BERTを用いた英文空所補充問題の一解法

三木　一弘† 　太田　学††

† 岡山大学工学部情報系学科　〒700–8530 岡山県岡山市北区津島中三丁目1番1号
†† 岡山大学大学院自然科学研究科　〒700–8530 岡山県岡山市北区津島中三丁目1番1号
E-mail: †pb7p3yvg@s.okayama-u.ac.jp, †† ohta@cs.okayama-u.ac.jp

あらまし　英文空所補充問題は，空所を含む英文に対して，空所にあてはまる語句を回答する問題であり，機械学習のモデルの文脈理解度の評価に用いられる．近年，英文空所補充問題のための大規模データセットが公開されたことで，様々なモデルが提案されている．一方，Bidirectional Encoder Representations from Transformers (BERT) は様々な自然言語処理タスクにおいてState-of-the-art (SOTA) を達成している汎用言語表現モデルである．本稿では，選択肢付き英文空所補充問題に対して，BERTのMasked Language Model (MLM) タスクを利用した解法と，それを中国の高校入試および大学入試問題から構成されるCLOTHデータセットでfine-tuningして用いる解法を提案する．MLMタスクを利用した手法は，国内の大学入試センター試験問題に対する正答率が92.3%であり，比較したkenLMの正答率である40.7%を大きく上回った．また，BERTをCLOTHデータセットでfine-tuningしたモデルはテスト用CLOTHデータセットの高校入試問題に対する正答率が88.8%であり，KenLMの正答率58.8%を大きく上回った．

キーワード　自然言語処理，BERT，英文空所補充

1 はじめに

空所補充問題は一部が空所となっている文章に対して，空所に当てはまる語句や文を予測するタスクであり，機械学習のモデルの評価に用いられる代表的なタスクの一つである．特に近年では，英文空所補充問題のための大規模データセットが公開されたことで，深層学習等を利用した様々なモデルが提案されている．

英文空所補充問題のための代表的な大規模データセットには，CNN/Daily Mail データセット [1]，Story Cloze Test データセット [2]，CLOTH データセット [3] などが挙げられ，それぞれ問題の形式が異なる．CNN/Daily Mail データセットはニュース記事と，その要約であるクエリのペアからなる．すなわち，クエリの一部が空所となっており，当てはまる語句を予測する．Story Cloze Test データセットは 4 文から構成される文章と，その後に続く後続文の選択肢が 2 つ与えられ，正しい後続文を予測する問題である．Story Cloze Test データセットは LSDSem 2017 Shared Task [4] における英文空所補充問題のデータセットとしても採用された．CLOTH データセットは複数の空所を含むフレーズがあり，空所ごとにそれぞれ 4 つの選択肢が与えられ，空所に当てはまる語句を予測する問題である．CLOTH データセットは中国の高校入試や大学入試における英語科目の長文問題をもとに作成されている．

本稿で取り扱う英文空所補充問題は，CLOTH データセットのような文章中の空所に対して当てはまる語句を選択肢から予測する形式の問題である．よって空所を含む英文に対して，空所に当てはまる語句を選択肢から予測する．CLOTH データセットの英文空所補充問題の例を表 1 に示す．表 1 の “___” は空所を表し，正答の選択肢は太字で表している．表 1 に示すように CLOTH データセットでは英文法や語彙，内容理解を問う問題が含まれている．

一方，Bidirectional Encorder Representations from Transformers (BERT) [5] は近年，広い範囲の自然言語処理タスクにおいて State-of-the-art (SOTA) を達成した言語表現モデルであり，大規模なテキストコーパスを用いて Transformer という自己注意機構を備えたニューラルネットワークの事前学習を行い，その後，個々のタスクに応じて fine-tuning を行う．BERT の事前学習では Masked Language Model (MLM) タスクと Next Sentence Prediction (NSP) タスクを用いて言語表現を獲得している．

本稿の構成は次の通りである．2 節で，関連研究について述べ，3 節では，BERT の構造と BERT を利用した英文空所補充問題の解法について説明する．4 節では，評価実験の内容と結果を示し，5 節で，その結果について考察する．6 節で，本研究のまとめと今後の課題について述べる．

