クローズドブック質問応答における言語モデルの知識強化

Yajima, Riho / 矢嶋, 梨穂

(出版者 / Publisher)
法政大学大学院情報科学研究科
(雑誌名 / Journal or Publication Title)
法政大学大学院紀要. 情報科学研究科編
(巻 / Volume)
19
(開始ページ / Start Page)
1
(終了ページ / End Page)
6
(発行年 / Year)
2024-03-24
(URL)
https://doi.org/10.15002/00030622

クローズドブック質問応答における言語モデルの知識強化
Enhancing Knowledge of Language Model in Closed-Book QA

矢嶋　梨穂
Riho Yajima
法政大学大学院情報科学研究科情報科学専攻
E-mail: 22t0020@cis.k.hosei.ac.jp

Abstract

Language models such as Chat-GPT generate new answers from the data they learn, improving work efficiency and generating ideas. On the other hand, incorrect answers may be generated, and this is likely due to the language model not being able to maintain accurate knowledge. In this research, we aim to provide language models with more accurate knowledge by devising a learning method rather than increasing the scale of the model. Specifically, we propose Title Prediction as an intermediate pretraining for Closed-book question answering (QA). Title Prediction learns how to output the title of an article from any two paragraphs on Wikipedia. By doing this, we hope to improve the relationship between separate paragraphs and embed accurate knowledge about titles into the model, even for paragraphs where titles are omitted. In my experiments, we confirmed that title prediction improves the performance of Closed-book question answering.

1. まえがき

近年，BERT や GPT などの Transformer 系モデルのパラメータ数は急激に増加しており，学習方法やモデル構造などの変更よりも，モデルを大規模化させることで，下流タスクの性能の向上を実現している．先行研究より，大規模言語モデルのパラメータ内に知識を暗黙的に保持していることが示されている[1]が，言語モデルが持つ知識を制御することは困難である．大規模言語モデルは，多くの GPU を必要とし，コストがかかるという問題もある．本研究では，言語モデルが正しい知識を保持していないと解けないタスクであるクローズドブック質問応答に着目し，学習方法を工夫することで，1 台の GPU でも言語モデル内部に正確な知識をより多く持たせることを目指す．

ニューラルネットワークを用いた質問応答では，機械読解やオープン検索質問応答などのタスクが研究行われている．機械読解[2][3]では，質問文と答えに関連する文書が与えられ，文書から質問文と関連度の高い文字列を答えとして抜き出す読解を行っている．オープン検索質問応答（ORQA: Open Domain Question Answering）[4][5]では，質問文のみが与えられ，答えに関連する文書は与えられない．従来分かれて行われていた 質問から答えに関連する文書の検索(Retriever)と 検索された文書から質問に関連する答えを抜き出す読解（Reader） の 2 つのステップが一連の流れで行われる．情報検索には BERT モデルを用いるため，より文脈に応じた文書が取り出される．読解は，機械読解と同様の手法で行われる．機械読解とオープン検索質問応答は共に，関連文書を参照して質問に答える読解を行うため，文書が知識の役割を担っていると捉えられる．そのため，モデルは文書の情報や知識を保持していなくても，文書を理解する読解能力があれば質問に答えることができる．一方，クローズドブック質問応答は，与えられた質問に対して，外部情報を参照しないで直接答えを求める．そのため，モデルは答えに関する正確な情報や知識を保持していること，蓄積された知識を質問応答で使えるようにすることが求められる．

クローズドブック質問応答に関する研究では，モデルの大規模化の他に，事前学習とクローズドブック質問応答のファインチューニングの橋渡しをする中間事前学習で性能が向上することが示されている．SSM （Salient span masking）[5][6] は固有名詞や日付などをマスクし，復元するように学習を行う．マスキング戦略学習 [7] は，中間事前学習でマスクすべき最適なスパン（=質問で問われる可能性の高いスパン）を抽出するマスキング方針を学習し，その後中間事前学習に学習したマスキング方針を展開することで，タスクに関連する知識を言語モデルのパラメータに詰め込む．BART モデルを使用した実験では，TriviaQA データセットにおいて，マスキング戦略で高い効果があることを示されている．

