

技術ナレッジ活用に向けた Retriever-Reader モデルの検証*

Evaluating the Retriever-Reader Model for Utilizing Knowledge from Technical Documents

蓬田 綾香
Ayaka YOMOGIDA

村瀬 文彦
Fumihiko MURASE

平野 徹
Toru HIRANO

三谷 陽
Akira MITANI

坂 一忠
Kazutada BAN

飯田 哲也
Tetsuya IIDA

岩堀 恵介
Keisuke IWAHORI

竹野 貴法
Takanori TAKENO

We evaluated the Retriever-Reader model, which is used for question-answering tasks. We aim to efficiently utilize the technological knowledge accumulated within the company and enable users to easily obtain necessary answers from a vast number of documents. The target documents are those related to material development accumulated within the company. For the Retriever, we evaluated whether the top-ranked documents based on the vectors of BERT or Sentence BERT contained answers to the questions. For the Reader, we assessed two methods: one that extracts the position of the answer from the document, and another that classifies each sentence in the document as either an answer or not. The results indicated that Sentence BERT had a higher accuracy in the Retriever phase. In the Reader phase, the method of determining the labels was more effective in appropriately identifying answers to user questions.

Key words :

Natural Language Processing, Large Language Model, Question-Answering, Information Retrieval

1. はじめに

日々の業務活動を通して社内には様々な文書が蓄積されており、これら大規模テキストデータに含まれるナレッジの効率的な活用によるデジタルトランスフォーメーションの需要が高まっている。適用事例の一つとして、材料開発業務におけるナレッジ活用を想定している。自動車部品は様々な材料で構成されており、

どの部品に使用されるか、その自動車がどの地域で使用されるかによっても求められる特性が異なるため、多種多様な材料開発が必要である。材料開発を統括する部署では、開発から製品量産までの材料に関するデータベースを保有している。事業部から材料開発に関する問い合わせがくると、担当者は各種データベースにアクセスし、検索ワードをもとに文書を絞り込み、文書の中から得たい技術ナレッジを取得し、事業部へ回

答する。具体的な事例を Fig. 1 に示す。質問は材料開発業務関連ではあるが、材料や製品は多岐に渡る。検索された文書の中からナレッジ記載箇所を特定するために多くの文書に目を通す必要があり、従来の検索方法では欲しいナレッジにたどり着くまでの負荷が大きいことが問題点であった。そこで、膨大な文書からユーザーの知りたいナレッジを直接特定できるように、質問応答タスクで活用されている Retriever-Reader モデルの検証を実施した。

Retriever-Reader モデルとは、大量の文書から質問に関連する文書を検索する Retriever と検索された文書から回答を抽出または生成する Reader で構成されているモデルである¹⁾²⁾。Retriever として、TF-IDF や BM25³⁾ のように質問文中に存在する単語を用いたマッチングや、質問文や文書を密なベクトルに変換し、類似度を用いて検索する方法が挙げられる。今回は、対象とするデータベースに表記ゆれが多く含まれていたため、表記ゆれに強いベクトルによる検索方法を用いた。ベクトルへの変換には BERT⁴⁾ を用いた。Reader は抽出型と生成型に大別される。生成型では複数の文書を基に要約した文章や、新たな文章を生成できるが、事実とは異なる意味・内容の誤った回答となる可能性がある。今回対象とする技術ナレッジの特定において、ユーザーに誤った情報を提示することは避けたいため、情報の正確性を重視して抽出型を用いた。

Question : [材料名]を使うときの注意点は?
Answer : [材料名]は水分、腐食環境下で著しく劣化が進行する。

Question : [不良名]が発生した原因と対策は?
Answer : キズが入った[部品名]を使用し、内面にキズを付けた。[部品名]に異常がないか目視検査。

