小論文の自動採点に向けたオープンな基本データの構築および現段階での自動採点手法の評価

竹内 孔一
岡山大学大学院自然科学研究科
koichi@cl.cs.okayama-u.ac.jp

大野 雅幸
岡山大学工学部情報系学科
pw2z9792@s.okayama-u.ac.jp

泉仁 宏太
岡山大学工学部情報系学科
pm9n6cei@s.okayama-u.ac.jp

田口 雅弘
岡山大学院社会文化科学研究科

稲田 佳彦
岡山大学院教育学研究科

飯塚 誠也
岡山大学全学教育・学生支援機構

阿保 達彦
岡山大学大学院自然科学研究科

上田 均
岡山大学大学院自然科学研究科

1 はじめに

本研究プロジェクトでは小論文採点の負担や評価のばらつきを軽減することを目標とした小論文の自動採点手法の開発を目指している．自動採点手法の開発を通して，将来的に記述のどこが悪いのか，どう直せば良いかといった指摘が行えるシステムの開発を視野に入れている．以下では，現在構築している小論文データとシステムの現状を報告する．

記述式による課題には大きく分けて，正解の文章が仮定できるものと，仮定できないものの 2 種類があると考えられる．ここでは，先行研究 [5] に従い前者のものを短答式タイプ，後者のものをエッセイタイプと呼ぶ．

短答式の場合，自動採点手法として模範解答を用意し，答案との適合度による比較が考えられる．一方で，エッセイタイプでは自分の考えを書くため模範解答を用意することが難しい．また字面上の答案は数百字にわたるため，模範解答を用意したとしても内容の一致度を測定するには含意認識技術が必要となる．

英語を中心とした記述式問題の答案に対する自動採点の研究は文献 [4] にまとめられている．一方，日本語文書に対する手法としては，短答式タイプのものについて機械学習を利用した手法 [6, 5] が試されている．一方，エッセイタイプは石岡の Jess[3] が統計的な異常値検出をベースに構築されている．

小論文課題を含めた記述式問題の自動採点手法の開発において困難である点の一つは，利用できる共通の小論文データが存在しないことである．先行研究 [6, 5] も学内試験の答案や模試の答案を利用しており，他の研究機関が利用できる見通しは無い．

そこで本研究では，小論文採点手法を構築するにあたって，公開できる模擬試験の小論文データの構築を始めている．受講者が講義を理解し，講義に対する複数の問題に対して小論文課題を制限時間内に作成する．現段階で 320 人分 (6 課題) の小論文を集めており，人手による採点スコアを付与している段階である．研究で利用できることから，小論文の自動採点システム作成の際のテストベッドとして利用できることを期待している．

そこで本論文では模擬試験データの設計と構築に関する議論，並びに，得られた小論文データを基に，構築している自動採点システムによる評価値が人手の採点スコアとどの程度の相関があるかについて述べる．

2 模擬試験データの構築

模擬試験データの構築は，単に自動採点手法だけに着目したのでは無く，現在行われている筆記による記述試験から電子的なデータへの変換可能性まで考慮して構築している．下記ではまず模擬試験によるデータ収集の枠組として，取り上げた検討事項ついて議論し，実施した模擬試験の内容，並びに，現在得られている小論文データの内容について記述する．

2.1 模擬試験による小論文収集の枠組

小論文を収集する上で，他の研究グループでも利用できることを前提とした．このため，実際の試験問題に対する答案の利用は想定せず，模擬試験により小論文を収集する．模擬試験の基本的な枠組として，受講者が講義を聴いた後，講義に関する課題を与えられて制限時間内に小論文を記述する．また，最終的な答案は電子ファイルであること，電子ファイルは後に他の研究グループでも利用できるように受講者から許諾をいただくこととした．

この基本枠組をもとに模擬試験内容を具体的に構築する際に下記のことを考慮した．

(a) 答案入力の問題

最終的に答案を言語処理可能なテキストデータで収集する必要がある．そこで問題となるのは，模擬試験の答案の入力方法である．現状では，小論文の試験は筆記による入力が一般的であると考えられる．これにより受講者の漢字の能力なども同時に評価できるが，一方で，電子化するためには OCR などで文字読み取りを行うか，人手による再入力が必要となる．