SSM，マスキング戦略学習を簡潔に説明すると，共に下流タスクのクローズドブック質問応答に効果的なトークンをマスクし，そのマスクされたトークンの予測をする，マスク予測を中間事前学習で行っている．マスク予測の中間事前学習は，BART の事前学習の Text Infilling タスクの形と近く，クローズドブック質問応答タスクとのギャップがまだ大きいのではないかと考える．さらに，モデルはマスクされたトークン付近の文脈に注目して予測している可能性が高い．よって，入力テキストがモデルの最大系列長に収まらない場合の分割されるコンテキ

Supervisor: Prof. Satoru Fujita

スト同士や離れた位置にあるトークン同士の関連性は弱いと考えられる．

言語モデルの事前学習データの一つに Wikipedia がある．Wikipedia の記事は基本的に節によって構成され，各節は複数の段落で構成されている．段落では，記事タイトルについての説明をしているが，タイトルである主語を省略して説明されたものは少なくない．タイトルが省略された段落では，何についての記事か不明確なまま，文法や文脈を学習している可能性がある．

本研究では，言語モデルにより多くの正確な知識を持たせるためのクローズドブック質問応答の中間事前学習として，Wikipedia の任意の 2 段落からその記事のタイトルを出力する，**タイトル予測（Title Prediction）**を提案する． 入力に任意の 2 段落を用いることで，離れた位置にある段落同士の関連性を高めること，出力を記事タイトルにすることで，Wikipedia のタイトルが省略された段落に対して，何についての記事か正しく学習できることを期待する．具体的には，事前学習モデルに BART 日本語 Pretrained モデルを使用し，中間事前学習で提案手法のタイトル予測の学習を行うことで，より正確な知識をモデルに埋め込む．そして，学習されたモデルをクローズドブック質問応答タスクでファインチューニングすることで，質問の答え方を学習させる．実験では，タイトル予測によってより多くの問題に対して，正しく解答できることを確認した．

2. クローズドブック質問応答

2.1 概要

マスキング戦略によるクローズドブック質問応答[7]に倣い，日本語の質問応答データを用い，日本語文書に対するクローズドブック質問応答モデルを構築する．具体的には，事前学習された BART モデルに対して，（質問，答え）のデータを使用して，ファインチューニングを行う．

2.2 BART モデル

BART（Bidirectional Auto-Regressive Transformer）は，Transformer ベースの Encoder-Decoder 型モデルである[8]．テキストをノイズ関数でノイズ化し，元のテキストを再構築する学習を行う．Text Infilling では，スパンの長さをポアソン分布で決め，各スパンを 1 つの<mask>に置き換える．これによって，任意の長さの文字列を予測することができるようになる．Sentence Permutation は複数の文からなる文書を分割し，文の順番をシャッフルする．Sentence Permutation と Text Infilling の 2 つのノイズ関数で学習された場合のパフォーマンスが最も高く，テキスト生成や理解タスクに有効である．

本研究では，事前学習モデルに BART 日本語 Pretrained モデル[1)]を使用する．入力テキストは日本語 Wikipedia 全て（約 1800 万文）である．入力テキストの前処理は，半角を全角に正規化し，Juman++（v2.0.0-rc3）で形態素に分割，さらに Sentence Piece でサブワードに分割したものを使用している．語彙数（サブワード数）は 32,000 である．BART-base モデルは 6 層の Encoder-Decoder で，隠れベクトルの次元数 768 次元で，パラメータ数は 140M である．

2.3 データセット

AI 王公式配布データセットは，日本語の質問応答を促進させることを目的としたクイズ問題を題材とした質問応答データセットである．本研究では，クローズドブック質問応答のデータセットに AI 王公式配布データセット Version 2.0[2)] 使用する．学習用データ（22,335 問）を分割した 21,218 問を学習用，1,117 問を検証用，開発用データ（1,000 問）をテスト用として使用する．データセットに含まれる“（正規化された）問題文”を入力の質問文，“（正規化された）正解のリスト”を出力の答えとして使用する．その他のデータ（クイズ問題の大会名，出題の基準となる日付，出題種別など）は使用しない．

