論文 / 著書情報
Article / Book Information

題目(和文)	事前学習済み言語モデルを用いた検索モデルに対する教師なしドメイン適応
Title(English)	
著者(和文)	飯田大貴
Author(English)	Hiroki Iida
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第12780号, 授与年月日:2024年3月26日, 学位の種別:課程博士, 審査員:岡崎 直観,井上 中順,徳永 健伸,宮崎 純,村田 剛志
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第12780号, Conferred date:2024/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	審査の要旨
Type(English)	Exam Summary

（博士課程）

論文審査の要旨及び審査員

報告番号	甲第　　号		学位申請者氏名		飯田 大貴	
論文審査 審査員		氏　名	職　名		氏　名	職　名
	主査	岡崎 直観	教授	審査員	村田 剛志	教授
	審査員	井上 中順	准教授			
		徳永 健伸	教授			
		宮崎 純	教授			

論文審査の要旨（2000 字程度）

本論文は、「事前学習済み言語モデルを用いた検索モデルに対する教師なしドメイン適応」と題し、和文 5 章から構成されている。深層学習に基づく情報検索では、文書とクエリを固定長のベクトルで表現し、ベクトル間の内積などで文書の適合度を計算する方法論（密ベクトル検索）が採用されている。この時、深層学習モデルとして、事前学習済み言語モデルが用いられる。密ベクトル検索では文書とクエリが文脈依存型の分散表現で表現されるため、単語の意味の違いや同義語を自然に扱えるという利点があるが、教師あり学習に基づくため、学習したドメイン以外で利用しにくいという課題がある。本論文では、密ベクトル検索などの事前学習済み言語モデルを用いた検索モデルを教師無しでドメイン適応させる研究に取り組んでいる。

第 1 章「序論」では、研究背景として、情報検索において事前学習済み言語モデルの利用方法が述べられている。その課題として、教師データが必要なため、学習したドメイン以外で精度が低下し利用が困難なことを指摘している。その後、この課題を解決するアプローチとして、対象データにおけるトークンの重要度を用いる手法と事前学習済み言語モデルに語彙を追加するアプローチの二つを挙げ、これらの差異および貢献を説明している。

第 2 章「準備と関連研究」では、まず従来の検索モデルである BM25 の導出を行っている。次に、事前学習済み言語モデルの代表として BERT の説明を行っている。事前学習済み言語モデルを用いた検索モデルである密ベクトル検索、SPLADE、ColBERT、COIL-tok を述べ、その学習方法について記載している。また、BM25・事前学習済み言語モデルを用いた検索モデルの双方で確率モデルとしての定式化も行っている。その後、検索モデルに対する教師なしドメイン適応手法の既存研究を述べた後、評価指標を説明している。

第 3 章「対象データにおけるトークンの重要度を用いるドメイン適応」では、密ベクトル検索の課題である、クエリ中のキーワードを含む文書を上位にできないという課題に対して取り組んでいる。この問題を解決する検索モデルとして COIL-tok が先行研究で提案されている。検索では、クエリと文書をトークンに分割して処理を行うが、COIL-tok はクエリと文書で一致したトークンに対して処理を行うため、キーワードの一致を考慮する。また、COIL-tok では BERT から各トークンのベクトルを得ることで、単語の意味の違いも考慮する。本研究では、COIL-tok を教師データのドメイン以外で精度を向上させるための手法として Contextualized BM25 (C-BM25)を提案している。C-BM25 は、対象データにおけるトークン重要度として、BM25 で重み付けを行っている。さらに、トークンのベクトルを密ベクトル検索で学習したモデルを用いることで、単語の意味をより高精度に判別させている。提案した C-BM25 を、検索分野で広く用いられている教師なしのベンチマークである BEIR にて評価を行い、既存手法を上回ることを示した。また、検索における応答時間が実用的であることを示している。

第 4 章「事前学習済み言語モデルに対する語彙追加を用いるドメイン適応」では、事前学習済み言語モデルに語彙を追加し継続事前学習を行う AdaLM を用いることを提案している。第 3 章の手法では、適用対象が密ベクトル検索に限られていた。しかし、SPLADE や ColBERT といった学習したドメイン以外でも高精度な検索モデルがある。そのため、これらに適用可能な教師なしドメイン適応手法であれば、更なる精度向上が見込まれる。さらに、教師なしで検索精度を向上させる要求はより専門的なドメインで強い。専門的なドメインでは、教師データを多く得られるドメインとは語彙や単語頻度が異なる場合が多い。この差異を事前学習済み言語モデルのドメイン適応で解消するのが、AdaLM である。AdaLM を用いることで、ドメイン特有の語がトークナイズされることを防ぎ、SPLADE においてキーワードを含む文書を検索結果の上位にすると共に、継続事前学習を用いてドメイン特有の語の同義語・関連語を得る。本研究では、AdaLM を BERT に適用した後 SPLADE の学習を

行っている。これに加えて、クエリ中のキーワードを含む文書を上位にできないという課題を解決するために、対象データにおけるトークンの重要度を合わせて用いている。BEIR のうち専門的なドメインであるバイオ・医療および科学技術ドメインのデータセットで評価実験を行ったところ、提案手法（AdaLM の適用および対象データにおけるトークンの重要度の考慮を行った SPLADE）は既存手法を上回る検索精度を示した。また、SPLADE に対して適用可能な既存の教師なしドメイン適応手法である疑似クエリを用いたアプローチとも比較を行い、AdaLM に優位性があることを確認している。

第 5 章「結論」では、本論文のまとめと今後の展望を述べている。

本論文では、検索での教師無しドメイン適応手法に向けて、対象ドメインにおけるトークンの重要度を用いる手法、事前学習済み言語モデルに対象ドメインの語彙を追加する手法の二つを提案した。いずれの手法も、事前学習済み言語モデルを用いた検索モデルを教師無しで用いる場合の精度低下要因の一つである「クエリ中のキーワードが完全一致する文書を上位にできない」という課題を解決するものであり、検索対象文書のドメインシフトに有効な手法であることを実験的に示した。本論文の成果は、すでに実用化されている密ベクトル検索の応用領域を広げることから、工学の発展にも寄与する。また、大規模言語モデルを用いた密ベクトル検索などへの発展も考えられる。よって、本論文は博士（工学）の学位論文として十分価値あるものと認める。