そこで，本研究プロジェクトでは OCR 読み取り装置による試験の可能性を模索するために，筆記による小論文の入力を実施する．さらに，受講者同士で解答終了後，人手で他人の答案を電子化テキストデータに入力する作業を行うこととした．これにより正しく電

子化された答案と，筆記による答案の画像，ならびにOCR に掛けた答案を得ることができるため，現状での OCR の精度を確認することができる．

(b) 講義と課題の種類

講義は 2 種類とし，一方は課題の提示方法として，講義内容のスライドや小論文課題を最初から印刷して配布した．他方は講義受講後に課題のみ配布し，受講中にどこが小論文として聞かれるかは受講生には知らせない方法をとった．難易度を変えることにより，幅広い質の小論文を得ることを試みた．

(c) 採点評価の枠組

得られた小論文に対して人手による採点を付与する．小論文の評価方法については既に先行研究 [1, 2, 3] が指摘しているように定常的な評価基準は存在しない．そこで後の自動採点システムの開発の手助けになるように内部的に次の 4 つの基準に分解して評点をつけることにした．(1) 設問に対する理解力，(2) 文章の論理性 (論述の展開の良さ)，(3) 妥当性 (論述の内容が妥当で説得力があるか)，(4) 文章力 (言葉の使い方，誤字脱字) である．これらの 4 項目について 1 から 5 までのスコアを付け，最終的な小論文の良さはその合計点で表す．現在，得られた小論文に対して上記の基準で採点中である．上記の 4 項目で多数の小論文を採点するのは容易ではないため，問題がでてきた場合は見直すことを想定している．

2.2 模擬試験の実施

模擬試験は 8 月と 12 月にそれぞれ 2 日間開催し，講義は 2 種類，各講義で 3 問の課題を与えた．4 日間とも講義と課題は同じである．受講生は 1 日で 2 講義 6 課題の小論文を記述した．午前と午後でそれぞれ 30 分の講義を受講し，その後 1 時間で 3 つの小論文を記述する．教室の関係から 1 日で 100 人程度の受講生を募集した．受講生は岡山大学内の学生に限定し，1 度受けた学生は受講させていない．これにより幅広く文書を集めた．課題内容は表 1 の通りである．

表 1: 講義の内容と課題の文字数制限

	講義 1	講義 2
内容	グローバリゼーションの光と影	自然科学の構成と科学教育
字数	(1)300 字以内， (2)250 文字以内， (3)300 字以内	(1)100 字以内， (2)400 字以内， (3)500〜800 字

2.3 現段階の小論文データ

講義のタイトル，講義内容の書きおこし (2000 字以下程度)，各課題の出題意図と評価のポイント，および答案 (筆記の画像データ pdf，書き起こし word ファイル，書き起こし excel ファイル) である．受講者は全体で 328 人．講義 1 の答案は各課題について 328 件，講義 2 の答案は各課題について 327 件である (1 人途中棄権による)．よって 1965 件の小論文が集められている．各小論文には，実施日と受講者 ID が付与されている．人手による採点スコア付けは講義 1 の前半 161 人分が現在終了している．次節ではこの一部を利用して構築中の小論文採点手法を評価する．

3 自動採点システム: 各モジュールの設計と簡易な実験結果

記述式問題の自動採点には様々な方法が考えられる．大きく分けて既に採点したデータをもとに評価する機械学習による方法 [6, 5] と，採点データは不要で採点の基準や評価方法を考慮しつつ採点する方法 [3] である．ここでは後者のアプローチをとる．

前節の評価で述べたように，小論文の採点手法として 4 つの評価軸を設定した．よって自動採点システムでは評価軸に即した 4 つの評価モジュール (理解力モジュール，論理性モジュール，妥当性モジュール，文章力モジュール) を仮定し，構築を進めている．全体像を図 1 に示す．

