センター英語試験の意見要旨把握問題に対する BERT の適用方法の検討

杉山弘晃　　成松宏美　　東中竜一郎
NTT コミュニケーション科学基礎研究所
hiroaki.sugiyama.kf@.hco.ntt.co.jp

1 はじめに

我々は「ロボットは東大に入れるか」プロジェクト [4] において，引き続き英語（特に，センター試験の英語問題）に取り組んでいる．本稿は，センター試験の英語問題で出題される意見要旨把握問題の解法について述べる．意見要旨把握問題は，複数人の議論において，ある発言者が提示した意見の要旨を最もよく表す文を 4 つの選択肢（要旨候補）から選ぶ問題である．具体例を図 1 に示す．[5] では，注意機構（Attention）を利用した深層学習ベースのモデル（Stanford Attentive Reader）を，RACE と呼ばれる大規模な英語試験問題データセット [3] で学習することで，40%程度の正解率を達成できることを報告した．

一方，近年世界的に Machine reading comprehension（MRC）タスクが注目を集めており，極めて速い速度で性能向上が進んでいる．特に，Transformer と呼ばれる自己注意機構を備えたニューラルネットワークを大規模なテキストコーパスを用いて事前学習し，個別の問題に対して転移学習する，OpenAI GPT（GPT）や BERT と呼ばれるアプローチが，MRC を含む様々な言語処理タスクにおいて SOTA を達成しており，大きく注目されている [1, 2]．BERT の転移学習によって，意見要旨把握問題についても正答率向上が期待できるが，一方で深層学習ベースのモデルは多様な入出力の関係を柔軟に学習できるため，意見要旨把握問題への適用方法には様々なアプローチが考えられる．本研究では，意見要旨把握問題における BERT の有用性を検証するとともに，意見要旨把握問題に適した適用方法について比較分析を行う．

2 手法：BERT による意見要旨把握問題の解答

2.1 Transformer を用いる自動解答手法

Transformer を用いる自動解答手法の先駆けとして，OpenAI GPT（以下 GPT）が提案されている．GPT は，図 2（左）に示す Transformer モデルを，大規模なテキストコーパスで事前学習し，個別の問題に対して転移学習する手法である [1]．事前学習と組み合わせることで，比較的小規模なデータセットでも効率よく学習することができ，様々な言語処理タスクで高い性能を示している．Transformer は，位置情報（Position embedding）付きのテキストを入力として，「自分と関係する周辺情報を集約する」機能を持つ自己注意機構を繰り返し適用することで，タスクに適した特徴ベクトルを計算するモデルである．GPT では Transformer の事前学習に，追加アノテーションが不要であり，かつ汎用的なタスクとして，Left-to-right 形式で次単語を予測する言語モデル的なタスクを用いている．また，個別のタスクに対する転移学習は，図 2（右）に示す形式を Transformer へ入力し，得られた特徴ベクトルをタスクに適した出力形式へ変換する，FFNN や Softmax 関数を学習することで実現している．

一方，GPT の事前学習では，Left-to-right の一方向の関係しか考慮できないという制約があった．また，GPT のもう一つの制約として，文のまとまりを表す情報が，Delimiter（図 2（右）の Delilm）を利用した

Stephen: Thank you, Dr. Ishii. I agree we are living in a time when technology will soon improve even more rapidly. Looking back at the 1900s shows us how people faced rapid changes in their societies. I think this has lessons for us today. One of the biggest changes of the 20th century was the rise of a global society. I believe airplanes made this possible. For the rst time, people could travel quickly to the farthest corners of the earth and experience life in other countries. Certainly telephones and the Internet had an impact as well. But there's no substitute for traveling to new places and actually meeting people.
Sue: I've heard this opinion before, Stephen. Are you saying [32] ?
選択肢:
(1) airplanes helped create our global society
(2) foreign travel was not possible before the 1900s
(3) technology will soon change more slowly
(4) telephones and the Internet were more important than airplanes

