| | |
|---|---|
| 氏　　名 | 佐々木（ささき） 翔大（しょうた） |
| 学位の種類 | 博　士（情報科学） |
| 学位記番号情 | 博　第　７８８　号 |
| 学位授与年月日 | 令和４年　９月２６日 |
| 学位授与の要件 | 学位規則第４条第１項該当 |
| 研究科、専攻 | 東北大学大学院情報科学研究科（博士課程）システム情報科学専攻 |
| 学位論文題目 | Analyzing the Effect of Cross-lingual Transfer, Compression and Whitening on Natural Language Encoding（自然言語符号化における言語横断的な転移学習、圧縮、白色化の効果の分析） |
| 論文審査委員 | （主査）東北大学教　授　乾　健太郎<br>東北大学教　授　篠原　歩　　東北大学教　授　大町　真一郎<br>東北大学教　授　鈴木　潤 |

## 論 文 内 容 の 要 旨

### Chapter 1: Introduction

A *text encoder* is an essential component of almost all deep Natural Language Processing (NLP) models. Its role is to convert the input text into embeddings, i.e., vectors that represent the meaning of the text. As a result of various refinements of the text encoder, deep NLP models now perform incredibly well, however, only on a limited number of tasks. In addition, recent models lack practical applicability as they are not controllable due to being too large.

This dissertation focuses on further improving text encoders from two perspectives: *quality* and *applicability*. A high-quality text encoder must provide suitable encodings that transform the information obtained from the text without any loss of information. A text encoder must also be widely applicable in the real world where language has ever-evolving usages and forms. Aiming for high-quality and widely applicable text encoders, we explore three aspects with potential for improvement: (i) the required amount of training data, (ii) memory requirements for embeddings and unknown words, and (iii) biases in embeddings.

We first propose a cross-lingual transfer learning method to address the data size problem and demonstrate that it improves model performance for low-resource languages.
Specifically, as an example of an application of NLP, we conducted experiments on information retrieval tasks. We also clarify the model architecture for information retrieval in which the transfer learning works.

We then propose a method that combines memory sharing and key-value-query (KVQ) operation idea in order to simultaneously reduce the memory requirement for static word embeddings and deal with the unknown word problem. The proposed method succeeded in reducing the amount of required memory while acquiring unknown word embeddings which performed well on word similarity tasks.

Finally, we examine the semantic effects of *whitening*, one of the isotropization methods for

anisotropic embeddings, i.e., reducing spatial bias. From our experiments on a sentence similarity task, we report that the effect on static word embeddings has significant overlap with the effect of removing word frequency bias, and that this is not the case for contextualized embeddings.

## Chapter 2: Background

The main theme of this dissertation is to improve a text encoder that is an essential component of almost all deep NLP models. This chapter explains the internal components of a text encoder and their evolution. A text encoder consists of two main components: (i) an embedding layer and (ii) a task-specific layer. An embedding layer is a general component that models for various tasks can share. Its role is to serve as an entry point of a text encoder, converting input texts into embeddings (vectors) that machines can handle. A task-specific layer has a different network architecture for each task and basically cannot be shared across tasks. A task-specific layer acts as a bridge between an embedding layer and the goal of a target task. The resultant vector from a text encoder is used to determine an output of a model.

An embedding layer, as the entry point to a text encoder, needs to transform input texts into embeddings that machines can handle. To achieve this, an embedding layer must capture languages' universal properties and usages.

"Word" has long been recognized as the most intuitive and smallest unit for constructing the meaning of a sentence. At the beginning of the deep learning era of NLP, *distributional representations* were used as the semantic representations of words. Following distributional representations, word embeddings from neural language models have received more attention. Skip-gram and CBOW, also known as Word2Vec, received a great deal of attention for their success in acquiring representations that capture relationships between words. With the advent of these word embeddings, the affinity between NLP and deep learning has increased dramatically.

