The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

「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

Solving the opinion summarization problem in English in the "Can a Robot Get into the University of Tokyo?" project

東中 竜一郎 *1 Ryuichiro Higashinaka
杉山 弘晃 *1 Hiroaki Sugiyama
成松 宏美 *1 Hiromi Narimatsu
磯崎 秀樹 *2 Hideki Isozaki
菊井 玄一郎 *2 Genichiro Kikui
堂坂 浩二 *3 Kohji Dohsaka
平 博順 *4 Hirotoshi Taira
喜多 智也 *4 Tomoya Kita
南 泰浩 *5 Yasuhiro Minami
風間 健流 *5 Takeru Kazama
大和 淳司 *6 Junji Yamato

*1NTT コミュニケーション科学基礎研究所 NTT Communication Science Laboratories
*2岡山県立大学 Okayama Prefectural University
*3秋田県立大学 Akita Prefectural University
*4大阪工業大学 Osaka Institute of Technology
*5電気通信大学 The University of Electro-Communications
*6工学院大学 Kogakuin University

We have been working on the English problems in the "Can a Robot Get into the University of Tokyo?" project. Focusing on the opinion summarization problem, we show that our method, which uses deep learning with a large number of English problems, can improve the accuracy over a Word2vec-based method. We perform qualitative analyses of the trained models and results, and discuss our findings with regards to machine comprehension.

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 [Arai 16] 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．[東中 17] では，英語問題を短文問題，複数文問題，長文問題に分け，問題文がほぼ一文からなる短文問題はおおむね解けるものの，複数の文からなる複数文問題は 30%〜40%程度，さらに長い長文問題についてはチャンスレベルの正解率であることを報告した．

本稿は，複数文問題のうち，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，各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．

[東中 17] では，我々は Word2vec[Mikolov 13] の類似度に基づく従来手法（提示意見と要旨候補の Word2vec に基づくコサイン類似度を用い，最も類似した要旨を正解とするもの）と深層学習ベースの手法を比較し，Word2vec の類似度に基づく手法のほうがよいとの報告を行った．本稿では，意見要旨把握問題の改善手法について述べる．具体的には，RACE と呼ばれる大規模な英語問題のデータセットを用いた深層学習の手法により，Word2vec の類似度に基づく手法よりも高精度に意見要旨把握問題が解けることを示す．

2. データセット

ここでは，意見要旨把握問題の自動解答に用いるデータセットについて説明する．本研究では，解法の性能を測るベンチマークデータとして，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 問を収集した．このうち 120 問を訓練・開発セット（dev），114 問をテストセット（test）として用いる．この意見要旨把握問題のベンチマークデータは，丁寧に作成された良質な問題である一方，そのまま機械学習の訓練データとして用いるには量が不足するという問題がある．そこで本研究では，RACE データセット [Lai 17] 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活用し，機械学習による解法を試みる．RACE データセットの特徴として，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．現在提案されている Reading Comprehension 関連のデータセットでは，分量自体は数十万問以上と非常に多いものの，そのほとんどが単語もしくは短い語句を解答させる問題であり，意見要旨把握問題との差異が大きい．そのため，こうした問題は，意見要旨把握問題を解く上での訓練データとして用いることは難しい．一方 RACE デー

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 first 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)．

連絡先: higashinaka.ryuichiro@lab.ntt.co.jp

タセットは，分量自体は 9 万 7 千問程度と中程度の大きさながら，選択肢に文を含む，作成が難しい問題を多数含んでいるため，意見要旨把握問題の自動解答に有用であると予想される．

3. 手法 : SAR/GAR を用いた意見要旨把握問題の自動解答

Stanford-Attentive Reader (SAR) モデル [Chen 16] および Gated-Attention Reader (GAR) モデル [Dhingra 17] 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．我々の RACE データセットと東ロボのベンチマークデータを用いた評価実験においてもこれまでにない高得点を得ることが確認されている [喜多 18]．本稿は同じ手法を用い，詳細な分析を行っているものである．

