読解能力テストに対するニューラル言語モデルを用いた自動解答及びその結果の分析

青木拓磨 [1*] 原田裕文 [1*] 三浦大輝 [1*] 新井紀子 [2] 松崎拓也 [1]

[1] 東京理科大学 理学部第一部 応用数学科 [2] 教育のための科学研究所

{1417002, 1417086, 1417100}@ed.tus.ac.jp, arai@s4e.jp, matuzaki@rs.tus.ac.jp

1 はじめに

リーディングスキルテスト（RST）とは，「読解」のプロセスとして 11 の段階を想定し，それらを実践する力を 7 つの問題タイプで多面的に測ることで，読解能力を測定するテストである [1, 2]．中高生を中心に，小学生・成人も含め，これまで延べ約 20 万人が受検している．RST は項目反応理論 [3] に基づき設計され，各問題の困難度が推定される．また，困難度が推定済みの問題への反応に基づき受検者の各問題タイプにおける能力値が推定される．

近年，多数開発されている機械読解や「自然言語理解」のベンチマークデータでは人間が解いた際の精度を測定し，機械による解答精度と比較することが多い．しかしほとんどの場合「人間が正解した」かどうかは「被験者の 80 %以上が正解した」，といった大まかな定義に基づいており，かつ，RST のように多数の人間による解答結果が存在するものは我々が知る限り存在しない．

そこで，言語処理による自動解答結果と人間の RST 受検者の結果を比較することで，言語処理技術の課題と人の読解の特性の両方についての知見が得られることが期待できる．Arai ら [1] は，RST の係り受け解析問題（DEP）を CaboCha [4] で解析し，出力された係り受け構造と整合する選択肢を人手で選ぶことで 66% の正答率を得ており，能力値に換算すると受検者の最頻値をやや下回るものとなったことを報告している．また，最も誤りの多い問題文の特徴として，動詞句の並列が含まれる事を挙げている．一方，本論文では，係り受け解析の他に，照応解決・同義文判定についても取り組む．また，全ての解答器は BERT [5] を用いており，解答の選択まで全てのプロセスを自動的に行う．

＊ 同等の貢献

2 RST の問題例

本論文では，RST の 7 タイプの問題のうち「係り受け解析」，「照応解決」，「同義文判定」を対象とする．係り受け解析は文の構造を正しく把握する能力，照応解決は指示詞やゼロ代名詞が指すものを正確に特定する能力，同義文判定は 2 つの文章が同じ意味かどうかを判定する能力を測る．図 1〜図 3 に，これら 3 タイプの問題の例を示す．問題例は全て公式サイト [6] からの引用である．図中の提示文・補助提示文・選択肢が解答器への入力である．

3 手法

本研究では，日本語テキストで Pre-training された BERT を，各タスクで Fine-tuning することで解答器を作成する．以下，各問題タイプに対する解答器の概要を述べる．

3.1 係り受け解析

RST の係り受け解析問題を，選択肢を空欄に挿入した補助提示文を提示文に連結したものを選択肢の数だけ作り，そのうち 1 つを選ぶ課題として定式化する．具体的には，提示文を text1，選択肢を挿入した補助提示文を text2 として [CLS] + text1 + [SEP] + text2 + [SEP] を各選択肢について用意してそれぞれにスコアを与え，そのうち最もスコアが高いものを解答とする．スコアは，それぞれを BERT に入力したときの [CLS] に対する出力ベクトルと，パラメータベクトルの内積とする．

既存の問題は学習データとしては少量であったため，以下のように疑似問題を生成した．まず，RST 問題は，1〜3 文で構成されているため，コーパスから連続する 3 文以下の部分を抜き出し，提示文として用いた．補助提示文は，提示文が例えば「…A が

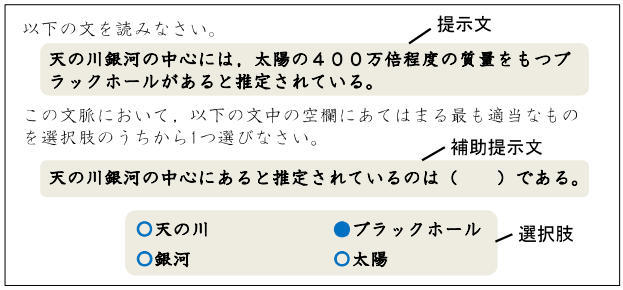

図 1　係り受け解析問題の例 (正解:ブラックホール)

