自動車免許試験自動解答における単語類似度の影響

的場 成紀 [1]　田邉 豊 [1]　小林 一郎 [2]　平 博順 [1]
[1] 大阪工業大学大学院 情報科学研究科　[2] お茶の水女子大学 基幹研究院 自然科学系
{m1m19a31, m1m20a21}@st.oit.ac.jp
koba@is.ocha.ac.jp
hirotoshi.taira@oit.ac.jp

1 はじめに

これまで我々は，与えられた文があらかじめ与えられた規則に適合するか否かを自動的に判定する規則適合判定技術の 1 つの例として，普通自動車免許試験問題の自動解答技術について研究を行ってきた [1][2]．機械読解の研究においては，BERT [3] をはじめとする大規模汎用言語モデルを用いた手法が高い正解率が得られ，広く用いられている．免許試験問題は，他の多くの機械読解タスクと比べて，与えられる文章が 1 文程度と短いのが大きな特徴である．BERT などの手法が適さない可能性もあったが，模擬問題を利用した BERT による学習手法では，免許試験問題においても一定の正解率が得られることを実験的に示されていた [1]．しかし，大規模汎用言語モデルを用いた手法では，出現単語の並びだけで正誤判定がされてしまっている可能性も考えられ，実用的な規則適合判定に BERT 等の手法をそのまま使用できるのか疑問があった．

そこで，本研究では，単語の類似度だけでは正誤判定がしにくい評価用データを作成し，自動解答手法の評価を行って，文を構成する単語が類似していても規則を正しく判定できるのか評価を行った．自動解答手法には，BERT [3]，self-attention + BiLSTM [4]，word2vec [5] の 3 つの手法を選び，性能の比較を行った．

2 普通自動車免許学科試験問題

普通自動車免許の学科試験は，自動車を運転するときに必要となる運転技術や交通規則に関する知識や運転者のマナーなどが問われる試験である．学科試験で問われる内容として文章問題とイラスト問題が出題され，文章問題は 90 問，イラスト問題は 5 問である．文章問題は，正誤の二択問題である．イラスト問題は，運転者から見た車外の様子が描かれたイラストが与えられ，危険予測などに関する問題に解答する問題である．各大問に対して二択問題 3 問が出題される．なお配点は，文章問題は各問 1 点，イラスト問題は各大問に対して完答で 2 点である．合計 100 点満点で，90 点以上が合格となる．

本研究では，学科試験の大部分を占める文章問題を実験の対象とした．また，実際の文章問題には標識や説明のイラストが付与された問題も含まれているが，本研究では主に文章のみで表された規則について扱い，イラストが付与された問題は今回は対象外とした．

3 評価用データセットの作成

単語の類似度だけでは規則適合判定が難しくなるように，1）オリジナルの問題文，2）オリジナルの問題文と単語類似度が高く，かつほぼ同じ内容で，正解もオリジナルと一致する問題 3) オリジナルの問題文と単語類似度が高く，ほとんど同じだが，正解がオリジナルと異なる問題の 3 種類のデータが含まれる評価用データセットを作成した．

具体的には, 以下の手順で作成を行った．まず，市販の問題集「試験によく出る普通免許 1000 題」[6] や「大事なとこだけ総まとめ　ポケット版　普通免許試験問題集」[7] を参考にし、問題セット A（7,992 問），問題セット B（103 問，解説文を含む），問題セット C（1,000 問）を作成した．

問題セット A，B, C は参考にした問題が異なるっている．また，問題セット A は，文末の言い回しのバリエーションを多くなるように作成され，問題セット B は，解説文を含んでいるという特徴がある．

つぎに、問題セット C の先頭 300 問から，図が含まれる問題を除いた 260 問を選び，テスト用問題

表 1　評価用データセットのサイズ

学習用データ	7,992 問
開発用データ	746 問
評価用データ	291 問（小問 3 問× 97 問）

の候補とした．テスト用問題の候補 260 問それぞれについて，問題セット A から文類似度が高い上位 20 問を抽出し，その中から人手で，テストセットの候補の問題と，a) 文意が同じ問題，b) 文意が極めて近いが正解（○×）が異なる問題，の 2 種類が存在するかどうかを確認した．その結果，問題セット B に a), b) の 2 種類とも存在する問題は，260 問中，97 問であった．この 97 問を最終的なテスト用問題とした．なお，確認に用いた文類似度は，日本語 wikipedia に基づく word2vec[1] を用いた単語類似度の平均で求めた．

表 2 に作成されたテスト用問題の問題例を示す．

表 2　テスト用問題の問題例

問題文	正解
光化学スモッグが発生するおそれがあるときは、運転は控える。	○
光化学スモッグは発生しそうなときの運転は、控えたほうがよい。	○
光化学スモッグが発生しているときや、発生するおそれのあるときでも、自動車の運転を控える必要はない。	×

訓練用データには，問題セット A と B を合わせた 8,095 問，開発用データには，問題セット C からテスト用データを除き，図を含む問題も除いた 746 問を用いた．

