ファクトイド質問応答における BERT の pre-trained モデルの影響の分析

関 直哉 [†‡] 水野 淳太 [‡] 門脇 一真 [‡*] 飯田 龍 [†‡] 鳥澤 健太郎 [†‡]
[†] 奈良先端科学技術大学院大学 先端科学技術研究科 情報科学領域
[‡] 国立研究開発法人 情報通信研究機構 (NICT)
[*] 株式会社日本総合研究所
seki.naoya.si8@is.naist.jp
{junta-m,kadowaki,ryu.iida,torisawa}@nict.go.jp

1 はじめに

本研究では，汎用言語モデル BERT の事前学習に用いるコーパスの種類や規模，事前学習時のバッチサイズが fine-tuning 後の性能にどのように影響するかをコーパスの種類・規模とバッチサイズの組み合わせが異なる 13 種類の事前学習モデルを用いて調査する．

調査対象としては日本語のファクトイド質問応答を対象とする．ただし，この評価を行うための大規模なアノテーション済みデータが存在しないため，我々は大規模情報分析システム WISDOM X[1] [1, 2] で収集されたファクトイド質問応答の回答候補に対してアノテーションを行い，約 22 万件の学習・評価用データを作成した．このデータを用いて評価実験を行い，各事前学習モデルを fine-tuning して結果を比較した．実験の結果，$\text{BERT}_{\text{BASE}}$ のモデルに関してはコーパスの種類を変更し，コーパスの規模を 2,000 万文から 22 億文に，バッチサイズを 50 から 4,096 に大きくすることで，F1 スコアが 70.36%から 77.09%に向上した．さらに，モデルサイズを $\text{BERT}_{\text{BASE}}$ から $\text{BERT}_{\text{LARGE}}$ に変更 (ただし，コーパスサイズは 12 億文) することで，F1 スコアが 78.76%まで向上することを確認した．

2 BERT の事前学習

Devlin ら [3] が提案した BERT は，大規模なテキストコーパスを用いて事前学習を行った後，タスクごとに fine-tuning する汎用言語モデルである．含意関係判定，感情分析，質問応答などの自然言語処理のタスクおいて，BERT は高い性能を発揮することが知られている．本研究では，表 1 に示す BERT のモデルサイズ，コーパス，バッチサイズの組み合わせを用いる．モデルサイズは Devlin ら [3] の $\text{BERT}_{\text{BASE}}$ と $\text{BERT}_{\text{LARGE}}$ を用いる．

[1] https://www.wisdom-nict.jp/

2.1 事前学習用のコーパス

本研究では，以下の 3 種類のコーパスを BERT の事前学習に利用する．Web ランダムコーパスと因果関係文コーパスの構築には NICT が収集した Web 文書を利用した．また，コーパスの単語分割には MeCab [4] を用いた（MeCab の辞書には Juman [5] を使用した）．

Wikipedia コーパス 2018 年 8 月時点の日本語版 Wikipedia のダンプデータを利用する．前処理として，記事のタイトルを取り除き，記事の本文のみを用いる．前処理後の文数は約 2,000 万文であり，今回はそれらを全て利用した（以下，Wikipedia2,000 万文と呼ぶ）．

Web ランダムコーパス Web40 億ページ中のテキストデータを利用したコーパスだが，コーパスの規模を Wikipedia コーパスと揃えるために約 2,000 万文を使用する（以下，Web ランダム 2,000 万文と呼ぶ）．

因果関係文コーパス 因果関係文コーパスは，Web ページ中の 7 文からなるテキストパッセージの集合であり，各パッセージは Oh らの因果関係検出器 [6] により自動検出された因果関係を少なくとも 1 つ含んでいる．Kadowaki ら [7] と同様に，$\text{BERT}_{\text{BASE}}$ の事前学習には Wikipedia コーパスと規模を揃えた 2,000 万文からなるコーパス（以下，因果関係 2,000 万文）と 4 億文からなるコーパス（以下，因果関係 4 億文），12 億文からなるコーパス（以下，因果関係 12 億文），22 億文からなるコーパス（以下，因果関係 22 億文）を，$\text{BERT}_{\text{LARGE}}$ の事前学習には因果関係 12 億文を用いる．

大規模なコーパスを用いた事前学習モデルには BERT の他に XLNet [8] や，RoBERTa [9]，T5 [10] などがあるが，それらのモデルの事前学習で用いられたコーパスと本研究で用いたコーパスの規模をバイト