2 関連研究

2.1 統計的言語モデルによる英文空所補充

2011 年から始まったプロジェクト “ロボットは東大に入れるか”[1] は，大学入試センター試験 [2] および東京大学入試において高得点をとることを目的としている．このプロジェクトの英語科目では，問題を一文問題，複数文問題（問題や選択肢が複数文からなるもの），長文問題の 3 つに大きく分類している．一文問題はさらに文法・語法・語彙問題，語句整序完成問題，発話文生成問題の 3 つに分類され，この文法・語法・語彙問題が我々の取り組む英文空所補充問題と一致する．

東中ら [6] は統計的言語モデルである KenLM [7] を用いて，

1 : https://21robot.org/
2 : https://www.dnc.ac.jp/

表 1　CLOTH データセットの英文空所補充問題の例

空所つき英文	選択肢			
She saw the rich girl ____ up the stairs and then turn back quickly.	runs	running	ran	**run**
That girl grew into a ____ woman with a husband.	poor	**rich**	homeless	good
Your clock is too old and too big. Why don't you buy a ____ like everyone else?	**watch**	house	computer	picture
Lisa is an English girl. She is ____ .	old	**young**	long	different

文法・語法・語彙を問う英文空所補充問題に取り組んだ．空所に選択肢を補充した英文に対して，KenLM で対数尤度を算出することで単語の並びとして尤もらしい選択肢を予測している．東中らは約 178GB の Common Crawl コーパス [3]を学習データとして 7-gram 言語モデルを構築した．また，大学入試センター試験の本試験および追試験の過去問，代ゼミセンター模試，ベネッセ模試，独自に収集したその他の問題を合わせた合計 552 問の 4 選択肢からなる英文空所補充問題をテストデータとした．東中らの手法では，テストデータにおける正答率は 82.2% であった．さらに，文長による対数尤度の正規化や，数詞を桁数のみ分かるように変換するなどの正規化を利用すると正答率は 85.7%まで向上した．

2.2　MPnet を拡張した言語モデルによる英文空所補充

玉城らは，Multi-Perspective Context Aggregation Network (MPNet) [8] を拡張したモデルによる英文空所補充問題の解法を提案した [9]．彼らは MPNet の中間層である多視点集約部において 2 つのモジュールを提案している．一つは，空所に各選択肢を補充した英文を用いて Attention を算出し，それを用いて空所から離れた語句の情報を選択肢の予測に反映させる空所補充文 Attentive Reader モジュールである．このモジュールは Chen らの提案した Attentive Reader [10] に基づくものである．もう一つは，文中に未知語が含まれる際にその算出方法を変更し，未知語以外の残りの語を用いて対数尤度を算出する KenLM Score モジュールである．このモジュールは *n*-gram(n は 1 から 5) に含まれる他の単語の情報を空所の予測に反映させるために，統計的言語モデルである KenLM を利用している．4 択の英文空所補充問題であるテスト用 CLOTH データセットでの高校入試問題に対する正答率は，Wang らの MPNet では正答率が 47.2%であったのに対して玉城らのモデルは 55.2%であった．

3　BERT による英文空所補充問題の解法

3.1　BERT

図 1 に BERT の概略図を示す．E_i が入力系列を表し，T_i が出力系列を表している．また，Trm は Transformer [11] を表している．Transformer は Attention 機構を用いたニューラル機械翻訳モデルである．特に BERT では self-attention を利用して関連する単語の重みを計算している．これまで OpenAI GPT [12] のような一般的な言語表現モデルでは left-to-right モデルが用いられ，左から右の一方向からのみ token を読み込み学習を行うが，BERT では図 1 のように入力系列を双方向から読み込む事によって注目する単語の周囲全体に基づいて文脈を学習することができる．

これまでには ELMo [13] に見られるように，left-to-raight と right-to-left を独立して LSTM で学習して得られる特徴量を足し合わせる事で双方向の読み込みを実現していたが，BERT では同一のモデル内で双方向から注目する単語の周囲の文脈を学習することを実現し，前後の関係性を考慮する．

BERT において入力シーケンスをそのまま双方向から読み込むと，学習時に予測すべき単語を先読みしてしまうために，事前学習の際に Masked Language Model (MLM) と Next Sentence Prediction (NSP) を用いて学習を行う事により予測する単語を先読みする事を防いでいる．