Despite the success of static word embeddings, several drawbacks have been pointed out due to their static nature. The methods of static word embeddings basically assign a vector to each word in a predefined vocabulary. The assigned vector retains the meaning of the word itself without considering its contexts. This property becomes problematic when the scope of the target is expanded to ``sentence''. Embedding methods with an extended scope to contexts of words have also been explored. With the evolution of a GPU, models with many layers and large hidden states have been explored such as BERT.

While performance improvements were reported, undesired biases in embeddings caused by using a large raw corpus as training data have become apparent. In addition, the huge internal parameter set and the fact that many high-performance GPUs is required for the use of the models have raised doubts about their applicability in the real world.

Following the establishment of an embedding layer, a task-specific layer has been explored in various ways on a task-by-task basis. A convolutional neural network (CNN) has been used in text classification models as a task-specific layer to compose word embeddings. RNN, long short-term memory (LSTM), and gated recurrent unit (GRU) incorporated into a

sequence-to-sequence model have had great success in translation tasks. These networks were established as a general architecture that views a sentence as a word sequence rather than a bag-of-words (BoW), and have been applied to a wide range of tasks.

**Chapter 3: Cross-lingual Transfer Learning for Information Retrieval**

Multilingual document collections are becoming prevalent. Thus an important application is cross-lingual information retrieval (CLIR), i.e., document retrieval which assumes that the language of the user's query does not match that of the documents.

There are two main approaches to building CLIR systems. The *modular approach* involves a pipeline of two components: translation (machine translation or bilingual dictionary look-up) and monolingual information retrieval (IR). The idea is to solve the translation problem separately, so that CLIR becomes document retrieval in the monolingual setting.

A distinctly different way to build CLIR systems is what may be called the *direct modeling* approach. This assumes the availability of CLIR training examples of the form $(q, d, r)$, where q is an English query, d is a foreign-language document, a r is the corresponding relevance judgment for $d$ with respect to $q$. One directly builds a retrieval model $S(q, d)$ that scores the query-document pair. While $q$ and $d$ are in different languages, the model directly learns both translation and retrieval relevance on the CLIR training data. Compared to the modular approach, direct modeling is advantageous in that it focuses on learning translations that are beneficial for retrieval, rather than translations that preserve sentence meaning/structure in bitext.

However, there exist no large-scale CLIR dataset that can support direct modeling approaches in a wide variety of languages. Here, we present a large-scale dataset that is automatically constructed from Wikipedia: it can support training and evaluation of CLIR systems between English queries and documents in 25 other languages.

To demonstrate the utility of the data, we further present experiments for CLIR in low-resource languages. First, we introduce a neural CLIR model based on the direct modeling approach. We then show how we can bootstrap CLIR models for languages with less training data by an appropriate use of paramater sharing among different language pairs. For example, using the training data for Japanese-English CLIR, we can improve the Mean Average Precision (MAP) results of a Swahili-English CLIR system by 5-7 points.

**Chapter 4: Subword-based Compact Reconstruction for Open-vocabulary Word Embeddings**

A recent trend is to embed word meanings into a vector space by using the rapidly developing neural word embedding methods, such as Skip-gram, GloVe, and fastText. These methods have successfully been proven to capture high-quality syntactic and semantic relationships in a vector space. Typical examples of large, well-trained word embeddings are those trained on the CC corpus with 600 billion tokens by fastText and with 840 billion tokens by GloVe, which we refer to as fastText.600B and GloVe.840B, respectively. Such word embeddings are often used to improve performance in many natural language processing (NLP) tasks.