Fig. 1 Examples of Utilization Knowledge in Material Development

2. 手法

本研究では、製品開発時の材料に関する懸念事項をまとめたデータベースを対象に実験を行った。

2.1 モデル

汎用の事前学習済みモデルに対して、タスクのドメインに特化したデータで追加学習を行うとタスク精度が向上したとの報告⁵⁾があることから、BERT base Japanese (unicid-lite with whole word masking, jawiki-20200831)⁶⁾ に対して、対象データベースのテキスト (約 7MB) で追加学習を行った。追加学習では Masked Language Model のみで学習を行った。追加学習したモデルを基に、Retriever および Reader へのファインチューニングを行った。

Retriever 2つのベクトル変換手法を比較した。1つ目は、追加学習済みの BERT をファインチューニングせず、BERT の各トークンの出力を平均で Pooling したベクトルを用いた検索を行った。2つ目は、効率的に精度の高いベクトル化が行えると報告された Sentence BERT⁶⁾ によるベクトルを用いた検索を行った。ファインチューニングの目的関数には triplet Objective Function を用いた。基準 (anchor) となる文章 a に対し、似ている (positive) 文章 p または似ていない (negative) 文章 n との埋め込みベクトル s の差を取り、式 (1) を最小化する目的関数である。anchor-positive, anchor-negative の差は文書ベクトル間のコサイン類似度を用いた。Pooling 手法は出力ベクトルの平均を使用した。2つの手法ともに、質問文およびデータベースの各文書を入力とし、それぞれのベクトルを取得した。

$$\max((s_a - s_p) - (s_a - s_n) + \epsilon, 0) \quad (1)$$

Reader 2つの抽出手法を比較した。Fig. 2 にそれぞれのモデルを示す。1つ目は、SQuAD 2.0⁷⁾ に対する BERT の既存手法⁴⁾ である回答位置抽出を用いた。質問文と抽出対象の文書を連結させた入力に対して、回答可能な文書では回答の開始と終了の位置を、回答不可能な文書では、開始と終了が文頭 [CLS] の位置を出力するようファインチューニングした。2つ目は、抽出対象の文書を句点で区切った各文章および質問文を入力し、入力した文章は質問文に対する回答か否か分類する手法を用いた。[CLS] の出力を用いて、回答である場合 1 を、回答でない場合 0 を出力するようファインチューニングした。

* (一社) 言語処理学会の了解を得て、言語処理学会第 29 回年次大会 (2023)、セッション Q8-3 (https://www.anlp.jp/proceedings/annual_meeting/2023/pdf_dir/Q8-3.pdf) より一部修正して CC BY4.0(https://creativecommons.org/licenses/by/4.0/) に従って転載
i https://huggingface.co/cl-tohoku/bert-base-japanese-v2
ii https://www.sbert.net/

Table 1 Number of Documents, QA Pair, and Training Data for Retriever and Reader

Number of Documents	QA Pair	Retriever Triplet	Reader	
			Start / End Position	Label Classification
2985	93	760	76	350

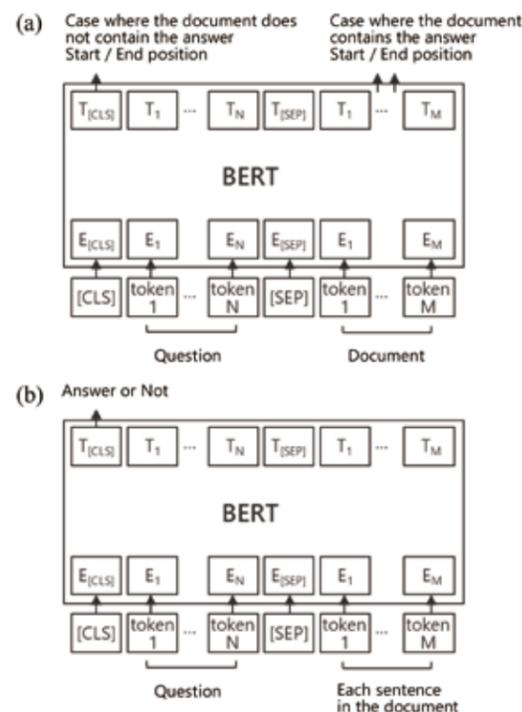


Fig. 2 Reader Model

- (a) Outputs the Start/End Position of the Answer
 (b) Outputs a Label Indicating Whether it is an Answer or Not

