CORRELATED TOPIC MODEL WITH TRANSFORMER EMBEDDINGS

by

Chun Wa Leung

A Master Thesis

Submitted to
the Graduate School of the University of Tokyo
on December 7, 2021
in Partial Fulfillment of the Requirements
for the Degree of Master of Information Science and Technology
in Computer Science

Thesis Supervisor: Akihiko Takano
Professor of Computer Science

ABSTRACT

Topic modeling is one of the most common information retrieval task in natural language processing. In particular, Correlated topic model(CTM) is a topic model which captures the correlation between topics associated. However, such a classic statistical approach was not able to capture positional information from sequential input. At that point, traditional topic models may perform poorly in generating words from large number of topics. In this research, we introduce Correlated Topic Model with Transformer embeddings, a generative model where combine the advantage of using positional information of words and topic correlation. Specifically, transformer embedding maps topic words into latent space and further assign to its assigned topic. We attempted to add a covariance prior to the topic model, LKJ correlation prior to logistic-normal distribution, which aims to fit the correlation information from the data. In addition, we extended our model to handle time-series data integrated with Gaussian Process latent variable model(GPLVM), which also capturing temporal information from words occurence of documents over time. The model was optimized using Stochastic Variational Inference (SVI), allows handling massive data sets with mini-batching. As compared to empirical results from experiments, our approach performs a better fit of the data than existing generative topic model and exhibit a better capability in obtaining high quality topics.