Journal of Computer Science

Original Research Paper

Topic-Transformer for Document-Level Language Understanding

Oumaima Hourrane and El Habib Benlahmar

Laboratory Information Technology and Modeling, Faculty of Sciences Ben M'sik, Hassan II University, Casablanca, Morocco

Article history
Received: 21-09-2021
Revised: 02-11-2021
Accepted: 12-11-2021

Corresponding Author:
Oumaima Hourrane
Laboratory Information Technology and Modeling, Faculty of Sciences Ben M'sik,
Email: oumaima.hourrane-etu@etu.univh2c.ma

Abstract: As long as natural language processing applications are considered prediction problems with insufficient context, usually referred to as a single sentence or paragraph, this does not reveal how humans perceive natural language. When reading a text, humans are sensitive to much more context, such as the rest or other relevant documents. This study focuses on simultaneously capturing syntax and global semantics from a text, thus acquiring document-level understanding. Accordingly, we introduce a Topic-Transformer that combines the benefits of a neural topic model that captures global semantic information and a transformer-based language model, which can capture the local structure of texts both semantically and syntactically. Experiments on various datasets confirm that our model has a lower perplexity metric compared to standard transformer architecture and the recent topic-guided language models and generates topics that are conceivably coherent compared to those of regular Latent Dirichlet Allocation (LDA) topic model.

Keywords: Neural Topic Model, Neural Language Model, Topic-Guided Language Model, Document-Level Understanding, Long-Range Semantic Dependencies

Introduction

Several research studies have been dedicated to developing efficient probabilistic models of documents. Topics or latent variables are often included in these models, whose purpose is to apprehend unique statistical patterns in conjunction with documents. From the topic models standpoint, most of them are bag-of-words models where the ordering of words is overlooked, which is an excellent way to gain the global semantics. In another case, this can be a good reason why bag-of-words models do not work efficiently on natural language understanding tasks. Topic models have another limitation which s that they withdraw the stop words wince they do not carry any semantic meaning.

In another dimension, language models are one of the foundational methods of the NLP field that is useful in several tasks, including sentiment analysis (Hourrane and Idrissi, 2019b), machine translation (Koehn, 2009; Hourrane and Idrissi, 2019a), citations analysis and bibliometrics (Beskaravainaja and Kharybina, 2020; Hourrane *et al*., 2018); plagiarism detection (Hourrane and Benlahmar, 2017; Hourrane and Benlahmer, 2019); and many more applications. Recently, much progress has been made by neural methods based on Recurrent Neural Networks (RNN) (Mikolov *et al*., 2010), Convolutional Neural Networks (CNN) (Gehring *et al*., 2017); and self-attention networks known as Transformers (Vaswani *et al*., 2017). These language models usually predict the probability of textual tokens, most of the time at the sentence-level, considering that independence of sentences between each other. However, these models lack the capacity of capturing long-term dependency. Instead, they concatenate all the sentences and add a unique token to mark sentence boundaries. The cost of this simple approach is that the model size may quickly increase with the input sequence length, making it hungrier for computation, memory and data size and more difficult to train.

Motivated by the limitations above, we propose the Topic-Transformer, a new approach to learn the topic and language models jointly. More specifically, we incorporate the advantages of a topic model that captures global semantic information and a transformer-based language model, which efficiently represents the local structure of text both semantically and syntactically, which grants us to both sensitize the language model predictions to the long-range document record using topic vectors and to produce topics that are coherent to the local context. We conduct some experiments and exhibit that our approach is indeed apt to remarkably decrease the language model's perplexity and practically detect coherent topics.

Related Works

Several previous methods have been proposed to ameliorate the results of language models.

A Sentence Level Recurrent Topic Model (Tian *et al*., 2016) assume that the words in the same sentence share

the same topic and that the generation of a single word relies on the historical words in the same passage. For that, they used an LSTM with word embedding as input and for the topic model, similar to LDA, they assume there is a k-dimensional Dirichlet prior distribution of the topic mixture weight of each document. So, the historical words and the sentence's topic jointly affect the LSTM hidden state and the next word. Finally, the authors adopt the mean-field variational inference method for the posterior approximation of hidden variables.