図 1: 意見要旨把握問題の例（ベネッセ模試 2016 年 6 月第 3 問 C）．正解は (1)

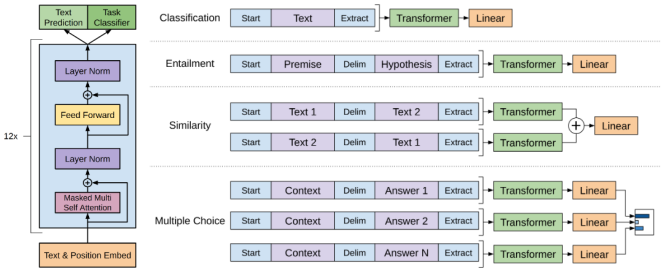

図 2: Transormer モデルの構造（右）と個別タスクに対する転移学習時の入出力・モデル構造（左）[1]

間接的な表現に限定されており，文単位の関係性を表現しにくいという問題があった．BERT は，入力情報に文のまとまりを表す Segment ID およびそのベクトル表現（Segment embeddings）を追加し直接的に文のまとまりを与えるとともに，事前学習を双方向の言語モデルタスクと 2 文の結束性を判定するタスクの 2 種類について行う手法である．

2.2 4 択問題への BERT の適用

BERT を 4 択の意見要旨把握問題に適用する場合，各選択肢を単独で評価する方法と，複数選択肢の比較に基づいて評価する方法が考えられる．本研究では，同時に考慮する選択肢の個数および考慮の仕方の影響を調べるため，以下の 4 パターンの方法を比較する．なお，以下の図の document/doc は，図 1 中の意見主張部分（ここでは Stephen の発話），query はまとめ役の発話部分（Sue の発話），option は選択肢を表す．

個別比較　GPT の Multiple Choice と同様，入力テキストに選択肢を一つだけ埋め込みその選択肢の正答らしさを推定し，4 択の推定値のうち最大のものを正答として出力する方法である．入力情報の模式図を図 3 に示す．入出力関係がシンプルで学習しやすい反面，選択肢間の関係性を考慮できない問題がある．

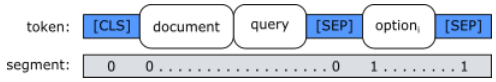

図 3: 個別比較

2 択埋込　入力テキストに 2 つの選択肢を [SEP] で区切って埋め込み，2 クラス判別問題として解答する方法である．模式図を図 4 に示す．2 つの選択肢を同時に比較することで，それらの選択肢間の比較に基づく解答が可能になると期待できる．Segment embeddings が 3 種類必要となるが，事前学習時には Segment embeddings は 2 種類しか学習されていないため，もう 1 種類のベクトルを用意する必要がある．本研究では，これら 2 種類のベクトルを Segment ID の 0, 2 に対応させ，それらを端点とする内分ベクトルを 1 に対応するベクトルとして利用する．こうして得られた未学習のベクトルを利用しても精度低下がないことを，上記個別比較条件を対象とする予備実験を通して確かめている．

図 4: 2 択埋込

4 択埋込　各選択肢を [SEP] で区切って入力に全て埋め込み，4 クラス判別問題として解答する方法である．模式図を図 5 に示す．全ての選択肢を同時に比較することで，より選択肢間の比較に基づく解答が得られると期待できる反面，学習が難しくなると予想される．

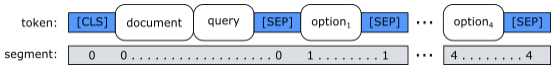

図 5: 4 択埋込

2 択連結　2 つの選択肢を別々に埋め込んだ 2 つのテキストを個々に BERT に入力し，得られた特徴ベクトルを連結して得られたベクトルを用いて 2 クラス判別問題として解答する方法である．模式図を図 6 に示す．

