社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

短答式試験における自動採点のための概念辞書を用いたデータ拡張手法の提案

加藤　博之† 　石岡　恒憲†† 　峯　　恒憲†††

† 九州大学大学院システム情報科学府，福岡県
†† 大学入試センター研究開発部，東京都
††† 九州大学大学院システム情報科学研究院，福岡県
E-mail: †kato.hiroyuki@m.ait.kyushu-u.ac.jp, ††tunenori@rd.dnc.ac.jp, †††mine@ait.kyushu-u.ac.jp

あらまし　自然言語処理に関する研究では，BERT をはじめとする大規模汎用言語モデルの登場によって，様々な種類のタスクにおいて処理精度の向上がもたらされているものの，答案の自動採点など，いまだ実用的な精度には達しておらず，さらなる精度の向上が望まれている．そこで，本研究では短答式試験の自動採点において，概念辞書を用いて答案中の単語を置換するデータ拡張を行い，自動採点の精度を向上させる方法を提案する．高校生の社会科と国語の模試データを用いて行った実験では，社会科ではデータ数の少ない範囲で，国語ではデータ数の多い範囲で精度の向上が見られた．また，精度の高い学習においては学習器が置換後のデータと元データを乖離して捉えていることが分かった．

キーワード　自然言語処理, 自動採点, データ拡張, 概念辞書, BERT

Automatic Short Answer Scoring using Thesaurus-Based Data Augmentation

Hiroyuki KATO†, Tsunenori ISHIOKA††, and Tsunenori MINE†††

† Graduate School of Information Science and Electrical Engineering, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka, 819-0395 Japan
†† Research and Development department, National Center for University Entrance Examinations, 2-19-23 Komaba, Meguro-ku, Tokyo, 153-8501 Japan
††† Faculty of Information Science and Electrical Engineering, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka, 819-0395 Japan
E-mail: †kato.hiroyuki@m.ait.kyushu-u.ac.jp, ††tunenori@rd.dnc.ac.jp, †††mine@ait.kyushu-u.ac.jp

Key words　NLP, AutoScoring, DataAugmentation, Thesaurus, BERT

1. はじめに

記述式試験は，受験者の思考力をよく反映する形式として，大学入学共通テストへの導入も検討されるなど注目を集めている．一方で，採点のコストや採点の揺れなどの問題が指摘されており，近年の BERT など大規模汎用言語モデルの登場などで急速に発展している自然言語処理技術を利用した記述式試験の自動採点技術に期待が寄せられている．信頼性のある記述式試験の自動採点システムが実現すれば，採点のコストの削減や，採点ゆれの軽減などに効果が見込まれる．しかし，完全な自動化の形で実サービスに用いられるほどの信頼性を持つモデルはまだ実現されておらず，さらなる技術向上が望まれている．

本研究では，短答式試験の自動採点における課題の一つである多様な語彙への対応に向けて，概念辞書を用いて生徒解答中の単語を置換するデータ拡張を行い，自動採点の精度を向上させる方法を提案する．高校生の社会科と国語の模試データを用いた実験では，社会科ではデータ数の少ない範囲で，国語ではデータ数の多い範囲で精度の向上を確認した．一方で，社会科のデータ数が多い範囲や国語のデータ数が少ない範囲においては，精度が悪化する場合が見られた．また，精度の高い学習においては学習器が置換後のデータと元データを乖離して捉えていることが分かった．

次節以降では，以下の構成で提案手法に関して説明する．2 節では関連研究について説明し，3 節では提案手法について説明する．4 節では実験の実施例の詳細を述べ，5 節では実験の結果とその分析を述べる．6 節はまとめである．