表 1: 事前学習に用いるコーパスの種類・規模とバッチサイズの組み合わせ

モデルサイズ	$\text{BERT}_{\text{BASE}}$						$\text{BERT}_{\text{LARGE}}$
	コーパス						
バッチサイズ	Web ランダム 2,000 万文	Wikipedia 2,000 万文	因果関係 2,000 万文	因果関係 4 億文	因果関係 12 億文	因果関係 22 億文	因果関係 12 億文
50	○	○	○	-	-	-	-
256	-	○	○	○	-	-	-
1,024	-	○	○	○	-	-	-
4,096	-	-	-	○	○	○	○

表 2: 事前学習に用いられるコーパスのバイト数

コーパス（モデル）	バイト数
BERT [3]	16GB
XLNet	158GB
RoBERTa	160GB
T5	750GB
Web ランダム 2,000 万文	2.5GB
Wikipedia2,000 万文	3.4GB
因果関係 2,000 万文	2.9GB
因果関係 4 億文	51GB
因果関係 12 億文	165GB
因果関係 22 億文	353GB

数で比較した結果を表 2 にまとめる．我々が使用した因果関係 22 億文は T5 で使用されたコーパスと比較して小さいものの，他のコーパスよりもバイト数で規模が大きい．また，Wikipedia2,000 万文と比較して約 100 倍もの規模があることがわかる．

2.2 パラメータ設定

事前学習のバッチサイズは 50, 256, 1,024, 4,096 を使用する（各コーパスの事前学習で使用したバッチサイズは表 1 を参照）．また，共通の設定として二段階に分けて事前学習を行った[2]．

3 ファクトイド質問応答

ファクトイド質問とは「誰・何・どこ」等の疑問代名詞を含む質問（例：地球温暖化は何を引き起こすか？）であり，これらの質問に対する回答は「異常気象」や「海面上昇」といった名詞である（「なぜ」,「どうやって」,「どうなる」をともなう質問はファクトイド質問には含まれない）．本研究では，ファクトイド質問応答を質問と回答候補，回答候補を抽出した元文（以下，元文と呼ぶ）の 3 つ組が与えられた際，回答候補が適切な回答か否かを分類する二値分類のタスクとして扱う．例えば，質問「地球温暖化は何を引き起こす」，回答候補「海水温上昇」，元文「地球温暖化は海水温上昇を引き起こすと言われています。」という 3 つ組が与えられた際，元文を参照して回答候補が事実として述べられているため「海水温上昇」を正解とみなす．

以降で，ファクトイド質問応答のデータセットを作成するのに利用した WISDOM X [1,2] の概要を紹介し，次にデータ作成の詳細を説明する．

3.1 WISDOM X

大規模情報分析システム WISDOM X [1,2] は，ユーザから質問が与えられた際に Web ページから回答候補を検索し，回答候補と回答候補を含む元文を提示する．WISDOM X はファクトイド質問,「なぜ地球温暖化が起こる」といったなぜ型の質問,「地球温暖化が進むとどうなる」といったどうなる型の質問など複数の形式の質問に対する回答をユーザに提供する．

このうち，ファクトイド質問の回答候補の検索ではバイナリパタンとユナリパタンと呼ばれる 2 種類のパタンを使ってテキストと照合を行うことで回答候補の検索を行っている．バイナリパタン [11] とは，係り受けで繋がっている 2 つの変数化した名詞をもつ「A は B を引き起こす」のようなパタンを指す．ユナリパタンは「A を引き起こす」のような形式で表されるパタンを指し，A には名詞が当てはまる．WISDOM X では，NICT が独自にクロールした Web40 億文書に対し，形態素解析などの処理を行い，バイナリパタンとユナリパタンを抽出し，検索用インデックスを構築する．質問「地球温暖化は何を引き起こす」が与えられた際，そこから抽出されるバイナリパタン「A が B を引き起こす」とその中の A もしくは B に当てはまる名詞 (ここでは「地球温暖化」) をこのインデックスに照合して回答候補を検索する．また，柔軟な照合を行うために，質問から得たバイナリパタンだけでなく，バイナリパタンと含意関係にある他のバイナリパタン [11] も利用する．例えば,「A が B を引き起こす」についてはそのバイナリパタンと含意関係にある「A で B が発生する」や「B をもたらす A」も用いて回答候補を検索する.「地球温暖化は海で何を引き起こす」のように質問に名詞が複数個含まれる場合には,（1）パタン「地球温暖化は B を引き起こす」と照合し，かつ「海」を含む文（例：地球温暖化は気候変動を引き起こし，世界中の海で水温が上昇している），（2）パタン「海で B を引き起こす」と照合し，かつ「地球温暖化」を含む文（例：地球温暖化が進行し，海で台風を引き起こす），のように照合する条件を変更して回答候補を検索する．バイナリパタンで回答候補が得ら