At a document level, the Topic-RNN model (Dieng *et al*., 2016) aims to capture the semantics connecting words in a text/document through latent topic vectors. More precisely, this model combines an RNN-based language model that captures syntactic or local dependencies and a topic model that captures the semantics and the global dependencies. At first, draw a topic vector for the document using a normal distribution. Then, they compute the hidden state for each word in the document and draw a stop word indicator using the sigmoid function to control how the topic vector influences the output. The topic vectors are utilized as a base instead of through the RNN's hidden states. For the model inference, the authors used the vocational objective function, and they chose to infer the topic vectors by employing a feed-forward neural network.

In the same context, another model (Lau *et al*., 2017), uses the convolution max-pooling encoder that takes as an input word representation and produces a single document vector, which is then blended with the topic vectors passing through an attention mechanism that calculates the weighted average of topic vectors. These vectors predict the next term in the text/document. The authors used an LSTM language model that includes the weighted topic vectors to predict the following terms; they added an extension that includes some other metadata and document labels.

Moreover, a topic compositional neural language model (Wang *et al*., 2018) represents a method that captures both and simultaneously the global semantics and the local structure of a document. First, to learn the global semantic meaning and parameterize the multinomial topic distribution, the authors used a Random Gaussian vector passing through a Soft max function. Moreover, to learn the local structure of the document, they used as a language model Mixture-of-Expert (MoE) where each expert is an LSTM model.

An alternative approach introduces a Topic-Guided Variational Autoencoder (TGVAE) (Wang *et al*., 2019) as a language model. This model uses a Gaussian mixture model GMM tuned by a topic model, which is learned jointly with the VAE model. In addition, the author used an inference method based on the Household flow that generally encourages the complexity and diversity of the learned topics.

While these topics-guided language models have shown potentiality, they have other drawbacks. For instance, some of these models employ only pre-trained topic models. Another critical limitation lies in combining the learned topics in the language model, primarily by adding the topic vectors as additional features of a neural network.

Preliminaries

This section summarizes the background behind building the Topic-Transformer model, including constructing the latent topic model and neural language model.

Probabilistic Topic Modeling

Probabilistic subject models are based on statistical methods to identify conceptual subjects to which parts of the text might be attached. To achieve this, they infer the latent semantic structure underlying long-range unstructured text data. Using statistical co-occurrence models among words in documents, probabilistic subject models and their changes are excellent for retrieving overall semantics. However, they generally tend to suffer from document word disorder due to an unreliable hypothesis-Word exchange in models.

One of the conventional topic models is Latent Dirichlet Allocation (LDA) (cite LDA). It gives a scalable and robust approach for text modeling by including latent variables of each word, which indicate the topic distribution. The generative process of LDA can be summed as follow:

$$t \sim Dir(\alpha_0), z_n \sim Discrete(t), w_n \sim Discrete(\beta z_n)$$

where, t expresses the topic proportion of a document d. For $n \in [0, N_d]$ where N_d is the number of words in d, z_n represents the topic assignment for word w_n, α_0 is the hyper parameter of the Dirichlet prior and b_{zn} denotes the distribution over words for topic z_n. The marginal likelihood is represented as follow:

$$p(d|\alpha_0, \beta) = \int_t p(t|\prod_n \sum_{z_n} p(w_n|\beta_{z_n} p(z_n|t) dt \quad (1)$$

Neural Language Modeling

Language modeling embodies a crucial part in various natural language processing applications. Several language models exist, starting from a regular n-gram model to the most popular Transformer models, which try to answer the problem of correctly predicting the next word starting from a sequence of historical words. A language model aims to learn a probability distribution across a sequence of words in a pre-defined vocabulary. Let us denote V as the vocabulary set and $\{y_1, ..., y_{N_v}\}$ as a sequence of words with each $y_n \in V$. The likelihood of the sequence is defined through a joint probability distribution as follows:

$$p(y_1, ..., y_{N_v} = p(y_1 \prod_{n=2}^{N_v} p(y_n|y_{1:n-1})) \quad (2)$$

Recurrent Neural Network-based language models (Mikolov *et al*., 2010) define the conditional probability of each word in through the hidden state h_n given all the previous words $y_{1:n-1}$:

$$p\left(y_n \left| y_{1:n-1}\right.\right) = p\left(y_n \left| h_n\right.\right) \quad (3)$$