3. 提案手法

3.1 タイトル予測（Title Prediction）

タイトル予測では，Wikipedia の任意の 2 段落からその記事のタイトルを予測するように学習する．段落全体からタイトルを予測するため，文脈理解に強くなると考えられる．BART の事前学習の Text Infilling や先行研究の中間事前学習のマスク予測は，マスクの前後の文脈に合うトークンを予測するというタスクになっており，文章全体というよりも局所的な文章理解や文章構成の学習に適していると思われる．一方，質問応答では，質問文の文脈全体から，中心となるキーワードを求めるようになっており，単純なトークン予測ではない．質問応答のデータで学習することで，質問応答の性能を上げることはできるが，学習データにはない質問に答えることは難しく，また，質問応答データの作成には問題作成のための知識やコストがかかる．大量の「文脈全体の予測」問題を提供できるデータとして，タイトル予測は適していると考える．

Wikipedia は，記事タイトルについて様々な視点の節からなる段落で説明している．各節のトピックは独立し，物語のような起承転結はあまりないため，節同士の順番はそれほど重要ではないと考える．任意の 2 段落を用いることで，記事タイトルに関する異なるトピック同士の関連性を高め，より構造的な知識を言語モデルに保持させることを期待する．

例えば，記事「ミシュランマン」の節は，“世界最大級のタイヤ会社”から始まり，“製造拠点”，“モータースポーツ”，“ビバンダム（ミシュランマン）について”，“ミシュランガイド”，“マネージング・パートナーの水死”と構成されている．この記事の構成では，節“世界最大級のタイヤ会社”と節“ミシュランガイド”の位置は離れている．タイトル予測の任意の 2 段落

1） https://huggingface.co/ku-nlp/bart-base-japanese

2） https://sites.google.com/view/project-aio/ competition2

によって，節“世界最大級のタイヤ会社”の段落と節“ミシュランガイド”の段落のペアを学習すれば，関連性を学習することができる．そして，AI 王の検証用データにある質問「1900 年に自動車での旅行を活性化するために始めたレストランのガイドブックが現在は世界的に広まっている，フランスのタイヤメーカーといえば何でしょう?」に答えることができると考える．クイズ問題は，上記の質問のように Wikipedia の異なる節の知識を必要とする質問もあり，そのような質問にタイトル予測は適していると考える．

タイトル予測の入力段落数の設定について，段落数が 1 の場合，テキストが短い段落からでは，情報が限られすぎてタイトルを予測するのが困難ではないかと考える．また，長い段落の場合，タイトル候補が多すぎて 1 つに絞れないのではないかと考える．1 段落ではなく，複数段落にすることで，それぞれの段落に共通するキーワード，つまりタイトルを予測できると考える．しかし，入力系列長が質問応答の質問文と比較して，長くなればなるほど，クローズドブック質問応答との形から離れてしまう．そのため，本研究では，入力段落数は 2 とする．

3.2 データ作成

3.2.1 Wiki40-B

Wikipedia の記事データには，Wiki-40B[3] の日本語データセットを使用する．Wiki-40B は，40 言語以上の Wikipedia 記事を前処理したデータセットである．前処理によって，Wikipedia に多く含まれる非コンテンツページを削除している．非コンテンツページとは，曖昧さ回避ページ（複数のトピックに関連付けられる，名前付きエンティティのリストで構成される記事），リダイレクトページ（同義語を使用したクエリを正規ページに自動送信するためのコンテンツが含まれない記事），削除されたページ，非エンティティページ（記事の大部分がリスト，情報ボックス，画像などで構成されるページ，例：日本の映画一覧）である．さらに，Wikipedia 記事のマークアップ言語で書かれたコンテンツのクリーンアップによって，関連項目や参考文献，画像キャプション，表，リストが削除されている．Wiki-40B の前処理済みのテキストには，ページタイトル，節，段落のマークアップが含まれている．タイトル予測では，このマークアップを利用して，記事タイトルと段落データを取り出したものを利用する．

3.2.2 タイトル予測データの前処理

Wiki-40B で取り出した記事タイトルと段落データに対して，さらに以下の処理を行う．

タイトルの正規化：同義語の記事タイトルは，曖昧さを回避するために括弧で補足されている．例えば，“マクベス（シェイクスピア）”や“マクベス（スコットランド王）”，“マクベス（交響詩）”がある．クローズドブック質問応答で用いる AI 王データセットの答えは，“マクベス”であり，括弧の補足は不要である．“マクベス”のように括弧を省略したタイトルを「正規化したタイトル」とする．クローズドブック質問応答とのギャップを小さくするために，タイトル予測では正規化したタイトルを予測させる．

