Topic Modeling using Jointly Fine-tuned BERT for Phrases and Sentences

ZHOU ZIKAI

Using topic modeling to analyze large collections of documents has been a functional approach for many years. However, traditional topic models such as Latent Dirichlet Allocation rely on the assumption of "Bag-of-Words", which ignore the connection and inner semantics between words in terms of phrases. Although phrases act as important grammatical units in human language, due to the restriction on vocabulary size and model complexity, less research has been conducted on phrase-level topic models. Some research takes phrases into account by simply discovering and adding phrases into the original vocabulary set, and learning those phrases just like normal words. This approach however would make an enormous vocabulary set, then lead to problems such as huge models and unstable learning.

An alternative way of topic modeling is to use distributed representations of words and documents. Top2vec introduced a new way of topic modeling by using jointly embedded documents and word vectors, which could be produced by a pre-trained BERT model. After creating semantic embeddings for both documents and words, Top2vec applies dimensionality reduction and clustering to document embeddings and generates topic vectors in the same semantic space. To conduct a phrase-level Top2vec model, phrases should also be embedded into the same semantic space as well as words and documents. However, there is no existing pre-trained BERT-based model designed to produce embedding for both sentences and phrases.

In this research, we propose a phrase-level topic model based on pre-trained distributed representations of words, documents, and phrases. We showed that previous BERT-based models focus on either sentence or phrase embeddings, which is not optimal for topic models based on the semantics of these different grammatical units. Due to the fact that there is no existing BERT-based model designed to produce embedding for both sentences and phrases, we propose a jointly fine-tuned BERT model for sentences and phrases embeddings. The experiment shows that our proposed joint model could produce high-quality embeddings for both sentences and phrases.

Furthermore, we construct a phrase-level topic model, which is based on the structure of Top2vec, using embedding produced by our joint model. The topic evaluation results show that the joint model outperforms other BERT-based models, and improves the topic quality in terms of phrase-level topic modeling. In addition, we discovered that proper nouns and phrases are successfully embedded into the correct topics, meaning that the joint model captures the semantics of those words and utilizes them into the topic model, which traditional statistical topic models failed to achieve.

Academic Advisors: Kei WAKABAYASHI