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)	論文要旨
Type(English)	Summary

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

系・コース： Department of, Graduate major in	情報工学 系 知能情報 コース	申請学位（専攻分野）： Academic Degree Requested	博士 Doctor of （ 工学 ）
学生氏名： Student's Name	飯田 大貴	審査員主査： Chief Examiner	岡崎 直観

要旨（英文 800 語程度）
Thesis Summary (approx.800 English Words)

Information retrieval (IR) is a widely used technique for retrieving data that matches a query request from a large amount of data. In particular, Web search, a service for retrieving documents on the Internet, has become so pervasive that it is now indispensable in our daily lives. Traditional IR methods rely on token matching between queries and documents. These models fail to capture contextual nuances and semantic similarities when the same concept is presented with different terminologies. Recent advances have seen the integration of pre-trained language models like BERT in retrieval models, enhancing the ability of search systems to interpret the meanings of queries and documents. A well-known example in these models is dense retrieval, which encodes queries and documents into dense vectors, computing relevance through the inner product of these vectors. This approach has shown a marked improvement in capturing semantic relationships over traditional token-based matching, mainly when substantial supervision data (query-document pairs), typically sourced from click logs, is available.

Despite these advancements, challenges persist in environments where ample supervised data is impractical or infeasible, such as in specialized domain searches or searches within private organizational databases. These settings often suffer from a scarcity of query-document pairs due to the requirement of domain expertise for data generation or security restrictions that prevent using log data. In addition, transferring a dense retrieval model to a target domain trained on a source domain for which a large amount of supervision data is available is also difficult because the distribution between the source domain and the target domain changes in terms of vocabulary, word frequencies, types of queries, and relevant documents to the queries.

Therefore, this study aims to improve the accuracy of zero-shot retrieval where no supervised data is available on the target domain. Specifically, we propose an unsupervised domain adaptation method for retrieval models using pretrained language models.

First, we propose a method that uses the importance of tokens in the target domain.
One of the causes of inaccuracy in dense retrieval on the target domain is the inability to rank documents higher with exact keyword matches in the query. On the other hand, IR models that use token matching, like BM25, cannot capture context. Therefore, we propose Contextualized-BM25 (C-BM25), a hybrid model combining the strength of keyword matching with contextual understanding, using context similarity of exact matching tokens between queries and documents. For calculating the context similarity, we use a dense retrieval model. Furthermore, weighting the exactly matched tokens with BM25 allows us to give more weight to keywords in the target domain. We evaluated C-BM25 on a benchmark dataset of zero-shot retrieval. The result showed the effectiveness of C-BM25.

One of the challenges of C-BM25 is that it can only apply to dense retrieval. However, it is challenging to use dense retrieval in highly specialized domains. This is because the accuracy of dense retrieval is greatly reduced since the vocabulary and word frequencies in specialized domains differ from those in the source domain. One retrieval model that solves this problem is SPLADE, which encodes queries and documents into sparse vectors and performs query expansion and document expansion. SPLADE is also a highly accurate retrieval model in domains other than the source domain. Therefore, an unsupervised domain adaptation method applicable to SPLADE is required to improve retrieval accuracy further in specialized domains.

For this reason, we propose using a domain adaptation method of a pretrained language model to improve the accuracy of SPLADE in highly specialized domains. We used AdaLM, a domain adaptation method for pretrained language models, which adds vocabulary to BERT and performs continual pretraining on a corpus from the target. In addition, we used the importance of tokens in the target data to rank documents higher, including the keywords in the query. Specifically, we weighted each element of the sparse vector encoded by SPLADE with IDF. In addition, we added the relevance scores

in BM25 together. Through experiments, we have shown that SPLADE with AdaLM is, on average, more accurate in zero-shot retrieval than existing methods in the bio-medical and scientific domains, where the vocabulary and word frequency differ significantly from the source domain. In addition, we showed that AdaLM is more accurate than pseudo-queries, which is an effective unsupervised domain adaptation method for retrieval. This verified the effectiveness of AdaLM as an unsupervised domain adaptation method. Finally, since AdaLM can be applied to all IR models that use pre-trained language models, we applied it to several IR models, including dense retrieval, and demonstrated its effectiveness. We also showed that applying AdaLM to multiple retrieval models and ensembling them further improved the retrieval accuracy.

Keywords: Information Retrieval, Zero-shot Retrieval, Pretrained Language Model, Unsupervised Domain Adaptation, Dense Retrieval