However, well-trained word embeddings have several limitations, and we focus on two issues surrounding them: (i) the massive memory requirement and (ii) the inapplicability of out-of-vocabulary (OOV) words. The total number of embeddings often becomes unacceptably large, especially in limited-memory environments, including GPUs, because the vocabulary size is more than 2 million words, requiring at least 2 gigabytes (GB) of memory for storage. The memory requirement could be straightforwardly reduced by merely discarding the less important words from the vocabulary. However, such a naive method worsens the drawback of the inapplicability of OOV words, whose handling is highly desirable in real systems because input words can be uncontrollably diverse. Therefore, at present, there is a trade-off between the number of embedding vectors and the applicability of OOV words. Our goal is to investigate and develop a method that simultaneously equips less memory requirement and high applicability of OOV words.

A popular approach for mitigating or solving OOV word issues is to use subword information. Conceptually, the subword-based approach covers all words that can be constructed by a combination of subwords. Thus, the subword-based approach can greatly mitigate the OOV word issue. We extend this approach to simultaneously reduce the total number of embedding vectors through the reconstruction of word embeddings by using subwords. The key techniques of our approach are twofold: memory-shared embeddings and a variant of the key-value-query (KVQ) self-attention mechanism. That is, our approach reconstructs well-trained word embeddings by using a limited number of embedding vectors that are shared by all the subwords with an effective weighting calculated by the self-attention mechanism.

We experimentally show that our reconstructed subword-based embeddings can successfully imitate well-trained word embeddings, such as fastText.600B and GloVe.840B, in a small fixed space while preventing quality degradation across several linguistic benchmark datasets from word similarity and analogy tasks. We also demonstrate the effectiveness of our reconstructed embeddings for representing the embeddings of OOV words. Finally, we confirm the performance of our reconstructed embeddings in several downstream tasks such as the named entity recognition task and the textual entailment task.

**Chapter 5: Examining the Effect of Whitening on Static and Contextualized Word Embeddings**

Static word embeddings (SWE) and contextualized word embeddings (CWE) are the foundation of modern natural language processing systems. However, while the aim of creating such embeddings is to provide accurate representations of word, phrase, and sentence meaning, they also reflect and sometimes amplify biases inherent in the training data, such as gender bias, social bias, and word frequency bias. For SWE, prior research has demonstrated that the embedding space exhibits a spatial frequency bias; namely, frequent words tend to concentrate along a particular direction. Generally, this anisotropy, i.e., the non-uniform angular distribution of word vectors, is undesirable because it leads to inefficient use of the embedding space. Furthermore, frequency-based anisotropy causes frequent words to be represented by similar vectors simply by virtue of their high frequency, although their meaning may not be similar.

Aiming to reduce the negative impact of anisotropy, several isotropization methods have been proposed. These methods make embeddings more isotropic, i.e., transform embedding vectors so that they have a more uniform angular distribution. Isotropization methods can be divided into supervised debiasing methods and unsupervised post-processing methods. The primary goal of supervised debiasing methods is to remove biases with respect to specific categories, such as gender, nationality, and word frequency. If the bias manifests itself as an uneven distribution of word embeddings, then debiasing results in a more isotropic embedding space.

The focus of this study is the arguably simplest and most common unsupervised isotropization method, namely whitening. Informally, whitening is a linear operation that transforms a set of spatially correlated (and therefore anisotropic) vectors into a set of uncorrelated (isotropic) vectors. Previous studies have demonstrated that whitening performs better than other isotropization methods for CWE. However, a major disadvantage of whitening and other unsupervised post-processing methods is that their impact on various forms of bias and other semantic properties of embeddings is not yet understood, although understanding bias in SWE and CWE is a prerequisite for their ethical use in real-world applications. In this chapter, we present an initial analysis of the semantic impact of unsupervised post-processing.

Our preliminary analysis indicates that the effect of whitening partially includes the effect of frequency debiasing. Our research question is thus whether the effect of whitening consists of frequency debiasing only. To increase the granularity of the effect of whitening, we employ a method whose effect is frequency debiasing only; specifically, we propose a reconstruction-based frequency debiasing (RFD), which focuses only on removing frequency bias in embeddings. We then compare the behavior of whitening with that of RFD. Our experimental results indicate that whitening removes word frequency bias in SWE as well as biases other than word frequency bias in CWE.