2. 関 連 研 究

近年自然言語処理の様々なタスクにおいて精度の向上をもたらしているニューラルネットワークを基盤とする技法として，大規模汎用言語モデルがある．大規模汎用言語モデルでは，大規模なコーパス，モデル，計算資源を用いて学習を行って品質の高い言語モデルを生成し，そのモデルに必要に応じたタスクを追加で学習させることで，計算量を追加の学習だけに抑えつつ，高品質な言語処理を実現する．また，大規模汎用言語モデルは文章をベクトルに変換する embedding としての性質も持つ．しかし，従来主流であった単語単位の変換と異なり，文中の単語の順番に強く依存した学習を行うモデルが多い．文脈情報を含む学習を行うことで，文章の特徴をより深く捉え，精度の改善に貢献していると考えられる．そのような大規模汎用モデルの代表的な例として，Jacob ら [1] の提案したモデル BERT がある．BERT は，入力した文章の一部を隠し，その隠れた文章を推定するタスクなどを学習することで高品質な言語モデルを生成する．BERT は目的のタスクに応じた追加の学習を行う方法を取ることで発表当時の自然言語処理の最先端技術 (State Of The Art: SOTA) を多数更新し，その後も BERT を元にした研究が活発に行われている．BERT は大規模汎用言語モデルの有用性を示す例の一つである．大規模汎用言語モデルは自動採点にも用いられており，Zichao ら [2] は BERT の学習に用いるコーパスに対して採点対象の問題の科目に応じたコーパスを追加することで，従来手法よりも採点精度を向上させている．また，この大規模汎用言語モデルは従来のモデルと組み合わせて用いる事も可能で，Reham ら [3] は文中の単語の相互作用を捉える補助として，時系列情報を双方向に学習する BiLSTM を BERT の出力に接続している．また，Yuqi ら [4] は医療に関する報告書の区分を推定するタスクで，文中のラベリングを行うために BiLSTM を BERT の出力に接続している．

機械学習において，データの規模が学習精度に大きく影響することが知られ，データを増やすことが採点精度の向上に繋がる．しかし，タスクによってはデータの生成コストが高いなどの理由でデータを簡単に増やせない場合もある．そこで，実際のデータを増やすのではなく，疑似データを生成することで見かけ上のデータを増やすデータ拡張が行われる．自動採点においても，データの生成には解答データの提供に協力してくれる解答者と専門の知識や技能を持った採点官が必要になるため，データの生成コストが比較的高く，自動採点はデータ拡張が有効に働くタスクの一つである．Lun ら [5] は，短答式試験の自動採点において，BERT を用いるほかに，モデルに入力として渡す模範解答文と受験者解答文のペアに対して，満点の解答を模範解答と同様に扱ったデータを追加することでデータを拡張を行い，採点精度を向上させている．また，データ拡張では他のデータを組み合わせて追加データを生成する方法も用いられる．William ら [6] は，twitter の怒りコメントを原因別に分類するタスクにおいて，GoogleNews など他の巨大なコーパスから生成した embedding を用いて，元のデータ中の単語を embdding 中の最も近い単語に置き換えたデータを追加することでデータ拡張を行い，分類精度を向上させている．また，WordNet などの概念辞書による単語の置き換えもデータ拡張手法の一つであり，Steven ら [7] は大規模汎用言語モデルの GPT-2 の学習において，コーパス中の単語を WordNet の上位語に置き換えてデータ拡張を行い学習を行うことで，GPT-2 の各種タスクでの精度を向上させている．Weikang ら [8] は，問いかけに対する解答抽出タスクで，直接的なデータ拡張ではないが，WordNet の上位語・類語から単語の分散表現情報を拡張し，抽出精度を向上させている．

3. 提 案 手 法

本研究では，自動採点タスクにおいて，学習器に与える文字列データ中の単語を概念辞書を用いて変換し，データ拡張を行うことで，採点精度を向上させる手法を提案する．

3.1 BERT

BERT は大規模汎用言語モデルの一つであり，その学習では，単語の位置関係を処理することで文章構造の特徴把握も行うが，同時にコーパスに出現する単語の意味獲得も行われている．そのため，BERT の embedding を利用することは BERT が学習を行ったコーパスのデータを活用することになり，学習データの拡充にも繋がると考えられる．よって，本研究では文章把握に加え学習データの拡充を狙い，単語をベクトル化する embedding として BERT を用いる．ただし，BERT は入力された文章をそのまま出力するように学習しているため，BERT の出力をそのまま用いるのではなく，出力直前のレイヤーの出力を embedding として用いる．

