# 大規模実データにおける記述式問題自動採点システムの検証

竹谷謙吾[1], 高井浩平[1], 清水杏奈[2], 早川純平[2], 森康久仁[3], 須鎗弘樹[3]

[1]千葉大学大学院融合理工学府，[2]千葉大学工学部，[3]千葉大学大学院工学研究院
k_taketani@chiba-u.jp

## 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予定であり，採点には多大なコストを要する．こうした背景から，採点コストの削減，つまり人間が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を行うシステムを提案する．

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結果付きのデータをもとに採点を行う研究[1]や，設定した条件をもとに採点を行う採点支援システムの研究[2]がある．しかし，大量の採点結果付きのデータが得難いといった課題や，精度が十分ではないといった課題がある．

本システムでは，人手により事前に設定された採点設定をもとに，採点結果付きのデータを必要としない高精度な採点を行う．解答と正答の文字列を解析し，自動で正解不正解の判断ができる問題のみ自動的に採点する．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. 自動採点システム

提案する自動採点システムについて，2.1 節ではシステム構成，2.2 節以降では採点アルゴリズムについて述べる．

### 2.1 システム構成

システムの構成図を図 1 に示す．あらかじめ人手により設定された採点設定をもとに自動採点を行う．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．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．採点結果付きのデータをテストデータとして用いる場合には，手動採点にまわった解答をクラスタリングし，特徴を可視化することで採点設定改善の支援を行う．

採点設定として，以下の項目を設定する．
・文字数制限(範囲内でなければ不正解とする)
・事前置き換え単語(表記を統一するために採点前に置き換える語)
・NG ワード(含まれていれば不正解とする語)
・NG エンドワード(文末に含まれていれば不正解とする語)
・キーワード(含まれていなければ不正解とする語)
・キーフレーズ(含まれていれば正解とする意味的なまとまりを持つ短い文)

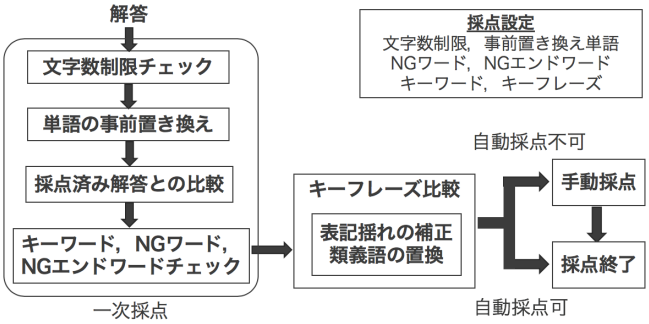


図 1　システム構成図

### 2.2 一次採点

採点設定をもとに，満たすべき条件を全て満たしているかを確認する．一つでも満たしていない場合，不正解であると判定し，採点を終了する．全て満たしている場合，キーフレーズ比較の処理へ移行する．

採点済み解答との比較では，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．採点が行われた解答は採点結果とともに採点済み解答に追加される．

### 2.3 キーフレーズ比較

Needleman-Wunsch Algorithm[3]というシーケンスアライメントアルゴリズムによって解答文とキーフレーズの比較を行う．このアルゴリズムは 2 つの配列において，一致する文字数を最大にし，不一致の数が最小になるよう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った後にシーケンスアライメントを行うことで，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．その様子を図 2 に示す．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，非共通部分を抽出し，比較を行うことができる．

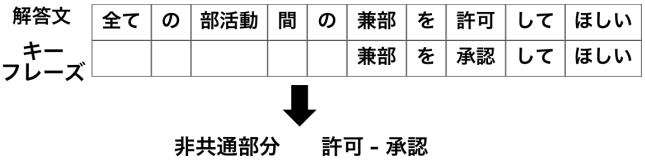


図 2　シーケンスアライメントの例

抽出された非共通部分について，表記揺れの補正と類義語の置換を行う．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に正しく採点を行うことができない．また，「人」と「人間」のように，同じ意味を持つ単語についても表記が異なるため正しく採点を行うことができない．そこで本システムでは，読み方や単語の意味カテゴリに着目して同一の意味であるかどうかを判断することで表記方法に依存しない採点を行う．

### 2.4 採点設定の改善支援

採点結果付きのテストデータがあり，採点設定を改善することができる場合，テストデータの採点終了後に手動採点にまわった解答をクラスタリングする．現在の採点設定では自動採点できない文章に共通する特徴が明確になるため，有効なキーフレーズや NG ワード を人手により設定することが容易となる．

## 3. 実験

提案する自動採点システムを用いた実験について，3.1 節では実験設定，3.2 節では結果について述べる．

### 3.1 実験設定

中学生を対象に行われた模試の記述式問題 3 科目分それぞれ約 1200 人分の解答データについて採点を行う．その後，採点設定の改善をして再度採点を行う．問題の特徴を以下に示す．

・国語　20 字以内　穴埋め形式
・社会　25 字以内　穴埋め形式
・理科　文字数制限無し　自由記述形式

評価の指標として，自動採点率と採点精度を算出する．自動採点率は，全解答文のうち本システムで自動採点の対象となった解答の割合を表し，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を表す．

### 3.2 実験結果

実験結果を表 1，表 2 に示す．

表 1　採点結果(採点設定改善前)

| | 国語 | 社会 | 理科 | 平均 |
|---|---|---|---|---|
| データ数(人) | 1198 | 1174 | 1195 | 1188 |
| 自動採点率(%) | 55.8 | 54.1 | 63.0 | 57.6 |
| 採点精度(%) | 99.4 | 98.6 | 100 | 99.3 |

表 2　採点結果(採点設定改善後)

| | 国語 | 社会 | 理科 | 平均 |
|---|---|---|---|---|
| データ数(人) | 1198 | 1174 | 1195 | 1188 |
| 自動採点率(%) | 64.5 | 63.2 | 76.4 | 68.0 |
| 採点精度(%) | 99.9 | 99.6 | 100 | 99.8 |

表 1,2 より，全ての問題に対して非常に高い精度で採点できていることが分かる．また，採点設定を改善することで，自動採点率・採点精度ともに向上していることが分かる．教科別で見ると理科が最も良い結果が得られていることが分かる．これは，理科は単語を答える問題に近いため，表現の幅が小さく，採点設定が容易であったことが理由だと考えられる．明確な答えが存在し，文字数制限や使用単語の指定などによって表現の幅が小さくなればなるほど自動採点率・採点精度は向上すると考えられる．

## 4. 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，日本語記述式問題の自動採点を行うシステムを提案した．国語社会理科の 3 教科それぞれ約 1200 人分の模試の解答データに対して実験を行った結果，100%に近い精度で約 60%の解答を自動採点することができ，採点にかかるコスト削減に繋がることを確認できた．従来の自動採点手法と比較すると，採点結果付きのデータを必要とせず，高い精度で自動採点が行える点において優位性があると考えられる．

今後の展望として，より自由度の高い文章においてどれだけ有効であるかを検証するとともに，100%の採点精度を維持しながら自動採点率の改善を行っていく必要がある．

## 謝辞

本システムを作成するにあたり，実験を行うためのデータを提供して下さった株式会社進学研究会に心から感謝申し上げます．

## 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]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).

[3] 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)