zhang_2019_learning_document_representation_via_topic_enhanced_lstm_model

Year

2019

Author(s)

Wenyue Zhang and Yang Li and Suge Wang

Title

Learning document representation via topic-enhanced LSTM model

Venue

Knowledge-Based Systems

Topic labeling

Fully automated

Focus

Secondary

Type of contribution

Novel approach

Underlying technique

Result of supervised topic modeling (Latent topic modeling layer)

Topic labeling parameters

\

Label generation

Table 6
Performance of topic detection.

Models	Topics	Words	C_v
lda2vec	Space	Astronomical, astronomy, satellite, planetary, telescope	0.556
	Encryption	Encryption, wiretap, encrypt, escrow, clipper	0.572
	X Windows	Mydisplay, xlib, window, cursor, pixmap	0.472
	Mid-East	Armenian, Lebanese, Muslim, Turk, Sy	0.200
	All 20 Topics		0.567(avg)
TE-LSTM	Space	Space, orbit, mission, astro, NASA	0.593
	Encryption	Encryption, cryptography, key, escrow, chip	0.610
	X Windows	Motif, faq, widget, window, windows	0.602
	Mid-East	Israel, Israeli, Lebanon, Armenia, Arab	**0.671**
	All 20 Topics		0.588(avg)
TE-LSTM+SC	Space	Orbit, space, shuttle, launch, mars	**0.600**
	Encryption	Cryptography, key, encryption, nsa, chip	**0.618**
	X Windows	Motif, faq, sunos, windows, widget	**0.621**
	Mid-East	Israel, Israeli, Armenia, Armenian, Jew	0.630
	All 20 Topics		**0.611**(avg)

Table 7
Performance of rest detected topics detected by the TE-LSTM+SC.