また，BERT の出力に含まれる時系列情報に対し柔軟に対応するため，時系列情報の処理が可能な再帰型ニューラルネットワークの一つである BiLSTM と Attention 機構を用い，embedding を BiLSTM に入力して，BiLSTM の重みを更新することで採点のタスクを学習するモデルを提案する．Attention 機構の導入に伴い，出力されるコンテキストベクトルから最終的な点数を推定するための全結合層も追加する．提案するモデルの概要を図 1 に示す．

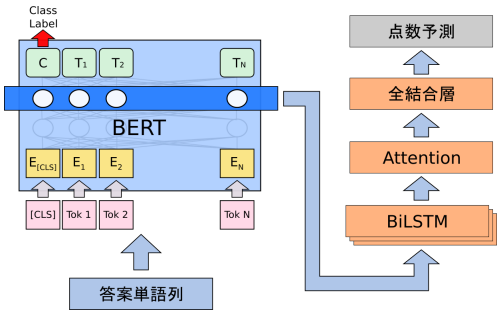

図 1　学習モデル概要図

3.2 単語の置換

データ拡張に用いられる手法として，データ中の単語を意味の近い単語に置換して，置換したデータを元のデータに追加する手法がある．文章の意味をなるべく変えずに表層的な表現を変更することで，データ中の表現が増え，より幅広い表現に対

応したデータを生成できる．

本研究では，概念辞書を用いて単語の置換を行い，データ拡張を行う．特に，概念辞書として日本語 WordNet [9] を用いる．日本語 WordNet は Princeton WordNet を翻訳した概念辞書であり，人手で構築された概念辞書であるため，信頼性の高いデータとなっている．信頼性の高いデータを用いることで，より品質の高いデータ拡張が行えると考える．単語の置換においては，形態素解析器で単語単位で区切った文章に対して，その単語が日本語 WordNet の語彙にあるかどうかを調べ，もし語彙があれば，関連を持つ単語を検索し，元の文章の単語と置換する．日本語 WordNet は多くの関連性が登録されているが，本研究では単語の類似性に注目し，単語の類語と上位語を用いた．また，単語には複数の単語と関連性を持つものもあり，関連語を索引した際に複数の候補が得られるが，本研究ではその中から一つを選択して用いた．複数の候補からの選択には，単語の Embedding を参照してコサイン類似度の最も高いものを選択する方法と，ランダムに選択する方法を用いた．

さらに本研究では，これらの単語の置換をより質の高いものにするため，ストップワードは単語の置換対象から除外した．データ拡張の意図から，出現頻度がきわめて高い単語は幅広い文章において使用され，対象データの特徴を捉えたものではないため，データを拡張する必要性が薄いためである．

また，単語の置換において，表層上の単語の処理によって文に合った語意を捉えられなかったり，品詞が異なる単語が選択されたりする場合が見られた．例えば，「〜しているため」といった文章があった場合に，接尾辞の「いる」を動詞の「いる」と捉え，さらに名詞の「存在」と置換するなどである．これは，WordNet が品詞の情報を持たず，単語の表層上の情報でしか索引出来ないことに起因する．この文意に合わない置換を防止するため，置換前後で品詞が変化していない単語のみを置換候補とした．単語の置換手順の概要を図 2 に示す．

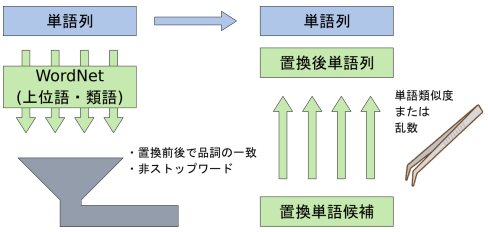

図 2　単語置換手順概要図

3.3　学　　習

採点タスクの学習は，元のデータと単語置換で得られた拡張データを組み合わせて 2 段階で行う．まず 1 段階目として，元のデータと拡張データを 1 セットずつ結合して入力し学習させ，2 段階目として元のデータのみを入力として再度学習を行う．この操作は，1 段階目で幅広い表現のデータを用いて，表層ではなくデータの深層の意味を捉え，2 段階目の操作で表層上の特性にも適応させる効果を狙い設定した．