[2] 一回目はステップ数=100 万，length=128，二回目はステップ数=10 万，length=512 で学習する．GPU は，バッチサイズ 50 では Tesla P100 を 1 枚，それ以外のバッチサイズでは Tesla V100 を複数枚（DataParallel で学習，最大で 64 枚）用いた．$\text{BERT}_{\text{LARGE}}$ の学習は Tesla V100 を一回目の事前学習で 64 枚，二回目で 128 枚使用した．$\text{BERT}_{\text{LARGE}}$ の事前学習には約 707 時間を要した．

表 3: データセットの統計値

	データ件数	正例件数（割合）
学習	174,765	56,358(32.2%)
開発	10,881	3,487(32.0%)
テスト	32,477	10,851(33.4%)

れない場合は，ユナリパタンを利用して回答候補を検索する．ユナリパタンでの検索は，質問中の名詞「地球温暖化」を含み，かつパタンに照合する文を回答を含む文の候補とする．上記の例では「地球温暖化が進み、海水温上昇を引き起こす」のような文から回答候補「海水温上昇」を得ることができる．バイナリパタンやユナリパタンを用いることで単純に名詞等のキーワードで検索するよりも回答候補を適切に絞り込むことができ，高速に検索することが可能となる．実際の検索では，指定した数の回答候補数を検索し終わるか，設定した検索時間の上限に達するまで検索が行われる．

また，WISDOM X は質問に答えるだけでなく，ユーザに質問を提案する質問提案機能を備えており，この機能によってトピックワードとなる名詞（例 : 地球温暖化）が与えられた際に，トピックワードに関連する質問（例 : 地球温暖化を何で防ぐ）の一覧を生成することが可能である．

3.2 データセット

本研究で利用するファクトイド質問応答データセットは Asao ら [12] が既にアノテーション済みのデータと，新規に WISDOM X のファクトイド質問応答の出力に対してアノテーションしたデータからなる．WISDOM X を用いたデータの収集では，(1) アノテータが考えたファクトイド質問を WISDOM X に入力し，回答候補と元文を得る方法，(2) アノテータが考えたトピックワードを WISDOM X に入力し，質問提案機能を用いて質問集合を得た後で，さらにそこから質問を選択してその質問を WISDOM X に入力することで回答候補と元文を得る方法の 2 通りを併用した．アノテータは質問と回答候補，元文の 3 つ組を参照し，質問に対して回答候補が元文中で事実として述べられているならその回答候補を適切な回答，それ以外は不適切な回答としてアノテーションを行った．

表 3 に Asao らのデータと WISDOM X を用いて作成したデータを統合し，学習用，開発用，テスト用に分割したデータの事例数を示す．

3.3 BERT を用いた fine-tuning

fine-tuning の際は，ファクトイド質問応答のデータを [SEP] トークンを用いて「⟨ 質問の単語列 ⟩[SEP]⟨ 回答候補の単語列 ⟩[SEP]⟨ 元文の単語列 ⟩」のフォーマットで BERT に入力する．例えば，質問「地球温暖化は何を引き起こす」，回答候補「海水温上昇」，元文「地球温暖化は海水温上昇を引き起こすと言われています。」という 3 つ組に対して，「地球 温暖 化 は 何 を 引き起こす [SEP] 海水 面 上昇 [SEP] 地球 温暖 化 は 海水 面 上昇 を 引き起こす と 言わ れて い ます 。」が BERT の入力として与えられる．

4 実験

表 1 に示した 13 種類の事前学習モデルをファクトイド質問応答の学習データでそれぞれ fine-tuning し，fine-tuning 後のモデルでファクトイド質問応答の分類性能の評価を行った．

