CE-001

日本語記述式問題の自動採点システムの提案
Automatic scoring system for Q&A in Japanese descriptive form

竹谷 謙吾† 高井 浩平† 森 康久仁‡ 須鎗 弘樹‡
Kengo Taketani Kohei Takai Yasukuni Mori Hiroki Suyari

1. はじめに

2020 年度から始まる大学入学共通テストには記述式問題が導入される．受験者数が数十万人規模の試験における採点には，多大な時間や人件費等の大きなコストがかかる等の問題がある．こうした背景をもとに，採点コストの削減，つまり人間の採点する問題数を減らすことを目的として，日本語記述式問題の自動採点および採点支援を行うシステムを提案する．

従来の記述式問題の自動採点手法には，機械学習により正解不正解の 2 値分類を行うもの[1]や，キーワードが含まれているか否かにより採点を行うもの[2]がある．しかし，機械学習による手法では，採点のための学習データセットを準備することが困難であり，精度も十分ではないという問題がある．キーワードを元に採点を行うものは，文章の流れを捉えていないため，こちらも十分な精度に達していないという問題がある．

そこで本システムでは，単語の並びを考慮したキーフレーズによる文字列比較をベースとして，学習データセットを必要としない採点を行う．解答文に対してシーケンスアライメントを用いることで，いくつかの正答の中から最も類似した正答を選択する．そして，キーフレーズを調べるとともに，類義語の置換や表記揺れの補正，文字数や必須キーワード，解答形式の適切さのチェックを行うことで，文章の持つ意味を考慮した採点を行う．さらに，採点結果を蓄積して学習し，正解不正解の予測を提示することで採点者の支援を行う．確実に正解不正解の判定ができる解答を自動的に採点し，確実な判断ができないものを人手による採点とすることで，信頼性の高いシステムを目指す．

2. 関連研究

近年の記述式問題の採点に関する研究として，ニューラルネットワークを用いて，採点済みの解答データをもとに二人目の採点者として正解不正解の分類を行う研究[1]や，文字列比較をベースとして独自の採点基準を元に，採点者に予想点数を提示する採点支援システムの研究[3]がある．しかしこれらには，採点済みのデータが必要であり，一度人間が採点する必要があるといった問題や，あくまで予想点数を提示するだけであり，人間が採点する問題数は変わらないといった問題がある．日本語の複雑さによる正解文と解答文の照合の難しさや，機械学習用のデータをあらかじめ準備することが難しいといった多くの課題も明らかになっており，大きな採点コストの削減に繋がるような結果は未だ得られていない．

† 千葉大学大学院融合理工学府, Graduate School of Science and Engineering , Chiba University
‡ 千葉大学大学院工学研究院, Graduate School of Engineering , Chiba University

3. 自動採点システム

提案する自動採点システムについて，3.1 節では取り扱う問題について，3.2 節以降ではシステム構成と採点アルゴリズムについて述べる．

3.1 問題例

問題として，千葉県公立高校入試の社会科の記述式問題 4 問を用いる．問題と正答の例を図 1 に示す．問題はいずれも使用しなければならない単語がいくつか指定されており，20～30 字以内の文字数制限も存在する．単語の指定と文字数制限は大学入学共通テストで想定されている問題と同様の形式である．

次の文は，冠位十二階の制度について述べたものである。文中の［　　］にあてはまる適当なことばを，「家柄」「才能」「功績」の三つの語を用いて，**20 字以内**（読点を含む。）で書きなさい。

この制度は［　　］を役人に取り立てることを目的としたもので，かんむりの色などにより朝廷での位を示した。

家	柄	に	関	係	な	く	、	才	能
や	功	績	を	重	視	し	た	人	物

図 1　問題と正答の例

3.2 システム構成

システムの構成図を図 2 に示す．あらかじめ設定された解答条件やキーフレーズ等の採点基準を元に自動採点を行う．自動採点部分により正解不正解の判断ができない場合，手動採点に移行する．手動採点部分では，これまでの採点結果を元に採点予測を提示することで採点支援を行う．デジタル化された文字列データを入力とし，正解・不正解の２値を結果として出力する．

システムの初期設定として，問題の必須キーワード，文字数制限，適切な文末の品詞，キーフレーズ等の採点基準を設定する．キーフレーズとは，重要な単語を含む短い文章のことであり，キーフレーズによる採点を行うことで，重要な単語とその並び，自然な文章かどうかを判定する．必須キーワード，キーフレーズの例を表 1 に示す．

表 1　初期設定例

キーワード	キーフレーズ 1	キーフレーズ 2
家柄 才能 功績	家柄にとらわれず 家柄に関係なく 家柄によらず	才能や功績のある人物 才能がある人や功績がある人