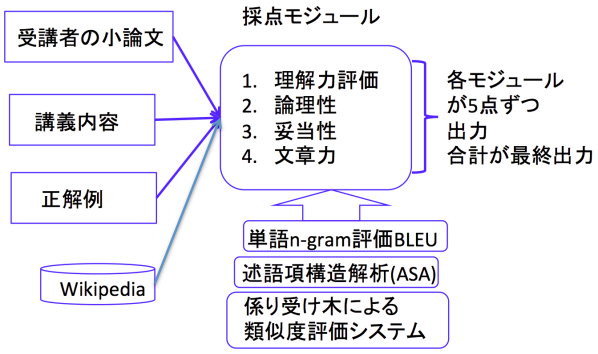

図 1: 自動採点システムの採点モジュールの構成

自動採点システムが利用できるデータとして，受講者の小論文，および講義内容テキスト，正解例がある．講義内容テキストとは，講義そのものはスライドを利用した発表形式であるが，その内容を書き言葉で整理した 2000 文字以下のテキストである．正解例は問題によって仮定できる場合には作成する．そうでない場合は，採点の結果から高いスコアを得ている答案を選び正解例にすることができる．

図 1 のモジュールの下に記述しているツール群は各モジュールで利用している処理システムである．現在，理解力評価モジュールと妥当性評価モジュールを構築をすすめており，下記に方針と手法について簡単に述べる．

3.1 理解力評価モジュール

理解力評価モジュールでは受講者が課題の指示に即した内容の小論文が書けているかを測定することが求められる．よって手法としては講義内容テキスト，および利用できる場合は正解例との単語を基にした類似度により評価する．

単語ベースの類似度として (A) 単純な内容語の一致数，(B)BLEU を利用した n-gram による類似度を試し

ている．(A) において文書 X と Y 類似度 $simA(X,Y)$ は各単語の内容語を x, y とすると

$$simA(X,Y) = \sum_{x \in X, y \in Y} I[x = y] \quad (1)$$

で表される．この際，I は [] 内の命題が成立したときに 1 を返す関数とする．ここで内容語とは形態素解析器の品詞が名詞，動詞，形容詞で，自立語のものを指す．

一方 BLEU による評価 $simB$ では 1-gram から 4-gram まで BLEU が出力する類似度を合算する．

$$simB(X,Y) = \sum_{i=1}^{4} bleu\_n\_gram(i) \quad (2)$$

ここで $bleu\_n\_gram(i)$ は BLEU が出力する i-gram のスコアを表している．さらにこれらを合算した類似度も利用する．

$$simAdd(X,Y) = simA(X,Y)/C + simB(X,Y) \quad (3)$$

ここで C は混合パラメータで現段階では 10 に固定している．

3.2 妥当性評価モジュール

妥当性評価モジュールでは受講者が小論文で展開している内容がどれだけ妥当であるかを評価するモジュールである．妥当かどうかを判断するのは容易ではないが，ここでは小論文の部分命題が妥当ならばその内容は既にどこかで知られている内容であると仮定した．例えば 3.3 節の課題 1 に対する回答例としてグローバル化に関係する多国籍企業 (講義では「マクドナルド」を取り上げている) を取り上げて具体的に論を展開するなどが可能であるが，こうした記述は Wikipedia の「マクドナルド」の項目にも見受けられる[1]．

そこで各小論文と Wikipedia との命題の類似度を計算する．文同士の命題的内容の類似度を計算する方法として，研究室で開発している述語項構造解析器 ASA[2] を利用した係り受け解析木を作成し，文同士の係り受け木が大きい物を採用する．ASA は係り受け解析器 CaboCha の出力を受けて，係り受けに対して意味役割を付与し，各文節の意味的な主辞，助詞，能動態や受動態などを識別する．例えば「太郎は，次郎が壁にボールを投げたのを見た。」は「太郎は，ボールを投げた。」を含意しないが，共通する単語の頻度で評価すると類似性は高い．しかし，図 2 に示すように係り受け木を作成することで，木構造では共通する部分が少なく評価が下がる仕組みである．