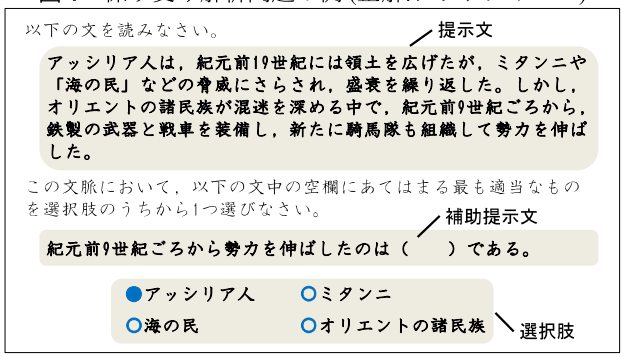

図 2　照応解決問題の例 (正解:アッシリア人)

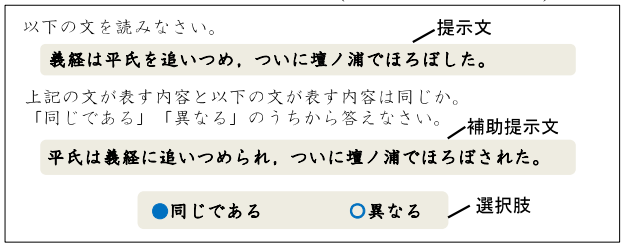

図 3　同義文判定問題の例 (正解:同じである)

B を C に V する．…」という形であるとき，

B を V するのは (　　) である．

C に V するのは (　　) である．

(いずれも正解は「A」) などを，ルールに基づいて生成した．負例となる選択肢は，提示文内の名詞から正解を除く 3〜7 個をランダムに選んで用いた．

3.2　照応解決

植田ら [7] の論文中の「ベースモデル」を用いて，ゼロ照応解析を含む述語項構造解析と照応解析の同時学習を行った．以下，その概要をまとめる．

まず，入力文書を形態素分割した後，WordPiece を用いてサブワードに分割する．次に，入力系列の先頭に [CLS]，末尾に [SEP] と 5 つの特殊トークン [著者]・[読者]・[不特定:人]・[NULL]・[NA] を挿入し，BERT に入力する．[著者]・[読者]・[不特定:人] はそれぞれ外界の照応先に対応し，[NULL] と [NA] はそれぞれ，述語項構造解析において項が存在しない場合と，照応解析において照応先が存在しない場合を表す．予測の際は，次の計算を行う．述語項構造解析では，述語と文書中の全サブワードに対し，ガ格・ヲ格・ニ格のスコアを計算し，スコア最大の項をその述語の項として出力する．より正確には，述語サブワード p_i に対し，項候補サブワード a_j が格 c の項になる確率を，

$$P(a_j|p_i,c) = \frac{\exp(s_c(a_j,p_i))}{\sum_k \exp(s_c(a_k,p_i))}$$

$$s_c(a_j,p_i) = \boldsymbol{v}^T \tanh(W_c \boldsymbol{a}_j + U_c \boldsymbol{p}_i)$$

と定義する．ここで，$\boldsymbol{p}_i, \boldsymbol{a}_j$ はサブワード p_i, a_j に対応する BERT 最終層のベクトル，$\boldsymbol{v}, W_c, U_c$ は Fine-tuning で新たに導入されたパラメータを表す．なお，述語サブワード及び項候補サブワードとしては形態素の先頭サブワードを用いる．照応解析の場合は，先行詞サブワードと照応詞サブワードに対して同様の計算を行う．

RST 問題に解答器を適用する際には，まず問題に対するパターンマッチによってゼロ照応かそれ以外かに分ける．ゼロ照応の場合は，はじめに補助提示文を係り受け解析し，空欄部分を項とする述語を抽出する．次に，各選択肢を形態素分割し，最右の名詞を抽出する．抽出した述語サブワードと各選択肢サブワードに対し，ガ格・ヲ格・ニ格のスコアを計算する．このとき，各スコアが [NULL] に対するものより小さい場合，候補から除外する．最後に，格ごとに softmax で正規化し，スコア最大の選択肢を解答とする．照応解析では，問題文から指示詞を抽出し，指示詞サブワードと各選択肢サブワードに対して同様の計算を行う．

3.3　同義文判定

同義かどうかを判定する 2 つの文章を text1，text2 とするとき，[CLS] + text1 + [SEP] + text2 + [SEP] を BERT に入力し，[CLS] に対する出力ベクトルとパラメータベクトルの内積が閾値以上であれば「同じである」，そうでなければ「異なる」と判定する．Fine-tuning に使用したデータセットは，京都大学テキストコーパス (KTC) ver. 4.0 と日本語 SNLI (Stanford Natural Language Inference) [8] [9] である．