段落の処理：箇条書きや改行のみなど，タイトルを予測するには情報が不十分な短い段落は，学習のノイズになってしまう．そのため，スペース（\u3000）を除いた文字数が 15 文字以下の段落は除外する．また，段落からタイトルを予測する際に，頻出単語や抜き出し問題になることを防ぐために，段落中に含まれるタイトルは<mask>に置き換える．

記事のフィルタリング：タイトル予測では，記事タイトルに関する異なるトピック同士や離れたコンテキスト同士の関連性を高めることを期待している．これらはテキストが長い記事の特徴である．テキストが短い記事に関しては，事前学習で十分に学習できていると考える．そのため，タイトル予測で対象とする記事データは，3 節以上 6 段落以上のものにする．タイトル予測で蓄積された知識を活用できるかを確認するのがクローズドブック質問応答であるが，クローズドブック質問応答で用いる AI 王データセット（学習用データ）では，答えの最大文字数は 29 文字である．すなわち，文字数が 30 文字以上の正規化したタイトルの記事を学習しても，クローズドブック質問応答で知識の有無を確認することができない．そのため，正規化したタイトルが 30 文字以上の記事もタイトル予測の学習データから除外する．

任意の 2 段落：タイトル予測の入力データの“任意の 2 段落”は，各記事の段落リストをシャッフルし，リストの先頭から 2 段落ずつ取り出したものとする．これを 2 回行い，2 段落の組み合わせが重複した場合は段落の順番を入れ替える．すなわち，1 つの段落に対して 2 通りの段落ペアを作成する．このように，段落ペアは機械的に作成をする．そのゆえ，作成された 2 段落のペアが必ずしも最適であるとは限らない．モデル入力時，2 段落の間に BART モデルの区切り文字“</s></s>”が挿入される．

3.2.3 データセット

Wik-40B は，記事毎に学習用（90%），検証用（5%），テスト用（5%）に分割されているが，この分割のままでは検証用とテスト用に充てられている記事をタイトル予測で学習することができない．そのため，Wiki-40B の学習用と検証用とテスト用の全ての記事を混ぜ，記事タイトルと段落データの前処理（**3.2.2**）を行い，シャッフルした後に，データを学習用（99%）とテスト用（1%）に分割したものをタイトル予測で使用する．記事フィルタリングによって，182,658 個の記事から，（入力，出力）＝（任意の 2 段落，正規化したタイトル）のデータが，学習用 2,982,668，テスト用 30,128 作成された．入力・出力テキストの前処理は BART の事前学習済みモデルと同様に，半角を全角に正規化し，Juman++で形態素に分割し，Sentence Piece でサブワードに分割する．