4.　デ　ー　タ

4.1　概　　要

提案手法の評価に使用したデータは，高校生の全国模試における記述式解答の文字起こし済みの解答本文と，その解答につけられた点数からなる．データは社会科から 3 教科の計 8 問と，国語から論説 5 問と小説 4 問の計 9 問，採点は専門家によって行われている．解答の要素ごとに部分点が定められ，その部分点を基準に採点されている．部分点は 0 か満点ではなく，内容に応じて部分点の一部が与えられる．誤字があった場合は 1 点減点される．また，日本史 B2_3，世界史 B2_3 では問題文で解答に関連するキーワードが指定されており，そのキーワードについての説明を行う形式となっている．

4.2　特　　徴

各問の統計情報を表 1，図 3，表 2，図 4 に示す．なお，図 3，図 4 はデータ数全体を 100% とし，データ数の割合を表す軸を点数ごとに横にずらして描画したグラフである．

分布から，社会科のデータでは日本史，世界史に 0 点の答案が割合が比較的多いことが分かる．これは，社会科の問題が知識を聞く問題が多く，要点となる知識を逃した場合に部分点が取得しにくい性質から来ていると思われる．

国語のデータでは，比較的均等に点数が分布していることが分かる．これは，用いたデータの配点が比較的高いため部分点が細かく設定され，粒度の細かい採点が行われる性質から来ていると思われる．

他にデータが科目固有に持つ特徴として，小説には登場人物の名前のような，その問題特有の固有名詞が出現するという特徴がある．

表 1　配点と文字数制限：社会

科目	配点	文字数制限 (以内)	サンプル数
地理 B_1	4	20	79
地理 B_4	6	60	79
日本史 B2_1	3	20	83
日本史 B2_3	6	60	83
世界史 B2_1	3	25	102
世界史 B2_3	5	60	102

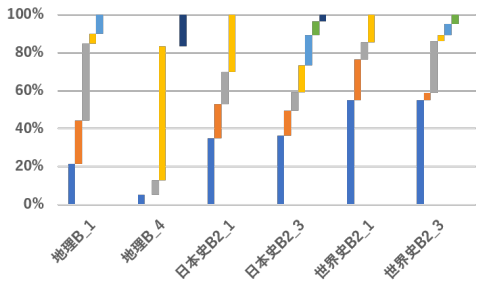

図 3　点数分布：社会

表 2　配点と文字数制限：国語

科目	配点	文字数制限 (以内)	サンプル数
Y14_1-2_1_3(論説)	16	70	2100
Y14_1-2_2_4(小説)	12	50	2100
Y14_2-1_1_5(論説)	15	70	2100
Y14_2-1_2_3(小説)	12	60	2100
Y14_2-2_1_4(論説)	15	70	2100
Y14_2-2_2_3(論説)	14	60	2100
Y15_2-3_1_5(論説)	16	80	2000
Y15_2-3_2_2(小説)	12	60	2000
Y15_2-3_2_4(小説)	14	80	2000

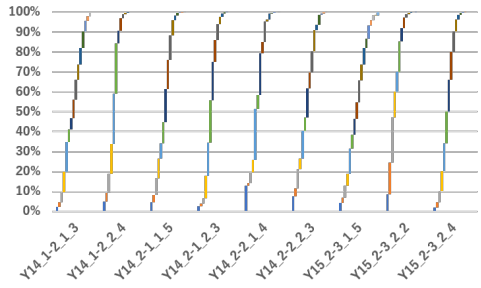

図 4　点数分布：国語

5. 実　　験

5.1 学習データの生成

学習データの生成にあたって，まず解答データを形態素解析器 JUMAN++ を用いて形態素解析・分かち書きをし，単語列に変換した．そして生成した単語列に対して 3 節で述べた単語置換を適用した．置換先の候補を取得する概念辞書は日本語 WordNet を用いた．ストップワードには日本語 wikipedia に含まれる単語の内出現頻度が最も高い 300 単語を設定し，その 300 語に含まれる語は置換対象から除外した．置換前後での品詞が変化するものを置換対象から除外するための品詞情報は，JUMAN++ のものをそのまま用いた．