3.1.1　入力データの前処理

BERT へ入力される英文はサブワードに分割される．サブワードに分割された語の中で語幹でないものには “##” が付与される．例えば “playing” は “play” と “##ing” に分割される．

3.1.2　Masked Language Model (MLM)

MLM の概略を図 2 に表す．MLM では入力シーケンスにある単語の 15%を [MASK] トークンに置き換える．そして，シーケンスにある単語のうちマスクされなかったものによって与えられる文脈に基づいて，[MASK] トークンに置き換えられた単語を予測するタスクである．例えば以下の文の場合，文中からランダムに選んだ単語 “ jump ” をマスクした文を作成する．その文に対して [MASK] トークンを正しく推測する．

（1）　They can jump so high and so far.

（2）　They can [MASK] so high and so far.

図 3 に BERT が MLM を実行した際の模式図を示す．W_i は各サブワードを表しており，BERT への入力となる．Transformer からの出力が O_i である．Transformer からの出力に分類レイヤを追加する．分類レイヤは全結合層，活性化関数 GELU とノルムにより構成される．次に，出力されたベクトルと埋め込み行列をかけて，出力を単語次元に変換する．その後，softmax 関数を使って正規化を行い，出力されたそれぞれの単語の確率 W_i' を計算する．

3.1.3　Next Sentence Prediction (NSP)

NSP の概略を図 4 に表す．NSP は入力として文のペアを受け取り，ペアにおいて 2 つ目の文章が元の文において後続の文になっているかを予測するタスクである．NSP の学習において BERT がペアとなっている 2 つの文を区別するために，ペアの文を入力として渡す前に図 4 のように埋め込み表現に置き換え

