

Neural Keyphrase Generation

Sibila M¹, Irshad M²

P G Student, CSE, Government Engineering College Palakkad, Kerala, India¹

Professor, CSE, Government Engineering College Palakkad, Kerala, India²

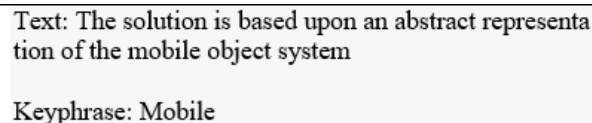
Abstract: In social media platforms like Twitter, You Tube etc. generates huge amount of user contents daily. In order to detect the user behaviour and interests keyphrases plays a crucial role. Keyphrases are short text pieces that can quickly express the key idea of source post. In case of extracting the main points from articles or documents the keyphrase generation is also important. Here proposes a methodology by generating keyphrases from the users post with the help of neural network representations and also generates the missing keyphrases which is the drawback of the previous systems. That is key phrase generation aims at predicting both present and absent keyphrases for user's posts. The proposed method is a sequence to-sequence (seq2seq) based neural keyphrase generation frame work. Also, this model is topic- aware for avoiding sparsity in social media languages. Here also discussing about key phrase generation using BERT which is a latest technology in today's world.

Keywords: Keyphrase, Seq2seq, Bert, Topic-Aware.

I. INTRODUCTION

Keyphrases are short pieces of text carrying the main or important topics reflecting in the source text. Keyphrase generation plays an important role in various text generation tasks. Humans find keyphrases by reading the text, then understand and get contextual information from the text, and finally summarize and write down the most meaningful phrases. In machine perspective the keyphrases are generated by using neural network representations. The recurrent neural networks are used for keyphrase generation which includes an encoder- decoder architecture incorporating with a copy mechanism.

Keyphrases are generated from documents for summarizing the documents or articles. In social media platforms the keyphrase generation aims to find the main topic conveyed in the source post. In this paper discussing about keyphrase generation for social media platforms using topic-aware neural networks and also keyphrase generation using BERT. BERT stands for Bidirectional Encoding Representations from Trans- formers which is a deep bidirectional neural network representation for keyphrase generation. The topic-aware neural network representation includes a neural topic model and a Sequence-to-Sequence based generation model. Neural topic model is for topic words generation and the Sequence-to- Sequence method for keyphrase generation. BERT for key phrase generation it includes two stages: 1) Loading pre- trained model 2) Fine tuning for keyphrase generation.

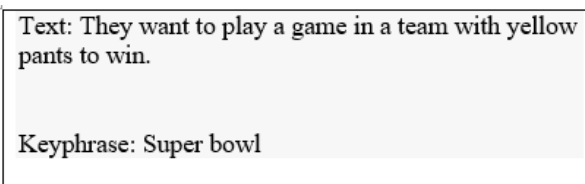


Text: The solution is based upon an abstract representation of the mobile object system

Keyphrase: Mobile

Fig 1: Example of keyphrase generation (Present keyphrase)

The keyphrases are short texts highly summarize the significant content of a source text. The keyphrases generated are either present in the source text are called present keyphrases or those absent in the source text are called absent keyphrases. Figure.1 shows an example of present keyphrase generation and Figure.2 shows an example of absent keyphrase generation. In the first example the keyphrase is mobile and it is present in the source text. The second example shows the generation of absent keyphrase. Using the topic-aware neural network the absent keyphrase generation is possible and the working of topic-aware neural networks is discuss in the section 3.



Text: They want to play a game in a team with yellow pants to win.

Keyphrase: Super bowl

Fig 2: Example of keyphrase generation (Absent keyphrase)

A. Applications

1. *Applications of keyphrase generation in case of documents:*
 - Knowledge mining (concept)
 - Information retrieval
 - Summarization
2. *Applications in case of keyphrase generation in case of social media platforms:*
 - Instant detection of trending events.
 - Finding user's interests
 - Summarizing public opinions

II. RELATED WORKS

This section deals with various neural methods for keyphrase generation. These methods mainly focus on the keyphrase generation in documents or articles.

A. Methodology-1 Chen et.al 2019

A neural keyphrase generation which includes an extraction and retrieval mechanism [3]. It is a novel approach that leverages a multi-task learning frame work. The proposed multi-task learning frame work jointly learns the extractive and generative model. It includes an extractive, generative and retrieval methods for keyphrase generation. A novel merging module combines these three methods for further improvement.