置換候補単語の中から，embedding のコサイン類似度が最も高い単語を選ぶ操作では，朝日新聞社が公開している word embedding データ [10] から glove を用いて計算を行った．

また，データ数による手法の効果の影響を確かめるために，社会科のデータは全件を 40 件，60 件，79 件に制限したデータを，国語のデータは全件を 200 件，400 件，600 件，800 件，1000 件に制限したデータを作成した．

5.2 学　　習

学習モデルの embedding として用いる BERT には，京都大学が公開している BERT 日本語 Pretrain モデルのうち, Japanese_L-12_H-768_A-12_E-30_BPE を用いた．また，BERT のレイヤーに接続する BiLSTM は，置換手法を行わない事前実験でのパラメータチューニングの結果から，隠れ層 300 次元の BiLSTM を 3 層重ねて用いた．Attention 機構のコンテキストベクトルを処理する全結合層は，BiLSTM と同じく 300 次元とした．

学習は社会科は 10 分割，国語は 3 分割の交差検証で行った．訓練データとして割り当てられたデータのうち，2 割を検証データとし，EarlyStopping を用いて，学習中に検証データの精度の向上が 10 周連続で見られなかったときに学習を止めた．また，単語置換手法では 2 段階に分けて学習を行うが，その両方に対して同様の EalyStopping を適用した．

5.3 結果の測定

結果はテストデータにおける推定点数と実際の点数の採点誤差を求めた．ただし科目ごとに配点が異なるので，各科目の精度を平等な基準で扱うため，評価指標には満点を 1 として正規化して採点誤差の二乗平均平方根を求める RMSPE(Root Mean Squared Percentage Error) を用いた．

また，置換データのモデル中での振る舞いを観測するため，訓練データに対して元データでの出力と置換データでの出力を比較する指標を 2 つ求めた．一つは，元データでの推定点数と置換データでの推定点数の差で，評価指標には満点を 1 として正規化して推定点数の平均絶対誤差を求める MAPE(Mean Absolute Percentage Error) を用いた．元データのみで行う再学習を経て，モデルの表層的な学習の変化の度合を観測できると考える．もう一つは，モデル中の Attention 機構での元データと置換データの Attention 値の分布差で，評価指標には系列データの乖離度を示す DTW(DynamicTimeWarp) を用いた．BiLSTM の学習をよく反映する Attention 値の分布の変化を観測することで，モデル深層部の学習の変化の度合いを観測できると考える．

実験は各設定につき 3 回行い，各交差検証ごとに評価指標を求め，その平均を結果とした．

6. 結果と考察

6.1 社会科データ

社会科のデータについて，各データ数，各単語置換手法について予測した点数と実際の点数との採点誤差をまとめたグラフを図 5 に示す．ただし，各科目での指標は平均を求め，一つの結果にまとめて用いている．

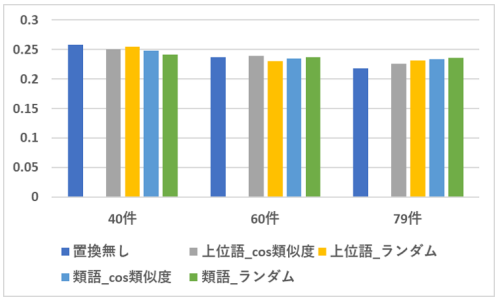

図 5　データ数，置換手法別 採点誤差 (RMSPE)：社会

図 5 の採点誤差から，40 件，60 件で行ったデータに関して，ほぼ全ての置換手法で精度が向上している事が分かる．特に，40 件でのデータでは類語をランダムに選択する手法では 6.7%，60 件でのデータでも上位語をランダムに選択する手法で 2.7%RMSPE を抑えることが出来ている．一方で，79 件でのデータでは全ての置換手法で精度が悪化している事が分かる．類語をランダムに選択する手法では，RMSPE が 8.1% 増大し