$$h_n = f\left(h_{n-1}, y_{n-1}\right) \quad (4)$$

where the function *f(.)* can be replaced by either a Short-Term Long Memory (LSTM) cell (Hochreiter and Schmidhuber, 1997), or a Gated Recurrent Unit (GRU) cell (Chung *et al*., 2014); While Recurrent Neural Networks achieved state-of-the-art performance on language modeling task, a recent neural network architecture called the Transformer which is based on the attention mechanism, also becomes very competitive (Vaswani *et al*., 2017); While the LSTM performed back-propagation through time by giving the model the last hidden state of the previous iteration, the Transformer passes all the previously hidden states to the current batch, to provide context to the first words in the batch. More specifically, the Transformer trains a neural network with parameter q to minimize the negative log-likelihood over a dataset $D = \{X_1,\ldots,X_{|D|}\}$:

$$L(D) = -\sum_{K=1}^{|D|} \log p_\theta\left(x_i^k \left| x_{<i}^k\right.\right) \quad (5)$$

The Topic-Transformer Models

We introduce our Topic-Transformer model as shown in Fig. 1 (a). Our model is a composition of a topic model and a transformer-based language model. The topic model tries to captivate the long-range semantics in the text, while the Transformer is intended to detect both the local semantic and syntactic relationships between words. This arrangement attempts to get better overall performance on document-level NLP tasks.

The Model

The generative process of the Topic-Transformer model is as follows:

1. Get the topic vector $\theta \sim N\left(\mu, \sigma_0^2\right)$.
2. Given the word $y_{1:n-1}$ for the *n-th* word Y_n in the document:
   (a) Perform the transformer encoding for each word in the sequence and get the output vector as follow: $z_n = z_1 \oplus ... \oplus z_{n-1}$.
   (b) Get the stop word indicator: $l_n \sim Bernoulli(\sigma(\Gamma^T z_n)\sigma$s With σbeing the Sigmoid function.
   (c) Get the word $y_n \sim p(y_n|z_n, l_n, \theta, B)$ where:

$$P\left(y_n = i \left| z_n, l_n, \theta, B\right.\right) \propto \exp\left(v_i^T z_n + (1 - l\,n) b_i^T \theta\right) \quad (6)$$

where, $N(\mu, \sigma_0^2)$ is an isotropic Gaussian distribution, with μ as the mean and σ_0^2 as the variance in each dimension. We pass the Gaussian vector into a Soft max function to fine-tune the multinomial topic assignments. Rather than utilizing the Dirichlet distribution, we prefer to use the Gaussian distribution, since it presents more flexibility in the next-word prediction task and also has benefits throughout the inference stage.

Then, we pass the input words to the transformer encoders, as depicted in Fig. 1 (b). An encoder takes each word's embeddings and positional encoding in the text sequence. The self-attention mechanism takes the resulting encoding vector and calculates their pertinence to form the output encoding, which an anticipatory neural network will then treat. This process is iterated based on the number of encoding layers from the model sittings.

Next, we add *ln* as the stop word indicator that manages how the topic vector q affects the output (Dieng *et al*., 2016). If $l_n = 1$, it indicates that y_n is a stop word, hence, the topic vector q has no supplement to the output.
Finally, a bias value is appended and calculated to favor more probable words to arise when crossing with$q$.

As shown in Fig.1 (a), we indicate all model parameters as $Q = \{G, V, B, W, W_d\}$, with W_d being a parameter for the inference network. The marks are the word sequences $y_{1:N}$ and the stop word indicators $l_{1:N}$. Therefore, the log marginal likelihood of the sequence $y_{1:N}$ is as follows:

$$\log p\left(y_{1:N}, l_{1:N} \left| Z_{1:N}\right.\right) = \log \int p(\theta) \prod_{n=0}^{N} p\left(y_n \left| z_n, l_n, \theta\right.\right) p\left(l_n \left| z_n\right.\right) d\theta \quad (7)$$

