lau_2017_topically_driven_neural_language_model

Year

2017

Author(s)

Lau, Jey Han and Baldwin, Timothy and Cohn, Trevor

Title

Topically Driven Neural Language Model

Venue

ACL

Topic labeling

Fully automated

Focus

Secondary

Type of contribution

Novel approach

Underlying technique

Result of supervised topic modeling (supervised tdlm), LSTM-based

Topic labeling parameters

LSTM size = “large”

Label generation

Approach 1

Result of supervised topic modeling (incorporating gold standard document labels)

Approach 2

Topics generated by topic models are typically in- terpreted by way of their top-N highest probabil- ity words. In tdlm, we can additionally generate sentences related to the topic, providing another way to understand the topics. To do this, we can constrain the topic vector for the language model to be the topic output vector of a particular topic

$$\begin{aligned}
\mathbf{z}_t &= \sigma(\mathbf{W}_z\mathbf{s} + \mathbf{U}_z\mathbf{h}_t + \mathbf{b}_z) \\
\mathbf{r}_t &= \sigma(\mathbf{W}_r\mathbf{s} + \mathbf{U}_r\mathbf{h}_t + \mathbf{b}_r) \\
\hat{\mathbf{h}}_t &= \tanh(\mathbf{W}_h\mathbf{s} + \mathbf{U}_h(\mathbf{r}_t \odot \mathbf{h}_t) + \mathbf{b}_h) \\
\mathbf{h}'_t &= (1 - \mathbf{z}_t) \odot \mathbf{h}_t + \mathbf{z}_t \odot \hat{\mathbf{h}}_t
\end{aligned} \tag{3}$$

where $\mathbf{z}_t$ and $\mathbf{r}_t$ are the update and reset gate activations respectively at timestep t. The new hidden state $\mathbf{h}'_t$ is connected to a dense layer with linear transformation and softmax output to predict the next word, and the model is optimised using standard categorical cross-entropy loss.

We present 4 topics from a APNEWS model (k = 100) and 3 randomly generated sentences conditioned on each topic in Table 8.

Topic	Generated Sentences
protesters suspect gunman officers occupy gun arrests suspects shooting officer	• police say a suspect in the shooting was shot in the chest and later shot and killed by a police officer . • a police officer shot her in the chest and the man was killed . • police have said four men have been killed in a shooting in suburban london .
film awards actress comedy music actor album show nominations movie	• it 's like it 's not fair to keep a star in a light , " he says . • but james , a four-time star , is just a ⟨unk⟩. • a ⟨unk⟩ adaptation of the movie " the dark knight rises " won best picture and he was nominated for best drama for best director of " ⟨unk⟩, " which will be presented sunday night .
storm snow weather inches flooding rain service winds tornado forecasters	• temperatures are forecast to remain above freezing enough to reach a tropical storm or heaviest temperatures . • snowfall totals were one of the busiest in the country . • forecasters say tornado irene 's strong winds could ease visibility and funnel clouds of snow from snow monday to the mountains .
virus nile flu vaccine disease outbreak infected symptoms cough tested	• he says the disease was transmitted by an infected person . • ⟨unk⟩ says the man 's symptoms are spread away from the heat . • meanwhile in the ⟨unk⟩, the virus has been common in the mojave desert .

Table 8: Generated sentences for APNEWS topics.

Motivation

The generated sentences highlight the content of the topics, providing another interpretable aspect for the topics. These results also reinforce that the language model is driven by topics.

Topic modeling

(Supervised) Neural topic model (associated with LSTM-based language model)

Baselines: (Supervised) nmt Cao et al., 2015 , sLDA (McAuliffe and Blei, 2008)

The architecture of the proposed topically-driven language model (henceforth "tdlm") is illustrated in Figure 1.
There are two components in tdlm:

- The language model is designed to capture word relations in sentences, while the topic model learns topical information in documents. The language model is a standard LSTM language model.
- The topic model works like an auto-encoder, where it is given the document words as input and optimised to predict them. The topic model takes in word embeddings of a document and generates a document vector us- ing a convolutional network. Given the document vector, we associate it with the topics via an attention scheme to compute a weighted mean of topic vectors, which is then used to predict a word in the document.