4.1 設定

fine-tuning 時のバッチサイズは 32 とした．また，学習率 1e-5，2e-5，3e-5，4e-5，5e-5，エポック数 1，2，3 のすべての組み合わせに対し個別に fine-tuning を行い，開発データでも最も F1 スコアが高いモデルをテストデータに適用した．

4.2 実験結果

各事前学習モデルを fine-tuning して作成したモデルをそれぞれテストデータに適用して得られた F1 スコアを表 4 にまとめる．この結果より，まず $\text{BERT}_{\text{LARGE}}$ を用いたモデルが最も良い結果を得ており，$\text{BERT}_{\text{BASE}}$ を用いたモデルで得られた全ての結果に対して有意に良い結果を得ていることがわかる (マクネマー検定，有意水準 5%で検定)．また，コーパスの規模を 2,000 万，バッチサイズを 50 に固定して，コーパスの種類のみを変更した場合の結果を比較すると因果関係文コーパスが他のコーパスを使用した場合よりも有意に良い結果を得ている．この理由としては，Web ランダムコーパスでは決まりきった宣伝文句等の有用でない文が多く，また，Wikpedia の記事は百科事典的な記述に偏っているため，相対的に因果関係文から事前学習したモデルが今回対象としたファクトイド質問応答のデータと相性が良かったと考えられる．さらに因果関係文コーパスを用い，バッチサイズとコーパスの規模を変更して得られた事前学習モデル間を比較すると，バッチサイズと規模を増やすことで 72.44%から 77.09%まで約 4.7%向上しており，大規模に事前学習することの効果を確認することができた．$\text{BERT}_{\text{LARGE}}$ をさらに大規模化することで性能向上が期待できるが，それについては今後の課題としたい．

5 回答の検索手法の分析

BERT 等の事前学習モデルの発展により分類タスクの性能が向上していくことで多くの処理がニューラルモデルで刷新されていく可能性があるが，一方でニューラルモデルはモデルの規模が大きくなるほど処理に時間がかかるために処理速度が問題となり，大量の候補が検索されても正確な回答をタイムリーに提供することが難しくなる．これはつまり，深層学習の

表 4: 各 pre-trained モデルに対する二値分類型のテストデータの F1 スコア

モデルサイズ	$\text{BERT}_{\text{BASE}}$						$\text{BERT}_{\text{LARGE}}$
	コーパス						
バッチサイズ	Web ランダム 2,000 万文	Wikipedia 2,000 万文	因果関係 2,000 万文	因果関係 4 億文	因果関係 12 億文	因果関係 22 億文	因果関係 12 億文
50	$70.54^{\dagger\ddagger*}$	$70.36^{\dagger\ddagger*}$	$72.44^{\dagger\ddagger}$	-	-	-	-
256	-	$72.01^{\dagger\ddagger}$	$73.41^{\dagger\ddagger}$	$75.80^{\dagger\ddagger}$	-	-	-
1,024	-	$73.66^{\dagger\ddagger}$	$74.09^{\dagger\ddagger}$	$75.64^{\dagger\ddagger}$	-	-	-
4,096	-	-	-	$76.92^{\dagger}$	$76.80^{\dagger}$	$77.09^{\dagger}$	**78.76**

† は因果関係 12 億文コーパスで事前学習した $\text{BERT}_{\text{LARGE}}$ を fine-tuning して得られたモデルの結果を他の結果と比較して，マクネマー検定 (有意水準 5%) で有意差があることを示す．‡ は因果関係 22 億文コーパスで事前学習した $\text{BERT}_{\text{BASE}}$ を fine-tuning して得られたモデルの結果を他の結果と比較して，マクネマー検定 (有意水準 5%) で有意差があることを示す．* はコーパスの規模を 2,000 万，バッチサイズを 50 に固定した場合に，因果関係コーパスで事前学習した $\text{BERT}_{\text{BASE}}$ を fine-tuning して得られたモデルの結果を Web ランダムコーパス，Wikipedia コーパスの結果と比較して，マクネマー検定 (有意水準 5%) で有意差があることを示す．