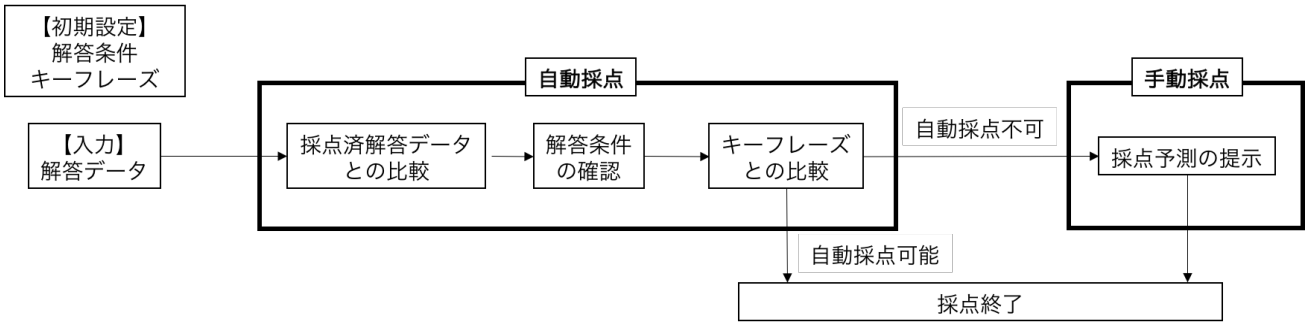

図 2　システム構成図

3.3 採点済み解答データとの比較

正解不正解の結果に基づいて蓄積された採点済み解答データと，新たに採点をおこなう解答データを比較し，同一の場合即座に正解不正解の判断を行い，採点を終了する．これにより，同一の解答を複数回採点する手間を省略することができる．

3.4 解答条件の確認

採点を行う解答データに対して，必須キーワードが全て含まれているか，文字数制限が満たされているか，文末の品詞が適切なものであるかなど，設定した解答条件が全て満たされているかを確認する．文末の品詞の確認とは，穴埋め形式の記述式問題において，次に続く文章と自然に繋がるために文末の単語が適切な品詞になっているかを調べることである．例えば図 1 の例では，名詞が適切な品詞であり，動詞や形容詞が文末にある場合は自然な文章とはならない．これらの解答条件が満たされていない場合，即座に不正解と判断する．

3.5 キーフレーズとの比較

解答文とキーフレーズを比較して類似するキーフレーズとの共通部分・非共通部分を抽出し，表記揺れの補正や類義語の置換を行うことによって自動採点を行う手法について，3.5.1 項以降で述べる．

3.5.1 シーケンスアライメント

解答文とキーフレーズの比較を行うためにシーケンスアライメントを行う．シーケンスアラインメントとは，主にバイオインフォマティクスの分野で利用され，2 つ以上の DNA などの配列において類似した領域を特定できるように並び替える技術である．本システムでは，シーケンスアライメントアルゴリズムの 1 つである Needleman-Wunsch Algorithm[4]を用いる．このアルゴリズムは 2 つの配列において，一致する文字の数を最大にし，不一致の数が最小になるようにスコア関数を用いて配列を並び替えることで最適な配列を取得する．2 つの配列を並べた時，位置 p にある 2 つの文字が一致する場合+1 点(マッチ)，位置 p にある 2 つの文字が不一致の場合-1 点(ミスマッチ)，位置 p において，片方の配列の文字に対して，もう片方の配列の文字が存在しない場合-1 点(ギャップ)とするスコア関数を用いる．2 つの配列の配列長が同じになるように欠損に対応するギャップ記号を挿入し，スコアが最大となるように配列を並び替えることで最適な配列を得る．

採点を行う解答文と各キーフレーズについて形態素解析を行い，文章を単語で区切るとともに，特定の品詞のみを抽出する．形態素解析を行なった文章に対してシーケンスアライメントを行うことで，図 3 のように，解答文とキーフレーズとの共通部分・非共通部分を抽出することができる．これにより，単語の並びに着目して解答文とキーフレーズとの一致率を計算し，設定されているいくつかのキーフレーズの中から最も類似した文章を選択することや，次節以降で述べるように，非共通部分について単語の意味レベルまで考慮した比較を行うことができる．

原文

解答文	家柄に関係なく、才能がある人や功績がある人
キーフレーズ	才能や功績のある者

形態素解析

解答文	家柄	関係	なく	才能	ある	人	功績	ある	人
キーフレーズ	才能	功績	ある	者					

シーケンスアライメント

解答文	家柄	関係	なく	才能	ある	人	功績	ある	人
キーフレーズ				才能			功績	ある	者

図 3　シーケンスアライメント

3.5.2 表記揺れの補正

漢字かな混じり文をローマ字文に変換するプログラム・辞書である KAKASI[5]，単語を表層系から基本形に変換することができる形態素解析器 MeCab[6]を使用し，シーケンスアライメントにより抜き出された非共通部分の表記揺れの補正を行う．単純な文字列の比較では，同一の単語であっても漢字とひらがなのように表記が異なる場合や形容詞・動詞の活用形により表記が異なる場合について正しく採点を行うことができない．KAKASI による読み方に着目したローマ字文への変換と，MeCab による基本形への変換により，表記方法に依存しない採点を行うことができる．これにより，少ない数の正答で広い範囲の解答を採点することができるようになる．