以下，本研究で使用した SAR/GAR モデルを用いた意見要旨把握問題の自動解答手法について GAR モデルを中心に述べ，後で SAR モデルとの違いについて説明する．

3.1 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 [Cho 14] 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．

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

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 $Q^{(k)}$ とを用いて，

$$\begin{aligned} \alpha_i^{(k)} &= \mathrm{softmax_i}(Q^{(k)T} d_i^{(k)}) \\ \tilde{q}_i^{(k)} &= Q^{(k)} \alpha_i^{(k)} \\ x_i^{(k)} &= d_i^{(k)} \odot \tilde{q}_i^{(k)} \end{aligned} \tag{1}$$

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 [Lai 17] と同じく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 を双線型アテンション機構 [Chen 16] を用いて次の式で計算する．

$$\begin{aligned} \alpha_i &= \mathrm{softmax_i}((X_i^{(k)})^T W_1 q^{(k)}) \\ s_d &= \sum_i \alpha_i X_i^{(k)} \end{aligned} \tag{2}$$

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．

SAR モデルの場合は, アテンションがゲート付アテンションではなく通常のアテンションであることが異なっている。

4. 結果

4.1 実験設定

3. 節で述べた GAR と SAR を用いて正解率を比較する．学習およびテストに用いるデータは，RACE の train データ 87860 問，大学入試センター試験の本試験と追試験，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 問である．RACE のデータについては，独自に表記誤り等を修正したデータを用いた．これらのデータを，学習データセット (train)，開発データセット (dev)，テストデータセット (test) に分け，開発データセットで正解率が最大になった epoch でのテストデータセットの正解率を比較する．

データセットの分け方は，次の 2 通りの設定を試した．1 つ目の設定では，RACE の train データを学習データセットに，意見要旨把握問題 234 問のうち，120 問を開発データセット，114 問をテストデータセットに分けた．2 つ目の設定では，RACE の train データ +120 問の開発データセットのうち 60 問を学習データセットとして，残りの 60 問を開発データセット，1 つ目と同じ 114 問をテストデータセットに分けた．

GAR および SAR におけるモデルの安定性を評価するために，いくつかのパラメータを変えて正解率を評価する．変動させたパラメータおよび設定した値は，隠れ層数 1 もしくは 2, ユニット数 128 もしくは 256，学習率 0.1 もしくは 0.3 である．それ以外のパラメータについては，固定値を用いた．dropout は一般に良く用いられる 0.5，ランダムシードは [Lai 17] においてデフォルトで設定されている 1013 を用いた．語彙サイズは，50000 とした．大学受験に必要な語彙数よりも多いが，意見要旨把握に良く出現する，人名等が未知語となり悪影響を与えることを懸念し，少し多めに設定した．また最適化に用いる勾配法は SGD を用い，epoch は 100 まで行った．ただし，2 層 256 ユニット数のモデルについては，時間の関係で epoch は 50 までで評価した．

また，従来手法として，意見文書の各文と選択肢を表す文を，Word2Vec [Mikolov 13](w2v) を用いて得られた単語ベクトルの集合を平均したベクトルで表現し，各文と選択肢の文のコサイン類似度を計算し，最も類似した選択肢を答えとする手法とのスコアを比較する．

4.2 結果

表 1 に結果を示す．dev acc は東ロボ dev の正解率，test acc は東ロボ test の正解率である．全体をみると，SAR において，1 層 256 ユニット数で学習率を 0.3 に設定した時には，正解率が 0.2 と著しく低い．本設定での epoch ごとの正解率をみると，RACE データおよび東ロボ dev/test 全てにおいて不動であり，学習が進んでいなかった．従って，学習できなかった本設定を除いて評価を行う．表下部の平均値においても，本設定を除いた正解率のみを用いて算出した．GAR および SAR に