活用が前提であっても，事前に回答候補を効率的に絞り込むことが必要となるということである．本節では WISDOM X で使用しているパタンによるファクトイド質問応答の回答候補の絞り込みが十分な量の適切な回答を得ることにどのように影響するかを調査する．パタンによる絞り込みを行わない場合の手法としてはファクトイド質問中の名詞を用いて回答の元文を検索する手法を採用する．表 3 の開発データ 10,881 件中の質問内の名詞のみで WISDOM X で使用している Web テキスト集合から検索を行い，そのうち 10,549 件 (全体の 97%) に対して元文を得た (各質問について平均で約 14 文を収集)．得られた元文と質問の対からランダムに 500 事例をサンプリングし，その事例に対して適切な回答を含むか否かをアノテーションしたところ，35 件 (全体の約 7%) のみが回答を含むことがわかった．一方，サンプリングした 500 件中の質問が開発データ内でもともと対応づけられている事例に関する正例 (適切な回答を含む) の個数は 155 件 (全体の 31%) であり，名詞のみで元文を検索した場合の約 4 倍である．つまり，同じ分量の事例を BERT 等で分類することを考えた場合，パタンによる絞り込みも併用することで約 4 倍の効率で適切な回答を得られることがわかる．一方で，パタンの絞り込みによって正解がフィルターされてしまう可能性もあるが，その影響の調査に関しては今後の課題とする．

6 おわりに

本研究では，日本語ファクトイド質問応答を対象に BERT の事前学習に利用するコーパスの種類，規模，バッチサイズがどのように性能に影響するかの調査を行った．この結果，コーパスの種類としては自動検出した因果関係を含む文とその前後文からなるテキストを事前学習用のコーパスとして利用した場合に最も性能が向上し，また，コーパスの規模とバッチサイズはそれらを増加させるごとに性能が向上することを確認した．さらに，モデルサイズが $\text{BERT}_{\text{LARGE}}$ である事前学習のモデルも構築し，そのモデルを fine-tuning したモデルを利用することで本研究で使用したファクトイド質問応答の評価データに対して F1 スコアで 78.76% という性能を得た．また，ファクトイド質問応答の検索については「A は B を引き起こす」といったパタンを用いた検索を BERT 等の分類モデルと併用することで適切な回答を効率的に検索できることについても示した．

参考文献

[1] Masahiro Tanaka, Stijn De Saeger, Kiyonori Ohtake, Chikara Hashimoto, Makoto Hijiya, Hideaki Fujii, and Kentaro Torisawa. WISDOM2013: A large-scale web information analysis system. In *Proc. of IJCNLP: System Demonstrations*, pp. 45–48, 2013.

[2] Junta Mizuno, Masahiro Tanaka, Kiyonori Ohtake, Jong-Hoon Oh, Julien Kloetzer, Chikara Hashimoto, and Kentaro Torisawa. WISDOM X, DISAANA and D-SUMM: Large-scale NLP systems for analyzing textual big data. In *Proc. of COLING*, pp. 263–267, 2016.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL:HLT*, pp. 4171–4186, 2019.

[4] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In *Proc. of EMNLP*, Vol. 4, pp. 230–237, 2004.

[5] Sadao Kurohashi, Toshihisa Nakamura, Yuji Matsumoto, and Makoto Nagao. Improvements of Japanese morphological analyzer JUMAN. In *Proc. of The International Workshop on Sharable Natural Language*, pp. 22–28, 1994.

[6] Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Motoki Sano, Stijn De Saeger, and Kiyonori Ohtake. Why-question answering using intra- and inter-sentential causal relations. In *Proceedings of ACL*, pp. 1733–1743, 2013.

[7] Kazuma Kadowaki, Ryu Iida, Kentaro Torisawa, Jong-Hoon Oh, and Julien Kloetzer. Event causality recognition exploiting multiple annotators' judgments and background knowledge. In *Proc. of EMNLP-IJCNLP*, pp. 5815–5821, 2019.

[8] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, and Quoc V Le. XLNet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pp. 5754–5764. 2019.

[9] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv e-prints*, arXiv:1907.11692, 2019.

[10] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv e-prints*, arXiv:1910.10683, 2019.

[11] Julien Kloetzer, Stijn De Saeger, Kentaro Torisawa, Chikara Hashimoto, Jong-Hoon Oh, Motoki Sano, and Kiyonori Ohtake. Two-stage method for large-scale acquisition of contradiction pattern pairs using entailment. In *Proc. of EMNLP*, pp. 693–703, 2013.

[12] Yoshihiko Asao, Ryu Iida, and Kentaro Torisawa. Annotating zero anaphora for question answering. In *Proc. of LREC*, pp. 3523–3528, 2018.