3.5.3 類義語の置換

日本語の概念辞書である Wordnet[7]を使用し，シーケンスアライメントにより抜き出された非共通部分について，単語の意味を比較する．Wordnet は単語を synset と呼ばれる類義関係のセットでグループ化しており，一つの synset が一つの概念に対応している．Wordnet における synset の例を表２に示す．

表 2　Wordnet における synset の例

synset	単語
person	人格者 , 人 , 人称 , 人間 , 方 , 者
individual	個人 , 人 , 人間

解答文のキーフレーズとの非共通部分について，それぞれの単語の synset を比較し，同じ概念の単語であると判断した場合に解答文の該当する単語をキーフレーズの該当する単語で置換し，再度比較を行う．単語の意味に着目して比較を行うため，少ない数のキーフレーズで広範囲の解答を採点することができる．

シーケンスアライメントにより抽出された非共通部分に対して，表記揺れの補正と類義語の置換を行う例を図 4 に示す．

3.6 採点予測の提示

自動採点部分において正解不正解の判断ができなかったものについて，手動採点の支援として，採点予測の提示を行う．

全解答データに含まれる助詞と記号を除く全ての単語を辞書に登録し，tf-idf を用いて，全ての解答データにおける単語の出現頻度に対する各解答データの単語の出現頻度をもとに各解答データの特徴量を取得し，ベクトル化する．tf-idf によるベクトル化では，ほぼ全ての文章に出現する必須キーワードなどの単語は重要度が低くなり，その他の単語の重要度が高くなるため，文章ごとの差異を捉えることができる．

その後，手動採点を行う解答データと，採点済の全解答データのベクトルのコサイン類似度を計算することにより，ベクトル間の類似度を算出する．採点済みの正解・不正解のどちらの解答に対する類似度が高いかにより，予測の提示を行う．ここで，類似する解答がない場合は採点予測を提示せず，類似する解答があった場合のみ採点予測の提示を行うこととする．

4. 実験

提案する自動採点システムを用いて行なった実験について，4.1 節では実験設定について，4.2 節では実験結果について述べる．

4.1 実験設定

3.1 節で述べた形式の問題 4 問について，それぞれの問題に対して，大学生と中学三年生を中心に 68 人分の解答データを用意し，システムの入力として用いる．

採点予測を使用する場合と使用しない場合に分けて実験を行い，自動採点率と採点精度を評価の指標として算出する．ここで，採点予測を使用する場合の結果は，30 番目以降の解答データの採点において採点予測を行い，その結果をもとに自動採点を行なったとした結果である．自動採点率は，68 の解答のうち自動採点を行なった解答の割合であり，採点精度は，自動採点を行なった解答のうち正しく採点できている解答の割合である．

4.2 実験結果

実験結果を表 3,4 に示す．

表 3　採点結果（採点予測なし）

	問 1	問 2	問 3	問 4	平均
採点精度(%)	100	100	100	100	100
自動採点率(%)	39.7	39.7	39.7	50.0	42.3

表 4　採点結果（採点予測あり）

	問 1	問 2	問 3	問 4	平均
採点精度(%)	97.1	95.6	95.6	100	97.1
自動採点率(%)	52.9	52.9	60.3	60.3	56.5

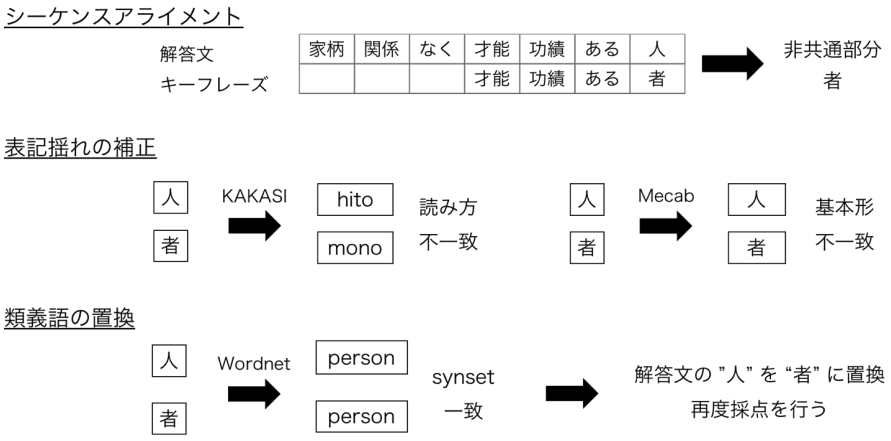

図 4　非共通部分に対する処理の例

採点予測なしの場合では，すべての問題に対して採点精度が 100%であり，誤った採点をすることなく，非常に信頼性の高い採点を行えていることが分かる．採点精度が 100%であれば一度採点した問題を再度採点する必要がないため，自動採点率がそのまま採点コストの削減に繋がることになる．採点予測なしの場合の自動採点率は平均して 42.3%である．