3 : https://registry.opendata.aws/commoncrawl/

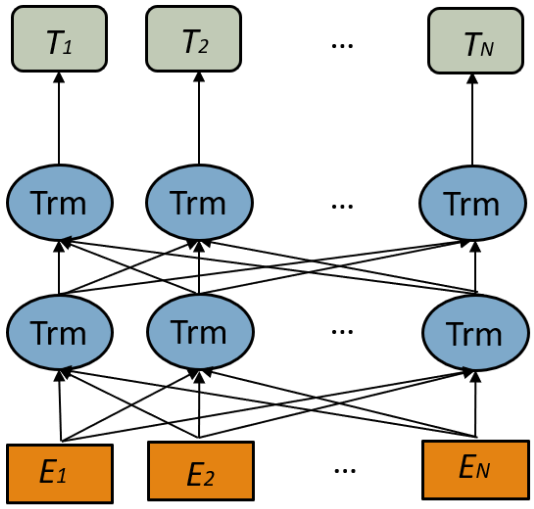

図 1　BERT の概略図

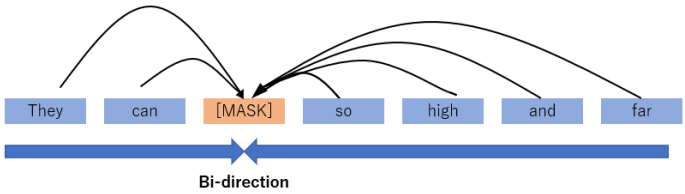

図 2　MLM の概略図

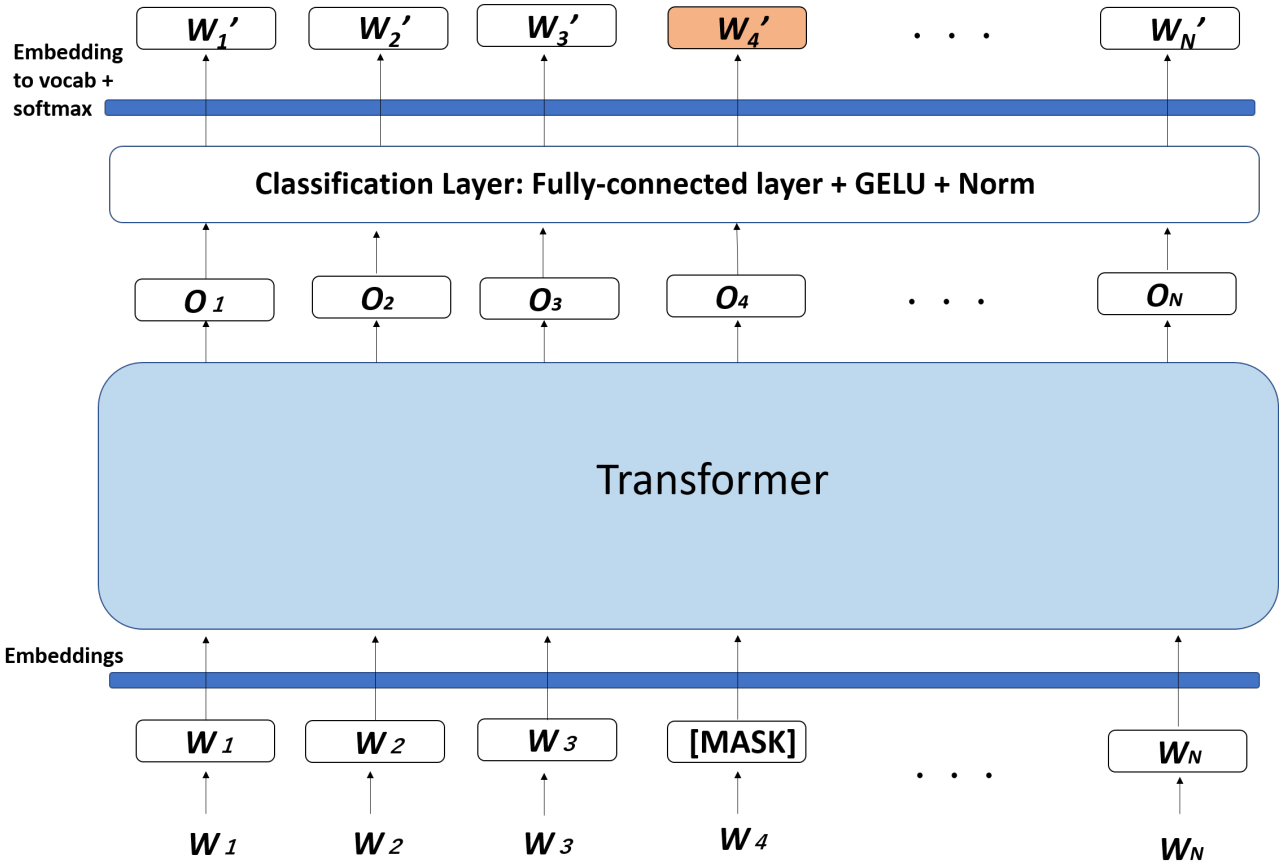

図 3　MLM 実行時の模式図

られる．すなわち図 4 中の Token Embeddings はトークンの ID を表し，Segment Embeddings は文のペアの区切れを表す．また Position Embeddings はシーケンス内の単語の位置を表している．それらを加算したベクトルのシーケンスが Transformer への入力になる．[CLS] は文の先頭を表すトークンであり [SEP] は文章の区切れを表す．事前学習において入力となる文のペアの 50%は，以下の (1) のようにオリジナルの文章中において自然に続く文であるが，残りの 50%のペアは後続文が以下の (2) のようにコーパスからランダムに選択された文である．

（1） [CLS] the man went to [MASK] store [SEP]
he bought a gallon [MASK] milk [SEP]

（2） [CLS] the man [MASK] to the store [SEP]
penguin [MASK] are casual birds[SEP]

