センター試験を対象とした高性能な英語ソルバーの実現

杉山弘晃[1] 成松宏美[1] 菊井玄一郎[2] 東中竜一郎[1]
堂坂浩二[3] 平博順[4] 南泰浩[5] 大和淳司[6]

[1]NTT コミュニケーション科学基礎研究所 [2]岡山県立大学
[3]秋田県立大学 [4]大阪工業大学 [5]電気通信大学 [6]工学院大学

1 はじめに

「ロボットは東大に入れるか」（以下，東ロボ）は，NII の新井教授をリーダーとし，センター試験や東京大学の 2 次試験の問題を解くことで，人工知能が，人間が実際に解く問題をどこまで解けるのかを明らかにすることを目的としたプロジェクトである [8].

東ロボ英語チームは，センター試験の英語問題を自然言語処理・知識処理の基礎研究を進めるベンチマークと捉え，センター試験に含まれる多様な英語問題に対する自動解答に関する知見を積み重ねてきた．英語チームが東ロボに参画した 2014 年にはセンター模試において 95 点を達成し，受験者平均を超える成績を達成した[1]．しかし，その後の点数は伸び悩んでいた．具体的には，一文程度の文章を対象とした短文問題は高精度で解けるが，5 文程度からなる複数文問題（例えば，不要文除去問題や意見要旨把握問題）や長文問題は，複雑な文脈情報を扱う必要などから，高得点を達成することが難しかった．

一方で，近年深層学習に基づく文書読解技術が急速に進展している．特に，BERT [1] や GPT [5] に代表される，極めて大規模なテキストデータによる事前学習されたモデルを，所定のタスクに Finetune する方法は多くの自然言語処理のタスクで高性能を収めるようになってきている．そういった技術の中でも XLNet [7] は近年多くのタスクで最もよい成績を収めている手法の一つである．XLNet は，RACE [3] と呼ばれる中国における中高生向けの英語問題のデータセットに対して，8 割を超えるスコアを達成している．

本稿では，東ロボ英語チームが開発した，高性能な英語ソルバーについて述べる．本ソルバーは XLNet（問題によっては BERT）を随所に用いている．加えて，問題（例えば，後述する不要文除去問題や段落タイトル付与問題）によっては，XLNet の単純適用では解くことが難しいものがある．我々は，そうした問題に対し，独自技術を適用することで，2019 年のセンター試験（本試験）において，185 点（200 点満点，偏差値は 64.1）という高得点を達成した[2].

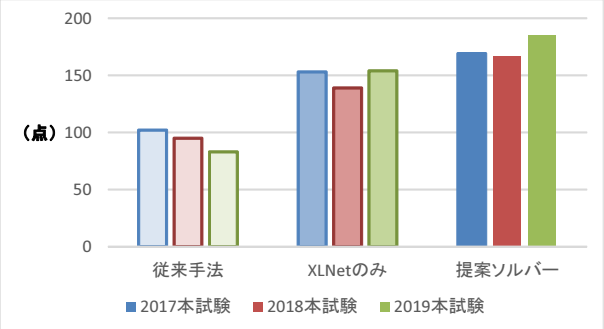

図 1: センター試験英語問題に対する成績変化

図 1に，ここ 3 年のセンター試験に対する，従来手法（英語チームの 2016 年度の手法 [12]），XLNet を単純適用した手法，提案ソルバーのスコアを示す．図から分かる通り，ここ 3 年のセンター試験で高得点が実現できている．偏差値についても，偏差値 60 以上をキープできている．問題別の点数については，表 1を参照されたい．

本稿では，東ロボ英語ソルバーにおける我々の工夫について述べる．具体的には，複数文問題および長文問題への BERT/XLNet の適用とその結果，疑似問題を用いた不要文除去問題の解法，組み合わせ最適化に基づく段落タイトル付与問題の解法について詳述する．

2 BERT/XLNet の適用

本節では，各問題（後述する不要文除去問題・タイトル付与問題は除く）に対する，BERT/XLNet の適用結果について述べる．表 2に，各問題について，提案ソルバーの用いるアルゴリズム，Finetune に用いたデータ，従来手法，提案ソルバーと従来手法の 2017 年本追試・2018 年本追試・2019 年本試の合計点を示す．