3) https://research.google/pubs/wiki-40b-multilingual-language-model-dataset/

表 1. 中間事前学習の学習データのイメージ

+ Random Mask **入力：**<s> 創設者の<mask> 兄弟<mask> いち早くモータリーゼーションの時代が到来することを確信し、同社の製品の宣伝をかねて自動車旅行者に有益な情報を提供するためのガイドブックとして、１９００年に3万5，０００部を無料で<mask> したのが<mask> ガイドの始まりである<mask> 二度<mask> 大戦<mask> を除いて毎年更新し、１９２０年から<mask><mask> なった。</s></s><mask> の起源は、１８６３年<mask> 設立されたＢａｒｂｉｅｒ，<unk> Ｄ<mask>ｕｂｒ<mask>ｅ<unk> ｅｔ<mask><unk> Ｃｉｅという株式合資会社<mask>。アンドレとエドゥアールの<mask> 兄弟の<mask> を冠したＭｉｃｈｅｌｉｎ<unk> ｅｔ<unk> Ｃｉｅとの名称は１８８９年に採用されたものである。１８９０年代<mask>、発明されたばかりの自動車用に耐える空気入りタイヤの実用化に取り組んで成功をおさめ、１９４９<mask> からは世界初の市販ラジアルタイヤ「<mask> Ｘ」を市販した歴史を持つ、世界<mask> タイヤ業界における老舗である。</s> **出力：**<s> 創設者のミシュラン兄弟がいち早くモータリーゼーションの時代が到来することを確信し、同社の製品の宣伝をかねて自動車旅行者に有益な情報を提供するためのガイドブックとして、１９００年に3万5，０００部を無料で配布したのがミシュランガイドの始まりである。二度の大戦中を除いて毎年更新し、１９２０年からは有料となった。</s></s> ミシュランの起源は、１８６３年に設立されたＢａｒｂｉｅｒ，<unk> Ｄａｕｂｒéｅ<unk> ｅｔ<unk> Ｃｉｅという株式合資会社である。アンドレとエドゥアールのミシュラン兄弟の名前を冠したＭｉｃｈｅｌｉｎ<unk> ｅｔ<unk> Ｃｉｅとの名称は１８８９年に採用されたものである。１８９０年代前期、発明されたばかりの自動車用に耐える空気入りタイヤの実用化に取り組んで成功をおさめ、１９４９年からは世界初の市販ラジアルタイヤ「ミシュラン Ｘ」を市販した歴史を持つ、世界のタイヤ業界における老舗である。</s>
+ Title Prediction **入力：**<s> 創設者の<mask> 兄弟がいち早くモータリーゼーションの時代が到来することを確信し、同社の製品の宣伝をかねて自動車旅行者に有益な情報を提供するためのガイドブックとして、１９００年に3万5，０００部を無料で配布したのが<mask> ガイドの始まりである。二度の大戦中を除いて毎年更新し、１９２０年からは有料となった。</s></s><mask> の起源は、１８６３年に設立されたＢａｒｂｉｅｒ，<unk> Ｄａｕｂｒéｅ<unk> ｅｔ<unk> Ｃｉｅという株式合資会社である。アンドレとエドゥアールの<mask> 兄弟の名前を冠したＭｉｃｈｅｌｉｎ<unk> ｅｔ<unk> Ｃｉｅとの名称は１８８９年に採用されたものである。１８９０年代前期、発明されたばかりの自動車用に耐える空気入りタイヤの実用化に取り組んで成功をおさめ、１９４９年からは世界初の市販ラジアルタイヤ「<mask> Ｘ」を市販した歴史を持つ、世界のタイヤ業界における老舗である。</s> 出力：<s>ミシュラン</s>
+ Random Mask &Title Prediction **入力：**<s> 創設者の<mask> 兄弟<mask> いち早くモータリーゼーションの時代が到来することを確信し、同社の製品の宣伝をかねて自動車旅行者に有益な情報を提供するためのガイドブックとして、１９００年に3万5，０００部を無料で<mask> したのが<mask> ガイドの始まりである<mask> 二度<mask> 大戦<mask> を除いて毎年更新し、１９２０年から<mask><mask> なった。</s></s><mask> の起源は、１８６３年<mask> 設立されたＢａｒｂｉｅｒ，<unk> Ｄ<mask>ｕｂｒ<mask>ｅ<unk> ｅｔ<mask><unk> Ｃｉｅという株式合資会社<mask>。アンドレとエドゥアールの<mask> 兄弟の<mask> を冠したＭｉｃｈｅｌｉｎ<unk> ｅｔ<unk> Ｃｉｅとの名称は１８８９年に採用されたものである。１８９０年代<mask>、発明されたばかりの自動車用に耐える空気入りタイヤの実用化に取り組んで成功をおさめ、１９４９<mask> からは世界初の市販ラジアルタイヤ「<mask> Ｘ」を市販した歴史を持つ、世界<mask> タイヤ業界における老舗である。</s> **出力：**<s>ミシュラン</s>

4. 実験

4.1 中間事前学習モデル

事前学習済みモデルにそのままクローズドブック質問応答のファインチューニングを行うモデルを Baseline モデルとし，Wikipedia の段落データを用いて，以下の 3 種類の中間事前学習を行い，その後クローズドブック質問応答でファインチューニングをする．表 1 に中間事前学習の具体的な入力・出力データのイメージを示す．

1. **Baseline**：中間事前学習を行わない．
2. **+ Random Mask (RM)**：中間事前学習でマスク予測を行う．入力は，各トークンが 10%の確率で動的マスクされている任意の 2 段落で，出力は，マスク部分を復元した 2 段落である．
3. **+ Title Prediction (TP)**：中間事前学習でタイトル予測を行う．入力は任意の 2 段落で，出力は正規化したタイトルである．
4. **+ Random Mask & Title Prediction (RM & TP)**：中間事前学習でマスクされた段落のタイトル予測を行う．入力は各トークンが 10%の確率で動的マスクされている任意の 2 段落で，出力は正規化したタイトルである．