3.2　MLM タスクを利用した手法

本稿では英文空所補充問題に BERT の事前学習済みのモデル

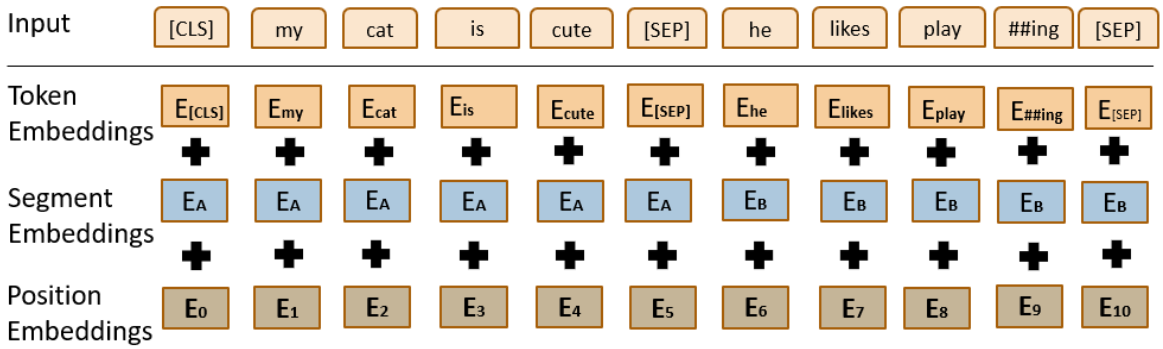

図 4　NSP の概略図

を使用して MLM を行い，英文空所補充問題の回答を試みる．BERT の事前学習済みモデルは，英語版 Wikipedia と BooksCorpus [14] を学習データとして事前学習を行ったものである．まず英文の空所を [MASK] トークンに置き換える．次に，各選択肢の単語を [MASK] トークンと置き換え BERT の事前学習済みのモデルを用いて，各選択肢ごとに出現確率を計算して正規化した後，最も確率が高い選択肢を英文空所補充問題への回答とする．

3.3　CLOTH データセットによる fine-tuning

本稿では 3.2 節で説明した BERT を CLOTH データセットで fine-tuning して用いる方法を提案する．このモデルを用いて各選択肢ごとに出現確率を計算して正規化した後，最も確率が高い選択肢を英文空所補充問題の回答とする．fine-tuning で使用する CLOTH データセットは，実験で用いるテストデータとは異なるデータセットである．BERT では，それぞれのタスクに応じた教師ありデータを用いて fine-tuning をする事により特定のタスクに特化したモデルを構成できるため，事前学習済みのモデルをそのまま適用するよりも正答率の向上が期待できる．

4　評 価 実 験

4.1　実験データ

ここでは実験に用いるデータについて説明する

4.1.1　学習データおよび検証データ

BERT の学習データおよび検証データには CLOTH データセットを用いる．CLOTH データセットは中国の高校入試および大学入試の長文問題から作られた英文空所補充問題であり，学習データは 5,513 の長文からの 76,850 問，検証データは 805 の長文からの 11,067 問からなる．検証データは学習データを用いて fine-tuning を行った後に，モデルの評価を行うためのデータである．なお CLOTH データセットでは空所に入る語句の選択肢が 4 つ与えられる．

4.1.2　テストデータ

テストデータには CLOTH データセット，大学入試センター試験の英文空所補充問題，Microsoft Research Sentence Completion Challenge データセット [15] の 3 つを用いる．本稿では Microsoft Research Sentence Completion Challenge データセットを “MSR データセット” と呼ぶ．

CLOTH データセットのテストデータは，中国の高校入試問題である 335 の長文からの 3,198 問，中国の大学入試問題である 478 の長文からの 8,318 問からなる．学習データおよび検証データで用いるデータはこれには含まれない．ここから無作為に選出した 3,000 問に対する Amazon Mechanical Turkers [4] による人手の正答率は，高校入試問題では 89.7%，大学入試問題では 84.5%であった [3].

日本の大学入試センター試験では大問 2A が英文空所補充問題であり，2000 年度から 2016 年度に行われた本試験および追試験のうち，空所が 1 箇所のみである 324 問を評価の際のテストデータとして使用する．大学入試センター試験では空所に入る語句の選択肢が 4 つ与えられる．324 問のうち選択肢のいずれも 1 単語であるものは 161 問ある．

MSR データセットは 1,040 問の英文空所補充問題であり，英文はアーサー・コナン・ドイルの小説シャーロック・ホームズシリーズ 5 冊から抽出したものである．英文中の 1 箇所が空所となっており，選択肢が 5 つ与えられている．ここから無作為に選出された 100 問に対する人手の正答率は 91%であった [15]. また，MSR データセットは選択肢が 5 つあるが，CLOTH データセットが 4 択の英文空所補充問題であることから難易度を揃えるために選択肢を 4 つにする必要がある．そこで正答でない選択肢を無作為に 1 つ除いたものを実験では用いる．