文書 X と Y があった場合，各書内の 1 文を xs, ys とし，各係り受け木を T_{xs}，T_{ys} とする．2 つの木に対して最大マッチする部分木を検出し，部分木内の単語

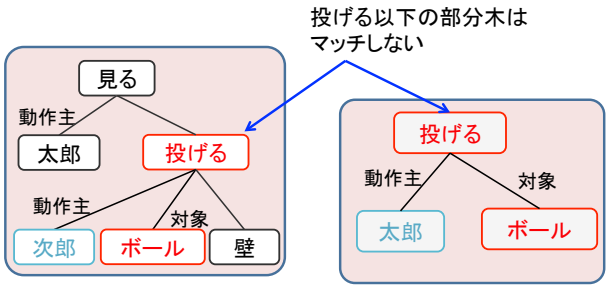

図 2: 述語項構造解析器 ASA を利用した係り受け木の比較

数を返す関数を $MSTree()$ とすると文書 X と Y に対する係り受け解析木による類似度 $simT(X,Y)$ は，

$$simT(X,Y) = \sum_{xs \in X} \sum_{ys \in Y} MSTree(T_{xs}, T_{ys}) \quad (4)$$

となる．現段階では部分木の直接的な影響が人手による評価スコアとどのような相関になるか調べるために，正規化は行わず $simT$ を妥当性モジュールの出力スコアとする．

$MSTree()$ の計算には含意認識タスクで構築している係り受け木比較器 [7] を利用する．係り受け木比較器には，木ではなく単語レベルでのマッチの場合に無視する足きり値が設定できる．大きい値ほど小さな木のマッチを無視する．そこで次節の実験では足きり値を変更して実験する．

Wikipedia の文書に対して小論文と上記の係り受け木による類似度を計算するが，Wikipedia の全文書を対象とすることは計算時間の問題と，不要な文書による不正確な適合が生じるため，一部の文書に候補を絞りたい．そこで (1) 正解例，もしくは (2) 学生の小論文から内容語を抽出し，Wikipedia の各ページでカテゴリー名が含まれてるページのみを比較対象とした．

3.3 評価実験

講義 1 の小論文課題の 1，2，3 について人手による採点スコアが付与されていることから，この部分集合である 30 件について，理解力モジュールならびに妥当性モジュールのスコア評価する．評価方法は各モジュールの出力値と採点スコアとの相関係数を利用する．ここで，講義 1 の内容および課題について記述する．

講義 1 の内容: グローバリゼーションの光と影

課題 1: グローバリゼーションは、世界、または各国の所得格差をどのように変化させましたか。また、なぜ所得格差拡大、または縮小の現象が現れたと考えますか。300 字以内で答えなさい。

課題 2: 多国籍企業は、グローバリゼーションの進展の中でどのような役割を果たしましたか。多国籍業の具体例をあげて、250 字以内で答えなさい。

[1] 2017 年 1 月 18 日アクセス.「マクドナルド」のページに「グローバリゼーション」の文字が掲載されている.

[2] http://cl.cs.okayama-u.ac.jp/study/project/asa/asa-scala

課題 3: 文化のグローバリゼーションは、私たちの生活にどのような影響を与えましたか。また、あなたはそれをどのように評価しますか。具体例をあげて、300 字以内で答えなさい。

エッセイタイプの課題ではあるが内容には幅がある．例えば課題 1 では字数は長いがある程度回答すべき内容は絞られている．よって正解例が仮定できる．一方で課題 3 は自分の考えを中心に記述するものであるため，正解例の仮定は難しい．

まず表 2 に理解力モジュールについて評価した結果を示す．ここで swb は単純な単語マッチを示し，swb+n-gram はさらに BLEU の結果を加算したものである．相関係数は人手で付与されたのスコアのうち，理解力に相当するスコアとの比較である．まず表 2 から単純な単語マッチによる方法が 0.5〜0.7 とある程度相関が出ることが分かる．一方で n-gram まで考慮した場合に相関が下がる．