4.2 詳細設定

1 台の GPU を用いる．言語モデルは大きいバッチサイズで学習すると精度が向上することが分かっているが[9]，大きいバッチサイズは GPU のメモリをより多く必要とするため，勾配累積（Gradient Accumulation）を用いる．勾配累積は，バッチ全体の勾配を一度に計算するのではなく，小さなバッチで勾配の反復計算を行う．バッチサイズが 16，勾配累積ステップ数が 32 の場合，メモリの使用量を節約しながら，バッチ数が 512 の場合と同等の計算を行うことができる．

中間事前学習のタイトル予測では，BART[8] や RoBERTa[9]モデルの事前学習のパラメータを参考に設定する．ただし，バッチサイズは BART 日本語 Pretrained モデルの事前学習時のバッチサイズの 512 に合わせるために，バッチサイズを 16，勾配累積ステップ数を 32 とする．最大学習ステップ数 300,000 で行い，50,000 ステップ毎にモデルを保存する．各モデルの学習には 30 時間以上要した．中間事前学習，クローズドブック質問応答のファインチューニングのハイパーパラメータの詳細は，表 2 と表 3 にそれぞれ示す．

モデルの出力をテキストに変換するデコード手法には，貪欲法とビームサーチをそれぞれ用いる．貪欲法は，各時刻で最も確率が高いトークンを選ぶ手法である．ビームサーチは，各時刻で確率の高い上位ビームサイズ分のトークンを記録し，終了記号の</s>を生成するまで繰り返す．そして，記録したトークンを対数確率によって，順位を付ける．ビームサイズが 1 のビームサーチは貪欲法と等しい．

表 2. 中間事前学習のハイパーパラメータ

Hyperparam	BART-base
Warmup Steps	10k
Peak Learning Rate	6.00E-04
Batch Size	16
Gradient Accumulation steps	32
Weight Decay	0.01
Train Samples	{50k, 100k, 150k, 200k, 250k, 300k}
Learning Rate Decay	linear
Gradient Clipping	0

表 3. クローズドブック質問応答のハイパーパラメータ

Hyperparam	BART-base
Warmup Steps	750
Peak Learning Rate	2.00E-04
Batch Size	16
Gradient Accumulation steps	16
Weight Decay	0.01
Max Epochs	100
Learning Rate Decay	cosine
Gradient Clipping	0

4.3 評価方法

BART 日本語 Pretrained モデルに対して，中間事前学習（**4.1**）を行い，Wikipedia 記事の知識をモデルに埋め込む．そして，AI 王データセットでクローズドブック質問応答のファインチューニングを行い，どれだけの質問に答えることができるかを評価することで，モデルが保持する知識量を確認する．クローズドブック質問応答の評価方法は完全一致用いる．モデルが予測した答えと AI 王データセットの答えが完全に一致した場合を正解とするが，表記揺れの影響を小さくするために，“・”，“=”，“\u3000”は無視して評価を行う．また，AI 王データセットのテスト用として使用したデータ（開発用，1000 問）には，別解データが含まれているため，別解データのどれかに当てはまれば正解とする．

4.4 学習サンプル数による精度の変化

中間事前学習で用いるデータ（任意の 2 段落，正規化されたタイトル）は，学習用データ数が約 3.0M と，数が多い．そこで，中間事前学習の学習サンプル数が，クローズドブック質問応答の精度に影響を与えるかを確認する．図 1 は，横軸が中間事前学習で学習した学習サンプル数，縦軸が AI 王データセット検証用の精度を表している．

グレーは，3 つの異なるシードでクローズドブック質問応答のファインチューニングをした Baseline モデルの平均値を表している．オレンジは Random Mask，緑は Title Prediction，青は Random Mask & Title Prediction で中間事前学習を行ったモデルで，中間事前学習とクローズドブック質問応答の学習はそれぞれ 1 つのシードで行っている．実線は，貪欲法 （greedy algorithm），点線はビームサーチ（beam search）のビームサイズ 10 の結果である．

