Topic Modelling

Machine Learning for Natural Language Processing I

Attention! Required answers are marked in blue !

December 10, 2022

The goal of the report is to figure out the trends in the computer science field up til now. The task aims to cluter a bunch of report topics in an ubsupervised fashion, get the words with high probabilities and provide an insight of the trend at the specific time period. By using Latent Dirichlet Allocation (LDA) and Combined Topic Modelling (Combined TM), the task can be solved. The trend coincides with the reality, meaning cyber security between 1990 and 2010, and neural network, computer vision after 2010.

1 Introduction

Most of the time, getting labelled data is extremely challenging. Thus, using unsupervised learning methods are very useful in such cases. To get the trend of the computer science papers, the titles of the paper between three periods, *Before 1990*, *From 1990 tp 2010* and *After 2010* are collected and clustered into five topics. Fifteen words with highest probabilites of each topic will be printed out for trend guessing. The evaluation of the models include coherence of the results and whether the trend conincides with the reality.

2 Method

The data is first downloaded from dblp. After preprocessing, the data is sent to Latent Dirichlet Allocation (LDA) and Combined Topic Modelling (Combined TM) for training. The data is separated into three periods, before 1990, from 1990 to 2010, and after 2010. Data for each period is trained by LDA(Panchal, n.d.) and Combined TM respectively.

2.1 Data

The dblp is a database with metadata on computer science related publications. For this task, the published year and the titles of the publications will be extracted for the training.

2.2 Preprocessing

Preprocessing for LDA is simple. The numbers are removed, and only English characters are allowed. In addition, all characters are lower cased.

For Combined TM, the stop words in the titles are removed. The preprocessed data are then used for training the Bag-of-Words (BOW)(Brownlee, n.d.) model. The unprocessed data is sent to all-mpnet-base-v2 model for generating contextualised embeddings. The contextualised representation, along with the output of BOW will be sent to Neural Topic Models, ProdLDA, for training.

2.3 Training & Evaluation

The data in three periods are all trained with LDA and Combined TM for 5 clusters. After training, the 15 most related words (words with higher probability in the probability distribution) to the topics will be printed out for topic guessing. For LDA, the probability of the top 15 words word will be printed out as bar charts. As for Combined TM, an additional visualisation offered by the package will be shown. In addition, topic names of the five topics will be guessed and the topic names from the two models will be discussed.

3 Results & Discussion

The results of LDA and Combined TM are listed in Table 2 and Table 3 respectively. The incoherent topic is marked as *None* in the result table.

The results of LDA is roughly coherent. The trends can be guessed by the top words. The trend for the period *after 2010* is quite good as most of the popular

Table 1: *LDA Result*

Period	Data Size
Before 1990	40'000
From 1990 to 2010	330'317
After 2010	824'411

subjects nowadays are detected. However, the trend before that is somehow not that clear. Although most topics are assigned with topic names, the guessed topic names are not specific and clear. This is because the result is not extremely coherent. This can be expected because LDA is a probabilistic model and can be poorly performed on short data like title names instead of contexts(Ipshita, n.d.)(Yan, n.d.).

The results of Combined TM is more coherent than the ones of LDA like the paper(Bianchi, Terragni, and Hovy, 2021) proposed. The trend is clear in both periods, *From 1990 to 2010* and *After 2010*. By clear it doesn't mean that all five topics are more or less similar, but it means that the trend does coincide with the actual trend in the world at that time. The result of Combined TM is quite a success.

By comparing the results of two models, it is clear that there's no doubt computer vision and neural network are popular in the period *After 2010*. However, the results for the other two periods are then debatable. Both models don't show clear trends for the period *Before 1990*. This might due to the fact that the data size for the three periods are very different (Table 1). The large data size of period *After 2010* might be the main reason of why it performs better.

However, among two models, Combined TM does generate more coherent results. This might because, as the paper suggested, neural topic models have improved the coherence and contextual embeddings have enhenced the performances of neural models. By combining these two, the model generates more meaningful results. Although the downside of this is the limitation of SBERT's[1] sentence-length limit, the length of the titles don't exceed the limit. Thus, this model is of great fit to our task.

Bibliography

Bianchi, Federico, Silvia Terragni, and Dirk Hovy (Aug. 2021). "Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language*

Table 2: *LDA Result*

Topic	Name
~1990	
1	Algorithm Optimisation
2	Graph Algorithm
3	Control System
4	Linear Computation
5	Pattern Recognition
1990~2010	
1	None
2	Dynamic Software Architectures
3	Control System
4	Network Efficiency
5	Robust Estimation
2010~	
1	Deep Learning
2	Neural Network
3	Control System
4	Simulation Modeling
5	Control System

Table 3: *Combined TM Result*

Topic	Name
~1990	
1	None (something about libraries)
2	None (something about libraries)
3	Dynamic Planning
4	Multi-Processing
5	Text Understanding
1990~2010	
1	Cryptography
2	Stochastic Optimisation
3	Wireless Sensor Network
4	Machines Translation
5	Cyber Security
2010~	
1	Neural Network
2	Neural Network & Bayesian
3	Cyber Security
4	Computer Vision
5	Cryptography

[1] SBERT is used for genrating contextual sentence embeddings in this model.

Processing (Volume 2: Short Papers). Online: Association for Computational Linguistics, pp. 759–766. DOI: 10.18653/v1/2021.acl-short.96. URL: https://aclanthology.org/2021.acl-short.96.

Brownlee, Jason (n.d.). *A Gentle Introduction to the Bag-of-Words Model*. URL: https://machinelearningmastery.com/gentle-introduction-bag-words-model/. (accessed: 10.12.2022).

Ipshita (n.d.). *Topic Modelling using LDA*. URL: https://medium.com/analytics-vidhya/topic-modelling-using-lda-aa11ec9bec13. (accessed: 10.12.2022).

Panchal, Sandeep (n.d.). *Topic Modeling with Latent Dirichlet Allocation (LDA)*. URL: https://medium.com/analytics-vidhya/topic-modeling-with-latent-dirichlet-allocation-lda-196c287e221. (accessed: 10.12.2022).

Yan, Xiaohui (n.d.). *What's the disadvantage of LDA for short texts?* URL: https://stackoverflow.com/questions/29786985/whats-the-disadvantage-of-lda-for-short-texts. (accessed: 10.12.2022).