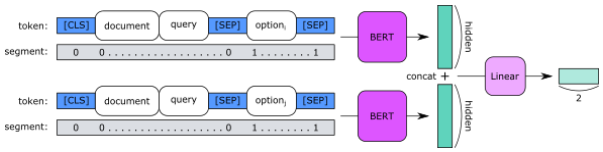

図 6: 2 択連結

2.3 入力における Option の配置

意見要旨把握問題を BERT への入力形式に変換する際，2.2 節で示したように doc, query から option を分離する方法と，query の空欄に option を埋め込んで完全な文として扱う方法の 2 つが考えられる．本節ではそれぞれの変換方法の詳細について説明する．

option 分離　図 3 同様，doc と query を連結し option と [SEP] で区切る方法である．doc+query の Segment ID として 0 を振り，option のみ Segment ID を 1 とする．RACE の学習でも利用されている区切り方で，選択肢の違いを際立たせる効果が期待できる．

doc 埋め込み　図 7 のように，query の空欄に option を埋め込んだものについて，Segment ID に 1 を設定し，doc の元の位置に埋め込んだものである．query と option を連結し 1 つの文とすることで，query に含まれる主語等の情報を活かして解答できるようになる．また，query および option の doc 中の位置を正しく反映できる利点がある．

図 7: doc 埋め込み

3 実験

3.1 データセット

本研究では，[5] と同様に，訓練データとして RACE データセット [3] と呼ばれる，中国人中高生向けの英語試験を大量に収集したデータを用いる．RACE データセットの特徴として，9 万 7 千問程度と中程度の大きさであることと，いわゆる長文読解に該当する問題であり，本文中の意見との含意関係を問うような，意見要旨把握に類似した問題を含む点がある．モデルの訓練には，RACE の train データ 87860 問（独自に表記誤り等を修正したもの）を利用する．解法の性能を測るベンチマークデータには，大学入試センター試験の本試験及び追試験の過去問，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 234 問を用いる．234 問中 120 問を開発データセット（dev），114 問をテストデータセット（test）として用いる．以下の実験では，dev で正解率が最大になった epoch での test の正解率を比較する．

3.2 実験設定

本研究では，[5] で示されている Stanford Attentive Reader（SAR）をベースラインとして，2 章で説明した GPT, BERT と正答率で比較する．GPT の事前学習モデルとして OpenAI が公開しているモデル[1]を用いる．BERT の事前学習モデルとして，OpenAI が公開しているモデルのうち，base と large の 2 種類のモデル[2]を利用する．base は GPT と同程度の規模のモデル（12 層，隠れ層 768 次元，12 ヘッド）であり，large はより大規模なモデル（24 層，隠れ層 1024 次元，16 ヘッド）である．各正答率の信頼区間の計算には二項分布を用い，多重検定の補正には BH 法を利用した．

GPT および BERT を転移学習する際の訓練パラメータを以下に示す．以降の結果では各パラメータで学習した中で最も dev の値が高かったモデルを用いる．Learning rate の初期値として，$5e-5, 5e-6, 5e-7$ を用い，バッチサイズは GPT は 32 と 64，BERT は 32 を用いた．Dropout は 0.1 と 0.3 を用いた．

3.3 結果と分析

3.3.1 手法間比較

図 8 に，適用手法として個別比較を，入力形式として option 分離を選択した場合の，SAR, GPT, BERT-base, BERT-large の正答率を示す．手法改良およびモデルサイズ大規模化に伴い正答率が向上しており，4 群間の χ^2 検定の結果，正答率の差異には有意傾向が認められた（$\chi^2(3) = 6.33, p = 0.096$）．また残差分析の結果，BERT-large の調整済み残差が 2.01 で最も大きく，$p = 0.044$ で有意に異なることが示された．