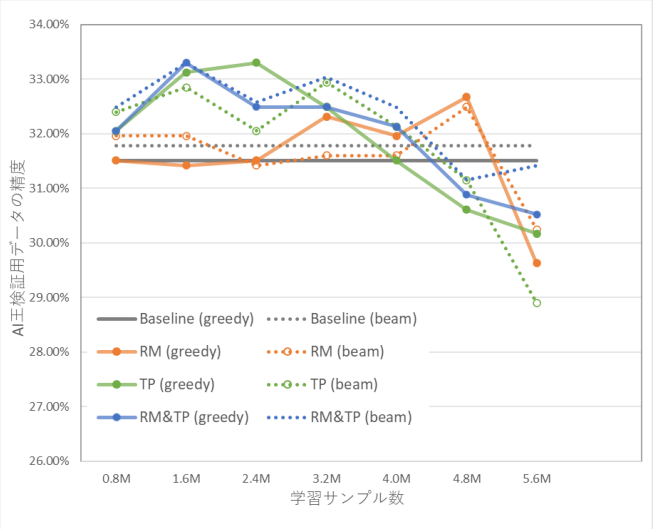

図 1. 学習サンプル数における精度の変化

表 4. AI 王データセットの実験結果

	Model		AI王	
		デコード戦略	検証用	テスト用
BART-base	Baseline	貪欲法	$31.51_{\pm.25}$	$33.10_{\pm.32}$
	Baseline	ビームサーチ (size 10)	$31.78_{\pm.14}$	$31.66_{\pm.16}$
	+ RM	貪欲法	$32.26_{\pm.24}$	$33.40_{\pm.20}$
	+ RM	ビームサーチ (size 10)	$31.15_{\pm.41}$	$32.33_{\pm.21}$
	+ TP	貪欲法	$33.06_{\pm.14}$	$34.56_{\pm.57}$
	+ TP	ビームサーチ (size 10)	$32.58_{\pm.45}$	$33.26_{\pm.55}$
	+ RM & TP	貪欲法	$33.12_{\pm.39}$	$34.63_{\pm.49}$
	+ RM & TP	ビームサーチ (size 10)	$33.24_{\pm.14}$	$33.90_{\pm.62}$

タイトル予測を行うモデル Title Prediction（TP）と Title Prediction & Random Mask（RM&TP）は，学習サンプル数が 1.6M から 3.2M のとき，Baseline に比べて，精度が高い傾向にある．また，学習サンプル数が 4.8M を超えると，Baseline よりも精度が落ちてしまう．いくつかの要因が考えられる．多くのデータを学習したことで，過学習が生じてしまったことが原因だと考えられる．学習の初期段階では，モデルが段落全体からタイトルを予測するように学習するため，質問文全体から答えを予測する質問応答の精度が向上したと考えられる．しかし，学習サンプル数が増えるにつれて，段落のペアの組み合わせが異なっていても，一方の段落を学習したことがあれば，その段落のタイトルを思い出せばタイトルを予測することができるというように，タイトル予測が文脈を読み取って予測する学習から段落のタイトルを思い出す暗記タスクに変換してしまったのではないかと考える．

4.5 評価結果

学習サンプル数による精度 （4.4）の結果を踏まえて，学習サンプル数が 3.0M（学習ステップ数 187,500）で中間事前学習を行ったモデルに対して，3 つの異なるシードでクローズドブック質問応答のファインチューニングを行ったモデルの精度平均と標準偏差を表 4 に示す．検証用データに比べて，テスト用データの精度が高いのは，

テスト用データでは，別解リストを用いて評価を行っているからである．クローズドブック質問応答のファインチューニングの学習設定は全て同じであるが，中間事前学習でタイトル予測を行ったモデル（TP と RM&TP）の精度が高いことが分かる．Random Mask で中間事前学習を行ったモデルは，Baseline に比べわずかに高いことから，学習データで用いた Wikipedia の任意の 2 段落のデータ構造が，クローズドブック質問応答の手助けをしたことがわかる．

