# Gated-Attention Readerを用いた英語意見要旨把握問題の自動解答

喜多智也　　平 博順
大阪工業大学 情報科学部
e1b14036@st.oit.ac.jp, hirotoshi.taira@oit.ac.jp

## 1 はじめに

「ロボットは東大に入れるか」(東ロボ) プロジェクト [1] において，我々は英語科目について自動解答システムの研究に取り組んでいる[1]．これまで主に取り組んできた1次のセンター試験問題英語科目においては，模擬試験で受験者平均を上回る自動解答システムを実現してきた [2]．2016年度のシステムでは，発音・アクセント問題ではほぼ全問正解が可能で，英語の文法・語法についての知識が問われる語句整除問題など英語の単文を対象とした問題については9割以上の正解率をマークしている．しかしながら，東大に合格するためには少なくとも英語全体で8割以上の正答率が必要であり，これまで正答率が伸び悩んでいる2文以上が含まれる対話文完成問題や長文読解問題においても高い正解率を得る必要がある．

本研究では，これまでの研究で正答率が伸び悩んできた問題形式の1つである「意見要旨把握問題」について，Gated-Attention Readerと呼ばれるニューラルネットワークの手法を用いることで，従来よりも10ポイント以上の大幅な正解率の向上が見られたため，これについて報告する．

## 2 英語意見要旨把握問題の概要と従来手法

英語科目における意見要旨把握問題は，複数人でやり取りされている議論の中で，特定の人物の意見について，要約となっている選択肢を選ぶ問題であり，センター試験では毎年3問程度出題されている．意見要旨把握問題の例を図1に示す．

この例では，ロジャーとケビンが「優れたリーダーが持っている特性」について議論を行っている．ケビンが意見を述べている7つの発話に対して，ロジャーがその意見を要約したものが空欄を含む文となっており，この空欄を埋める選択肢を選ぶ問題になっている．この問題では，ケビンの意見が要約された「リーダーはチームメンバのスキルを向上させるようにチームを助けるべきである」という意見が書かれた1番の選択肢が正答となる．

単語だけに注目すると，誤答の選択肢の方がむしろ本文中で使用されている単語が使用されているため，単語だけに注目した単純な手法では正解を得るのが困難な問題となっている．

これまで，我々が精度評価のベンチマークに使用している，センター試験および模擬試験の過去問に対する評価実験では，リカレントニューラルネットワークによる手法でも高い精度は得られず，word2vec[3] を用いて，1文単位で本文と選択肢の間の文間類似度を計算する手法が最も高い精度が得られていた．しかし，このword2vecを用いる手法でも高々正解率は40%で更なる性能向上が求められていた [2]．

[1]2017年度から全体精度の報告会は開催されないことになったが，プロジェクトが終了した訳ではなく，各教科で研究は着実に進められている．

## 3 提案手法

上記のword2vecを用いた手法では，たまたま複数文の内の1文に選択肢の文とよく似た文が含まれていて，単語レベルでの言い換えが行われている場合は正答を得られやすいが，意見が複数の文に分散しれ述べられていたり，単語を越えるレベルでの言い換えが行われている場合には，正答するのが難しいという問題があった．また，様々なトピックが出題される入試問題について，複数文の文脈を学習するには，センター試験や模擬試験の過去問データだけでは圧倒的に訓練データが不足しているという問題があった．

そこで本研究では，複数文の文脈を捉えるためにGated-Attention Reader (GAR) [4] と呼ばれるニューラルネットワークのモデルを用いるとともに，学習のための訓練データとして，既存のベンチマークのデータセットに加えて，RACEデータセットと呼ばれる大規模な英語テスト問題のデータセットを用いて学習を試みた．

### 3.1 Gated-Attention Readerを用いた英語意見要旨把握問題の自動解答

Gated-Attention Reader (GAR) モデル [4] は，アテンション機構を持つリカレントニューラルネットワーク (RNN) の一種であり，いくつかの穴埋め型のReading Comprehensionタスクにおいて，高い精度が得られており注目されている．以下，GARモデルを用いた意見要旨把握問題の自動解答手法の概要を説明する．

#### 3.1.1 問題本文と質問文のエンコード

まず，問題本文 $D$ と質問文 $Q$ について，文を構成する単語を単語ベクトルに変換し，それぞれ $X^{(0)} = (x_1^{(0)}, x_2^{(0)}, ..., x_{|D|}^{(0)})$，$Y = (y_1, y_2, ..., y_{|Q|})$ とする．ここで $|D|$ と $|Q|$ は問題本文および質問文の単語長である．