KTC から 2 つのタイプの疑似問題を作成した．1 つ目は，文末の動詞の文節に係る複数の文節を，それらを根とする部分木とともにランダムに入れ替えたものを元の文とペアにし，正解を「同じである」としたものである．さらに入れ替え後の文からランダムに名詞を 2 つ選択して交換したものを元の文と

ペアにし，正解を「異なる」とした問題を作成した．

2 つ目のタイプの問題を作る際は，最初に「A は B を V した」のように，文末の述語がヲ格を持つ動詞である文を抽出した．そして 1 つ目と同様，文末の文節に係る文節をランダムに入れ替えた．最後に「A は B を V した」を「B は A によって V された」と書き換え，元の文とペアにしたものを正解「同じである」の問題とした．さらに 1 つ目と同様に，受動態に変換した上で，さらにランダムに名詞を交換したものを正解が「異なる」の問題とした．

日本語 SNLI は含意関係認識の英語のデータセットである SNLI を機械翻訳によって日本語化したデータである．このデータは多数の文のペアから成り，各ペアに対して Entailment (含意する)，Contradiction (矛盾する)，Neutral (どちらでもない) の 3 つのラベルのいずれかが付与されている．Entailment は，正確には同義とは異なるがそれに近いものと考え「同じである」の例とした．Contradiction は，必ず「同じでない」を意味するため「異なる」の例とした．

4 実験設定

全ての解答器で，東北大学によって作成された日本語 BERT 事前訓練済みモデル[1]を利用した．

係り受け問題に対する解答器の訓練データとして，NAIST テキストコーパス ver. 1.5（NTC1.5）[10] から 3.1 節で説明した方法で，129,256 個の疑似問題を生成した．そのうち 9 割を訓練データ，残り 1 割を開発データとして用いた．開発データのスコアは 0.8255 となった．

照応解決問題に対する解答器の訓練にも NTC1.5 を使用した．データは既存研究で広く用いられている Taira ら [11] の分割に従って訓練・開発・評価データに分割した．述語項構造解析は，用言及び事態性名詞を対象とする述語として訓練した．照応解析は，全ての指示詞を対象とした．NTC1.5 を用いた評価の際には，システムが出力したサブワードが，正解サブワードと同じ共参照クラスタのいずれかに含まれる場合は正解とした．損失関数には Cross-Entropy を使用し，異なる初期値で 20 エポック学習させた．表 1 に，既存モデルとの F1 スコアでの比較を示す．評価対象とした述語の数が若干異なるため厳密な比較ではないが，格解析・ゼロ照応解析ともに既存研究と比べ精度が向上している．な

1 https://github.com/cl-tohoku/bert-japanese

表 1 NTC1.5 評価セットにおける既存研究との比較

	格解析	ゼロ照応解析	（指示表現を含む）照応解析
本研究	94.32(±0.06)	63.06(±0.31)	52.25(±1.01)
松林ら [12]	90.07	54.53	-
今野ら [13]	87.72	47.72	-

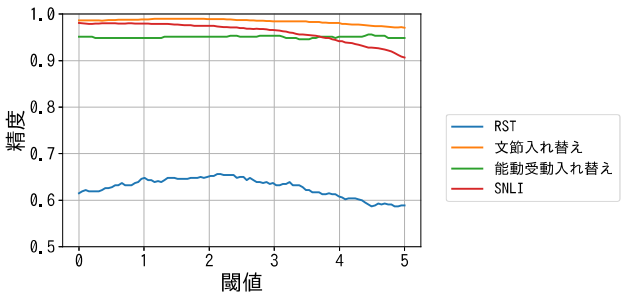

図 4 同義文判定の閾値と開発データに対する精度

お，照応解析については，指示表現のみを扱う先行研究がなかったため，ここでは比較しない．RST 問題を用いたテストには，異なる初期値による結果のうち NTC1.5 の評価データに対する精度が最も高いものを用いた．

同義文判定問題に対する解答器の訓練には，3.3 節で述べた通り KTC 及び日本語 SNLI を用いた．KTC から作成した疑似問題は，1 つ目のタイプが 57,155 問，2 つ目のタイプが 5,954 問であり，日本語 SNLI の Entailment が 177,257 問，Contradiction が 179,842 問である．これらの 99% を訓練データ，残りの 1% を開発データに用いた．図 4 は閾値と疑似問題及び RST 問題（いずれも開発データ）に対する精度の関係を示す．テストには RST の開発データに対する精度が最も高かった閾値 2.2 を用いた．

