氏　　名	ATTAPORN WANGPOONSAP
博士の専攻分野の名称	博士（情報科学）
学位記番号	医工農博甲第５０号
学位授与年月日	令和３年３月２３日
学位授与の要件	学位規則第４条第１項該当
専　　攻　　名	人間環境医工学専攻　生命情報システム学コース
学位論文題目	A Study on Detecting Domain-Specific Senses and its Application to Text Categorization (分野依存語義の抽出と文書分類への適用に関する研究)
論文審査委員	主査　教　授　福本 文代
	教　授　鈴木　良弥
	教　授　服部　元信
	教　授　渡辺　喜道
	准教授　西崎　博光
	准教授　鈴木　智博

学位論文内容の要旨

This thesis consists of three works. The first work presents an unsupervised method for identifying a proper sense in a document called "Domain-specific senses (DSS)" to resolve the traditional methods do not perform well to transform documents to vectors. The work consists of three procedures following as 1) Pre-processing, 2) Calculating sense similarity, and 3) Ranking sense scores. The goal of pre-processing is to make a sense list for each category. The method utilizes PageRank to identify domain specific senses. Its input is a graph whose node is a sense extracted from the WordNet and the edge is the semantic similarity between two senses. The author calculated the semantic similarity by utilizing Word Mover's Distance. The method was tested on WordNet 3.1 and the Reuters corpus. The result of the sense assignment on Reuters corpus obtained by the method attained 1.5 improvements on the Inverse Ranking Score (IRS) over the Cosine approach.

In the second work, a semi-supervised method is proposed to identify DSS. Because most results from unsupervised learning methods are not effective and they have room

to improve the overall performance. This work addresses the problem and proposes an approach for improving performance based on deep learning. It consists of four procedures following as 1) Pre-processing, 2) Producing embedding 3) Building a graph with an adjacency matrix, and 4) Predicting categories and propagation. Pre-processing is to extract senses and glosses in the WordNet from given categories. Each word extracted from the documents of RCV1 is annotated for POS and is lemmatized using Stanford CoreNLP. Noun and verb words are chosen and used to find their senses and gloss texts from the WordNet. For each category, noun and verb words are extracted. The sense embedding is learned by using bert-as-service. Bidirectional Encoder Representations from Transformers (BERT) is a type of neural network model for pre-training language embeddings developed by Devlin et al. BERT is applied to learn the feature representation of gloss texts to build sense embeddings as an input for the prediction stage. In this work, the pre-trained BERT model. BERT-Base is used, Uncased (L=12, H=768, A=12) which represents a model consisting of 12 layers, 768 hidden units, and 12 attention heads. BERT's input formatting has two important special tokens consisting of [SEP] that is used to indicate the end of a sentence or used as a separator between two sentences. [CLS] is used to indicate the beginning of a sentence. Both tokens are inserted into a sentence after that token embeddings, the segment embedding, and the position embedding are summed up to build input embeddings. The author used input embeddings that have four dimensions including the number of layers, the number of batches, the number of tokens, and the number of features for creating sense embeddings. The pooling strategy is the average pooling on the second-to-last layer to obtained sense embeddings. The third step is Building a graph with an adjacency matrix. The co-occurrence matrix between senses is created, each of which meets three criteria: firstly, target senses and their POS are found in RCV1 documents. Secondly, they have the same category. Thirdly, each document contains more than one sense. By utilizing sense relationships, an adjacency matrix is created by converting the non-zero value in a co-occurrence matrix equal to one. The final step for determining domain-specific senses to predict each sense on a graph using a neural network model. The APPNP model is utilized in this work because it is based on the GCN model that is a very powerful neural network even 2 layers of GCN can generate useful feature representation of nodes on graphs and it also solves the lost

focus issue with PPR. Sense embedding is used which is the result from the second step as an input and then training with the APPNP model that predicts a proper category for each sense. The experimental results show that this approach works well and attain a 0.647 Macro F1-score.