次に，双方向GRU[5] を用いて，問題本文および質問文の単語ベクトルをエンコードする．

$$\begin{aligned} D^{(k)} &= \mathrm{BiGRU}_{\mathrm{D}}^{(\mathrm{k})}(X^{(k-1)}) \\ Q &= \mathrm{BiGRU}_{\mathrm{Q}}(Y) \end{aligned}$$

このときk層目の $D^{(k)}$ に関しては $k-1$ 層目のベクトル $X^{(k-1)}$ をエンコードする．

そして，この $D^{(k)}$ と $Q$ とを用いて，

$$\alpha_i = \mathrm{softmax_i}(Q^T d_i) \quad (1)$$

Roger: Thank you, Brian. That's a useful piece of information. Do you have anything to add, Kevin?
Kevin: Sure. Another key feature is that a good leader always tries to help others improve themselves. What I mean is, a leader breaks down work projects into tasks that encourage and challenge team members to develop existing or new skills. Thus, the team members control their own professional development and take pride in their own work. One technique leaders can employ to support workers is to ask questions about the task, rather than telling them how to perform the task. By doing this, workers can notice what is needed to complete a task efficiently and feel proud of their own success.
Roger: So , your main point is that a leader should [ ] .
Roger: OK. Let's take a few questions from the audience before we move on to the next part of the discussion. Does anyone have something to ask the speakers?

選択肢 (1) **assist the team to develop their own skills**
(2) encourage the team to ask questions
(3) feel controlled by the team members
(4) take pride in his or her questioning skills

図 1: 意見要旨把握問題の例 (2015 年度センター試験，追試験，問 3C より)

$$\tilde{q}_i = Q\alpha_i$$
$$x_i^{(k)} = d_i \odot \tilde{q}_i$$

を求める．この機構は「ゲート付アテンション」と呼ばれており，

$$X^{(k)} = \mathrm{GA}(D^{(k)}, Q^{(k)})$$

で表す．このようにして第 $k$ 層の双方向 GRU の入力である $X^{(k)}$ を得，順次，次層のパラメータを計算する．ここで，$d_i^{(k)}$ は $D^{(k)}$ の $i$ 番目の要素，$x_i^{(k)}$ は $X^{(k)}$ の $i$ 番目の要素を表す．また，$\odot$ はベクトルの要素同士の積を表す．

最後に，選択肢 $i$ が正解である確率 $p_i$ の計算は, Lai らが 4 択 RC タスクで用いた方法 [6] を参考にして以下のように計算した．

まず，4 つの選択肢それぞれについて，双方向 GRU でエンコードする．そして最終時刻でのこの双方向 GRU の隠れ層をそれぞれ，$h^{o_1}, h^{o_2}, h^{o_3}, h^{o_4}$ と書く．

次に，質問文に特化した問題本文のベクトル表現 $s^d$ を Chen らの研究 [7] にならい，双線型アテンション機構を用いて次の式で計算する．

$$\alpha_i = \mathrm{softmax_i}((X_i^{(k)})^T W_1 q^{(k)}) \quad (2)$$
$$s_d = \sum_i \alpha_i X_i^{(k)}$$

ここで，$q^{(k)}$ は $Q^{(k)}$ の最終時刻での隠れ層である．

最後に，選択肢 $i$ が正解である確率 $p_i$ を次の式で計算する．

$$p_i = \mathrm{softmax_i}(h^{o_i} W_2 s^d)$$

ここで，$W_1, W_2$ は学習可能なパラメータである．

このように，GAR モデルは，各隠れ層で毎回アテンション機構を利用する「マルチホップ機構」と，アテンションにゲート機構が組み込まれた「ゲート付アテンション」を持っているのが特徴であり，複数の箇所にアテンションを考慮することが可能となっている．

### 3.2 RACE データセット

問題本文が与えられ，それに対するいくつかの質問に答えるタスクは，Reading Comprehension (RC) タスクと呼ばれている．RACE データセット [6] は，最近公開された大規模な RC タスクのデータセットである．

中国の中高生向けの英語のテストを大量に収集したもので，約 28,000 の本文文章と 100,000 問を超える質問文で構成されている．質問文は，学生の言語能力を測る目的で，英語の専門家によって作成されており，問題の質が高いと言われている．各本文は，平均 300 語程度の単語から構成されており各本文に対しては，4 択の質問が 5 問出題されている．