### Chapter 6: Conclusion

The key contributions of this dissertation are summarized as follows.

#### Investigation of the effectiveness of cross-lingual transfer learning:

We proposed a cross-lingual transfer learning method and demonstrated that it improved model performance for low-resource languages. Specifically, as an example of an application of NLP, we conducted experiments on information retrieval tasks. We also clarified the model architecture for information retrieval in which the transfer learning works.

#### Reducing memory requirements and handling unknown words:

In order to simultaneously reduce the memory requirement for static word embeddings and deal with the unknown word problem, we proposed a method that combines memory sharing and key-value-query (KVQ) operation idea. The proposed method succeeded in reducing the amount of required memory while acquiring unknown word embeddings which performed well on word similarity tasks.

**Examining the effect of whitening on word embeddings:**

We examined the semantic effects of whitening, one of the isotropization methods for anisotropic embeddings, i.e., reducing spatial bias. From our experiments on a sentence similarity task, we reported that the effect on static word embeddings had significant overlap with the effect of removing word frequency bias, and that this was not the case for contextualized embeddings.

# 論文審査結果の要旨

深層学習を用いた言語処理モデルにおいて、「自然言語エンコーダ」は入力テキストをベクトル表現へ符号化する役割を果たす。自然言語エンコーダは、近年深層学習の発展に伴って目覚ましい進化を遂げてきたが、残された課題も多い。言語処理モデルにとって必須の構成要素となった自然言語エンコーダを改善することは、タスクに関係なく多様な言語処理モデルを改善することに繋がり、極めて重要性が高い。本論文では、自然言語エンコーダをその「質」と「応用可能性」の観点から分析し、言語処理モデルの実用に向けた改善手法の提案を行っている。具体的には、低資源言語モデルの性能改善のための転移学習、単語埋め込みに介在するバイアス緩和手法である「白色化」の効果の検証、自然言語エンコーダのモデルサイズ問題と未知語問題の同時対処を行った。本論文は一連の成果をまとめたもので、全編 6 章からなる。

第 1 章は序論である。

第 2 章では、自然言語エンコーダの構成と変遷、各パラダイムにおいて象徴的な研究課題を体系的に整理することで、本研究の背景知識を説明している。

第 3 章では、言語処理モデルのタスク特化埋め込みを学習する際の、データ量問題に関して論じている。大規模データが入手しにくい低資源言語のための言語横断的な転移学習手法を提案し、情報検索モデルの性能改善に成功している。情報検索タスクにおいて、言語横断的な転移学習が有効であることを初めて示した貢献は、高く評価できる。

第 4 章では、単語埋め込みに介在するバイアスの緩和手法である白色化の効果を検証している。白色化には、単語埋め込みに介在する単語頻度バイアスを緩和する効果を含むことを示唆した。その上で、単語埋め込みが静的であるか、動的であるかによって、白色化の効果の全容が異なることも示した。一連の知見は同分野の研究者にとって示唆に富むもので、高く評価できる。

第 5 章では、単語埋め込みの実応用において問題となる、必要メモリ量問題と、未知語問題に同時に対処する手法を提案している。サブワード埋め込みに基づくフレームワークにおいて、メモリ共有手法、Key-value-query 演算を組み合わせる提案手法によって、単語埋め込みのモデルサイズを大幅に削減した上で、性能の低減を非常に小さく抑えることに成功するとともに、良質な未知語埋め込みを獲得することに成功しており、高く評価できる。

第 6 章は結論である。

以上、本論文は言語処理モデルの実用を考えたときに生じる 3 つの課題に焦点を当て、自然言語エンコーダの改善のための研究基盤の確立および今後の方向性を示したものであり、情報科学の発展に寄与するところが少なくない。よって、本論文は博士（情報科学）の学位論文として合格と認める。