[1]https://www.ntt.co.jp/news2014/1410/141030a.html

[2]https://www.ntt.co.jp/news2019/1911/191118a.html

2017 本試	中問	得点	満点	配点	正答数	問題数
発音	1AB	14	14	2	7	7
平叙文	2A	18	20	2	9	10
語句整序	2B	12	12	4	3	3
発話文生成	2C	8	12	4	2	3
会話文完成	3A	4	8	4	1	2
不要文除去	3B	15	15	5	3	3
意見要旨	3C	18	18	6	3	3
統計資料	4A	15	20	5	3	3
生活資料	4B	5	15	5	1	3
物語	5	24	30	6	4	5
論説	6A	30	30	6	5	5
タイトル付与	6B	6	6	6	1	1
計		169	200	（偏差値: 60.1）		

2018 本試	中問	得点	満点	配点	正答数	問題数
発音	1AB	14	14	2	7	7
平叙文	2A	20	20	2	10	10
語句整序	2B	12	12	4	3	3
発話文生成	2C	10	15	5	2	3
会話文完成	3A	0	0	4	0	0
不要文除去	3B	15	15	5	3	3
意見要旨	3C	12	18	6	2	3
統計資料	4A	15	20	5	3	3
生活資料	4B	15	20	5	3	4
物語	5	18	30	6	3	5
論説	6A	30	30	6	5	5
タイトル付与	6B	6	6	6	1	1
計		167	200	（偏差値: 60.5）		

2019 本試	中問	得点	満点	配点	正答数	問題数
発音	1AB	14	14	2	7	7
平叙文	2A	20	20	2	10	10
語句整序	2B	12	12	4	3	3
発話文生成	2C	15	15	5	3	3
会話文完成	3A	0	0	4	0	0
不要文除去	3B	15	15	5	3	3
意見要旨	3C	18	18	6	3	3
統計資料	4A	15	20	5	3	4
生活資料	4B	10	20	5	2	4
物語	5	30	30	6	5	5
論説	6A	30	30	6	5	5
タイトル付与	6B	6	6	6	1	1
計		185	200	（偏差値: 64.1）		

表 1: 2017 年〜2019 年のセンター本試験の得点内訳

平叙文完成（2A），語句整序（2B），発話文生成（2C）の 3 つの問題は短文問題であり，自然な単語の並びを答える問題である．そのため，従来手法 [12] では大規模言語モデルに基づく手法で解答していた．具体的には，CommonCrawl や Gigaword corpus 等から集めた 19 億単語からなる文章を用い，7-gram の言語モデルを単語頻度による閾値なしで学習していた．提案ソルバーでも基本的にこれを踏襲し，BERT の Head を言語モデルとして（LMHead を用いて）Finetune したものを用いている．Finetune には wikitext-2 の test データ[3]を利用し，過去のセンター試験や模試等から構成されるベンチマークデータを開発データとして用いた．語句整序問題では単純な N-gram でも完答していたが，平叙文完成および発話文生成では，BERT を用いることで性能を向上できている．

会話文完成問題は，会話中の発話の並びの自然さを答える問題である．従来手法 [10] では，大規模会話コーパスから SVM を用いて学習した，連続する 2 発話間の隣接発話らしさと，会話全体の感情極性の変化量に基づいて解答していた．提案ソルバーでは，Web から収集した会話文完成問題 360 問を用いて，BERT を 4 択問題で直接 Finetune している．本問題は 2018 年以降出題されていないため，表 2では 2017 年の 4 問の結果のみを示している．本問題については，現時点では BERT 等を利用しても正答率は 0.6 程度である．会話文の流れをより適切に捉えるためには，さらに多くの会話文を利用した学習が必要であると考えられる．

意見要旨把握問題，および，長文問題（統計資料（4A），生活資料（4B），物語（5），論説（6A））は，文書中の内容と選択肢の整合性を問う問題である．従来手法では，Word2vec や Gated Attention Reader によって本文中の各文と各選択肢の類似度を計算し，最も類似度が高くなる選択肢を正解とする手法で解答していた [11]．しかし，解答の根拠が複数文にまたがっている場合や，多様な言い換えがなされている場合などではほとんど正答することができず，チャンスレベルでしか解答できていなかった．提案ソルバーでは，これらの問題の解答に，XLNet を RACE で Finetune したモデルを適用している．これにより，テキストのみで問われる，意見要旨把握問題（3C），物語問題（5），論説問題（6A）は，それぞれ 0.85 を超える高い正答率を示している．一方，グラフや表の読解が求められる統計資料問題（4A）や，チラシや広告等からの計算が求められる生活資料問題（4B）は，通常の XLNet では答えられないため，正答率は 0.65 程度である．こうした問題に解答するには，表読解を行う手法 [6] や計算を行う手法 [2] が必要となる．