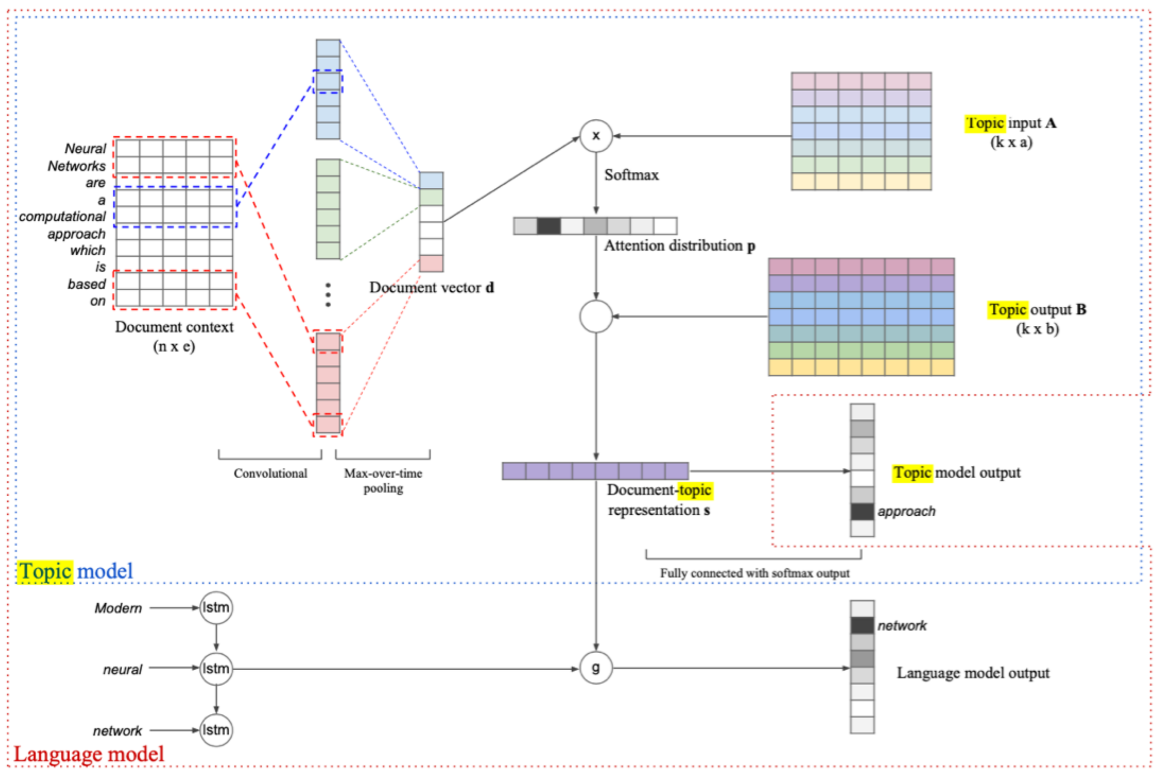

Figure 1: Architecture of `tdlm`. Scope of the models are denoted by dotted lines: blue line denotes the scope of the topic model, red the language model.

In datasets where document labels are known, supervised topic model extensions are designed to leverage the additional information to improve modelling quality. The supervised setting also has an additional advantage in that model evaluation is simpler, since models can be quantitatively assessed via classification accuracy.
To incorporate supervised document labels, we treat document classification as another sub-task in tdlm. Given a document and its label, we feed the document through the topic model network to generate the document-topic representation s, and connect it to another dense layer with softmax out- put to generate the probability distribution over classes.

To incorporate supervised document labels, we treat document classification as another sub-task in tdlm. Given a document and its label, we feed the document through the topic model network to generate the document-topic representation s, and connect it to another dense layer with softmax output to generate the probability distribution over classes.

Topic modeling parameters

Most hyper-parameter values for tdlm are similar to those used in the language and topic model experiments;

Hyper-parameter	Value	Description
m_1	3	Output sequence length for topic model
m_2	30	Sequence length for language model
m_3	300,150,500	Maximum document length
n_{batch}	64	Minibatch size
n_{layer}	1,2	Number of LSTM layers
n_{hidden}	600,900	LSTM hidden size
n_{epoch}	10	Number of training epochs
k	100,150,200	Number of topics
e	300	Word embedding size
h	2	Convolutional filter width
a	20	Topic input vector size or number of features for convolutional filter
b	50	Topic output vector size
l	0.001	Learning rate of optimiser
p_1	0.4	Topic model dropout keep probability
p_2	0.6	Language model dropout keep probability

Table 1: `tdlm` hyper-parameters; we experiment with 2 LSTM settings and 3 topic numbers, and m_3 varies across the three domains (APNEWS, IMDB, and BNC).

the only exceptions are:

- a: 80
- b: 100
- nepoch: 20
- m3:150
- LSTM: 1 layer + 600 hidden

Nr. of topics

50, 100, 150

Label

Approach 1

One of 20 gold standard labels from the 20Newsgroup dataset

Approach 2

A descriptive sentence generated by the LSTM language model

Label selection

\

Label quality evaluation

\

Assessors

\

Domain

Paper:

Dataset: News

Problem statement

We present a neural language model that incorporates document context in the form of a topic model-like architecture, thus providing a succinct representation of the broader document context outside of the current sentence.

Corpus

Origin: Various news sources (20News dataset)

Nr. of documents: 18.846

Details:

Partition	#Docs	#Tokens
Training	9314	2.6M
Development	2000	0.5M
Test	7532	1.7M

Table 5: 20NEWS preprocessed statistics.

Origin: Various news sources (APNews dataset)

Nr. of documents: 54.000

Details:

Collection	Training		Development		Test	
	#Docs	#Tokens	#Docs	#Tokens	#Docs	#Tokens
APNEWS	50K	15M	2K	0.6M	2K	0.6M

Document

APNews
news article contains additional document metadata, including subject classsification tags, such as "General News", "Accidents and Disasters", and "Military and Defense".

Pre-processing

- tokenise words and sentence
- lowercase all words
- filter low/high frequency word types and stopwords.

```
@inproceedings{lau_2017_topically_driven_neural_language_model,
    title = "Topically Driven Neural Language Model",
    author = "Lau, Jey Han  and
      Baldwin, Timothy  and
      Cohn, Trevor",
    booktitle = "Proceedings of the 55th Annual Meeting of the Association for
Computational Linguistics (Volume 1: Long Papers)",
    month = jul,
    year = "2017",
    address = "Vancouver, Canada",
    publisher = "Association for Computational Linguistics",
    url = "https://aclanthology.org/P17-1033",
    doi = "10.18653/v1/P17-1033",
    pages = "355--365",
    abstract = "Language models are typically applied at the sentence level,
without access to the broader document context. We present a neural language
model that incorporates document context in the form of a topic model-like
architecture, thus providing a succinct representation of the broader document
context outside of the current sentence. Experiments over a range of datasets
```

```
demonstrate that our model outperforms a pure sentence-based model in terms of
language model perplexity, and leads to topics that are potentially more
coherent than those produced by a standard LDA topic model. Our model also has
the ability to generate related sentences for a topic, providing another way to
interpret topics.",
}
```

#Thesis/Papers/Initial