4.2　実験データの前処理

CLOTH データセットにおいては 1 文中に空所が複数含まれる英文も存在するが，BERT への入力にそのような英文が含まれる場合には，空所が一つとなるように残りの空所を正答の語句で補充した複数の文章へと分割する．例えば，“He gets up very early. After having his ____ , he goes to ____ .” のような空所を含む英文があるとする．正答の選択肢はそれぞれ “breakfast” と “school” である．この場合では “He gets up very early. After having his ____ , he goes to **breakfast** .” のように一方の空所を正答の語句で補充し，BERT への入力とする．

4.3　BERT の事前学習済みモデル

BERT の事前学習済みモデルには，中間層の transformer が 12

4 : https://www.mturk.com/

表 2　4 択のテストデータ対する各モデルの正答率 (%)

モデル	テストデータ			
	CLOTH-M (3,198 問)	CLOTH-H (8,318 問)	センター試験 (324 問)	MSR (1,040 問)
BERT_MLM (base)	75.3	69.4	91.0	77.6
BERT_MLM (large)	77.0	73.1	**92.3**	80.7
BERT_fine-tuning (base)	85.3	80.9	75.0	76.6
BERT_fine-tuning (large)	**88.8**	**85.0**	75.6	**85.0**
kenLM (7-gram)	58.8	47.6	40.7	31.4
玉城らの手法	55.2	45.1	40.4	28.9

層重なっている BERT の base モデルと 24 層重なっている BERT の large モデルの 2 種類がある．その中にそれぞれ Uncased モデルと Cased モデルが存在し，Uncased はテキストを小文字化してから WordPiece [16] によるトークン化を行う．例えば “John Smith” は “john smith” に変換される．またアクセント符号の削除なども行う．Cased モデルは元のテキストの大文字やアクセント符号などもそのまま維持されており，大文字小文字の情報が重要な固有名詞認識や品詞のタグ付けなどでは Cased モデルが優れているが，一般的には Uncased モデルの方が良い結果が出ることが知られている．

4.4 比 較 手 法

統計的言語モデルである kenLM，および Wang らの MPNet を拡張した玉城らの手法を比較対象とし，提案した 2 つのモデルと比較する．両モデルともに CLOTH データセットを用いて学習と検証を行う．また，提案手法と同じテストデータでテストを行う．

KenLM は n-gram 言語モデルであり，n-1 単語を入力として次の 1 単語の確率分布を予測する．n-gram の n については東中らと同様に KenLM の 7-gram モデルを用いる．

4.5 実 験 結 果

3.2 節や 3.3 節で述べたように英文空所補充問題に BERT の事前学習済みモデルを利用して MLM タスクを利用した手法と CLOTH データセットで fine-tuning を行った手法で実験を行いそれぞれのモデルを評価する．評価尺度には英文空所補充問題の正答率を用いる．BERT の出力結果から与えられた選択肢 4 つに対する確率を出力し，その確率が最も高い選択肢を空所に補充すべき選択肢として予測する．

実験結果を表 2 に示す．表 2 中の CLOTH-M，CLOTH-H はそれぞれ CLOTH データセットにおける高校入試問題，大学入試問題を表す．センター試験は大学入試センター試験の英文空所補充問題，MSR は MSR データセットを表す．BERT_MLM (base)，BERT_MLM (large) は，MLM タスクを利用した手法を用い，事前学習済みモデルとしてそれぞれ BERT の base モデル，large モデルを使用している．BERT_fine-tuning (base)，BERT_fine-tuning (large) は CLOTH データセットを学習データおよび検証データとして fine-tuning した手法であり，それぞれ BERT の base モデル，large モデルを使用している．

表 2 では各テストデータに対して最も高い正答率を太字で表している．表 2 より，MLM タスクを利用した手法，fine-tuning した手法の両方が，比較手法である玉城らの手法の正答率を大きく上回った．BERT_fine-tuning (large) はセンター試験以外で最も高い正答率となった．センター試験のみ MLM 拡張モデルの 92.3%が BERT_fine-tuning (large) の 75.6%を上回った．

CLOTH データセットの大学入試問題である CLOTH-H，は高校入試問題である CLOTH-M と比べて正答率が低い傾向にあり，問題難易度の影響を受けていると言える．MSR データセットでは BERT_MLM (large), BERT_fine-tuning (large) の両手法が比較手法である KenLM の正答率 31.4%と比べて大きく正答率が向上した．これは BERT の事前学習時に学習データである BookCorpus [14] にシャーロックホームズシリーズの小説が含まれており，その小説を用いて BERT の事前学習が行われたからである．