3 不要文除去問題の解法

不要文除去問題は，所定の文章から，それを取り除くことで全体のまとまりが良くなる文を一つ選ぶという問題である．図 2に例を示す．

これまでに我々は，BERT を不要文除去問題に適用する手法を提案した [9]．今回，BERT での結果で最も正答率が高かった入力形式を用い，自動で作成する擬似不要文除去問題（疑似問題）を用いて XLNet を Finetune する．

[3]https://blog.einstein.ai/the-wikitext-long-term-dependency-language-modeling-dataset/

問題名	中問	提案ソルバー	Finetune データ	従来手法	提案ソルバー正答率	従来手法正答率
平叙文完成	2A	BERT	wikitext2-test	7-gram w/ 1.9G corpus [12]	0.98 (49/50)	0.78 (39/50)
語句整序	2B	BERT	wikitext2-test	7-gram w/ 1.9G corpus	1.0 (15/15)	1.0 (15/15)
発話文生成	2C	BERT	wikitext2-test	7-gram w/ 1.9G corpus	0.86 (13/15)	0.53 (8/15)
会話文完成	3A	BERT	Web 収集 360 問	隣接発話らしさ+感情極性の一貫性 [10]	0.5 (2/4)	0.25 (1/4)
意見要旨	3C	XLNet	RACE [3]	単語意味ベクトルの類似度	0.86 (13/15)	0.20 (3/15)
統計資料	4A	XLNet	RACE	単語意味ベクトルの類似度	0.65 (13/20)	0.05 (1/20)
生活資料	4B	XLNet	RACE	単語意味ベクトルの類似度	0.61 (11/18)	0.22 (4/18)
物語	5	XLNet	RACE	単語意味ベクトルの類似度	0.88 (22/25)	0.24 (6/25)
論説	6A	XLNet	RACE	単語意味ベクトルの類似度	1.0 (25/25)	0.36 (9/25)

表 2: 各問題への BERT/XLNet の適用方法と従来手法との性能比較

Wearing proper shoes can reduce problems with your feet. Here are some important points to think about in order to choose the right shoes.
(1) Make sure the insole, the inner bottom part of the shoe, is made of material which absorbs the impact on your foot when walking. (2) The upper part of the shoe should be made of breathable material such as leather or cloth. (3) Some brand-name leather shoes are famous because of their fashionable designs. (4) When you try on shoes, pay attention not only to their length but also to their depth and width. Wearing the right shoes lets you enjoy walking with fewer problems.

図 2: 不要文除去問題の例 (2017 年センター英語試験問題)．正解は (3)．

不要文除去問題は，平均 8 文程度で完結する文書である．このことから，10 から 20 文程度で完結するような文書を用いて，擬似不要文除去問題を作成するのがよいと考えられる．そこで，まず所定の文書から連続する 7 文を抜き出し，7 文以外の一文を選択して不要な文として挿入し，その後，挿入した文を含むように選択肢とする文を決定し，挿入した文を正解の（不要となる）選択肢とした．これにより，任意の文書から不要文除去問題を機械的に大量に作ることができる．

元となる文書には，文書の長さおよび試験問題に出現する話題としての近さに鑑み，RACE の本文箇所を用いることにした．本文の長さが 10 文以上からなる文書を対象として，各文書から少なくとも 1 つ以上の擬似問題を作成した．これにより，670,540 の問題を作成でき，XLNet の Finetune に用いた．

XLNet への入力形式には，選択肢を除いた前 N 文全てと後 N 文全てを用いた．これにより前後の文書の結束性を判断できるモデルを学習できることが期待できる．実際に試験問題を解く際も同様の入力形式によりスコアを算出し，4 つの選択肢のうち，そのスコアが最大となるときの選択肢を解答とした．

提案ソルバーは 2017 年から 2019 年のセンター試験における問題 15 問に対し，すべて正答できた．従来法について述べると，XLNet をそのまま不要文除去問題に適用し，言語モデルの尤度を文全体の流れの自然さとして解いた場合の結果の正答率は 0.4 であり，Word2vec で最も遠い選択肢を選んだ場合の正答率は 0.47 であった．提案ソルバーが大幅に点数を向上できていることが分かる．なお，評価用に作成した 120 問を用いて，BERT と XLNet での正答率を比較した結果，BERT は 0.63 であり XLNet は 0.87 であり，XLNet の方が精度が高かった．