ている．類語をランダムに選択する手法に限って見るとデータ数が大きくなる程に精度が向上する傾向はあるが，置換無しのデータにおいては，より鋭敏に精度が向上している．

次に，学習の 2 段階目の元データでの再学習を終えた段階での元データと置換データでの推定点数の差をまとめたグラフを図 6 に，元データと置換データでの Attention 機構の Attention 値の分布差をまとめたグラフを図 7 に示す．

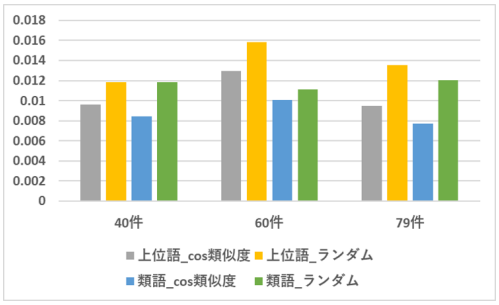

図 6　データ数，置換手法別 再学習後の置換有無での推定点数差 (MAPE)：社会

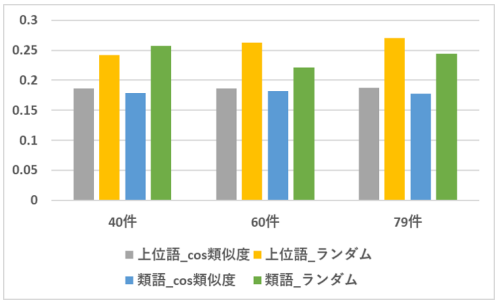

図 7　データ数，置換手法別 再学習後の置換有無での Attention 値の分布の乖離度 (DTW)：社会

図 6 置換有無の推定点数差から，類語を cos 類似度で選択する手法が一番点数差が少なく，元データのみでの再学習後も置換有りの文章が元の文章と似た文章として扱われていることが分かる．逆に，点数差が最も高い上位語をランダムに選択する手法では元データと置換有りのデータと乖離して扱われていることが分かる．また，次項の国語のデータに見られるが，再学習後の置換有無の点数差はデータ数が大きくなるほど開く傾向にある．社会データにおいても 40 件と 60 件ではその傾向が見られるが，79 件のデータでは点数差が大きく出ていない事が分かる．また，精度が向上した類語をコサイン類似度で選択する手法では，点数差が大きくなり，置換の有無で文章が乖離して扱われていることが分かる．

図 7 の Attention 値の分布の乖離度からは，値の大小が推定点数差とほぼ同等の傾向を示しているが，点数差においてやや高い値を示していた上位語をコサイン類似度で置換する手法が類語をコサイン類似度で置換する手法と同様に低い値を示している．このことから，類語をコサイン類似度で選択する手法では再学習を経ても深層部の変化が起こりにくいことが分かる．

以上をふまえ，採点誤差と置換有無での推定点数差から，社会科のデータでは置換の有無で文章が別の文章として認識されている方が精度が高く出る傾向が見られる．これは，社会科の問題に知識を問う傾向があり，問題の要であるキーワードを捉え置換先の単語と別物であると学習できたモデルの方が問題の特性をよく掴み，採点精度が高い事が考えられる．よって，社会科のデータにおいては，単語を置換する手法がキーワードの獲得を阻害し，悪影響をもたらすと考えられる．データ数が少ない範囲においては，置換手法が精度の向上に繋がっている事から，キーワードを乱す効果よりもデータを拡張する効果が大きく出て，総合的に良い効果が出たと考えられる．

6.2　国語データ

国語のデータについて，社会のデータと同様に，各データ数，各単語置換手法について予測した点数と実際の点数との採点誤差をまとめたグラフを図 8 に示す．ただし，各科目での指標は平均を求め，一つの結果にまとめて用いている．