5 実験結果

表 2 に，RST 問題に対する各解答器の正答率を示す．開発データとテストデータは作成時期の異なる問題セットを用いており，これらに対する正答率の差は問題傾向の違いによるものと考えられる．図 5 に，問題の困難度と解答器の正答率の関係を示す．グラフ上の点 (x, y) は，困難度が x に近い問題に対する解答器の正答率が，おおよそ y であることを表す．正確には以下の様にグラフを作成した．まず，問題セットを困難度の順に並べたものを $q_1, q_2, \ldots, q_N$ とする．そこから連続する w 問のウィンドウ $q_i, q_{i+1}, \ldots, q_{i+w-1}$ を取り出し，これらの問題の困難度の平均値を x_i，これらに対する解答器の正答率を y_i とする．これを $i = 1, 2, \ldots$ につい

表 2　RST 問題に対する解答器の正答率

	係り受け	照応解析	同義文判定
開発	0.72 (225/314)	0.72 (140/195)	0.66 (300/457)
テスト	0.65 (127/196)	0.60 (114/189)	0.49 (36/ 73)

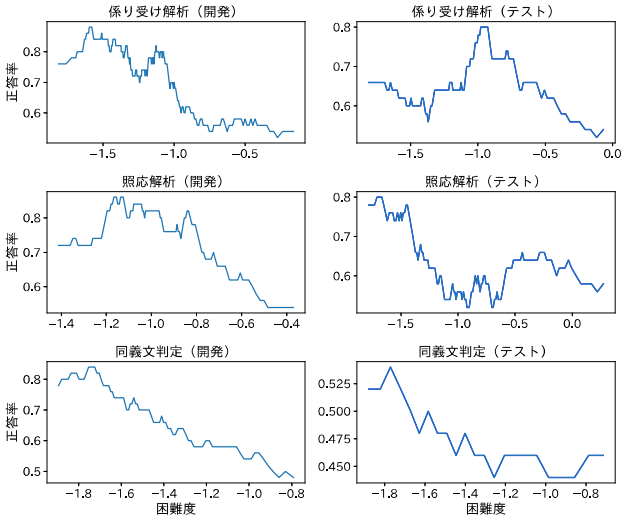

図 5　問題困難度と解答器の正答率の関係

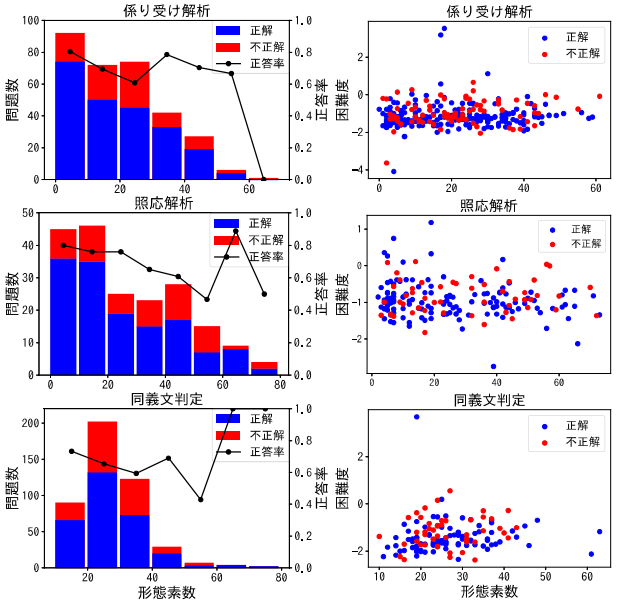

図 6　問題の形態素数と解答器の正答率及び困難度の関係

て行って得た座標 $(x_1, y_1), (x_2, y_2), \ldots$ を折れ線でつないで表示した．ウィンドウ幅 w は同義文判定のテストデータについては $w = 25$，それ以外については $w = 50$ とした．図から，係り受け解析のテストセットを除き，困難度が大きいほど解答器の正答率が低い傾向があること，すなわち人間にとって難しい読解課題は解答器にとっても難しいことが分かる．係り受け解析のテストデータでは困難度 -1.00 付近に正答率のピークが見られ，より易しい問題に対して正答率が低くなっている．原因として，訓練に用いた疑似問題ではカバーされないタイプの問題で，人にとっては易しいものがテストデータに多く存在したことが予想されるが，正確な理由は不明である．