5 考　　察

MLM タスクを利用した手法とそれを CLOTH データセットで fine-tuning した手法について考察する．表 2 より，MLM タスクを利用した手法と CLOTH データセットで fine-tuning をした手法のどちらでも BERT の large モデルの方が正答率が高かったため，BERT の large モデルについて考察する．

5.1 MLM タスクを利用した手法の有効性

本実験における MLM タスクを利用した BERT の large モデルの大学入試センター試験に対する正答率は 92.3%であった．これは KenLM を用いた東中らの実験の正答率である 85.7%を大きく上回った．ただし，彼らの実験のテストデータは大学入試センター試験の本試験および追試験の過去問，代ゼミセンター模試，ベネッセ模試，独自に収集したその他の問題を合わせた合計 552 問がテストデータである．以下では MLM タスクを利用した手法で不正解だった問題を分析し考察する．

5.1.1 MLM タスクを利用した手法の誤答例

MLM 誤答例を以下に示す．回答は BERT により算出された出現確率が高いものから順に並べている．

- 問題: The coffee shop opens at 7:30 and serves breakfast ____ 10 o’clock.”
  選択肢: [within, by, for, till]
  BERT の回答: [by, till, for, within]

選択肢はすべて時に関係する前置詞である．誤答例では “till”

表 3　センター試験問題 (324 問) において MLM タスクを利用した手法が誤答した問題

問題が問う内容	MLM タスクを利用した手法が誤答した問題数
内容理解	6
語彙	12
時制	4
人称	1
名詞の単複	1

表 4　MLM タスクを利用した手法と CLOTH データセットで fine-tuning を行った手法の正解と不正解

テストデータ	両モデルが正解	MLM タスクを利用した手法のみ正解	fine-tuning のみ正解	両モデルが不正解
CLOTH-M (3,198 問)	2,286(71.5%)	185 (5.8%)	530(16.6%)	197(6.2%)
CLOTH-H (8,318 問)	5,722 (68.8%)	365(4.4%)	1,372(16.5%)	859(10.3%)
センター試験 (324 問)	224 (69.1%)	76 (23.5%)	21 (6.5%)	3 (0.9%)
MSR (1,040 問)	716 (68.8%)	123 (11.8%)	168 (16.2%)	33 (3.2%)

が正答であるが，MLM タスクを利用した手法は “by” を回答した．“by” は期限を表す前置詞であり，“till” は継続を表す前置詞である．一般的に混同されやすい 2 つの選択肢が回答の上位 2 つに位置しており，人間の感覚に近いと言える．

5.1.2　MLM モデルの誤答の種類

センター試験問題 324 問の中で不正解だった 24 問の種類を表 3 に示す．センター試験には内容理解を問う問題や語彙を問う問題，時制や人称といった英文法を問う問題などが含まれている．表 3 の “人称” は主語に対応する動詞の活用形の種類や，代名詞の格の種類に関する問題を表し，“名詞の単複” は名詞の単数形や複数形に関する問題や，名詞につく冠詞に関する問題を表す．表 3 より語彙問題での誤答が他と比べて多い事がわかる．また前後の文脈から判断することが可能な人称や名詞の単複では誤答が少なく，MLM モデルが空所の前後の文脈を読み取り，予測に反映できていることがわかる．

5.2　BERT の fine-tuning の有効性

表 2 に示したように CLOTH データセットで fine-tuning をするとそれをしないよりも正答率が向上する問題が多かった．

表 2 より BERT_fine-tuning (large) は，CLOTH-M, CLOTH-H, MSR のデータセットで BERT_MLM (large) を上回り，及ばなかったのはセンター試験のみであった．ここでは MLM タスクを利用した手法と CLOTH データセットで fine-tuning を行った手法の正解と不正解の数を調べ，それらについて考察する．