B. Methodology-2 Gao et.al 2019

This methodology is based on an encoder-decoder architecture developed for keyphrase generation in documents [4]. In the proposed method the title is used to guide the keyphrase generation. The set of keyphrases from documents are generated by giving the title as an extra-query like input. That is here replacing the simple concatenation of input document and title by the title is given as extra-query like input to guide the context encoding.

C. Methodology-3 Zhang et.al 2018

It is a Seq2Seq based generation model for keyphrase generation which includes a coverage mechanism and reviewing mechanism [2]. The proposed Seq2Seq architecture models correlation among multiple keyphrases in an end-to-end fashion by incorporating these two mechanisms. The coverage vector is to check whether the word in the input document has been summarized by previous phrases. The reviewing mechanism includes to eliminate the duplicate keyphrases.

D. Methodology-4 Meng et.al 2017

This method proposes a deep keyphrase prediction using an Encoder-Decoder architecture [10]. It is a RNN based generative model for keyphrase generation. In this deep learning method the keyphrase generation is performed by capturing the deep semantic meaning of the content.

III. NEURAL KEYPHRASE GENERATION FOR SOCIAL MEDIA USING NEURAL TOPIC MODEL

The Social media platforms produces large amount of massive posts daily. To find the public opinions and trending events from these posts keyphrase generation plays an important role. Here discussing a deep learning based method for keyphrase generation by jointly training two neural networks [1]. The two neural networks are:

1. Neural topic model
2. Seq2Seq based generation model

The neural topic model generates the topic words and the Seq2Seq based keyphrase generation model generates the final keyphrase. Topic modelling using LDA (Latent dirchlet allocation) had certain limitation in topic word generation hence replacing it by neural topic model. It is a keyphrase generation in social media posts using a topic-aware neural network representation and finding the important points from user's post. Using this special network architecture, the absent keyphrases are also produced which is the drawback of the previous systems.