5. 考察

表 4 に示したタイトル予測の中間事前学習をおこなったモデル（TP, RM&TP）は，答えが「ミシュラン」の質問（**3.1**）に正しく回答することができた．タイトル予測の学習を行わない Baseline のモデルは，学習時のシードの値によって，答えが「ミシュラン」の質問が正解になることもある．Baseline の異なるシードによる 3 つのモデルの全てで不正解であり，タイトル予測によって，新たに正解になる質問はみられなかった．それでも，表 4 より，タイトル予測を行ったモデルの精度は Baseline よりも向上している．このことから以下の要因が考えられる． 本研究で使用した事前学習済みモデルは，Wikipedia の記事全文をマスクの穴埋めである Text Infilling で学習したモデルである．それゆえ，事前学習によって言語モデルが Wikipedia の記事に関して，ある程度局所的な知識を保持していたと考えられる．本研究の中間事前学習のタイトル予測によって，Wikipedia の段落の文脈理解を行った．これにより，Wikipedia の記事に関する新たな情報や知識をモデルに持たせるというよりも，Baseline モデルで不正解になりやすい知識の補強を行ったと考えられる．その結果，事前学習と同じ Wikipedia の記事データでも学習方法を変えることで，Baseline モデルに比べて，クローズドブック質問応答の性能が向上することを確認した．

また，Wikipedia のテキストからでは，答えるための根拠が不十分な質問には答えることができなかった．その観点からも，Baseline モデルで答えられなかった問題を新たに答えられるようにするためには，Wikipedia 以外の文書を学習データに与え，モデルを学習させる必要があると考える．

6. むすび

本研究では，事前学習とクローズドブック質問応答のファインチューニングの間で行う中間事前学習として，タイトル予測のデータセット作成と学習を行った．実験では，タイトル予測の学習を行った場合に，クローズドブック質問応答でより多くの質問に答えられることを確認した．このことから，タイトル予測の学習によって，モデルが正しく活用できる知識が増えたと言える．また，タイトル予測の学習を過剰にすると，タイトル予測の過学習によって，中間事前学習を行わないモデルよりクローズドブック質問応答の精度が低下することがわかった．

タイトル予測は Wikipedia だけではなく，百科事典や国語辞典のような辞書でも，用法，語法，用例からその単語を予測するという学習によって，応用することができると考える．Wikipedia 以外の文書データを学習することで，新たな情報や知識をモデルに持たせることができると考える．

文　献

[1] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller, “Language Models as Knowledge Bases?”, In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

[2] Pranav Rajpurkar, Jian Zhang and Konstantin Lopyrev, Percy Liang, “SQuAD: 100,000+ Questions for Machine Comprehension of Text”, arXiv:1606.05250v3 [cs.CL] 11 Oct 2016.

[3] Pranav Rajpurkar, Robin Jia, Percy Liang, “Know What You Don’t Know: Unanswerable Questions for SQuAD”, arXiv:1806.03822v1 [cs.CL] 11 Jun 2018.

[4] Kenton Lee, Ming-Wei Chang, Kristina Toutanova, “Latent Retrieval for Weakly Supervised Open Domain Question Answering”, arXiv:1906.00300v3 [cs.CL] 27 Jun 2019.

[5] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat , Ming-Wei Chang, “REALM: Retrieval-Augmented Language Model Pre-Training”, arXiv:2002.08909v1 [cs.CL] 10 Feb 2020.

[6] Adam Roberts, Colin Raffel,, Noam Shazeer, “How Much Knowledge Can You Pack Into the Parameters of a Language Model?”, arXiv:2002.08910v4 [cs.CL] 5 Oct 2020.

[7] Qinyuan Ye, Belinda Z. Li, Sinong Wang, Benjamin Bolte, Hao Ma, Wen-tau Yih, Xiang Ren, Madian Khabsa, “Studying Strategically: Learning to Mask for Closed-book QA”, arXiv:2012.15856v2 [cs.CL] 1 Jan 2021.

[8] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer, “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension”, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871–7880 July 5 - 10, 2020.

[9] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov, “RoBERTa: A Robustly Optimized BERT Pretraining Approach”, arXiv:1907.11692v1 [cs.CL] 26 Jul 2019.

発表リスト

1. 矢嶋梨穂，藤田悟，“構造化知識を内包する自然言語理解システムの構築”，情報処理学会　第 85 回全国大会，2023.3.
2. 矢嶋梨穂，藤田悟，“タイトル予測を用いたクローズドブック質問応答の知識強化”，電子情報通信学会　総合大会，2024.3（発表予定）．