Model Inference

We use variational inference for our model inference (Blei and Jordan, 2006). We denote $q(\theta)$to be the variational distribution on the marginalized variable θ. As a result, we build the variational loss function, also named the Evidence to Lower Bound (ELBO), which can be inscribed as follows:

$$L\left(y_{1:N}, l_{1:N} \left| q(\theta), \Theta\right.\right), \triangleq \mathrm{E}_q(\theta)$$
$$\left[\sum_{n=1}^{N} \log p\left(y_n \left| z_n, l_n, \theta\right.\right) + \log p\left(l_n \left| z_n\right.\right) + \log \frac{p(\theta)}{q(\theta)}\right] \quad (8)$$

$$\leq \log p\left(y_{1:N}, l_{1:N} \left| z_n \Theta\right.\right) \quad (9)$$

where, $q(\theta)$ is a fully-connected neural network with batch normalization and dropout. The topic distribution $q(\theta|X_d)$ for each document *d* is written as follows:

$$q(\theta \left| X_d\right.) = N\left(\theta \left| g_\mu\left(X_d\right), diag\left(\exp\left(g_\sigma\left(X_d\right)\right)\right)\right.\right) \quad (10)$$

where, $g_\sigma(.)$ and $g_\mu(.)$ are feed-forward neural network implementations and X_d is bag of words of document *d*. We then apply stochastic samples from $q(\theta/X)$ and the re-parameterization trick of (Kingma and Welling, 2013) to create an low-variance and unbiased gradient estimator.

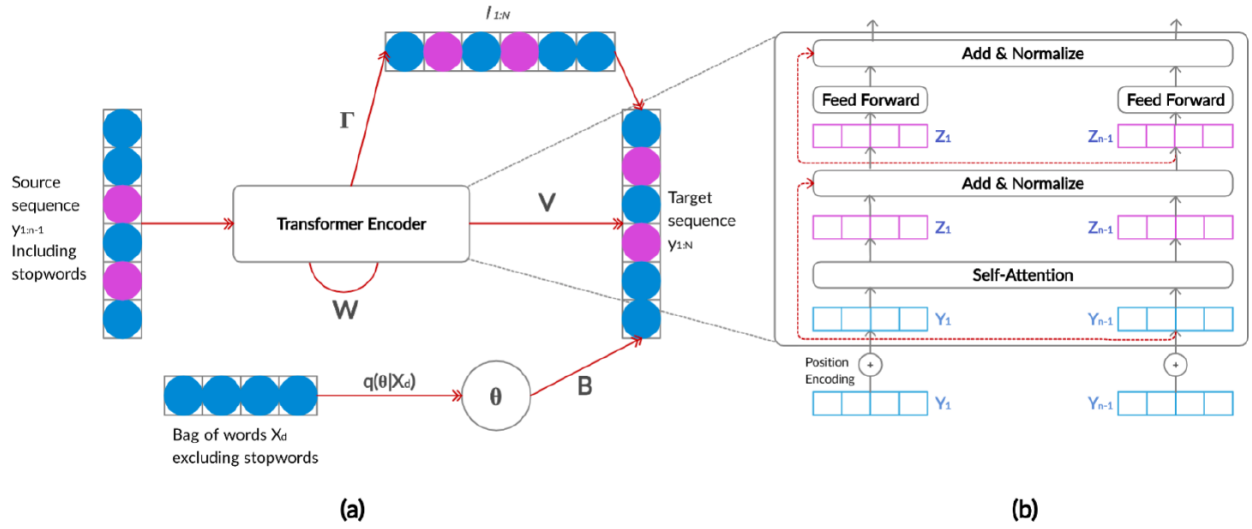

Fig. 1: (a) The Topic-Transformer joint learning framework, (b) The transformer encoder that takes as inputs word embeddings and gives as outputs transformer-based word encodings

Computation Complexity

Our model complexity is of $O(t^2 \times L \times N \times V \times K + W_d)$ The transformer part accounts for O $(t^2 \times L \times N)$ where t is the sequence length, L is the number of layers and N is the number of neurons in each layer. In addition, V is considered the dimension of the vocabulary, K is the size of the topic vector and W_d indicates the number of parameters of the inference neural network.

Experiments and Results

In this section, we evaluate our Topic-Transformer model on both language modeling and topic modeling tasks. Then, we employ this model on document clustering and interpret the overall results both quantitatively and qualitatively.

Language Model Evaluation

Datasets

We assess our method on conventional benchmark datasets for language modeling: WikiText-2 (Merity *et al*., 2017); and the Penn Tree-Bank (PTB) (Marcus *et al*., 1993); Table 1 describes the statistics. The vocabulary of the two datasets includes the special token [eos] that symbolizes the end of a sentence and [unk] symbolizes uncommon words. Since the inference network necessitates as input the bag-of-words, the vocabulary dimension of the inference neural network is reduced to 32314 after eliminating 964 stop words from WikiText-2 and 9551 after eliminating 449 pre-defined stop words for the PTB dataset.