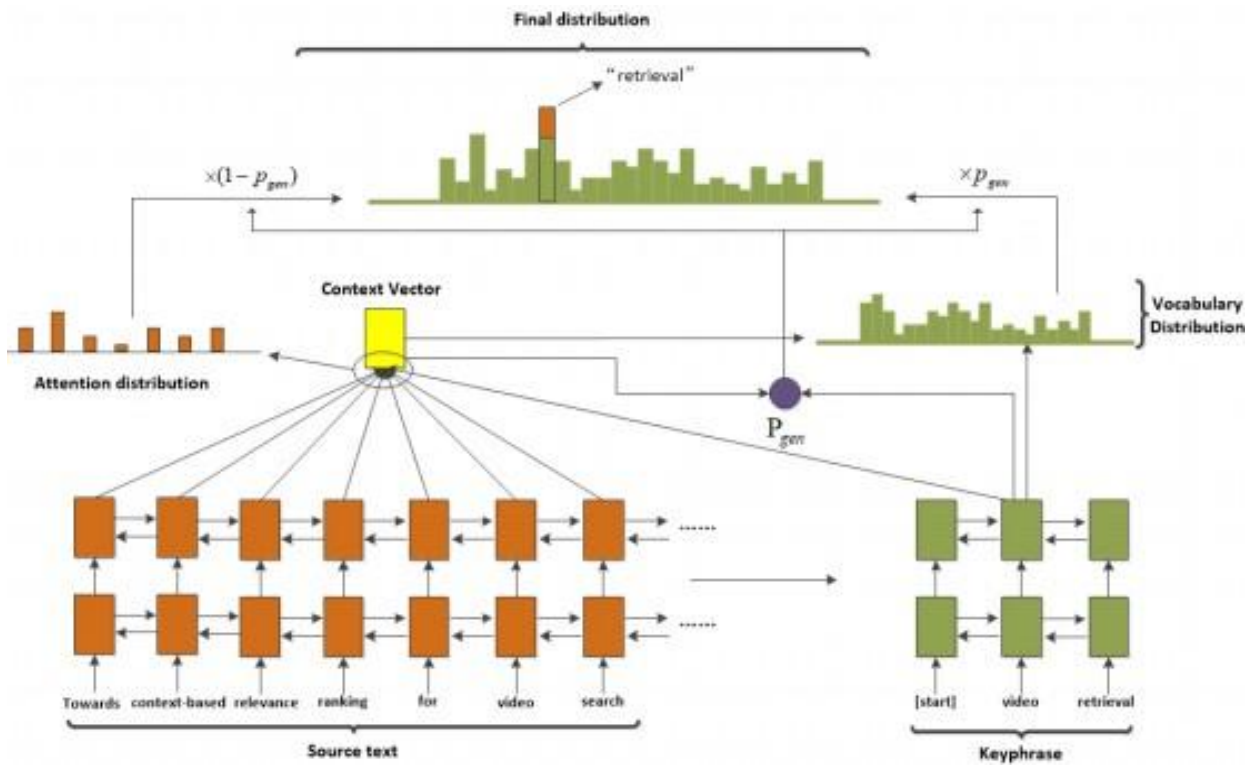


Fig 3: Seq2Seq model for keyphrase generation

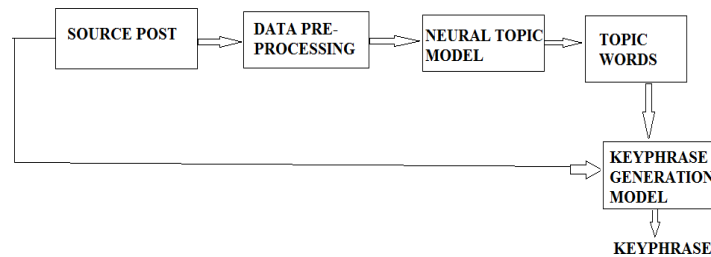


Fig 4: System Overview

A. *Neural Topic Model*

It is based on an encoder-decoder architecture. The encoder is Bow encoder and the decoder is Bow decoder to resemble the data reconstruction process. The topic words are generated from the user’s post using this model.

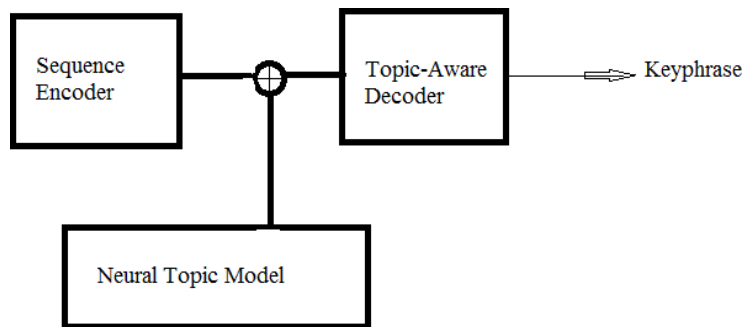


Fig 5: Overview

B. *Seq2Seq based Keyphrase generation model*

It is a Seq2Seq based generation model based on recurrent neural network. It includes a bidirectional GRU encoder and bidirectional GRU decoder for keyphrase generation. The sequence encoder finds the indicative features from the source post and the topic aware decoder generates keyphrases based on the encoded features and the latent topic generated from neural topic model.

IV. BERT FOR KEYPHRASE GENERATION

BERT stands for Bidirectional Encoder Representations from Transformers. It achieves good results on 11 NLP tasks. BERT is a contextual model. There are two types of embedding contextual embedding and context-free embedding. Word2Vec and Glove embeddings are context-free embedding, while BERT uses bidirectional contextual embedding.

BERT tasks mainly includes two stages

- 1) Loading pre-trained model
- 2) Fine-tuning for a specific task.

There are many BERT pretrained model trained on large corpus.

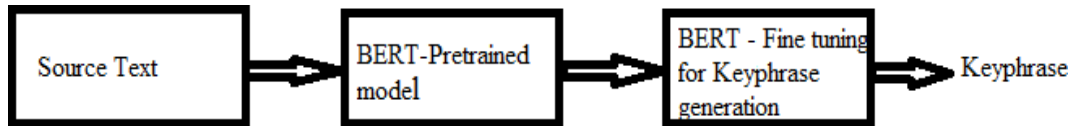


Fig 6: BERT for keyphrase generation

The basic building block of BERT is a transformer replacing the RNN variations like Bidirectional GRU and Bidirectional LSTM [5]. Transformers are faster than RNN and simple and parallelizable.

A. Pre-training

The pre-training of BERT is a method to train a general-purpose language model on a large corpus like Book corpus or Wikipedia. For pre-training it requires a large amount of data and also it is a time-consuming task. There are many BERT pre-trained models are available and fine-tuning these models for specific NLP tasks.

B. Fine-tuning

For fine-tuning it needs a training dataset which contains the source text and corresponding target keyphrases. Here uses Bert-uncased-base model as pre-trained model and fine-tuning it with the training dataset.