4　評価実験

3 章で作成したデータセットを使用して，免許試験の正誤判定問題について自動解答手法の性能評価を行った．自動解答手法には大きく，word2vec, selfattention+BiLSTM，BERT の 3 つの手法を比較した．word2vec について日本語モデル[2] を利用した．形態素解析器には MeCab を使い，辞書には NEologd を使った．word2vec を使った正誤問題の解き方として，先に訓練用データと評価用のデータのそれぞれを word2vec で文ベクトルに変換した．文ベクトルの変換方法は 1 文を MeCab で単語分割を行ったあとに，word2vec に流して単語ベクトルに変換し，文書に出現する全ての単語ベクトルの加算平均をとる．このときの加算平均を文ベクトルとして扱う．その後，評価用データを 1 文ずつ，訓練用データの各文書と cos 類似度で類似度を測定して，一番類似度の高い問題文を検索する．最後に一番高い問題文の正誤をモデルの出力として扱い，評価用データのラベルと合致するか確認をする．

BERT は日本語 Wikipedia コーパスで事前学習モデルを転移学習した．形態素解析器には MeCab+WordPiece [8] を使用した．

self-attention + BiLSTM では埋め込み層に BERT の特徴量を取得する方法をとった．BERT の特徴量を得るモデルは転移学習しておいたモデルと事前学習のみのモデルを用意した．比較実験として BERT の特徴量を使わずに学習データから得られる単語の特徴量で学習させた．

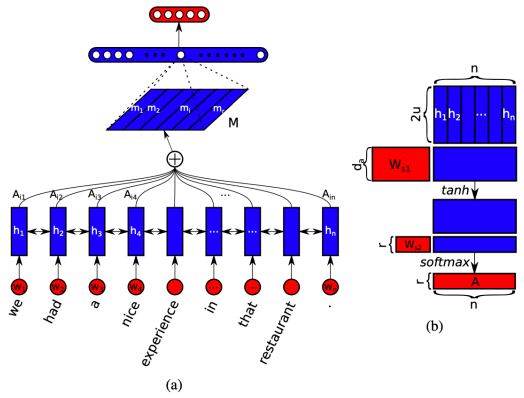

図 1　self-attention + BiLSTM（文献 [4] より引用）

4.1　実験結果

表 3 は評価用データ全体の正解率と小問 3 問で完答した正解率を表している．表 3 より最も正解率が高かったのは，self-attention + BiLSTM であり，次いで word2vec で正解数が多かった．

5　手法による正解した問題の違い

ここでは，手法の違いによってを正解した問題の違いを調べた．

5.1　word2vec の手法で正解が多かった問題

表 4 は word2vec で有意な問題の例である．問題の例では，数字が「800」か「750」の違いによって正誤が反転している．問題例のように数字の違いによって問題が正誤が反転している問題では完答ができていなかった．

1）　https://github.com/singletongue/WikiEntVec/releases
2）　https://github.com/singletongue/WikiEntVec/releases

表 3　評価用データの正解率

手法	正解率（オリジナル）	正解率（3 問完答）（正解数/問題数）
word2vec	0.80	0.49（48 / 97）
BERT	0.71	0.33（32 / 97）
self-attention + BiLSTM	**0.82**	**0.59（57 / 97）**
self-attention + BiLSTM + BERTembed w/o FT	0.76	0.48（47 / 97）
self-attention + BiLSTM + BERTembed w/ FT	0.63	0.21（20 / 97）

表 4　word2vec の手法で正解が多かった問題の例

問題文	正解
普通免許を受ければ、けん引装置のある車両総重量 **800** キログラムの被けん引車をけん引して運転できる。	×
けん引装置のある車両総重量 **800** キログラムの被けん引車をけん引しての運転は、普通免許でできる。	×
普通免許があれば、けん引免許がなくても、普通自動車で車両総重量 **750** キログラムの車をけん引できる。	○

5.2　言語モデルの手法で正解が多かった問題

表 5 は言語モデルで解けて，word2vec で解けなかった問題の例である．この問題の例の「車に乗る前には、車の前後に人がいないかどうかを確かめればよく、車体の下まで確かめる必要はない。」で，word2vec で最も類似度の高い問題では，類義語まで見ることができずに答えることができなかった．表 6 は類似度の高い順から検索された問題である．このように言語モデルは学習させる際に問題の類義語などが読み取れていることが分かる．

表 5　言語モデルの手法で正解が多かった問題の例

問題文	正解
運転者は、**車の前後に人がいないか、車の下に子供がいないかなど**、周囲の安全を確かめてから乗車する。	○
運転者は、車に乗る前に、**車の前後や車の下に人がいないかを確かめるようにする**。	○
車に乗る前には、車の前後に人がいないかどうかを確かめればよく、**車体の下まで確かめる必要はない**。	×

5.3　BERT の単語を埋め込む手法で正解が多かった問題

表 7 は BERT の単語埋め込みをした self-attention+BiLSTM で有意な問題である．この問題の例では「交差点の手前」と「交差点の直前にさしかかったところ」でといった意味は同じであるが，表現が違う問題も正解していることが分かる．