4 段落タイトル付与問題の解法

段落タイトル付与問題とは第 6 問で提示される論説文において，指定された 4〜5 個の段落 $p_1, \ldots, p_N (N = 4, 5)$ のそれぞれに対してタイトルとしてふさわしい言語表現を解答対象の段落と同数の選択肢 $t_1, \ldots, t_N$ から重複なく選ぶ問題である．全ての段落について完答した場合にのみ加点されることからランダムに答えた場合の正答率は $N = 4$ のとき $1/4! = 4\%$ である．

解答手法は，大きく 2 つのステップに分けられる．一つ目は全ての段落と選択肢の組 (p_i, t_j) について，関連性（類似性）スコア $score(p_i, t_j)$ を計算するステップである．二つ目は得られたスコアを用いて，段落と選択肢を 1 対 1 に対応づけるステップである．

ステップ 1 では段落 p_i ごとに 4 つの選択肢それぞれのスコアを計算する．この計算には出力部に四択問題解答用のネットワークを付加した BERT を用いる．通常の四択問題は本文，質問，選択肢から構成される．今回の問題においては，本文は p_i，選択肢は t_j, 質問は空文字列とする．質問は本来「この段落に最もふさわしい選択肢はどれか」という内容の英文とすべきものであるが，本文と選択肢の関連性を計算する上で全く情報がないため空文字にした．これを段

図 3: 段落タイトル付与問題の入力形式

Model	BERT	RoBERTa	Word2vec
最大和	.80	.90	.45
貪欲法	.48	.86	-
最小値最大	.55	.71	-

表 3: 段落タイトル付与問題の正答率 (42 問中)

落 (p_i) を固定して t_j ごとに図 3のように構成し，モデルを通すと，$score(p_i, t_j)(j = 1, \ldots, N)$ が得られる．これを各段落 (p_i) に対して実行することにより，$score(p_i, t_j)(i, j = 1, \ldots, N)$ が得られる．

ステップ 2 では得られたスコアを用いて各段落に対して重複しないように選択肢を割り当てる．先行研究 [8] に従って次の 3 つの方法を試した．なお，各段落 p_i に対して選択肢 t_j を重複なく割り当てた時の (i, j) の集合を 1 つの解候補 c とする．

最大和: $\sum_{(i,j)\in c} score(p_i, t_j)$ が最大の c

貪欲法: 全ての (i, j) のうち $score(p_i, t_j)$ が最大のものを選んで確定し，i, j が重複しないようにこれを繰り返す

最小値最大: $\min_{(i,j)\in c} score(p_i, t_j)$ が最大となる c

この手法の評価結果を表 3に示す．対象データは 2016 年以前のセンター試験，および，予備校の模擬試験からなる 42 問であり，BERT および RoBERTa [4] の large の事前学習モデルに対して RACE の学習セットの問題全てを用いて Finetune した．Word2vec を使った従来法 [8] との比較も示す．

表から，Finetune モデルにより，段落とタイトルの関連性がより正確に数値化できていることが分かる．ステップ 2 の解答選択については従来と同様にスコアの和が最大の候補を選ぶ方法が最良であった．

5 おわりに

本稿では，東ロボ英語チームが開発した，センター試験の英語問題における高性能な英語ソルバーについて述べた．センター試験におけるほとんどの問題は高精度で解くことが可能となったが，まだ解くことが困難な問題（生活資料，グラフや表の読解，会話の流れの理解）も明らかになった．今後はこのような，言語以外の情報や実世界の常識的知識が強く関わるタイプの問題に取り組んでいきたい．

謝辞

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．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に感謝いたします．

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL*, pp. 4171–4186, 2019.

[2] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proc. NAACL*, 2019.

[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] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[5] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[6] Jianshu Chen Yunkai Zhang Hong Wang Shiyang Li Xiyou Zhou Wenhu Chen, Hongmin Wang and William Yang Wang. TabFact : A large-scale dataset for table-based fact verification. In *Proc. ICLR*, 2020.

[7] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Proc. NeurIPS*, 2019.

[8] 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.

[9] 成松宏美, 杉山弘晃, 菊井玄一郎, 平博順, 的場成紀, 東中竜一郎. センター英語試験の不要文除去問題に対する bert の適用方法の検討. 人工知能学会全国大会論文集, pp. 3C4–J–9–01, 2019.

[10] 堂坂浩二, 坂本祐磨, 高瀬惇. 隣接発話らしさを利用した英語会話文完成問題の回答手法. 人工知能学会全国大会論文集, pp. 1K3–4, 2016.

[11] 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. 言語処理学会年次大会論文集, pp. D5–3, 2018.

[12] 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, pp. 2H2–1, 2017.