The third work of the thesis focuses on the influence of DSS on text categorization. The experiments are divided into two sections. The first section is to apply DSS from unsupervised learning to the text categorization task to examine how well the automatically acquired sense contributes to categorization accuracy by comparing one of the WSD techniques, Context2vec. Whereas the second section is also to apply DSS from semi-supervised learning to the text categorization task and is compared the results between semi-supervised learning and unsupervised learning. The text categorization task with unsupervised learning has empirically proven that DSS can achieve a better performance than the WSD method. The DSS results attained at F-score 0.745 for 442 senses, furthermore, when applying the text categorization, The result of the categorization accuracy at the Macro F-score is 0.832 that exceed the WSD method 0.053. Whereas the text categorization task with semi-supervised learning has also proven that DSS can reach a better performance than unsupervised learning. The DSS results show that the method can improve text categorization performance as it achieved a 0.918 Macro F-score and 0.142 improvements compare with the CNN baseline model.

論文審査結果の要旨

令和 3 年 2 月 3 日 16:00 からコンピュータ理工学科実習室にて Wangpoonsarp さんの博士論文 A Study on Detecting Domain-Specific Senses and its Application to Text Categorization (分野依存語義の抽出と文書分類への適用に関する研究)の学位論文審査を実施した。本博士論文は、3 つの課題に注目し、その解決方法を提案している。1 点目は、単語の語義がスポーツや経済などの分野に依存して決まる、例えば、単語 Court がスポーツ分野に頻出する文書では、Court はテニスコートの意味でよく用いられ、司法分野に頻出する文書では、裁判所の意味で決まる。そこで、WordNet と呼ばれる単語のシソーラス辞書に記載されている単語がどの分野で頻出した場合に、その語義が一意に決定できるかを自動的に判定する手法を提案した。具体的には、シソーラスをノードを単語語義、エッジを単語語義間の類似度からなるグラフ構造で表現し、PageRank と呼ばれる手法を適用する

ことにより、各分野に頻出する単語語義を推定する手法を提案している。単語語義間の類似度には、Word Mover Distance と呼ばれる、深層学習ベースの距離尺度法を用いた計算手法を提案した。ロイター記事コーパスを用いた実験、及び比較実験により有効性を検証した。

2 点目は、より高精度な分野依存語義の判定を行うために、深層学習を用いたグラフ構造の学習を適用する手法を提案している。BERT と呼ばれる単語の埋め込み表現手法により得られた単語の意味表現を入力とし、Approximate personalized propagation of neural predictions と呼ばれる深層学習手法を利用することで、単語語義を抽出する手法を提案した。実験では、ロイター記事に頻出する名詞と動詞単語に対し、それらの分野依存語義が高精度で抽出可能かどうかをベースラインとの比較により検証している。

3 点目は、得られた分野依存語義の評価である。具体的には、外的評価として文書分類に着目し、PageRank による分野依存語義と Approximate personalized propagation of neural predictions による分野依存語義が文書分類の精度にどの程度貢献するかを検証している。文書分類手法は、分類のベースとして用いられている Convolutional Neural Networks を用い、ロイター記事の精度を分野依存語義なし、PageRank による分野依存語義あり、Approximate personalized propagation of neural predictions による分野依存語義ありとの比較により手法の有効性を示している。

学位論文審査では、3 章の研究に関する新規性について、また統計手法の問題点がどこにあるのか、さらに関連研究との比較をどのように実施したのかに関する質疑、議論を行った。続いて、5 章と 3, 4 章との関係に関する質疑においてその位置付けの説明を再度行った。また実験結果の精度に関する質問については、finance 分野が極端に悪い理由とその解決方法について議論した。

続く最終試験では、5 章の比較実験において、教師なしと半教師付き学習との fair な実験が実現できているかに関する質問、率を変更しての再実験の必要性について議論した。その後に、博士論文について、図の解像度についてもう一度見直すこと、また学位論文審査会での質疑に関してその回答を博士論文の説明に加えることの助言があった。さらに今後の課題に関する方向性についても議論した。

学位論文審査、及び最終試験において、3 つの研究を実施していること、及び 2 編の論文が採録済であり、博士修得に必要な本数を満たしていること、さらに現在 1 本投稿中であることから、論文の質及び量の両面にわたり本学の博士号にふさわしいと判断しこれを授与することとした。