5.4　どの手法でも解けなかった問題

表 8 は，どの手法でも解けなかった問題である．この問題では，「中央」と「右端」のみの違いだけで，問題の正誤が反転している．このように単語のみの変化で細かいところまでは読み解くことができなかった．

5.5　言語モデルにおける出力による違い

評価用データで文意と正誤が同じ問題と正誤が違う問題で，言語モデルによって出力結果が正解か不正解に関わらず意図していない問題を調べた．

5.5.1　言語モデルの出力が反転していた問題

表 9 は，文意と正誤が同じ問題でありながら言語モデルの出力結果が異なっている問題の例である．この問題の特徴として文章の構成が違うことがあげられる．この問題を解く点として「交通整理が行われていない交差点」、「道の幅」、「優先道路」が鍵となるが出現する順番が入れ替わっている．このような問題は言語モデルの出力結果が反転していた．

5.5.2　言語モデルの出力が同じになった問題

表 10 は，文意が同じで正誤が反転しており，言語モデルの出力結果が同じである問題である．上の問題の解説を確認してみると「安全地帯がないときは、1 人もいなくなるまで、後方で停止します。」となっており，安全地帯の有無によって徐行してよいかどうかが判断基準となっている．このように問題の単語のみの違いで，問題の構成が類似している問題はモデルの出力結果が反転していなかった．

6　おわりに

本研究では，普通自動車免許試験の自動解答技術について，規則適合性判定技術の観点から性能評価を行うため，単語の類似度だけでは正誤判定がしにくい評価用データを作成し，自動解答手法の評価を

表 6　「車に乗る前には、車の前後に人がいないかどうかを確かめればよく、車体の下まで確かめる必要はない。」のときの word2vec の上からの検索結果

問題文	類似度	正解
運転者は車に乗る前に、車の前後に人がいないか、車の下にこどもがいないかを確かめなければならない。	0.990	○
車に乗るときは、車の前後に人がいないか、車の下に子どもがいないかを確かめるようにする。	0.989	○
乗車する前に、人が車の前後にいないか、車の下に子供が潜んでいないかなどの周囲の安全確認を行う。	0.982	○
車に乗る前の安全確認では、車の前後の人の有無を確認するだけで十分なので、車体の下までの確認はしなくてよい。	0.981	×

表 7　BERT の単語を埋め込む手法で正解が多かった問題の例

問題文	正解
交差点の**手前で**、信号が青色から黄色に変わったときは、加速して一気に通過する。	×
交差点に**進入する手前で**信号が青色から黄色に変わった場合は、スピードを上げて一気に通過してしまう。	×
交差点の**直前にさしかかったところで**信号が黄色に変わったものの、停止線のすぐ手前だったため、そのまま通過することにした。	○

表 8　どの手法でも解けなかった問題

問題文	正解
自動車が一方通行路で右折するときは、あらかじめ道路の**中央**に寄って、交差点の中心のすぐ内側を徐行しなければならない。	×
一方通行の道路で右折するときは、あらかじめ道路の**中央**に寄り、交差点の中心の内側を徐行しなければならない。	×
一方通行の道路で右折するときは、あらかじめ道路の**右端**に寄り、交差点の中心の内側を徐行しなければならない。	○

表 9　言語モデルの出力が反転していた問題

問題文	正解
交通整理が行われていない道幅の同じ交差点 (優先道路は除く) では、車よりも路面電車が優先する。	○
優先道路を除いて、道の幅が違わない交差点では、交通整理がなされていない限り、路面電車のほうが車よりも優先される。	○

表 10　言語モデルの出力が同じ問題

問題文	正解
安全地帯のない停留所で、乗客の乗り降りのため停車している路面電車に追いついたときは、その横を徐行して通過する。	×
安全地帯があれば、停車して乗客の乗降を行っている路面電車に追いついても、徐行して走行してよい。	○

行った．自動解答手法として，BERT，self-attention + BiLSTM，word2vec などの手法を選び性能評価を行ったところ，BERT や word2vec による手法は，単語類似度に依存して規則適合性判定を行っている傾向がみられ，self-attention を使った BiLSTM を用いた場合が，最も正解率が高くなった．

謝辞

本研究は JSPS 科研費 18K11452 の助成を受けたものである．

参考文献

[1] 的場成紀, 古賀雅樹, 大塚基広, 小林一郎, 平博順. 運転免許試験で使用される語彙と省略語句の分析. 人工知能学会全国大会論文集 一般社団法人 人工知能学会, pp. 1N4J904–1N4J904. 一般社団法人 人工知能学会, 2019.

[2] 的場成紀, 古賀雅樹, 吉村優志, 田邉豊, 小林一郎, 平博順. 運転免許試験自動解答における問題解説文の利用. 言語処理学会第 26 回年次大会 (NLP2020) 発表論文集, pp. 121–124, 2020.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[4] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017.

[5] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. pp. 3111–3119, 2013.

[6] 倉宣昭. 試験によく出る普通免許 1000 題. 高橋書店, 2007.

[7] 学科試験問題研究所. ポケット版　普通免許試験問題集. 永岡書店, 2015.

[8] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner,

Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation, 2016.