技術ナレッジ活用に向けたRetriever-Readerモデルの検証

蓬田綾香[1]　村瀬文彦[1]　平野徹[2]　三谷陽[1]
坂一忠[1]　飯田哲也[1]　岩堀恵介[1]　竹野貴法[1]
[1]株式会社デンソー
[2]DENSO INTERNATIONAL AMERICA, INC.
{ayaka.yomogida.j6n, fumihiko.murase.j6b, akira.mitani.j5g, kazutada.ban.j4x,
tetsuya.iida.j6e, keisuke.iwahori.j7p, takanori.takeno.j5x}@jp.denso.com
{toru.hirano}@na.denso.com

概要

社内に蓄積された技術ナレッジを効率的に活用することを目指して，膨大な文書からユーザが必要な回答を簡易に得られるように，質問応答タスクで用いられている Retriever-Reader モデルの検証を実施した．デンソー社内で実際に蓄えられた材料開発業務の文書を対象とした．Retriever では，BERT と Sentence BERT の各々のベクトルに基づくランキング上位の文書に，必要な回答が含まれる文書があるか否かを評価した．Reader では，文書から回答位置を抽出する手法と文書中の各文が回答か否かのラベルを判定する 2 つの手法を評価した．Retriever では Sentence BERT の正解率が高く，Reader ではラベルを判定する手法の方がユーザの質問に対する回答を適切に判断できることが分かった．

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，これらの大規模テキストデータに含まれるの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．具体的な事例を図 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章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 BM25[3]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 BERT[4]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，ユーザに誤った情報を提示することは避けたいため，情報の正

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

質問：[不良名]が発生した原因と対策は？
回答：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。

図 1 材料開発業務における具体的事例

確性を重視して抽出型を用いた.

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 (unidic-lite with whole word masking, jawiki-20200831) [i]に対して，対象データベースのテキスト(約 7 MB)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 に質問文を入力し, BERT の出力ベクトルの平均を用いて Pooling し，質問文のベクトルとした. データベースの各文書も同様に，出力ベクトルの平均を用いて Pooling し，各文書のベクトルとした．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 BERT[6][ii]を用いて質問文および文書のベクトルを得た．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.

$$\max\left(\left(s_a - s_p\right) - \left(s_a - s_n\right) + \varepsilon, 0\right) \qquad (1)$$

Reader　2 つの抽出手法を比較した．図 2 にそれぞれのモデルを示す. 1 つ目は位置抽出で，SQuAD 2.0 [7]に対する BERT の既存手法[4]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学習した．2 つ目はラベル判定で，質問文とともに抽出対象の文書の各文をモデルに入力し，質問文に対する回答か否か[CLS]を用いてラベルを判定させた．各文は，文章を句点で区切った.

2.2 学習データ

表 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人手で作成した．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 として，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.

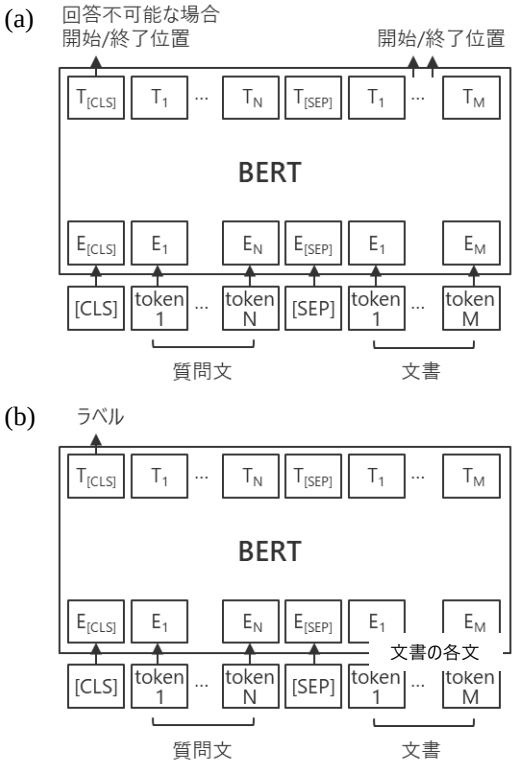

図 2 Reader モデル(a)開始/終了位置を出力 (b)回答か否かのラベルを出力

[i] https://huggingface.co/cl-tohoku/bert-base-japanese-v2

[ii] https://www.sbert.net/

表 1 文書数，QA セットおよび Retriever，Reader の訓練データ数

文書数	QA セット	Retriever	Reader	
		Triplet データ	位置抽出	ラベル判定
2985	93	760	76	350

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上位に来ても評価は変わらないこととした．

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読点で句切り，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 の結果を表 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．

表 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，今回の質問応答においては，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．

表 2 Retriever の結果

モデル	上位 1 件 正解率	上位 10 件以内 正解率
BERT	0.059	0.235
Sentence BERT	0.294	0.764

表 3 Sentence BERT の類似度上位 10 件以内で回答を含まない文書例

質問	**フェノール樹脂**部品の留意点は？
検索結果	樹脂の劣化により各種特性が低下する。熱、水分、により **PBT** が劣化し、強度低下する可能性がある。…

表 4 Reader の結果

モデル	正解率
位置抽出	0.760
ラベル判定	0.780

4.2 Reader

Reader の結果を表 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ており．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. 質問応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．

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．

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.