TOPICRNN: A RECURRENT NEURAL NETWORK WITH LONG-RANGE SEMANTIC DEPENDENCY

Adji B. Dieng *
Columbia University

Chong Wang
Microsoft Research

Jianfeng Gao
Microsoft Research

John Paisley
Columbia University

ABSTRACT

In this paper, we propose TopicRNN, a recurrent neural network (RNN)-based language model designed to directly capture the global semantic meaning relating words in a document via latent topics. Because of their sequential nature, RNNs are good at capturing the local structure of a word sequence – both semantic and syntactic – but might face difficulty remembering long-range dependencies. Intuitively, these long-range dependencies are of semantic nature. In contrast, latent topic models are able to capture the global semantic structure of a document but do not account for word ordering. The proposed TopicRNN model integrates the merits of RNNs and latent topic models: it captures local (syntactic) dependencies using an RNN and global (semantic) dependencies using latent topics. Unlike previous work on contextual RNN language modeling, our model is learned end-to-end. Empirical results on word prediction show that TopicRNN outperforms existing contextual RNN baselines. In addition, TopicRNN can be used as an unsupervised feature extractor for documents. We do this for sentiment analysis on the IMDB movie review dataset and report an error rate of 6.28%. This is comparable to the state-of-the-art 5.91% resulting from a semi-supervised approach. Finally, TopicRNN also yields sensible topics, making it a useful alternative to document models such as latent Dirichlet allocation.

1 THE TOPICRNN MODEL

TopicRNN is a generative model. For a document containing the words $y_{1:T}$,

1. Draw a topic vector[1] $\theta \sim N(0, I)$.
2. Given word $y_{1:t-1}$, for the tth word y_t in the document,
   (a) Compute hidden state $h_t = f_W(x_t, h_{t-1})$, where we let $x_t \triangleq y_{t-1}$.
   (b) Draw stop word indicator $l_t \sim \text{Bernoulli}(\sigma(\Gamma^\top h_t))$, with σ the sigmoid function.
   (c) Draw word $y_t \sim p(y_t|h_t, \theta, l_t, B)$, where

$$p(y_t = i|h_t, \theta, l_t, B) \propto \exp\left(v_i^\top h_t + (1 - l_t) b_i^\top \theta\right).$$

The stop word indicator l_t controls how the topic vector θ affects the output. If $l_t = 1$ (indicating y_t is a stop word), the topic vector θ has no contribution to the output. Otherwise, we add a bias to favor those words that are more likely to appear when mixing with θ, as measured by the dot product between θ and the latent word vector b_i for the ith vocabulary word. As we can see, the long-range semantic information captured by θ directly affects the output through an additive procedure. Unlike Mikolov and Zweig (2012), the contextual information is not passed to the hidden layer of the RNN.

The model is learned via amortized variational inference.

*Work was done while at Microsoft Research.

[1]Instead of using the Dirichlet distribution, we choose the Gaussian distribution. This allows for more flexibility in the sequence prediction problem and also has advantages during inference.

2 EMPIRICAL EVIDENCE

We assess the performance of our proposed TopicRNN model on word prediction and sentiment analysis.

Table 1: TopicRNN and its counterparts exhibit lower perplexity scores across different network sizes than reported in Mikolov and Zweig (2012). **??** shows per-word perplexity scores for 10 neurons. Table 1a and Table 1b correspond to per-word perplexity scores for 100 and 300 neurons respectively. These results prove TopicRNN has more generalization capabilities: for example we only need a TopicGRU with 100 neurons to achieve a better perplexity than stacking 2 LSTMs with 200 neurons each: 112.4 vs 115.9)

(a)

100 Neurons	Valid	Test
RNN (no features)	150.1	142.1
RNN (LDA features)	132.3	126.4
TopicRNN	128.5	122.3
TopicLSTM	126.0	118.1
TopicGRU	118.3	**112.4**

(b)

300 Neurons	Valid	Test
RNN (no features)	–	124.7
RNN (LDA features)	–	113.7
TopicRNN	118.3	112.2
TopicLSTM	104.1	99.5
TopicGRU	99.6	**97.3**

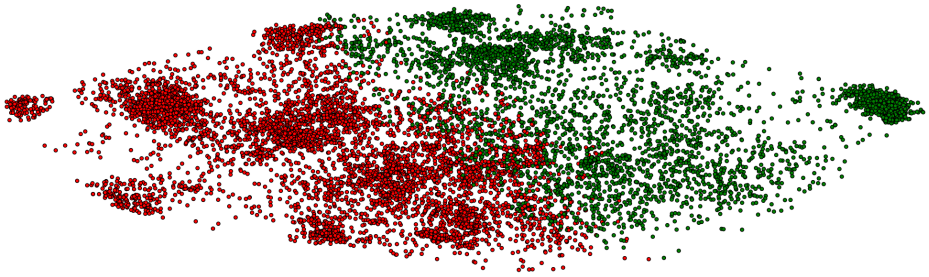

Figure 1: Clusters of a sample of 10000 movie reviews from the IMDB 100K dataset using TopicRNN as feature extractor. We used K-Means to cluster the feature vectors. We then used PCA to reduce the dimension to two for visualization purposes. red is a negative review and green is a positive review. This corresponds to a classification error rate of **6.28%** which is comparable to the SOTA error rate (**5.91%**) that uses a semi-supervised method.

REFERENCES

Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin. A neural probabilistic language model. *journal of machine learning research*, 3(Feb):1137–1155, 2003.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. In *ICML*, volume 14, pages 1188–1196, 2014.

Y. Miao, L. Yu, and P. Blunsom. Neural variational inference for text processing. *arXiv preprint arXiv:1511.06038*, 2015.

T. Mikolov and G. Zweig. Context dependent recurrent neural network language model. In *SLT*, pages 234–239, 2012.