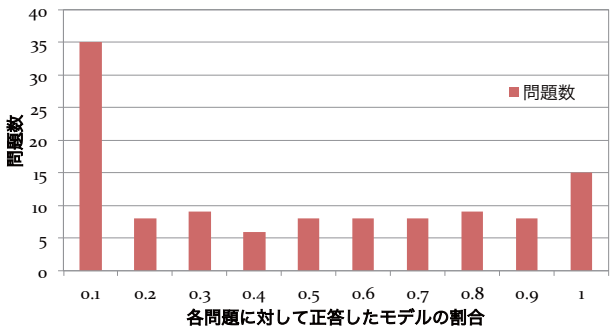

図 2: 正答したモデルの割合と問題数

おいて，それぞれ 16 個のモデルの結果を比較すると，test の正解率はいずれも 0.42±0.05 程度，0.44±0.05 程度に収まっており，モデルの違いによる正解率の差はほとんど見られなかった．従って，正解率の観点では，パラメータの変動によらず安定した結果が得られていると言える．

また意見要旨把握問題の dev の半分を学習データセットに含めた場合の結果を比較した場合，GAR では，意見要旨把握問題の dev を含めると，dev と test ともに 0.01 ポイント正解率が低下した．一方，SAR では，意見要旨把握問題の dev を含めない方が，dev で 0.03 ポイント，test で 0.01 ポイント正解率が向上した．このことから，学習に含めた意見要旨把握問題のデータセットは，RACE の train データの総量に対して，0.1% にも満たないが，SAR においては，意見要旨把握問題特有の特徴を何らか学習し，正解率の向上に寄与できた可能性がある．なお，w2v を用いた手法での同テストセットの正答率は，0.3 前後にとどまっていたことから，GAR/SAR で RACE データを学習することで，約 15 ポイント正解率が向上することがわかった．

4.3 モデルごとの性能比較

アテンションを用いる手法と w2v との正誤の分布を評価する．アテンションを用いる GAR および SAR については，モデルによる正答率のばらつきが小さい問題を対象として，正誤問題数の分布を評価した．図 2 は，横軸に正答したモデルの割合，縦軸にその問題数を表す．なお，正答したモデルの割合は，0.1 のとき，0 から 0.1 まで，0.2 のとき，0.1 から 0.2 までの正解率を含む．0.1 および 1 において，正答したモデルの割合が他と比べて高く，全体的な分布が凹型になっていることから，正答する傾向にある問題と誤答する傾向にある問題があることが分かる．

そこで．図 2 において，7 割以上のモデルが正答した問題を GAR および SAR で正解した問題，0 から 3 割のモデルが正答した問題を，GAR および SAR で誤答した問題として分類し，w2v との正答誤答問題数を調べた．その結果を表 2 に示す．GAR/SAR の 7 割のモデルが正答する問題 32 問のうち 20 問が w2v のモデルにおいて不正解であり，w2v で正答する 33 問のうち 16 問が GAR/SAR で不正解であった．我々は，これらの 20 問，16 問を分析することで，GAR/SAR におけるアテンションの効果および，アテンションのモデルに足りない特徴を調べることができると考えた．次節に分析を示す

5. 分析

5.1 定量評価

アテンションの効果を分析するために，アテンション機構の利用有無による正解率を比較する．アテンション機構を利用しない場合には，(2) の算出において $q^{(k)}$ を削除する．学習とテストの両者において利用する (有)，利用しない (無) のそれぞれの組み合わせについて評価する．分析の対象は表 1 において test の正解率が最も高い 0.48 であった 2 つのモデルとし，それぞれの正解率で評価する．結果を表 3 に示す．test の正解率は，All に示す．また，アテンションに寄与する質問文（意見要旨把握問題において空所のある文）に，正答する上で必要な情報が含まれているかどうかで問題を分類し，その正解率も比較した．質問文の重要度は，質問文が So, you say _. のように特に情報が含まれていない場合に「低」，それ以外を「高」とした．