Metrics and Settings

We used these preprocessed datasets for fair comparison using the perplexity metric which is surprising for a language model. It is calculated as the exponential of the average negative log-likelihood.

Regarding the topic modeling part, we adopted a 2-layer neural network to learn the function $q(\theta|X_d)$, within each layer 200 hidden units and the ReLU Nair and Hinton (2010) as the loss function and we experiment with different numbers of topics; 30, 50 and 100 respectively. Regarding the baseline and the transformer encoder of our model, we consider 2 settings: (i) 2-hidden layers with 2 attention heads in each layer, (ii) 4-hidden layers within each layer four attention-heads. The dimension of both input embedding and the feed-forward network model is 200.To mitigate overfitting, we used a dropout function with a rate of 0.2 in both the topic model and the transformer's hidden layers. Furthermore, adaptive Softmax is used to speed up the training process. Throughout the training, all the hyper-parameters are fine-tuned based on the performance of the validation dataset. We empirically find that these settings are logically robust and optimal over the two datasets. All the experiments were carried out using Pytorch.

Table 1: Statistics of PTB and WikiText-2

	PTB	WikiText-2
Vocabulary	10 000	33 278
#Train Tokens	929 590	2 088 628
#Validation Tokens	73 761	217 646
#Test Tokens	82 431	245 569

Baseline

We examined our proposed model with diverse quantities of topics and diverse quantities of layers and heads and compare it with 6 different methods: (1) A regular Transformer (Vaswani *et al*., 2017); with the same settings we set for our model; (2) the "Topic-RNN" which is a mutual learning model that learns concurrently an RNN-based language model and a topic model (Dieng *et al*., 2016); (3) the "TDLM" which is a joint learning model that learns concurrently a language model and a convolution-based topic model (Lau *et al*., 2017); (4) the "TGVAE" which is a joint learning framework that learns concurrently a VAE-based neural sequence model and a topic model (Wang *et al*., 2019).

Results

Table 2 shows the results of our Topic-Transformer model which achieves superior performance across the two datasets than the baselines, demonstrating the practical merit of our model. Our Topic-Transformer achieves a lower perplexity score among a relatively greater quantity of layers and heads. It is important to regard that lately numerous large transformer models have been suggested as language models, like BERT (Devlin *et al*., 2018); Transformer-XL (Dai *et al*., 2019); and T5 (Raffel *et al*., 2019). In this study, we aimed at a comparably shallow transformer model for a reasonable judgment. Therefore, we are apt to presume that the higher results metrics are primarily a result of our topic-guided language modeling approach. Furthermore, by adding more layers and heads to the transformer, we get a reduced perplexity. Therefore, we conclude that if we increase the size of the transformer encoder in our model in a comparable way with the large transformers mentioned before, this can lead to even more improved results, showing the efficiency of including global semantics such as topics. We additionally perceive steady advances while lowering the number of topics, the thing that proves the performance of our Topic-Transformer.

Topic Model Evaluation

Dataset

We conduct further experiments and analysis on a NeuIPS scientific papers dataset (Perrone *et al*., 2016); specifically to evaluate the global semantic coherence of the documents driven from the topic model side of our joint learning framework. This dataset includes around 9717 extracted text for all NeurIPS papers ranging from the first 1987 conference to the 2016 conference. In the preprocessing steps, we lowercase all characters; tokenize words and sentences using Stanford CoreNLP (Manning *et al*., 2014) and filter unique words that occur less than ten times. For the topic model, we also remove stop words and the top 0.2% most frequent words.

Metrics and Settings

We assess the quality of the learned topics by examining the coherence of insinuated topics (Mimno *et al*., 2011). In fact, we try to average the coherence of topics upon the topmost 10/30/50 topic words. For quantitative comp*arison, we use the following baseline models: The TF-IDF model (Ramos, 2003); the LDA model (Blei et al*., 2003); and a standard Transformer model with 4-layers 16-heads. For the settings, we used the same setting as in the previous section, we chose the 4-layers 16-heads option since it gives us better results in language modeling evaluation. Additionally, as we perform the document clustering task, we also evaluate our method by using the silhouette value (Aranganayagi and Thangavel, 2007), which is a measure of how similar an object is in its own cluster compared to other clusters, to demonstrate the logic of our method. The result of the clustering is shown in Fig. 2 where we compare our models with the TF-IDF model and a basic transformer for qualitative analysis.