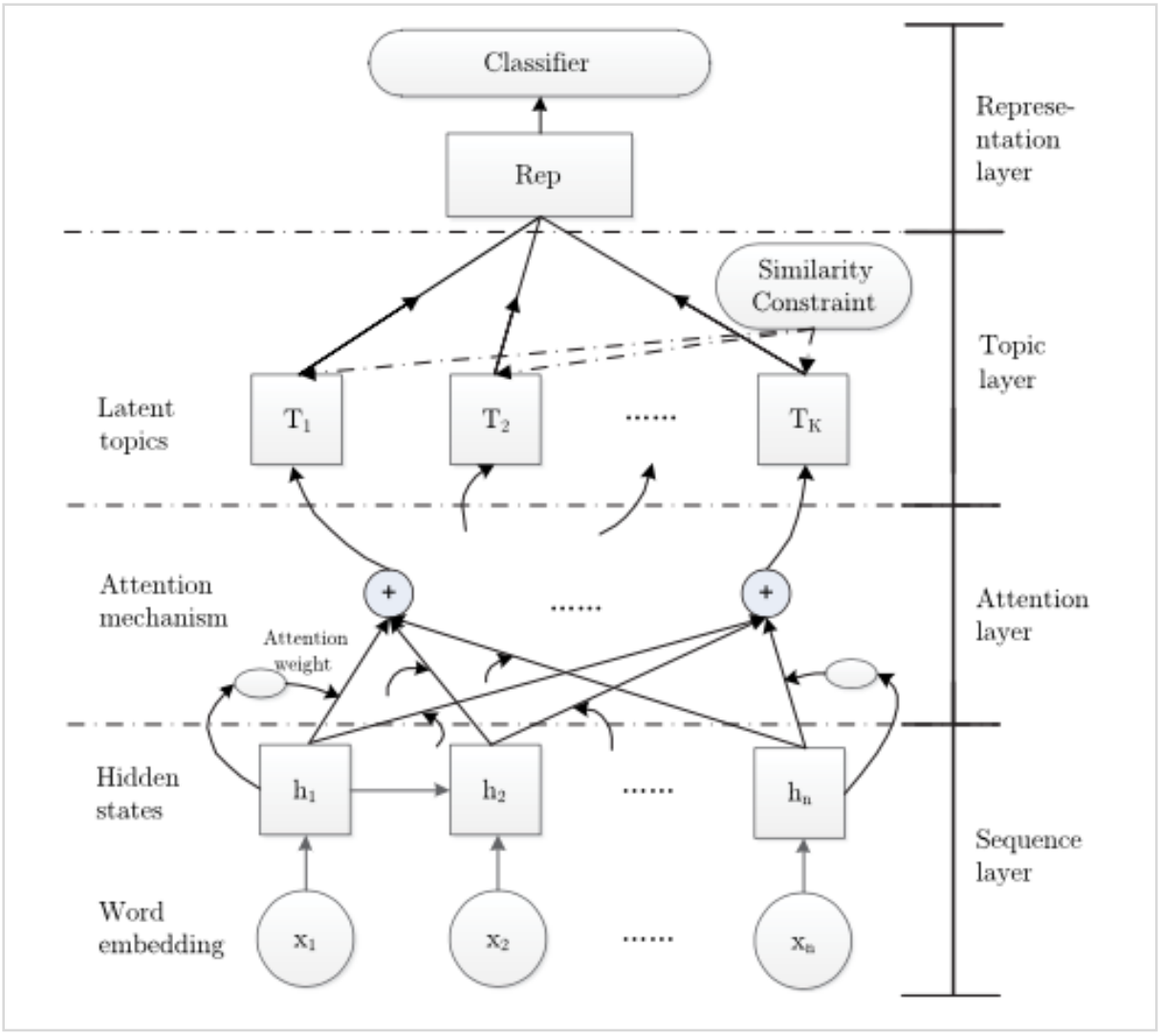
Topics	Words	C_v
Atheism	Atheism, god, murder, atheist, car	0.547
Graphics	Graphics, polygon, algorithm, image, windows	0.498
PC.hardware	Windows, isa, gateway, bios, vlb	0.611
Mac.hardware	Mac, centris, apple, quadra, iisi	0.664
Forsale	Sale, shipping, offer, sell, obo	0.607
Autos	Car, dealer, honda, toyota, wheel	0.589
Motorcycles	Bike, dod, ride, motorcycle, rider	0.723
Baseball	Baseball, season, team, hitter, player	0.641
Hockey	Hockey, nhl, lindros, selanne, team	0.659
Electronics	Car, electronics, circuit, project, joystick	0.494
Med	Drug, med, treatment, medical, patient	0.623
Christian	Christian, Jesus, Christ, God, Christianity	0.703
Guns	Gun, handgun, firearm, weapon, baseball	0.616
Politics	Homosexual, sexual, gay, windows, economic	0.641
Religion	Christian, gay, Christianity, God, religion	0.584
Ms-windows	Windows, Indiana, Microsoft, version, site	0.551