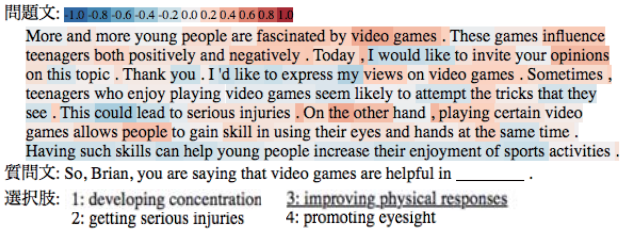

図 3: GAR で正答し，w2v で誤答した問題の例（2007 年のセンター試験 3B-1）．アテンションのかかり具合を濃淡で表している．

結果より，アテンション機構を用いて学習したモデルでは，アテンション機構を用いないと正解率が低下するが，アテンション機構を用いずに学習したモデルの正解率は，アテンション機構を用いて学習およびテストした結果と比較して若干の低下にとどまっていた．このことから，アテンション機構の効果は限定的と言える

5.2 定性評価

限定的な効果のアテンション機構であるが，今回若干の改善が見られてはいるので，それがどのように寄与しているかを，代表的な問題を 2 問抽出して確認する．アテンションを分析する対象として，GAR のモデルを用いる．図 3 は，w2v では不正解であったが GAR で正答した問題での GAR のアテンションを可視化したものである．この問題では，w2v では 2 を，GAR では正解の 3 を解答している．2 の選択肢には，問題文中と全く同じフレーズが出現しており，語の一致や表層的な近さで選ぶと誤答となる問題である．一方で，アテンションを用いた場合，質問文に含まれる video games が問題文中に出現する場合に高いアテンションがかかっており，いずれも正答の選択肢を選ぶ上でヒントになる箇所であることがわかる．従って，GA での正答は複数文をみることができたことによる効果である可能性が高い．

次に，w2v では正解であったが GAR で誤答した問題のアテンションを図 4 に示す．この問題では，w2v では正解の 2 を，GAR では 3 を解答している．表層的に近い文を選択するだけで正答できる問題であるが，GAR では good leader に関わる語にアテンションがかかっており，それらは本問題文中で頻繁に出現する語の一つであることから，アテンションが悪い影響を与えているように見える．

以上より，質問文中に，問題文中の注目箇所を制限するための重要な単語が含まれる問題に対しては，アテンションを用いる GAR 等の手法の効果が見られるが，問題文のトピックや問題文中に高頻度で出現する単語が含まれる問題に対しては，逆に悪影響となる可能性もあるとわかった．

表 1: 様々なパラメータで実験したアテンションモデルでの正解率の比較

パラメータ			GAR/RACE		GAR/RACE+dev		SAR/Race		SAR/Race+dev	
層数	ユニット数	学習率	dev acc	test acc	dev acc	test acc	dev acc	test acc	dev acc	test acc
1	128	0.1	0.52	0.4	0.53	0.45	0.54	**0.48**	0.6	0.4
1	128	0.3	0.52	0.43	0.53	0.38	0.54	0.45	0.63	0.46
1	256	0.1	0.51	0.39	0.55	0.37	0.52	0.46	0.57	0.47
1	256	0.3	0.52	0.43	0.55	**0.48**	0.25	0.2	0.25	0.2
2	128	0.1	0.54	0.42	0.52	0.43	0.58	0.43	0.57	0.44
2	128	0.3	0.57	0.38	0.60	0.46	0.54	0.44	0.52	0.42
2	256	0.1	0.48	0.42	0.51	0.46	0.51	0.39	0.52	0.4
2	256	0.3	0.46	0.38	0.47	0.36	0.52	0.43	0.52	0.45
Average			0.52	0.41	0.53	0.42	0.53	**0.44**	0.56	0.43