2.2 学習データ

Table 1 に学習に使用したデータ数を示す。データベース内の文章が回答となるような質問文を手で作成した。今回はデータベースの中でも特定の材料に関する QA セットを作成した。QA セットの 2 割を評価用として使用した。

Retriever 質問文を anchor、質問に対する回答を含む文書を positive、回答を含まない文書を negative とし、1 件の質問につき、10 件の triplet データを作成した。negative データは回答を含まない文書の中からランダムに 10 件抽出した。

Reader 位置抽出では、質問文、回答を含む文書、回答の開始位置および終了位置を 1 セットとしたデータを作成した。ラベル判定では、質問文、句点で句切った後の文およびその文が回答か否かのラベルを 1 セットとしたデータを作成した。

3. 評価方法

3.1 Retriever

評価用の質問文に対して、データベースから関連する文書を検索した。質問文のベクトルとデータベースの各文書の文書ベクトルとのコサイン類似度を用いて、類似度が高い順に文書をランキングした。上位 1 件または 10 件以内に質問に対する回答を含む文書があるか判定し、正解率で精度を評価した。1 件の質問につき複数件の正解文書が存在する場合もあるが、10 位以内の正解文書数やどの文書がより上位に来るかは考慮していない。

3.2 Reader

評価用の質問文に対して、3.1 の Retriever で類似度上位 10 件以内の回答を含む文書を用いた。位置抽出では、質問文と回答を含む文書を入力し、回答の開始／終了位置を得た。抽出部分に回答が含まれているかを判定し、正解率で精度を評価した。ラベル判定では、回答を含む文書を句点で句切り、質問文と分割後の各文を入力し、回答か否かのラベルおよびその確率を得た。回答である確率が最も高い文書中の 1 文が実際の回答と合っているかを判定し、正解率で精度を評価した。

4. 結果と考察

4.1 Retriever

Retriever の結果を Table 2 に示す。上位 1 件、10 件以内の正解率ともに Sentence BERT の精度が高くなった。BERT を用いた検索結果では、質問文中の重要な単語（材料名、加工方法など）を見逃している例も多かった。一方 Sentence BERT では、BERT での検索結果ほど見逃している例は少なく、質問文と文書で異なる表記で記載されていても類似度上位になる文書も確

認された。BERT で得た文書ベクトルは、追加学習後の各トークンの出力ベクトルの平均を取ったものであり、各トークンのベクトルは Masked Language Model に適したベクトルであるが、Pooling した場合に文書の特徴が得られるようなベクトルではないことが推測される。そのため、Sentence BERT でファインチューニングをした方が、似ている文書同士の Pooling したベクトルが類似するように各トークンのベクトルが出力されるため、正解率が向上したと考えられる。

Table 3 に Sentence BERT のランキング上位 10 件以内で、質問への回答を含まない文書例を示す。質問では“フェノール樹脂”に対する留意点を尋ねているが、異なる樹脂である“PBT”の留意点が文書には記載されていた。樹脂材料の留意点という観点では類似文書であるが、今回の質問回答においては、不適切な文書となる。人手で QA セットを作成した関係から訓練データが非常に少なく、樹脂の留意点としての類似文書の傾向は学習できたが、材料種の違いは学習できなかったと推測される。回答として不適切な文書を triplet の negative として加え訓練データを拡充することで、類似文書の中でも回答として適切な文書の傾向を学習でき、精度向上できると考えられる。

Table 2 Results of the Retriever

Model	Top 1 Accuracy	Top 10 Accuracy
BERT	0.059	0.235
Sentence BERT	0.294	0.764

Table 3 Example of a Document Not Containing an Answer Within the Top 10 Similarity Scores of Sentence BERT

Question	フェノール樹脂部品の留意点は？
Retriever	樹脂の劣化により各種特性が低下する。
Result	熱、水分、により PBT が劣化し、強度低下する可能性がある。...