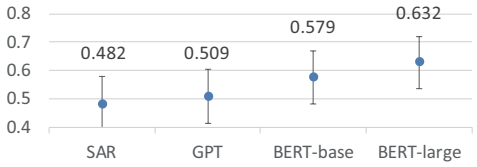

図 8: 手法間の比較．エラーバーは二項分布による 95%信頼区間を表す．

3.3.2 4 択問題への適用方法による比較

表 1: 4 択問題への適用方法による比較

適用方法	個別比較	2 択埋込	4 択埋込	2 択連結
正答率	0.578	0.578	0.333	0.394

図 1 に 4 択問題への適用方法による比較の結果を示す．個別比較と 2 択埋込が同一の正答率となっており，2 択埋込による選択肢比較の有用性は確認できなかった．一方，2 群間 χ^2 分析の結果，4 択埋込および 2 択連結は個別比較および 2 択埋込と比較して有意に性能が低下していた（4 択埋込：$\chi^2(1) = 12.889, p = 3.3e-4$，$BH$ 補正後 $p = 9.9e-4$，2 択連結:$\chi^2(1) = 7.022, p = 8.0e-3, BH$ 補正後 $p = 0.012$）していた．これらの方法は推定する入出力関係が複雑であるため，学習が困難になったものと考えられる．

3.3.3 意見要旨把握問題の入力形式による比較

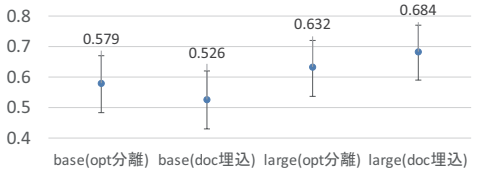

図 9: 入力形式による比較．エラーバーは二項分布による 95%信頼区間を表す．

option 分離と doc 埋込の 2 つの入力形式について，base, large の両モデルでの正答率の変化を図 9 に示す．base モデルにおいては option 分離が，large モデルにおいては doc 埋込が高い正答率を示しており，特に large モデルの doc 埋込は，option 分離と比較して有意ではないものの，本稿の施行中最も高い正答率（**0.684**）を示した．一方，入力形式の違いによる base・large モデル間の変化率の差異については，有意な差は認められなかった（$\chi^2(1) = 0.389, p = 0.532$）．

4 誤答タイプと正答率の関係分析

あるモデルが誤答の選択肢を誤って選んだ場合，そのモデルは誤答が誤りである根拠を認識できていないと考えられる．本章では，誤答の根拠を認識するために必要となる推論のタイプを人手でアノテーションし，その分類を基に，意見要旨把握問題中のどのような問

[1] https://github.com/openai/finetune-transformer-lm
[2] https://github.com/google-research/bert

表 2: 誤答判定に必要な推論の分類

ID	ラベル	説明
E1	否定	否定語によって doc と option が反対の意味になっている.
E2	非主要部	option の内容が doc の主要ではない部分を述べている.
E3	組合せ誤り	option が doc に含まれる語義の組み合わせで構成されているものの，doc と異なる意味になっている.
E4	doc 外語義	option の一部に doc に含まれない語義が用いられ，doc と異なる意味になっている.
E5	その他	その他.

題に対し BERT が有用であったか，改善できなかったかについて分析する．推論タイプのアノテーションは著者らのうち 2 名が合議で行った．

表 2 に誤答判定に必要な推論タイプの分類項目を示す．複数の推論タイプが付与された場合には，その他（E5）を除きラベルの番号が大きいほど判定が難しい問題であると考え，大きい番号のラベルに名寄せすることとする．E5 は名寄せしないため，総数は 342 問（114×3）よりも多くなっている．

表 3: 誤答判定に必要な推論タイプごとの誤答選択率

ラベル	個数	SAR	GPT	base	large
否定 (E1)	25	0.160 (4/25)	0.200 (5/25)	0.080 (2/25)	0.040 (1/25)
非主要部 (E2)	28	0.107 (3/28)	0.142 (4/28)	0.107 (3/28)	0.142 (4/28)
組合せ誤り (E3)	89	00.179 (16/89)	0.191 (17/89)	0.202 (18/89)	0.213 (19/89)
doc 外語義 (E4)	186	0.193 (36/186)	0.161 (30/186)	0.129 (24/186)	0.102 (19/186)
その他 (E5)	25	0.080 (2/25)	0.000 (0/25)	0.080 (2/25)	0.080 (2/25)