表 2: 手法毎の正答する問題数の分布

	正答モデルの割合	w2v		GAR/SAR の正誤割合
		正解	不正解	
GAR SAR	7 割以上	12	**20**	0.28 (32/114)
	3~7 割	5	25	0.26 (30/114)
	3 割以下	**16**	36	0.46 (52/114)
w2v の正誤割合		0.29 (33/114)	0.71 (81/114)	

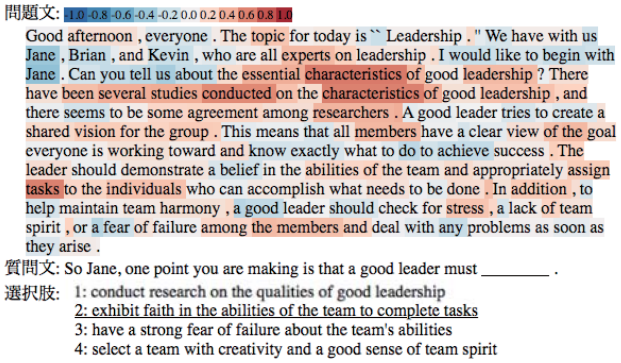

図 4: GAR で誤答し，w2v で正答した問題の例（2015 年のセンター試験追試 3C-1）

6. まとめ

本稿では，「ロボットは東大に入れるか」の英語における意見要旨把握問題の解法について説明した．深層学習と大規模な英語問題のデータセットを用いることで，Word2vec に基づく手法よりも高い正解率を達成した．具体的には，Word2vec に基づく手法は 30%の正解率だったところを，44%まで引き上げることができた．今後は，定性的分析やアテンション分析を通し，この正解率をさらに改善させる手法について検討していきたい．

謝辞

本研究を推進するにあたって，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に感謝いたします．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．

表 3: アテンションの効果の比較

	アテンション		dev	test		
	学習	テスト		All	質問文の重要度	
					高	低
GAR	有	有	0.55 (33/60)	0.48 (55/114)	0.45 (18/40)	0.5 (37/74)
	有	無	0.37 (22/60)	0.42 (48/114)	0.48 (19/40)	0.39 (29/74)
	無	無	0.53 (32/60)	0.45 (51/114)	0.43 (17/40)	0.46 (34/74)
SAR	有	有	0.54 (65/120)	0.48 (55/114)	0.45 (18/40)	0.5 (37/74)
	有	無	0.39 (47/120)	0.27 (31/114)	0.2 (8/40)	0.31 (23/74)
	無	無	0.54 (65/120)	0.46 (52/114)	0.43 (17/40)	0.49 (36/74)

参考文献

[Arai 16] Arai, N.: Todai Robot Project., in *NTCIR* (2016)

[Chen 16] Chen, D., Bolton, J., and Manning, C. D.: A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task, in *Proc. of ACL 2016*, pp. 2358–2367 (2016)

[Cho 14] Cho, K., Van Merrinboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y.: Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation, in *Proc. of EMNLP 2014*, pp. 1724–1734 (2014)

[Dhingra 17] Dhingra, B., Liu, H., Yang, Z., Cohen, W., and Salakhutdinov, R.: Gated-Attention Readers for Text Comprehension, in *Proc. of ACL 2017*, pp. 1832–1846 (2017)

[Lai 17] Lai, G., Xie, Q., Liu, H., Yang, Y., and Hovy, E.: RACE: Large-scale ReAding Comprehension Dataset From Examinations, in *Proc. of EMNLP 2017*, pp. 785–794 (2017)

[Mikolov 13] Mikolov, T., Chen, K., Corrado, G., and Dean, J.: Efficient estimation of word representations in vector space, *arXiv preprint arXiv:1301.3781* (2013)

[喜多 18] 喜多 智也, 平 博順 : Gated-Attention Reader を用いた英語意見要旨把握問題の自動解答, 第 24 回言語処理学会年次大会, D5-3 (2018)

[東中 17] 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, 2017 年度人工知能学会全国大会予稿集, pp. 2H2–1 (2017)