RACE データセットに含まれる問題を分析したところ，5 割弱の問題が意見要旨把握問題と似た問題になっていた．RACE データセットに含まれる問題には，5W1H を問うような問題をはじめ，文章に適切なタイトルを付与する問題 (Title)，文章中で言っていないことを答える問題 (False)，“How many” 質問される量を答える問題 (Quantity)，単語や句の意味を答える問題 (Meaning) など様々なタイプの問題が含まれている．図 2 にタイプ別の内訳を示す．

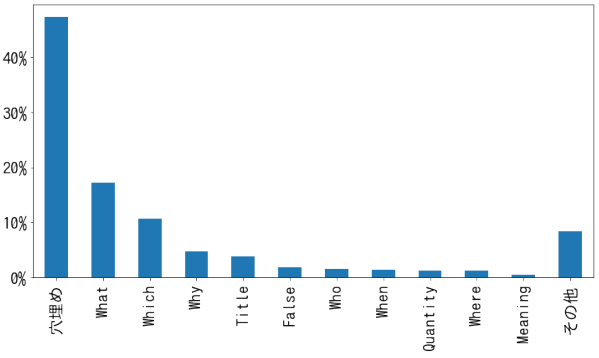


図 2: RACE データセットの問題形式の比率

データが大規模であるため，RACE データセットの訓練データを, 意見要旨把握問題の自動解答についてのモデル学習の訓練データの一部として用いることにした．

## 4 評価実験

我々がこれまでベンチマークとして用いてきたセンター試験およびセンター模試のデータの内，意見要旨把握問題に対し，Gated Attentive Reader (GAR) を用いた意見要旨把握問題自動解答手法について評価実験を行った．また，比較手法として従来法である word2vec を用いる方法，学習に GAR の代わりに非ニューラルネットワークの手法としては最近 SVM を上回る性能が得られるといわれている GBDT を用いた手法についても実験を行った．

### 4.1 実験データ

モデルの訓練にあたって RACE データセットの訓練セットを用いた．以下これを RACE と書く．

我々がこれまでベンチマークとして用いてきたセンター試験およびセンター模試のデータの内，意見要旨把握問題のみを抽出したデータを作成した．以下，本稿ではこのデータセットを「TR データセット」と呼ぶ．また，TR データセットを訓練とパラメータチューニングに用いる開発セット (TRdev) と精度評価のためのテストセット (TRtest) の 2 つに分けた．また，訓練データには上で述べた RACE データセットの訓練データも用いた．

これらのデータセットの問題数の内訳および統計情報を表 1，表 2 に示す．

表 1: 各データセットの問題数

| | RACE | TRdev | TRtest |
|---|---|---|---|
| 問題数 | 87,866 | 120 | 114 |

表 2: 各データセットの統計情報

| | RACE | TRdev | TRtest |
|---|---|---|---|
| 問題本文の文数 | 20.5 | 8.4 | 9.1 |
| 問題本文の単語数 | 369.0 | 128.1 | 141.0 |
| 質問文の単語数 | 11.7 | 9.3 | 9.2 |
| 選択肢文の単語数 | 7.0 | 9.8 | 10.4 |

表から分かるように，TR データセットと RACE データセットでは問題本文の文数と単語数に大きな開きがある．

### 4.2 実験設定

#### 4.2.1 提案手法（GAR モデル）

TRdev の半分の 60 問と RACE データを訓練データ，TRdev の残り半分の 60 問をパラメータチューニング用の開発データとして学習を行った．

GAR モデルの GRU の隠れ層は 128 次元，1 層とし，Dropout 確率は 0.5 に設定した．学習は SGD で行い，学習率は 0.3，gradient clipping は 10 に設定した．単語埋め込み層は，事前学習された Glove 単語埋め込み [8] を利用した 100 次元とした．100 エポックの学習を行い，開発セットの正解率が最も大きいモデルを選択した．

また，多言語ニューラル機械翻訳において，異なる言語データが混在している場合に，データの先頭にデータの種類を示すタグを付与することによって，翻訳精度の向上が見られている [9]．このテクニックが本研究でも有効であるかどうかを試すため，訓練データで，RACE 由来のデータと TR 由来のデータを区別するタグを先頭に付与したデータについても，実験を行った．