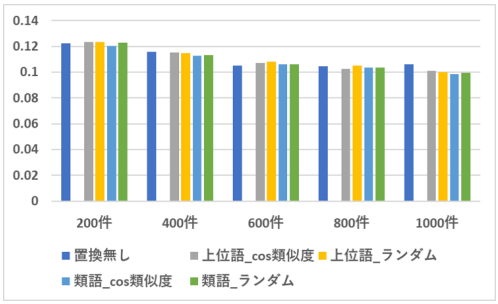

図 8　データ数，置換手法別 採点誤差 (RMSPE)：国語

図 8 の採点誤差から 200 件と 600 件のデータにおいては多くの精度で置換手法によって精度が悪化しているが，それ以外の件数においてはほぼ全ての置換手法で精度が向上していることが分かる．また，置換手法での結果のみに着目すると，全ての置換手法において，データ数が増えるにしたがって精度が向上していることが分かる．一方で通常手法では，600 件まではデータが増えるごとに順調に精度が向上しているが，800 件では 600 件の物と比較して精度がほとんど変化せず，1000 件のデータでは僅かに精度が悪化している．データ数の増加が精度の悪化をもたらしたとは考えにくいが，少なくとも精度の向上に鈍化が見られる．置換手法の中では，多くのデータ件数において，類語をコサイン類似度で選択する手法が最も良い精度を示している．

次に，社会科と同様に元データと置換データでの推定点数の差をまとめたグラフを図 9 に，元データと置換データでの Attention 値の分布差をまとめたグラフを図 10 に示す．

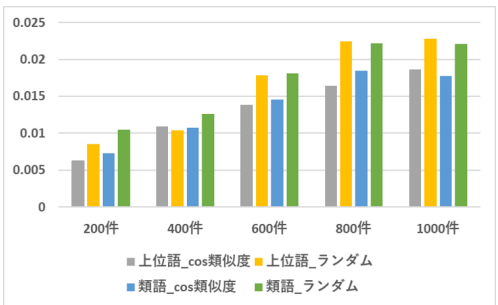

図 9　データ数，置換手法別 再学習後の置換有無での推定点数差 (MAPE)：国語

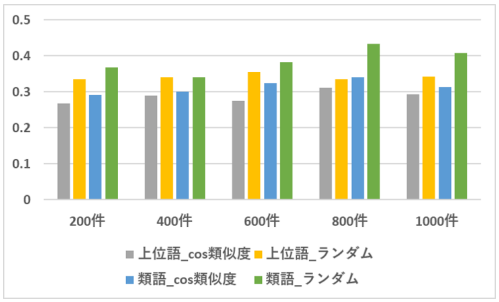

図 10　データ数，置換手法別 再学習後の置換有無での Attention 値の分布の乖離度 (DTW)：国語

図 9 の置換有無の推定点数差から，データ数が増えるにしたがって点数差が大きくなっていることがわかる．また，社会科のデータとほぼ同様に置換語をランダムに選択する手法では点数差が大きくなる傾向がある．

図 10 の Attention 値の分布の乖離度では，社会科のデータと比較して，類語で置換語を選択する方法の乖離度が高くなっている．また，データ件数と Attention 値の乖離度には明確な相関は無いように見られるが，社会科のデータと比較して乖離度が高くなっている．上位語をコサイン類似度で選択する手法に限って見ると，精度の向上が大きかった 800 件のデータに関して Attention 値の乖離度が高くなっていることがわかる．また，全体を通してみても，採点誤差が小さいほど Attention 値の乖離度が大きい傾向が見られる．

以上をふまえ，採点誤差と Attention 値の分布の乖離度に関連性が見られたことから，モデルの深い部分において，置換の有無で文章が別の文章として認識されている方が精度が高く出る傾向が見られる．これは，国語の問題でも具体的なキーワードが問題の要になっている場合があることから，社会科と同様にキーワードをしっかりと捉え，置換先の単語と別物であると捉えられたモデルの方が問題の要点を抑えられ，精度が高い事が考えられる．ただし，社会科と異なり，国語に出題されるキーワードは登場人物の名前など問題の文脈にそって用いられる特殊な意味を含んだものが多く，それらの特殊な意味を含んだ単語はそもそも学習器で学習する事が難しいため，社会科に置ける結果と異なり，キーワードが乱される悪影響が少なく，データ拡張の効果が阻害されずに発揮されたと考えられる．

