| | | | |
|---|---|---|---|
| 氏名 | DANG, Tran Binh | | |
| 学位の種類 | 博士（情報科学） | | |
| 学位記番号 | 博情第 488 号 | | |
| 学位授与年月日 | 令和 4 年 12 月 23 日 | | |
| 論文題目 | INTERACTION OF TOPIC AND CONTEXT INFORMATION FOR TEXT REPRESENTATION | | |
| 論文審査委員 | Nguyen Le Minh | JAIST | Professor |
| | Satoshi Tojo | JAIST | Professor |
| | Kiyoaki Shirai | JAIST | Assoc. Prof. |
| | Ken Satoh. | NII | Professor |
| | Shinobu Hasegawa. | JAIST | Professor |

**論文の内容の要旨**

Modern living is becoming more and more convenient with machines thanks to the rapid advancement of science and technology. Mainly, computers and the internet are the key elements that enable people to communicate with one another by storing, exchanging, and looking for knowledge in any field. Recently, machine learning and deep learning have developed incredibly quickly, especially in the field of NLP. With their capacity to calculate words and text, semantic tasks continuously advance by enormous leaps and bounds. But human language is highly flexible, inconsistent, and complex. It poses significant difficulties, such as semantic ambiguity, synonyms, contextual words and phrases, and homonyms, which have not yet been fully resolved. So, this study explores semantic challenges in Natural Language Processing (NLP) that improve the efficiency of task-solving techniques.

In this dissertation, we propose effective knowledge injection techniques for NLP semantic problems. We concentrate on using the Transformer architecture, the pre-trained language model, and the topic knowledge from the topic model to complete these tasks in light of the most recent state-of-the-art (SOTA) results. Semantic textual similarity and summarization are two specific tasks where the usefulness of our methods is demonstrated. In order to do this, we suggested a technique to enhance topic information coherence and took into account how it impacts the injection of subject and context information.

The first challenge is the semantic textual similarity task. In most applications, text understanding and representation are essential, especially in automatic processing. Together with the surface features of words, topic information is significant and necessary to provide the context meaning in the text representation. Recently, the integration of linguistic features and topic information has not received close critical attention. To take advantage of topic information, we propose a novel approach to integrate the topic features into the most popular language models called the Sub-word Latent Topic and Sentence Transformer (SubTST). Inspired by Sentence-BERT and tBERT, our proposed architecture has a significant chance to learn and incorporate topic information with linguistic features. The strength of our proposed approach comes from the delicate combination between latent topic information and linguistic features of language models instead of

only utilizing topic information in the previous works. The comparison in experiments and ablation studies against competitive baselines proves the strength of our proposed approach in most benchmark datasets.

The topic information has helped to direct semantics in text summarization, which is the second issue we consider. As a result, we offer research on the tBART, an innovative and effective way for incorporating topic information with the BART model for abstractive summarization. The suggested model incorporates the benefits of the BART, learns latent topics, and uses an align function to translate the token topic vector into context space. The experimental results demonstrate the potency of our suggested approach, which significantly outperforms existing methods on two benchmark datasets.

Finally, we focus on improving topic coherence. Topic coherence is the primary measure of topic modeling. The more accurately the latent topic is exploited, the higher the topic coherence value. In this study, we proposed a novel method for latent topic refinement called Support Learning for Topic Model (SupLeT). The method is based on non-negative matrix factorization and combined with distance metric learning to increase the quality of topic modeling. We used the learned latent topics during the training process as the "soft label" for the teaching of distance metric learning (DML). The idea of using this learning is that it brings the same topic words closer and tries to keep others as far away as possible. With the learning distance metric process iteratively, we can refine the word-document and word-word relations in each step of the training process. Our experiments show that the SupLeT outperforms baseline Latent Dirichlet Allocation and the base models (Non-negative Matrix Factorization and Semantics-assisted Non-negative Matrix Factorization) on the topic coherence metric and accuracy on topic-based document classification, and semantic similarity detection tasks on benchmark datasets.

To summarize, the focus of our research is on solving fundamental issues relating to the interaction between topic information and context information. The efficiency of the suggested methodologies and their potential for domain adaptation was demonstrated by the experimental findings and thorough analysis. The presented models and solution ideas have the potential to be widely applicable to different types of semantic representations of numerous NLP tasks in further studies.

**Keywords**: Knowledge injection, topic model, transformer, bi-encoder, BART, distance metric learning, semantic similarity detection, summarization.

論文審査の結果の要旨

This thesis proposes two novel methods to incorporate latent topics into pre-trained language models and its application to semantic tasks in NLP. The first model is SubTST, where latent topics of sub-words are injected into the BERT model. Experimental results showed that SubTST attained a good performance compared with other models in the semantic textual similarity task. His work's additional main contribution is applying topic models for the BART framework. The proposed model tBART shows that it outperforms all baseline models in the benchmark data of the text summarization task. The advantage of the tBART model is that it is built based on the popular architecture in deep learning – the BART architecture so that it can be widely used in NLP applications. In the final chapter, the thesis presents a new method to infer latent topics using distance metric learning and shows its effectiveness against the baseline model using several standard datasets. The candidate has published his works in a good journal and many international conference papers. Overall, this is an excellent dissertation, and we approve of awarding a doctoral degree to Mr. DangTran Binh.