#### 4.2.2 従来手法

これまでの研究で最高精度が得られている word2vec による方法および Gradient Boosting Decision Tree (GBDT) を用いる方法を比較手法として実験した．GBDT は，ブースティングの一種で近年 Support Vector Machine を上回る性能を示すことで注目されている分類学習手法である．本研究では，非ニューラルネットワークの手法の例として，特徴量として Bag of Words, tf-idf の一致度, word2vec の類似度を用いた GBDT による分類学習についても実験を行った．なお，GBDT の実装には XGBoost[10] を用い，訓練データを RACE，開発セットとして TRdev 全部を用いて，最適なハイパーパラメータのグリッドサーチを行った．

### 4.3 実験結果

実験結果を表 3 に示す．+TRdev は意見要旨把握問題の開発セット 60 問を学習データに追加したことを表している．追加の学習データがないものは，開発セットとして TRdev の全ての問題を使用している．

表 3: 実験結果

| 手法 | 学習データ | TRdev | TRtest |
|---|---|---|---|
| word2vec | なし | 40.0 | 29.4 |
| GBDT | RACE | 37.5 | 29.0 |
| GAR | RACE | 50.8 | 45.6 |
| | RACE + TRdev | 55.0 | **55.3** |
| | RACE + TRdev (tag 付) | 58.3 | 49.1 |

学習データとして TRdev の 60 問を追加した時の効果を有意水準 5%でマクネマー検定した結果，有意差があった．学習データに追加する TRdev にタグを付与すると TRtest の正解率は 55.3 % から 49.1 % に悪化したが，有意水準 5%のマクネマー検定を行ったところ有意差は認められなかった．

## 5 アテンション機構の可視化

アテンション機構が有効に機能しているかを調べるため，式 (1) および式 (2) のアテンション機構についての可視化を行った．可視化の対象の問題は図 1 の 2015 年度センター試験の追試験の大問 3C 問 3 である．この問題はテストセットに含まれており，GAR モデルを使用して正解だった問題である．

### 5.1 Gated-Attention の可視化

式 (1) の Gated-Attention を可視化したものを図 3 に示す．横方向が問題本文，縦方向が質問文で，文頭はそれぞれ左と上である．白に近いほど強い注意がかかっていることを表している．これを見ると質問文の文頭と文末に多く注意が集まっていることが分かる．

質問文の文頭に対しては，問題本文の文頭に注意が集まっているが，意見要旨把握問題の性質上，問題本文の始めは対話のトピックについての紹介や，前問との繋ぎの会話が入ることがほとんどであり，解答に必要のない場合が多い．

一方，質問文の文末の空欄やピリオドに対しては問題本文の主要な部分に多く注意が集中している．

### 5.2 BiLinear-Attention の可視化

式 (2) の BiLinear-Attention の可視化したものを図 4 に示す．赤に近いほど強い注意が集まっていることを表している．これをみると 解答の根拠となる “encourage”, “develop”, “existing”, “skills” に注意が集中しているのがわかる．このことから，この問題に関して