図 6 に問題の形態素数と解答器の正答率及び困難度の関係を示す．ここでの問題の形態素数とは，係り受け解析と照応解析では正解選択肢と述語の距離，同義文判定では 2 つの文章の形態素数の平均を意味する．図 6（左）から，係り受け解析を除き，形態素数が多いほど正答率が低い傾向があること，すなわち解答器にとっての難しさの 1 つの要因は形態素数であることが示唆される．一方，図 6（右）から，形態素数と人にとっての困難度には明確な相関がないことがわかる．すなわち人にとっての困難度は形態素数に依存しておらず，本研究で用いた解答器と人の能力の特性の違いの 1 つだと言える．

6　おわりに

本研究では，RST の問題に対する自動解答結果と人間の受検者の結果を比較することで，言語処理技術の課題と人間の読解の特性について検討した．その結果，人間にとって難しい読解問題は解答器にとってもおおむね難しい傾向があることが分かった．一方で，形態素数と解答器の正答率には負の相関があるのに対し，人にとっての困難度とは相関がみられなかった．このことから，人に比べ BERT に基づく解答器はより表層的な手掛かりに依存する度合いが大きいことが示唆される．係り受け解析と同義文判定では，疑似問題と RST 問題に対する精度に乖離がある．これは，疑似問題が RST の問題をカバーしきれていないことと，RST に比べ簡単すぎることが理由と考えられる．疑似問題の生成方法を改良し，解答器の精度を向上させることで，人の読解とのより詳細な比較が可能になることが期待されるが，難しい疑似問題を生成すること自体が挑戦的な研究課題となるだろう．一方，照応解決は，解答器の精度が NTC1.5 に対する現在の最高精度に近く，かつ，RST 問題に対する精度との乖離も比較的小さいことから，照応解決技術自体の精度を向上させることが主要な課題となるだろう．

謝辞

本研究は JST，さきがけ，JPMJPR175A および JSPS 科研費 JP16H01819 の支援を受けたものである．

参考文献

[1] Noriko H. Arai, Naoya Todo, Teiko Arai, Kyosuke Bunji, Shingo Sugawara, Miwa Inuzuka, Takuya Matsuzaki, and Koken Ozaki. Reading skill test to diagnose basic language skills in comparison to machines. *Proceedings of the 39th Annual Cognitive Science Society Meeting (CogSci 2017)*, pp. 1556–1561, 2017.

[2] RST パンフレット. https://www.s4e.jp/wysiwyg/file/download/1/1157.

[3] F. M. Lord and M. R. Novick. *Statistical theories of mental test scores*. Addisom-Wesley, 1968.

[4] 工藤拓, 松本裕治. チャンキングの段階適用による日本語係り受け解析. Vol. 43, No. 6, pp. 1834–1842, 2002.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019.

[6] 教育のための科学研究所, 2020-12 閲覧. https://www.s4e.jp.

[7] 植田暢大, 河原大輔, 黒橋禎夫. BERT と Refinement ネットワークによる統合的照応・共参照解析. 言語処理学会 第 26 回年次大会 発表論文集, pp. 1101–1104, 2020.

[8] 吉越卓見, 河原大輔, 黒橋禎夫ほか. 機械翻訳を用いた自然言語推論データセットの多言語化. 研究報告自然言語処理 (NL), Vol. 2020, No. 6, pp. 1–8, 2020.

[9] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *EMNLP*, 2015.

[10] 飯田龍, 小町守, 井之上直也, 乾健太郎, 松本裕治. 述語項構造と照応関係のアノテーション:NAIST テキストコーパス構築の経験から. 自然言語処理, Vol. 17, No. 2, pp. 25–50, 4 2010.

[11] Hirotoshi Taira, Sanae Fujita, and Masaaki Nagata. A Japanese predicate argument structure analysis using decision lists. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 523–532, 2008.

[12] 松林優一郎, 乾健太郎. 複数の述語間の関係を考慮した End-to-End 日本語述語項構造解析. 言語処理学会 第 24 回年次大会 発表論文集, pp. 101–104, 2018.

[13] 今野颯人, 松林優一郎, 大内啓樹, 清野舜, 乾健太郎. 前方文脈の埋め込みを利用した日本語述語項構造解析. 言語処理学会 第 25 回年次大会 発表論文集, pp. 53–56, 2019.