表 2: 理解力モジュールの評価 (相関係数)

	swb	swb+n-gram
課題 1	0.523	0.480
課題 2	0.549	0.471
課題 3	0.726	0.716

ここで，石岡ら [5] で提案されるように，採点基準に基づくスコアの修正を入れてみる．例えば課題 1 では「ジニ係数」についての表現が課題の意図として提示されているため,「ジニ」を含む n-gram 部分のスコアを 2 倍にする処理を行った所，swb+n-gram は 0.547 となり，上記の相関を上回る結果を得た．このことから課題意図をスコアにうまく取り込むことによって精度の向上が見込まれる．

次に，妥当性モジュールについてであるが，1 文書の解析に時間がかかるため現時点では課題 1 を対象にした評価を表 3 に示す．

表 3: 妥当性モジュールの評価 (相関係数，課題 1 のみ)

	足きり値 (0.3)	足きり値 (0.5)
正解例での Wiki	0.348	0.276
小論文での Wiki	0.333	0.263

表 3 では妥当性モジュールで基盤となる Wikipedia の文書の取り出し方で，正解例を利用した場合と各受講者の小論文を利用した場合で相関係数を求めている．正解例は課題 1 のみ仮定できるもので，これにより約 14 万文の Wikipedia の文書が獲得され，各小論文との係り受け木の比較を行っている．一方で，各小論文から Wikipedia 文書を獲得した場合はキーワードに依存するため，文書量が異なってしまう (獲得された Wikipedia の文書は約 5 万文から 23 万文)．妥当性モジュールの出力は正規化していないため，文書量が異なるとその影響を受ける．表 3 で正解例を利用した場合，足きり値に関係無く相関係数が勝っている原因の 1 つと考えられる．

また足きり値の影響であるが，大きい値にすると相関係数が顕著に減少している．大きい値は，大きな部分木の一致を意味するが，これは係り受け木レベルで同じ単語によるマッチを行っているため，柔軟に言い換えに対応しておらず，精度が下がったのではないかと考えられる．よってより幅広い言い換えを考慮した文の含意関係を同定する手法の開発が求められることが明らかになった．

4 おわりに

本論文では自動採点手法で利用可能なオープンな小論文データの構築について現状を報告した．また現段階の小論文データを利用して簡易な小論文採点手法を評価した．プロジェクトの状況に依存するが，小論文データは今後 2 年構築する予定である．採点が完了した段階で順次公開する予定である．

5 謝辞

本研究の進めるに当たり大学入試センター石岡恒憲先生には貴重なご意見，ならびに Jess の利用を許諾頂きました．また研究の遂行にあたり岡山大学学務部にご協力いただきました．深く感謝いたします．

参考文献

[1] E. V. Steedman, M. Tillema, G. Rijlaarsdam, and H. van den Bergh, editors. *Measuring Writing Recent Insights into Theory, Methodology and Practices (Studies in Writing)*. Brill Academic Pub, 2012.

[2] 石川巧. 「いい文章」ってなんだ? 入試作文・小論文の思想. ちくま新書, 2010.

[3] 石岡恒憲. 日本語小論文の自動採点および作文支援システムの開発. 科学研究費補助金研究成果報告書, 2007.

[4] 石岡恒憲. コンピュータ上で実施する記述式試験—エッセイタイプ，短答式，マルチメディア利用について—. 電子情報通信学会誌, Vol. 99, No. 10, pp. 1005–1011, 2016.

[5] 石岡恒憲, 亀田雅之, 劉東岳. 人工知能を利用した短答式記述採点支援システムの開発. 電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション, pp. 87–92, 2016.

[6] 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. 第 22 回言語処理学会年次大会発表論文集, pp. 370–373, 2016.

[7] 齋藤彰, 竹内孔一. コピュラ文を考慮した述語項構造解析器による含意認識. 電子情報通信学会言語理解とコミュニケーション研究会, 2017. (to appear).