採点予測ありの場合では，採点予測なしの場合に比べて自動採点率は向上しているが，採点精度は低下している．採点精度の低下は，tf-idf によるベクトル化では単語の出現頻度を素性にベクトル化を行なっているため，ほぼ同じ単語を使っているがその順番は異なるという文章を判別することができないことが原因であると考えられる．しかし，採点予測による支援として活用する場合には十分な精度であると考えられる．

問題によって自動採点率にばらつきが生じているのは，正答の設定に問題があることや，問題の難易度による正解不正解数のばらつきが原因であると考えられる．人手で設定したキーフレーズの数が約 10 文と少なかったことや，用意した解答データが 68 人分と少なかったことを考えると，設定する正答の数を増加させることや，より多くの解答データを用意することにより，自動採点率や採点精度は大きく向上すると考えられる．特に，採点済解答データとの比較のプロセスや採点予測の提示のプロセスでは，採点を行なった問題数が多ければ多いほど，より多くの問題を自動採点し，高い精度の採点予測を行うことができるため，解答データ数の増加による自動採点率や採点精度の向上が期待できる．

また，正解・不正解文ごとの採点結果を見ると，正解文に対しての平均自動採点率は 46.7%，不正解文に対しての平均自動採点率は 34.0%となった． 不正解文に対する自動採点率が低くなっている理由は，解答条件を満たしている不正解文について，不正解と判断するためのキーフレーズ等の判断基準が設定されていないことが原因であると考えられる．より自動採点率と採点精度を向上させるためには，不正解と判断するための採点基準を新たに設ける必要がある．

5. おわりに

本稿では，大学入学共通テストに記述式問題が導入されることを背景として，シーケンスアライメント等を用いた文字列比較をベースとして，日本語記述式問題の自動採点及び採点支援を行うシステムを提案した．約 270 の社会科の記述式解答文に対して 100%の精度で 42.3%の解答を自動採点，97.1%の精度で 56.5%の解答の採点予測を行うことができ，採点にかかるコスト削減に繋がることを確認できた．従来の自動採点及び採点支援の手法と比較すると，精度が非常に高く，採点の信頼性が高いという点や，学習データを必要とせず，一人目の採点者として機能し，人間の採点する問題数を削減することができている点において優位性があると考えられる．学習データセットを必要としない手法であるため，大学入学共通テストのような大規模なテストだけでなく，比較的小規模な解答データに対する採点支援システムとしての利用も見込める．

今後の展望として，さらなる自動採点率の向上や，より大規模なデータでの評価が挙げられる．今回は使用した解答データの数が少なく，採点精度や自動採点率の向上が期待できる提案手法に関しても満足な検証が行えていない．大学入学共通テストでの使用を考えると，数万～数十万の解答データでの検証が望ましい．実際の採点のことを考慮すると，採点精度は 100%であることが前提であるため，採点精度を落とすことなく自動採点率を上げていく必要がある．

謝辞

本システムを作成するにあたり，実験を行うための解答データの提供に協力をしていただいた皆様に心から感謝申し上げます．

参考文献

[1] 寺田 凛太郎, 久保 顕大 , 柴田 知秀, 黒橋 禎夫, 大久保 智哉, “ニューラルネットワークを用いた記述式問題の自動採点”, 言語処理学会 第 22 回年次大会 発表論文集, pp.370-373 (2016).
[2] “文章自動採点システム – クラウドゼミ”, http://www.cloudsemi.com/Mtext.pdf
[3] 亀田 雅之, 石岡 恒憲, 劉 東岳, “担当記述式問題解答文の採点支援システム JS4 の試作”, 言語処理学会 第 23 回年次大会 発表論文集, pp.1137-1140 (2017).
[4] Needleman, Saul B. and Wunsch, Christian D, “A general method applicable to the search for similarities in the amino acid sequence of two proteins.” Journal of Molecular Biology, 48, pp.443-453 (1970)
[5] “KAKASI – 漢字→かな(ローマ字)変換プログラム”, http://kakasi.namazu.org
[6] “MeCab: Yet Another Part-of-Speech and Morphological Analyzer”, http://taku910.github.io/mecab/
[7] “日本語 WordNet”, http://compling.hss.ntu.edu.sg/wnja/index.ja.html