V. IMPLEMENTATION

A. Dataset

As mentioned before, this proposed system aims at developing a keyphrase generation system from social media platforms using topic-aware neural network and BERT used for general domain. The social media dataset used are twitter and stack exchange. Each dataset split into train set, test set and validation set. BERT uses pre-trained BERT base uncased model and fine tuning with scientific articles done.

B. Data-preprocessing

Texts are preprocessed to remove any content that does not have any useful information. The workflow of preprocessing module includes:

1. Load the dataset
2. Posts with irrelevant main phrases (e.g. single character post) have been filtered out; non-alphabetic and retweet-only texts (e.g., RT) have also been deleted.
3. Second, Mentions (@username), links, and digits have been replaced with generic URL, MENT and DIGIT.
4. A vocabulary is created called BoW dictionary.

C. Implementation and Tools used

Bag of words approach is used for the embedding of words. These embedded words are then passed to the encoder module. Table shows the implementation details of the deep learning model. The proposed deep learning model is implemented using Pytorch.

Table: Encoder-Decoder model specifications

Parameter	Value
Layers	Input, Embedding, GRU
Embedding Size	100
Encoder size	300
Decoder size	150
Activation	Tanh,softmax
Epochs	100

Table: Transformer specifications

Parameter	Value
Transformer block	12
Hidden size	768
Self-attention heads	12
Optimizer	Adam

VI. RESULTS

For testing how well the system generates the possible keyphrases and how much similar is this keyphrase with the context mentioned in the post. Topic-aware neural networks produces present and absent keyphrases. BERT produces keyphrases present in the input text. Training time was greater for topic-aware neural as compared to BERT. BERT have high validation accuracy and less loss.

Compared with sequence-to-sequence model, incorporation of latent topics gives better results.

- Using topic-aware neural networks absent and present keyphrases are generated.
- The precision, recall and F1-score are calculated based on precision, recall@n and F1-score@n concepts.
- It is calculated based on micro average and macro average methods.
- BERT keyphrase generation extend to general domain and having high validation accuracy of 98.48% at learning rate 3e-05 and number of epochs 4.

VII. CONCLUSION AND FUTURE SCOPE

The major area of application of keyphrase generation is summarizing the source text. Seq2Seq model could prove its highest performance in the area of keyphrase generation. Apart from that, topic-aware neural networks which includes a neural topic model with Seq2Seq model for better performance. The proposed methods also includes BERT keyphrase generation which replaces RNN variants with transformers. An encoder-decoder module of topic-aware neural network is used for generating keyphrases of a user post by taking embedded words from the original text as inputs. The existing neural methods only deals with present keyphrase generation, while the topic-aware neural methods also deals with absent keyphrase generation. The proposed method of topic-aware neural keyphrase generation performed on social media datasets. BERT keyphrase generation is extended to general domain using a pretrained Bert-base-model.

1. The proposed methods can produce keyphrases using neural method and BERT.
2. The topic-aware neural methods also deals with absent keyphrase generation which is the draw-back of previous systems.
3. Proposed neural methods deals with social media domain and BERT used for general domain

REFERENCES

- [1]. Yue Wang Jing Li HouPong Chan Irwin King Michael R.Lyu Shuming Shi.2019. "Topic-Aware Neural Keyphrase Generation for Social Media Language". In ACL
- [2]. Jun Chen, Xiaoming Zhang, Yu Wu, Zhao Yan, and Zhoujun Li. 2018. "Keyphrase generation with correlation constraints." In Proceedings of Empirical Methods in Natural Language Processing
- [3]. Wang Chen, Hou Pong Chan, Piji Li, Lidong Bing, and Irwin King. 2019a."An integrated approach for keyphrase generation via exploring the power of retrieval and extraction." In Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies
- [4]. Wang Chen, Yifan Gao, Jiani Zhang, Irwin King, and Michael R. Lyu. 2019b." Title-guided encoding for keyphrase generation." In Proceedings of AAAI Conference on Artificial Intelligence.
- [5]. Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova.2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In ACL.
- [6]. Zhiyuan Liu, Peng Li, Yabin Zheng, and Maosong Sun. 2009. Clustering to find exemplar terms for keyphrase extraction. In Proceedings of Empirical Methods in Natural Language Processing.
- [7]. Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. ArXiv preprint arXiv:1703.01488.
- [8]. Xiaojun Wan and Jianguo Xiao. 2008. Single document keyphrase extraction using neighborhood knowledge. In Proceedings of AAAI Conference on Artificial Intelligence.