表 3 に，誤答判定に必要な推論の分類結果（名寄せ済み），および各誤答タイプの選択肢が選ばれた割合を示す．なお，BERT の base と large については，いずれも option 分離の結果を示している．E1, E2, E5 に分類された誤答はそれぞれ 30 個以下であり，大半の判定に必要な推論タイプが組合せ誤り（E3）と doc 外語義（E4）のいずれかであった．E3 と E4 について 4 群間の χ^2 検定を行ったところ，組合せ誤り（E3）には有意な差は認められなかった（$chi^2(3) = 0.355, p = 0.949$）が，doc 外語義（E4）については有意傾向のある差が見られた（$chi^2(3) = 6.99, p = 0.071$）．残差分析を行うと，SAR の補正済み残差が 2.09（補正済み $p = 0.096$）が最も大きく，次いで, large の補正済み残差が 1.97（補正済み $p = 0.096$）となっており，この 2 つで有意傾向のある差が見られた．これにより Transformer モデルは SAR よりも有意に誤答率が下回ること，および Transformer の中でも large モデルの誤答率が低いことが示された．

問題全体との正答率との関係性を考えると，正答率が高くなるに従って，選択肢中の一部が doc に含まれない単語で置き換えられた選択肢（doc 外語義）が誤って選ばれる頻度が減少している．BERT の事前学習により，表現間の大まかな意味的距離の推定性能が向上していると考えられる．一方，doc 中に含まれる単語句で置き換えられた選択肢（組合せ誤り）が誤って選ばれる頻度は減少しなかった．position embedding のみでは，単語間の係り受け関係など，やや深い関係性の表現は難しいと考えられる．

5 おわりに

本研究では，センター英語試験で出題される意見要旨問題に対し，Transformer ベースの手法の適用方法について比較を行った．実験の結果，BERT-large モデルの正答率が比較したモデル・手法の中で有意に高いことが示された．また，query と option をを doc 中の正しい箇所に埋め込んだテキストを入力として，BERT-large モデルで転移学習を行う方法が，本稿の試行中で最も高い正答率（0.684）を示していた．誤答と判定する根拠と正答率の関係を調べたところ，正答率が高くなるに従って，選択肢中の一部が doc に含まれない単語で置き換えられた選択肢が誤って選ばれる頻度が減少している一方，doc 中に含まれる単語で置き換えられた選択肢が誤って選ばれる頻度は減少しないことが示された．後者の問題を解決するため，position embedding で表現できる単純な位置情報に加え，係り受け関係などの利用を検討していく．合わせて類似タスクの利用や疑似問題の作成による，転移学習に利用可能なデータの拡充に取り組む．

謝辞

本研究の推進にあたり，大学入試センター試験問題のデータをご提供下さった独立行政法人大学入試センターおよび株式会社ジェイシー教育研究所，ならびに実験データをご提供くださいました学校法人高宮学園，株式会社ベネッセコーポレーションに感謝いたします．

参考文献

[1] Radford Alec, Narasimhan Karthik, Salimans Tim, and Ilya Sutskever. Improving Language Understanding by Generative Pre-Training. *arXiv:1802.05365*, 2018.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805*, 2018.

[3] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. RACE: Large-scale ReAding Comprehension Dataset From Examinations. In *Proc. EMNLP*, pp. 785–794, 2017.

[4] 新井紀子, 東中竜一郎（編）. 人工知能プロジェクト「ロボットは東大に入れるか」: 第三次 AI ブームの到達点と限界. 東京大学出版会, 2018.

[5] 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, 喜多智也, 南泰浩, 風間健流, 大和淳司. 「ロボットは東大に入れるか」プロジェクトの英語における意見要旨把握問題の解法. 人工知能学会全国大会, pp. 2C1–02, 2018.