Results

The results are depicted in Table 3. The Topic-Transformer gains encouraging results, with the best coherence across the scientific papers corpus. Additionally, the advantage of our Topic-Transformer over a standard transformer and LDA indicates that our method provides more robust topic guidance. Subsequently, to fully comprehend the topic model and confirm that the Topic-Transformer learns perceivable topic-based priors, multiple samples as drawn from each mixture element and visualize them with t-SNE (Van der Maaten and Hinton, 2008). As shown in Fig.2 we have learned a group of separable clusters, which are more distinguishable comparing with a simple TF-IDF model and a basic transformer. Each cluster maintains semantic meaning in the latent space. We additionally draw some inferred topic assignments of a sample of documents using the Topic-Transformer; 10 topics for 10 documents as for instance in Fig. 3. Same as the regular topic models, those distributions are too approximately sparse. Those qualitative analyses moreover confirm that the presented method completely gathers the sense of topic.

Table 2: Comparison of perplexity on PTB and WikiText-2 datasets

	PTB	WikiText-2
Topic-RNN	179.23	270.13
TDLM	171.01	244.65
TGVAE	160.84	250.03
Standard Transformer (2-layers 4-heads)	162.75	263.37
Standard Transformer (4-layers 16-heads)	157.53	239.29
Topic-Transformer (2-layers 4-heads, 30 Topics)	154.84	276.72
Topic-Transformer (2-layers 4-heads, 50 Topics)	156.02	232.69
Topic-Transformer (2-layers 4-heads, 100 Topics)	179.48	253.98
Topic-Transformer (4-layers 16-heads, 30 Topics)	108.73	213.38
Topic-Transformer (4-layers 16-heads, 50 Topics)	160.63	198.93
Topic-Transformer (4-layers 16-heads, 100 Topics)	151.30	227.62

Table 3: Silhouette and topic coherence comparison over NeurIPS papers dataset

Method	Silhouette	Coherence
TF-IDF	0.012	0.517
LDA	None	0.476
Transformer	0.061	0.536
Topic-Transformer	0.205	0.561

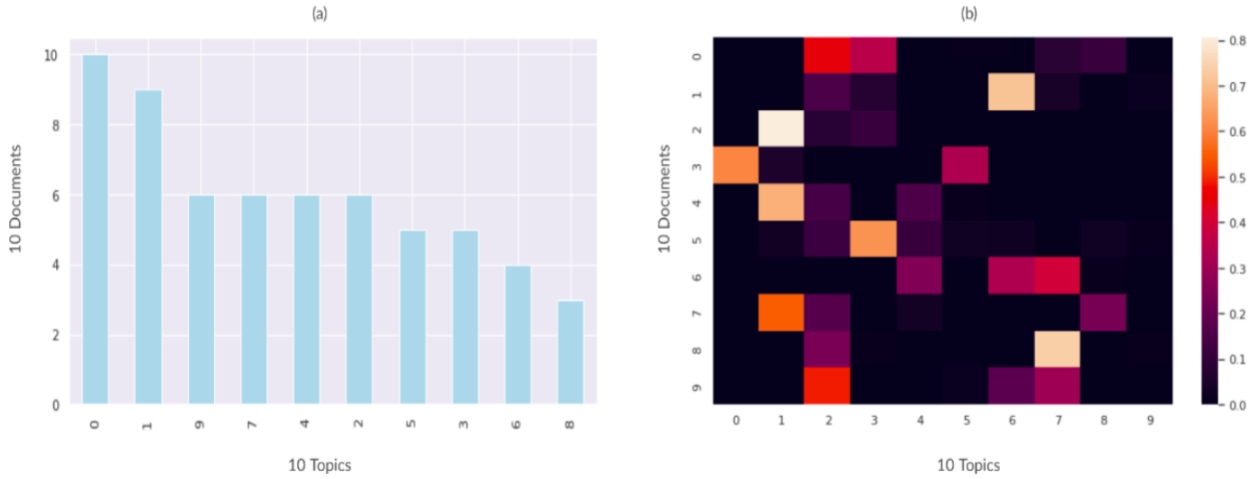

Fig. 2: Inferred distributions using Topic-Transformer on 10 different documents and 10 topics