4.2 Reader

Reader の結果を Table 4 に示す。4.1 の Sentence BERT のランキング上位 10 件以内で回答を含む文書 50 件に対して評価を行った。位置抽出およびラベル判定どちらも正解率はほぼ同等であった。しかし、位置抽出では入力した文書がそのまま出力されていた例が多く、必要な部分のみを抽出することができていなかった。ラベル判定による文単位での抽出では、正解した文書で回答でない部分を確実に除外することができ、平均して 50% 以上の文を削減することができた。回答として抽出した文は、文書からそのまま抽出しているため、質問には含まれていない製品、部品、工程などの固有名詞が含まれている例が見られた。質問回答の回答として提示する場合、該当箇所の抽出後に固有名詞を取り除くなどの後処理が必要になると考えられる。さらに、質問への回答として自然な形に文を成形する処理を加えることで、正確性を保った情報かつ適切な回答を提示できると考えられる。

Table 4 Results of the Reader

Model	Accuracy
Start / End Position	0.760
Label Classification	0.780

5. おわりに

本稿では、技術ナレッジ活用に向けた Retriever-Reader モデルの検証を行った。対象とするデータベースに表記ゆれが多いことから BERT を用いて Retriever および Reader のファインチューニングを実施した。Retriever では、質問文と質問への回答を含む文書のベクトルの類似度が高くなるよう Sentence BERT でファインチューニングすることで、正確率が向上した。Reader では、回答を含む文書を句点で各文に分割し、回答か否かのラベルの判定を行うことで質問に対する回答を適切に判断できることが分かった。

今後、Retriever では質問中の材料種の違いも考慮できるような訓練データセットの拡充、Reader では、質問への回答として不要な固有名詞の除外、回答として

自然な語尾への変換という取り組みが必要である。近年、大規模言語モデルの急速な発展により、Retrieverと生成型のGeneratorを組み合わせたRAG(Retrieval Augmented Generation)⁸⁾という手法が主流となってきている。この動向も踏まえながら、ユーザに正確かつ分かりやすい情報を提供するためのシステム構成の検討およびモデル改良に取り組んでいく。

参考文献

- 1) Kenton Lee, Ming-Wei Chang and Kristina Toutanova. Latent Retrieval for Weakly Supervised Open Domain Question Answering, In ACL, 6086–6096, 2019.
- 2) Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen and Wen-tau Yih. Dense Passage Retrieval for Open-Domain Question Answering, In EMNLP, 6769–6781. 2020.
- 3) S. Robertson and H. Zaragoza. The probabilistic relevance framework: Bm25 and beyond. Found. Trends Inf. Retr., Vol. 3, 333–389 2009.
- 4) Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, In NAACL, 4171–4186, 2019.
- 5) Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey and Noah A. Smith. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks, In ACL, 8342–8360. 2020.
- 6) Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks, In EMNLP-IJCNLP, Association for Computational Linguistics, 3982–3992. 2019.
- 7) Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD, In ACL, 784–789 2018.
- 8) Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela.: Retrieval-augmented generation for knowledge-intensive NLP tasks, NIPS'20: 34th International Conference on Neural Information Processing Systems, No. 793, 9459–9474 2020.

著者



蓬田 綾香
よもぎだ あやか

AI 研究部
自然言語処理や AI 技術の開発に従事



村瀬 文彦
むらせ ふみひこ

AI 研究部
自然言語処理や AI 技術の開発に従事



平野 徹
ひらの とおる

DENSO International America, Inc.
North America R&D 博士 (工学)
北米 R&D の技術統括。人工知能、
自然言語処理の研究開発に従事



三谷 陽
みたに あきら

AI 研究部 博士 (理学)
データ分析や AI 技術の開発に従事



坂 一忠
ばん かずただ

技術開発推進部
サーキュラーエコノミーにおけるリサイクル
材活用技術開発に従事



飯田 哲也
いいた てつや

材料技術部
Web システム開発に従事



岩堀 恵介
いわほり けいすけ

材料技術部 博士 (工学)
CAE を活用した DX 推進に従事



竹野 貴法
たけの たかのり

材料技術部 博士 (工学)
材料開発の DX 推進に従事