各テストデータに対する MLM タスクを利用した手法および fine-tuning した手法の予測結果の正解と不正解の数を表 4 に示す．表 4 中の () 内の値は各テストデータの全問題数に対する割合を表す．表 4 より CLOTH-M，CLOTH-H の両方のテストデータに対しては，MLM モデルよりも CLOTH データセットで fine-tuning を行った手法の方が 10 ポイント以上正答率が高い．この事から CLOTH データセットに対する fine-tuning の有効性が確認できる．

センター試験の問題では MLM タスクを利用した手法が CLOTH データセットで fine-tuning を行った手法に対して 20 ポイント以上高かった．また両方の手が共に不正解となった問題が全 324 問中 3 問と少なかった．また MLM タスクを利用した手法では正解したが，CLOTH データセットで fine-tuning を行った方法では不正解だったものが全体の 23.5 %を占めた．その多さからセンター試験の問題においては，CLOTH データセットでの fine-tuning は有効ではなかったことがわかる．

MSR データセットでは，4 ポイントほど CLOTH データセットで fine-tuning を行った手法が MLM タスクを利用した手法を上回っている．MSR データセットは小説の内容要約文の英文空所補充問題であり，語彙を問う問題のみである．そのため，CLOTH データセットによる fine-tuning が語彙を問う問題の正答率の向上に寄与したと考えられる．

6　ま　と　め

本稿では英文空所補充問題の解法として，BERT の事前学習で用いられている MLM タスクを利用した手法と，それを CLOTH データセットで fine-tuning して用いる方法を提案した．実験では，中国の高校入試および大学入試を集めた CLOTH データセットで fine-tuning し，fine-tuning 用いたものとは別の CLOTH データセット，大学入試センター試験，Microsoft Research Sentence Completion Challenge データセットの 3 種類のテストデータで評価した．BERT の MLM タスクを利用した方法では 4 つの選択肢から空所の語句を選ぶセンター試験問題に対する正答率が 92.3%となり，比較対象とした KenLM の 40.7%を上回った．CLOTH データセットで fine-tuning して BERT を用いる方法では，テスト用 CLOTH データセットの高校入試問題に対する正答率が 88.8%となり，これも比較手法である KenLM の正答率

の 58.8%を大きく上回った．

今後の課題としては，BERT 以降に発表され，言語処理タスクにおいて近年 BERT を上回る成果を挙げている言語表現モデルである RoBERTa や ALBERT の英文空所補充問題への適用が挙げられる．

文　献

[1] Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pp. 1693–1701, 2015.

[2] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 839–849, 2016.

[3] Qizhe Xie, Guokun Lai, Zihang Dai, and Eduard Hovy. Large-scale cloze test dataset created by teachers. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2344–2356, 2018.

[4] Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. Lsdsem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pp. 46–51, 2017.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[6] 東中竜一郎, 杉山弘晃, 成松宏美, 磯崎秀樹, 菊井玄一郎, 堂坂浩二, 平博順, 南泰浩, 大和淳司. 「ロボットは東大に入れるか」プロジェクトにおける英語科目の到達点と今後の課題. 人工知能学会全国大会論文集, Vol. JSAI2017, p. 2H21, 2017.

[7] Kenneth Heafield. KenLM: faster and smaller language model queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pp. 187–197, 2011.

[8] Liang Wang, Sujian Li, Wei Zhao, Kewei Shen, Meng Sun, Ruoyu Jia, and Jingming Liu. Multi-perspective context aggregation for semi-supervised cloze-style reading comprehension. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 857–867, 2018.

[9] 玉城悠仁, 新妻弘崇, 太田学. Multi-Perspective Context Aggregation Network の拡張による英文空所補充問題の一解法. 第 11 回データ工学と情報マネジメントに関するフォーラム, B1-1, 2019.

[10] Danqi Chen, Jason Bolton, and Christopher D. Manning. A thorough examination of the cnn/daily mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 2358–2367. Association for Computational Linguistics, 2016.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pp. 5998–6008. Curran Associates, Inc., 2017.

[12] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[13] Matthew E. Peters, Mark Neumann, Mohit lyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. arXiv preprint arXiv:1810.04805, 2018.

[14] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Alignings books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pp. 19–27, 2015.

[15] Geoffrey Zweig and Chris J. C. Burges. A challenge set for advancing language modeling. In *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*, WLM '12, pp. 29–36, 2012.

[16] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, Vol. abs/1609.08144, , 2016.