CORRELATED TOPIC MODEL WITH TRANSFORMER EMBEDDINGS

by

Leung Chun Wa

An Interim Report (Abstract)

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

Thesis Supervisor: Takano Akihiko
Professor of Computer Science

ABSTRACT

Topic Modeling is one of the most common information retrieval task in natural language processing. Such as latent Dirichlet allocation (LDA). However, as a classic statistical approach, which was not able to capture positional information from sequential input. At that point, traditional topic models 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, the transformer embedding maps topic words into latent space and further assign to its assigned topic. Moreover, 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. The model was optimized using stochastic variational inference(SVI). As result, our approach performs a better fit of the data than existing generative topic model and exhibit a better capability in obtaining high quality topics.