Motivation

\

Topic modeling

Latent topic modeling layer implemented as part of the proposed LSTM model



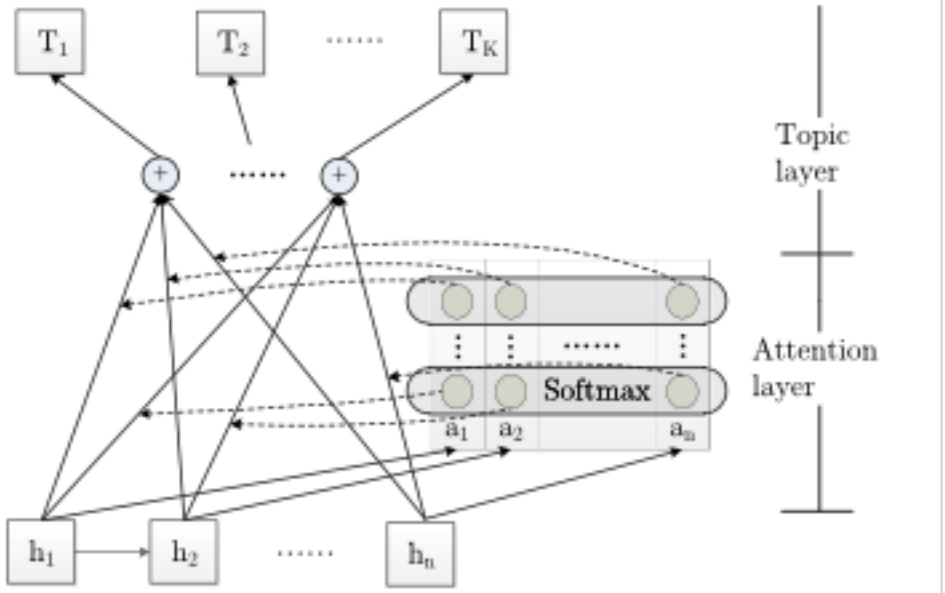


Fig. 2. Latent topic modeling layer. Taking as input hidden states and normalized attentive relation vectors, the layer then produces latent semantic topics (Ts and TCs).

BASELINE: lda2vec

Topic modeling parameters

Nr. of topics

of classes
20
25
2
6

Label

Class labels originally assigned to documents belonging to the four datasets.

Label selection

\

Label quality evaluation

\

Assessors

\

Domain

Paper: Document representation
Dataset: News, Miscellaneous (Wikipedia, SemEval2007), Online store reviews

Problem statement

In this work, we propose a new topic-enhanced LSTM model to deal with the document representation problem.
We first employ an attention-based LSTM model to generate hidden representation of

word sequence in a given document.
Then, we introduce a latent topic modeling layer with similarity constraint on the local hidden representation, and build a tree-structured LSTM on top of the topic layer for generating semantic representation of the document.
We evaluate our model in typical text mining applications, i.e., document classification, topic detection, information retrieval, and document clustering.

Corpus

20Newsgroup
wiki10
amazon reviews - amazon reviews dataset consists of 4 different domains, where each document is classified into 1 out of 2 sentiment polarities
SemEval2007 - contains 1250 documents that are labeled with the following emotion categories, i.e., anger, disgust, fear, joy, sadness, and surprise

Table 2
Statistics of datasets.

Datasets	# of documents	Avg length of doc	# of classes
20Newsgroup	18,848	93	20
Wiki10+	17,325	936	25
Amazon	8000	108	2
SemEval	1250	6	6

Document

Pre-processing

documents that contain less than 6 words are removed

```
@article{zhang_2019_learning_document_representation_via_topic_enhanced_lstm_model,
  abstract = {Document representation plays an important role in the fields of text mining, natural language processing, and information retrieval. Traditional approaches to document representation may suffer from the disregard of the correlations or order of words in a document, due to unrealistic assumption of word independence or exchangeability. Recently, long--short-term
```

memory (LSTM) based recurrent neural networks have been shown effective in preserving local contextual sequential patterns of words in a document, but using the LSTM model alone may not be adequate to capture global topical semantics for learning document representation. In this work, we propose a new topic-enhanced LSTM model to deal with the document representation problem. We first employ an attention-based LSTM model to generate hidden representation of word sequence in a given document. Then, we introduce a latent topic modeling layer with similarity constraint on the local hidden representation, and build a tree-structured LSTM on top of the topic layer for generating semantic representation of the document. We evaluate our model in typical text mining applications, i.e., document classification, topic detection, information retrieval, and document clustering. Experimental results on real-world datasets show the benefit of our innovations over state-of-the-art baseline methods.},
  author = {Wenyue Zhang and Yang Li and Suge Wang},
  date-added = {2023-03-26 19:37:09 +0200},
  date-modified = {2023-03-26 19:37:09 +0200},
  doi = {https://doi.org/10.1016/j.knosys.2019.03.007},
  issn = {0950-7051},
  journal = {Knowledge-Based Systems},
  keywords = {Document representation, Deep learning, Long--short term memory, Topic modeling},
  pages = {194-204},
  title = {Learning document representation via topic-enhanced LSTM model},
  url = {https://www.sciencedirect.com/science/article/pii/S0950705119301182},
  volume = {174},
  year = {2019}}

#Thesis/Papers/Initial