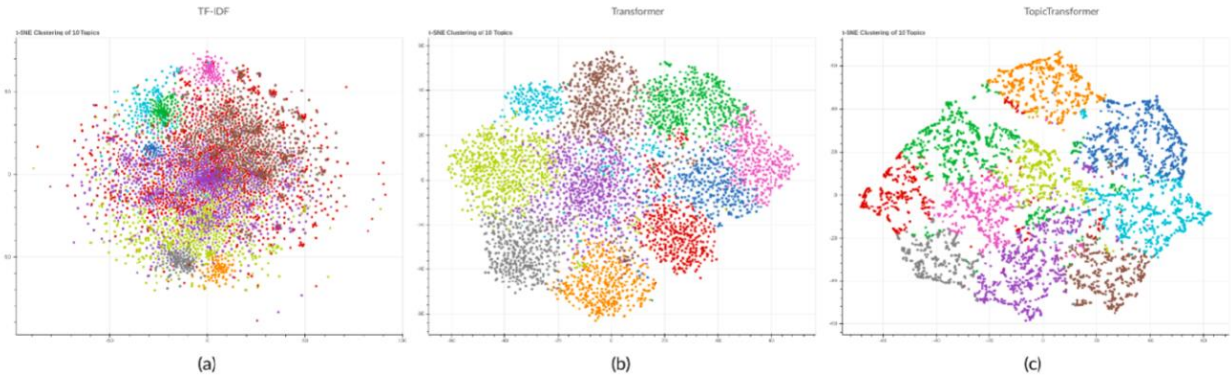

Fig. 3: The t-SNE visualization of NeurIPS papers drawn from TF-IDF, basic transformer and the learned Topic-Transformer models

Conclusion

This study introduces the Topic-Transformer model, a novel approach to jointly learn a language model and atopic model for a better document-level language understanding. The topic model captures global semantic information and the transformer-based language model captures the local structure of a sequence both semantically and syntactically. Topic-Transformer produces rival perplexity on the WikiText-2 and PTB datasets as opposed to a basic Transformer model and the existing topic-guided language models. We have also demonstrated the ability of Topic-Transformer to produce coherent topics. In future work, we may extend the Topic-Transformer parameters and fine-tune them for various natural language understanding downstream tasks.

Acknowledgement

Appeciation to the members of Laboratory Information Technology and Modeling for their support.

Author's Contributions

Oumaima Hourrane: Participate in all experiments, coordinate the data-analysis and contribute to the writing of the manuscript.

El Habib Benlahmar: Supervise, design the research plan and organized the study.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Aranganayagi, S., & Thangavel, K. (2007, December). Clustering categorical data using silhouette coefficient as a relocating measure. In International conference on computational intelligence and multimedia applications (ICCIMA 2007) (Vol. 2, pp. 13-17). IEEE. https://ieeexplore.ieee.org/abstract/document/4426662/

Beskaravainaja, E. V., & Kharybina, T. N. (2020). Analysis of the Factors That Affect the Citability of Research Articles. Scientific and Technical Information Processing, 47(2), 119-125. https://link.springer.com/article/10.3103/S0147688220020070

Blei, D. M., & Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures. Bayesian analysis, 1(1), 121-143.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichl*et al*location. The Journal of machine Learning research, 3, 993-1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf?TB_iframe=true&width=370.8&height=658.8

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555. https://arxiv.org/abs/1412.3555

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., & Salakhutdinov, R. (2019). Transformer-xl: Attentive language models beyond a fixed-length context. arXiv preprint arXiv:1901.02860. https://arxiv.org/abs/1901.02860

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. https://arxiv.org/abs/1810.04805

Dieng, A. B., Wang, C., Gao, J., & Paisley, J. (2016). Topicrnn: A recurrent neural network with long-range semantic dependency. arXiv preprint arXiv:1611.01702. https://arxiv.org/abs/1611.01702

Gehring, J., Auli, M., Grangier, D., Yarats, D., & Dauphin, Y. N. (2017, July). Convolutional sequence to sequence learning. In International Conference on Machine Learning (pp. 1243-1252). PMLR. http://proceedings.mlr.press/v70/gehring17a.html

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780. https://ieeexplore.ieee.org/abstract/document/6795963/

Hourrane, O., & Benlahmar, E. H. (2017, March). Survey of plagiarism detection approaches and big data techniques related to plagiarism candidate retrieval. In Proceedings of the 2nd international Conference on Big Data, Cloud and Applications (pp. 1-6). https://dl.acm.org/doi/abs/10.1145/3090354.3090369

Hourrane, O., & Benlahmer, E. H. (2019). Rich style embedding for intrinsic plagiarism detection. International Journal of Advanced Computer Science and Applications, 10(11).