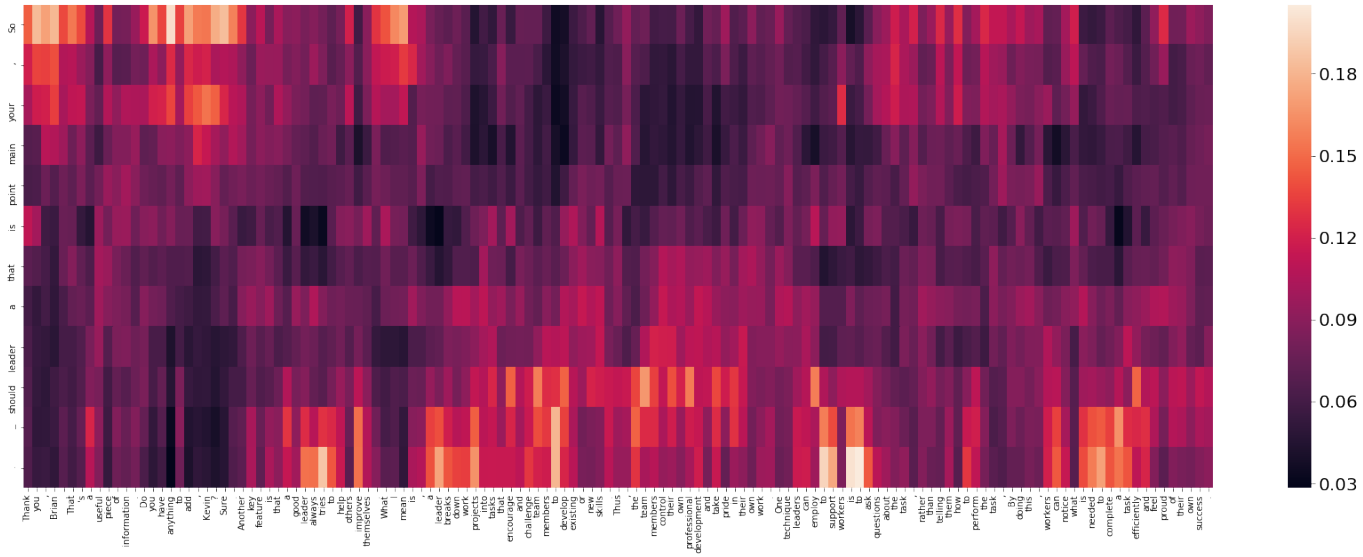


図 3: Gated-Attention の可視化

はアテンション機構が有効に働き学習が行われていると考えられる．

Thank you , Brian . That 's a useful piece of information . Do you have anything to add , Kevin ? Sure . Another key feature is that a good leader always tries to help others improve themselves . What I mean is , a leader breaks down work projects into tasks that encourage and challenge team members to develop existing or new skills . Thus , the team members control their own professional development and take pride in their own work . One technique leaders can employ to support workers is to ask questions about the task , rather than telling them how to perform the task , By doing this , workers can notice what is needed to complete a task efficiently and feel proud of their own success .

図 4: BiLinear-Attention の可視化

## 6 おわりに

本稿では，大学入試センター試験の英語科目においてこれまで正答率が低かった問題の一つである意見要旨把握問題について，Gated-Attention Reader モデルを用いた自動解答手法について述べた．大規模な英語のテスト問題から構築された RACE データセットを活用しながら，Gated-Attention Reader モデルにより文脈を詳細に考慮することで従来の手法より 10 ポイント以上の正答率が向上することを確認した．ニューラル翻訳の性能向上で有効であった訓練データ中の異なるデータ種類のタグ付与については，本研究では効果は見られなかった．今後は，GAR モデルについての詳細な実験を進め，より詳細な正解率改善の機構について解明したいと考えている．

## 謝辞

本研究を遂行するにあたり「『ロボットは東大に入れるか』大学入試センター試験関連オンラインタスクデータ」を利用しました．ご提供下さった「独立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」に感謝いたします．また，模擬試験データをご提供下さっ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，「ロボットは東大に入れるか」を推進している新井紀子教授をはじめ，国立情報学研究所の方々に深く感謝いたします．また，本研究の一部は以下の各氏（組織）との共同研究として行われました．東中竜一郎，杉山弘晃, 成松宏美 (以上 NTT), 菊井玄一郎，磯崎秀樹（岡山県立大），堂坂浩二（秋田県立大），南泰浩（電気通信大），大和淳司（工学院大学）．熱心な議論に感謝いたします．

## 参考文献

[1] Noriko Arai and Takuya Matsuzaki. The impact of AI on education - can a robot get into the university of tokyo? In *Proc. The International Conference on Consumer Electronics*, pp. 1034–1042, 2014.

[2] 東中竜一郎, 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. 2017 年度人工知能学会全国大会, 2017.

[3] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[4] Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William Cohen, and Ruslan Salakhutdinov. Gated-attention readers for text comprehension. In *Proc. of ACL 2017*, pp. 1832–1846, 2017.

[5] K. Cho, B. Van Merrinboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proc. of EMNLP 2014*, pp. 1724–1734, 2014.

[6] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. Race: Large-scale reading comprehension dataset from examinations. In *Proc. of EMNLP 2017*, pp. 785–794, 2017.

[7] Danqi Chen, Jason Bolton, and Christopher D. Manning. A thorough examination of the CNN/Daily mail reading comprehension task. In *Proc. of ACL 2016*, pp. 2358–2367, 2016.

[8] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, pp. 1532–1543, 2014.

[9] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, and M. Hughes. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association of Computational Linguistics*, Vol. 5, pp. 339–351, 2017.

[10] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785–794. ACM, 2016.