また，200 件のデータでは置換手法では採点精度が向上しておらず，データ数が増えた状況では置換手法で採点制度が向上していることから，データ数が一定程度確保された状況では，元データも多く，再学習時にキーワードの再獲得が可能になり，キーワードを乱す悪影響が軽減されている可能性も考えられる．

7. おわりに

本研究では，自動採点システムの精度向上の手法として，概念辞書を用いて生徒解答中の単語を置換するデータ拡張を行う方法を提案した．

高校生の社会科と国語の模試データで実験を行ったの結果，社会科のデータではデータ数の少ない範囲において精度の向上が確認でき，国語のデータではデータ数の大きい範囲において精度の向上が確認できた．ただし，逆に社会科のデータ数の多い範囲や国語のデータ数が少ない範囲においては精度が悪化する場合が見られた．また，単語を置換したデータ精度が向上している学習において，元データと単語を置換したデータを乖離して捉える学習が行われていることが分かった．

今後の課題としては，本研究で行った比較的類似度の高い上位語や類語への置換ではなく類似度の低い単語で置換した場合の精度への影響分析や，置換手法で精度が悪化した科目とデータ数の条件で採点精度を改善する方法の模索が挙げられる．

謝辞

本研究では，株式会社学研ホールディングスとの共同研究の一環で同社より提供された解答データを使用した．本研究の一部は科研費 JP18K18656, JP19KK0257, JP20H04300, および JP20H01728. の支援を受けた．深く感謝します．

文　献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutova, “BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding”, arXiv:1810.04805, 2018.

[2] Zichao Wang, Andrew S. Lan, Andrew E. Waters, Phillip Grimaldi, Richard G. Baraniuk, “A meta-learning augmented bidirectional transformer model for automatic short answer grading”, Proceedings of the 12th International Conference on Educational Data Mining, pp.667-670, Janualy. 2019.

[3] Reham Osama, Nagwa El-Makky, and Marwan Torki, “Answering Using Hierarchical Attention on Top of BERT Features”, Proceedings of the 2nd Workshop on Machine Reading for Question Answering, D19-5825, November. 2019.

[4] Yuqi Si, Jingqi Wang, Hua Xu, and Kirk Roberts, “Enhancing Clinical Concept Extraction with Contextual Embeddings”, Journal of the American Medical Informatics Association, Volume 26, Issue 11, pp.1297–1304, November. 2019.

[5] Lun, J. Zhu, J., Tang Y., & Yang M, “Multiple Data Augmentation Strategies for Improving Performance on Automatic Short Answer Scoring”, Proceedings of the AAAI Conference on Artificial Intelligence, Vol.34, No.09 13389-13396, April. 2020.

[6] William Yang Wang and Diyi Yang, “That’s So Annoying!!!: A Lexical and Frame-Semantic Embedding Based Data Augmentation Approach to Automatic Categorization of Annoying Behaviors using #petpeeve Tweets”, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp.2557-2563, September. 2015.

[7] Steven Y. Feng, Varun Gangal, Dongyeop Kang, Teruko Mitamura, Eduard Hovy, “GenAug: Data Augmentation for Finetuning Text Generators”, Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, pp.29-42, November. 2020.

[8] Weikang Li, and Yunfang Wu, “Exploiting WordNet Synset and Hypernym Representations for Answer Selection”, Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, pp.106-115, December. 2020.

[9] F. Bond, H. Isahara, K. Uchimoto, T. Kuribayashi, K. Kanzaki, “Extending the Japanese WordNet”, 言語処理学会第 15 回年次大会発表論文集, C1-4, pp.80-83. Janualy. 2009

[10] 田口雄哉, 田森秀明, 人見雄太, 西鳥羽二郎, 菊田洸, “同義語を考慮した日本語単語分散表現の学習”, 情報処理学会第 233 回自然言語処理研究会, Vol.2017-NL-233, No.17, pp.1-5, October. 2017.