

Study on Text Representation Learning and Its Applications

王, 博

<https://hdl.handle.net/2324/7363816>

出版情報 : Kyushu University, 2024, 博士 (工学) , 課程博士
バージョン :
権利関係 :



氏 名 : 王 博 (WANG BO) オウ ボウ

論 文 名 : Study on Text Representation Learning and Its Applications
(文書表現学習とその応用に関する研究)

区 分 : 甲

論 文 内 容 の 要 旨

Text representation learning, or text feature learning, is a fundamental preprocessing step in natural language processing (NLP). Because the raw text consists of discrete sequences of varying lengths, it needs to be transformed into continuous feature representations that capture meaningful patterns, making it suitable for machine learning algorithms.

Recently, text representation methods have evolved from hand-crafted features to deep learning-based automatic feature extraction. Modern large pre-trained language models, such as BERT (Bidirectional Encoder Representations from Transformers), efficiently encode text into semantic-rich embeddings. These embeddings can then be fine-tuned for specific downstream tasks, demonstrating strong performance across various NLP applications.

However, recent research highlights two main limitations in current text representation learning methods: first, traditional methods like pre-training and cross-entropy-based fine-tuning often provide weak supervision signals; second, these methods lack mechanisms to separate representations based on contrasting meanings (i.e., negative example learning). As a result, text representations of different meanings or categories often remain scattered, hindering downstream task performance.

To address these limitations, this study proposes a novel metric-based approach to optimize text representation. The objective of this method is straightforward: to pull similar examples closer together in the representation space while pushing dissimilar ones further apart. This approach strengthens supervision and explicitly incorporates negative example learning, enhancing the distribution of the text representation space and thus facilitating downstream tasks. To validate its effectiveness, this study applied the method across various real NLP applications, including:

1. Short Answer Scoring: Short answer scoring is widely used in automated examination systems. Traditional methods typically treat it as a text classification task, using {answer, score} pairs to fine-tune a language model. However, this often results in suboptimal representation space and unsatisfying accuracy. Additionally, previous approaches rarely considered the relationships between answers. To address these issues, this study proposes a similarity-based metric learning method for scoring. This method first defines similarity between answers based on their scores and rubric criteria, then uses this similarity to guide representation learning. By ultimately employing a simple MLP (Multi-Layer Perceptron)

network, this approach achieves significantly higher accuracy than traditional methods (7.4% in English and 6.8% in Japanese dataset on average in RMSE), while reducing the training and computation time by up to 1/300. This efficient and accurate method holds promising implications for enhancing automated short answer scoring.

2. **Out-of-Domain (OOD) Detection in Question-Answering (QA) Systems:** Traditional classification-retrieval QA systems are trained on an initial, fixed dataset and cannot handle unexpected, unanswerable inquiries in real situations, leading to forced or inappropriate responses. To address this limitation, a novel method based on Supervised Contrastive Learning (SCL) is developed, which can effectively detect OOD inquiries using only known in-domain data. This approach first conducts a data oversampling, then refines text representations by pulling embeddings of the same category texts closer and pushing those of different texts apart, creating compact clusters for each category. This enables a simple, distance-based algorithm to easily distinguish between answerable and unanswerable inquiries accurately. Experiments on three widely used datasets show that the proposed method surpasses current state-of-the-art methods in detection accuracy, achieving over 99% AUROC in average on three datasets. It also significantly reduces the model size to 1/3 and training time to 1/5, while still performing strongly in limited-data conditions that uses 20% training data to realize an AUROC over 98%.
3. **Building a Comprehensive QA System:** Traditional QA systems are restricted to predefined questions and lack adaptability to new data, limiting their functionality. Additionally, the common cross entropy-based training is often inefficient in training time and performance. To overcome these issues and develop a robust, efficient, and adaptable QA system, this work first expands the basic QA system with three additional tasks - out-of-domain question detection, new intent discovery, and continual learning - integrated within a cohesive workflow. After that, using supervised contrastive learning as the unified text representation optimization method, it achieves streamlined and effective training for each task. Experimental results and representation visualizations confirm that the new framework gives surpassing results than state-of-the-art methods across all four tasks, highlighting its efficiency.

Through these applications, it has been demonstrated that a well-designed metric learning approach can effectively optimize text representation for diverse NLP tasks. Its superior training performance simplifies downstream implementation, contributing to advancements in text representation learning.