to the i-th sentence Si of D.

We inserted token [CLS] before each sentence and appended token [SEP] at the end.

For a sentence S in the document D, the corresponding vector XS is represented by the token [CLS] from the top BERT output embeddings.

Extracting Layer

Not strictly necessary, but to generate a better Relevance, Coverage, and Discrimination topic label, we should provide the topic-related features of each candidate sentence to the summarization layer.

That can help to select sentences accurately and control redundancy easily.
So we propose a specific extracting layer to learn how to extract the useful features from the dense vector of each sentence based on the attention mechanism. Training a feature extractor in advance can accurately extract the important features related to a given topic from the embedding of sentences. This has a significant effect on improving the accuracy of binary classifiers in the summarization layer.
[FURTHER DETAILS IN PAPER]

Summarization Layer
We used a paired-attention to evaluate each sentence for bonus and penalty points separately in our summarization layer. The two opposite attention simulate the human behavior of considering two aspects while making decisions. The positive and negative differences can be added up to combine the overall difference in each value range, and more accurate assessment results can be obtained.
[FURTHER DETAILS IN PAPER]

Motivation

To reduce the cognitive overhead of understanding these topics for users, the application of automatic topics labeling technique to generate meaningful labels is receiving more attention and has become a very challenging work.

Topic modeling

LDA

Topic modeling parameters

Nr of topics: 25

Nr. of topics

25

Label

summary topic label with at most 100 words

Label selection

\

Label quality evaluation

According to Wan and Wang [3] and Mei et al. [4], the topic summary generated for each topic should satisfy the three criteria of higher Relevance, Coverage, and Discrimination.

Relevance

The higher relevance and lesser the redundancy, the better the summary topic labels are.
For each topic labeling model, we first computed the KLD value between each generated topic label and the corresponding discovered topic, and then averaged all the KLD values in the case of AP and SIGMOD.

TABLE I. COMPARISON OF THE AVERAGE OF KL DIVERGENCE BETWEEN SUMMARIZATION TOPIC LABEL AND CORRESPONDING DISCOVERED TOPIC

model	Relevance	
	AP	*SIGMOD*
LexRank	4.08042	3.18463
TextRank	3.92700	3.25402
Submodular	3.30045	2.73337
TLRANK-C	2.85353	2.04545
TLRANK-G	2.86216	2.03687
BERTSum-NP	3.20660	2.39269
TLPA-NEL	3.04770	2.05666
TLPA	**2.63031**	**1.82040**

Coverage

According to Wan and Wang [3], Coverage is defined as the ratio of words that appeared in the top 20 topic terms. It chooses the top-20 terms instead of 500, because the top-20 terms are more significant and representative than the rest. In general, higher Coverage denotes more top topic terms covered by the topic label and is more comprehensive.

We average all the Coverage values in the case of AP and SIGMOD.

TABLE II. A COMPARISON OF THE MEAN RATIO OF THE WORDS COVERED OUT OF TOP 20 TOPIC TERMS FOR EACH TOPIC

model	Coverage	
	AP	*SIGMOD*
LexRank	0.15	0.18
TextRank	0.17	0.13999
Submodular	0.14	0.14
TLRANK-C	0.23	0.25
TLRANK-G	0.23	**0.25**
BERTSum-NP	0.19	0.19
TLPA-NEL	0.19	0.21
TLPA	**0.26**	**0.23**

Discrimination

The same sentence may have different probability distributions for each topic. Therefore, we should try our best to avoid the same or similar sentences appearing in different topic labels. Thus, the smaller the similarity between the topics labels, the better the quality of the topics labels is. To obtain Discrimination, we compute the cosine similarity between two different topic labels and then average all the similarity values.

TABLE III. A COMPARISON OF THE MEANS OF ALL COSINE SIMILARITY VALUES AMONG DIFFERENT TOPICS LABELS

model	Discrimination	
	AP	*SIGMOD*
LexRank	0.03007	0.05963
TextRank	0.03744	0.07735
Submodular	**0.00729**	**0.01661**
TLRANK-C	0.02711	0.04743
TLRANK-G	0.0277	0.04988
BERTSum-NP	0.03416	0.05411
TLPA-NEL	0.02119	0.05599
TLPA	**0.01849**	**0.04599**

Assessors

\

Domain

Paper: Topic labeling

Dataset: News

Problem statement

The automatic topic labeling model aims at generating a sound, interpretable, and meaningful topic label that is used to interpret an LDA-style discovered topic, intending to reduce the cognitive load of end-users while browsing or investigating the topics.
In this study, we first introduced the pre-trained language model BERT to topic labeling tasks. It exploits the contextual embedding of the pre-trained language model to improve the quality of encoding sentences. To generate a topic label with higher Relevance, Coverage, and Discrimination, we propose a novel summarization neural framework.
Specifically, it exploits the paired-attention to model the relationship between the candidate sentences first and then decides which sentences should be included in the final summarization topic label.
Moreover, we expected that high-quality sentence encoding representation could improve our model's performance. So, for each discovered topic, we trained a specific layer to extract the important topic-related features from the sentence embeddings as well as filter the noise information. The experimental results showed that our model significantly outperforms the state-of-the-art and classic topic labeling models.

Corpus

Origin: SIGMOD and APNews

Nr. of documents: 3016 and 2246

Details:

- 24,000 training documents and 6,000 test documents for AP and SIGMOD collections respectively

Document

each training or test document D had 100 sentences, and the oracle sentences were tagged

Pre-processing

- punctuations, digital numbers and stop words have been removed
- words that are not nouns, verbs, adjectives, adverbs, or pronouns have been filtered out

---

```
@inproceedings{he_2021_automatic_topic_labeling_model_with_paired_attention_based_on_pre_trained_deep_neural_network,
    author = {He, Dongbin and Ren, Yanzhao and Khattak, Abdul Mateen and Liu, Xinliang and Tao, Sha and Gao, Wanlin},
    booktitle = {2021 International Joint Conference on Neural Networks (IJCNN)},
    date-added = {2023-04-23 21:14:46 +0200},
    date-modified = {2023-04-23 21:14:46 +0200},
    doi = {10.1109/IJCNN52387.2021.9534093},
    pages = {1-9},
    title = {Automatic Topic Labeling model with Paired-Attention based on Pre-trained Deep Neural Network},
    year = {2021}}
```

#Thesis/Papers/FS