Hourrane, O., & Idrissi, N. (2019a), October). An Empirical Study of Deep Neural Networks Models for Sentiment Classification on Movie Reviews. In 2019 1st International Conference on Smart Systems and Data Science (ICSSD) (pp. 1-6). IEEE. https://ieeexplore.ieee.org/abstract/document/9003171

Hourrane, O., & Idrissi, N. (2019b), October). Sentiment Classification on Movie Reviews and Twitter: An Experimental Study of Supervised Learning Models. In 2019 1st International Conference on Smart Systems and Data Science (ICSSD) (pp. 1-6). IEEE. https://ieeexplore.ieee.org/abstract/document/9003118/

Hourrane, O., Mifrah, S., Benlahmar, E. H., Bouhriz, N., & Rachdi, M. (2018, April). Using Deep Learning Word Embeddings for Citations Similarity in Academic Papers. In International Conference on Big Data, Cloud and Applications (pp. 185-196). Springer, Cham.https://link.springer.com/chapter/10.1007/978-3-319-96292-4_15

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.https://arxiv.org/abs/1312.6114

Koehn, P. (2009). Statistical machine translation. Cambridge University Press.

Lau, J. H., Baldwin, T., & Cohn, T. (2017). Topically driven neural language model. arXiv preprint arXiv:1704.08012. https://arxiv.org/abs/1704.08012

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of machine learning research, 9(11). https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf?fbclid=IwA

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014, June). The Stanford CoreNLP natural language processing toolkit. In Proceedings of 52nd annual meeting of the association for computational linguistics: System demonstrations (pp. 55-60). https://www.aclweb.org/anthology/P14-5010.pdf

Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. https://repository.upenn.edu/cis_reports/237/

Merity, S., Keskar, N. S., & Socher, R. (2017). Regularizing and optimizing LSTM language models. arXiv preprint arXiv:1708.02182. https://arxiv.org/abs/1708.02182

Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., & Khudanpur, S. (2010, September). Recurrent neural network based language model. In Interspeech (Vol. 2, No. 3, pp. 1045-1048). http://www.fit.vutbr.cz/research/groups/speech/servite/2010/rnnlm_mikolov.pdf

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011, July). Optimizing semantic coherence in topic models. In Proceedings of the 2011 conference on empirical methods in natural language processing (pp. 262-272). https://www.aclweb.org/anthology/D11-1024.pdf

Nair, V., & Hinton, G. E. (2010, January). Rectified linear units improve restricted boltzmann machines. In Icml. https://openreview.net/forum?id=rkb15iZdZB

Perrone, V., Jenkins, P. A., Spano, D., &Teh, Y. W. (2016). Poisson random fields for dynamic feature models. arXiv preprint arXiv:1611.07460. https://arxiv.org/abs/1611.07460

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683. https://www.jmlr.org/papers/volume21/20-074/20-074.pdf

Ramos, J. (2003, December). Using tf-idf to determine word relevance in document queries. In Proceedings of the first instructional conference on machine learning (Vol. 242, No. 1, pp. 29-48). https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.121.1424&rep=rep1&type=pdf

Tian, F., Gao, B., He, D., & Liu, T. Y. (2016). Sentence level recurrent topic model: Letting topics speak for themselves. arXiv preprint arXiv:1604.02038. https://arxiv.org/abs/1604.02038

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... &Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008). http://papers.nips.cc/paper/7181-attention-is-all-you-%0Aneed.pdf

Wang, W., Gan, Z., Wang, W., Shen, D., Huang, J., Ping, W., ... & Carin, L. (2018, March). Topic compositional neural language model. In International Conference on Artificial Intelligence and Statistics (pp. 356-365). PMLR. http://proceedings.mlr.press/v84/wang18a.html

Wang, W., Gan, Z., Xu, H., Zhang, R., Wang, G., Shen, D., ... & Carin, L. (2019). Topic-guided variational autoencoders for text generation. arXiv preprint arXiv:1903.07137. https://arxiv.org/abs/1903.07137