小論文自動採点データ構築と理解力および妥当性評価手法の構築

大野 雅幸
pw2z9792@s.okayama-u.ac.jp
泉仁 宏太
岡山大学大学院自然科学研究科
pm9n6cei@s.okayama-u.ac.jp
竹内 孔一
koichi@cl.cs.okayama-u.ac.jp
小畑 友也
岡山大学工学部
pbgn8vxd@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

本研究では小論文を自動採点するシステムの構築を目指している [1]．小論文の採点は既に先行研究で指摘している通り [2, 3, 4]，標準的な評価基準は存在しない．そこで課題の理解力，表現の論理性，内容の妥当性，文書の誤字脱字や文法性など 4 つの基準を設定し，人手による採点データとともに評価手法を構築する [1]．

英語における小論文の自動採点では実用システムが存在する [5] 一方で，日本語では研究段階であり，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, 7] が提案されている．記述式タイプに対しては近年，評価型ワークショップ NTCIR-13 の QA Lab-3 [1]において，東京大学の世界史の 2 次試験の小論文が課題として取り上げられ，機械学習を利用しないルーブリックに基づくパターンベースの採点手法が提案されている [8]．

このように研究で利用できる小論文と採点データが整備されれば自動採点の研究がより進むことが予測される．そこで，小論文採点手法を構築するにあたって，研究利用可能な模擬試験の小論文データと採点データの構築を引き続き行っている．本論文ではこれらの現状と理解力評価および妥当性評価において先行研究 [1] を上回る結果が得られたので報告する．

2 小論文データの現状

現在構築している小論文データは，受講者がまず講義を受け，講義内容に対する課題を制限時間内に記述するものである．講義は 2016 年度 2 種類行い，各課題について 3 問の設問を設定した (321 人分)．2017 年も異なる 2 種類の講義とそれぞれ 3 設問を設定し受講者に小論文を書いていただいた (180 人分)．小論文は筆記で行い，人手により電子データを作成した．OCR 文字読み取り誤りデータも同時に収集している．

現在，得られた小論文を人手でスコア付けする作業を進めている．採点基準であるルーブリックに基づき 2 人以上で採点を進めている．現段階で 2 つの講義データの各設問 1〜3 について，理解力，論理性，妥当性，文字誤りに関する 2 名の相関係数の平均は 0.55 である．引き続きデータを整理しつつ，見直しを進めているこれらのデータは研究利用について受講者から許諾をいただいており，整備ができ次第，公開する予定である．

[1]http://research.nii.ac.jp/qalab/

次節以降ではこの人手による採点データを利用して理解力評価手法および妥当性評価手法の各モジュールについて評価する．

3 理解力評価モジュール

講義の内容に対して高い理解力を持った受講生によって書かれた回答には，講義の内容とよく似た文章であると考えられる．そのため，回答と講義内容との類似度を測ることによって評価を行う．

3.1 節では形態素類似度評価について述べ，3.2 節で各単語に *idf* 重みをつけた場合について述べる．ここで形態素 N-gram の類似度による評価を使わず，単純な形態素の一致数としているのは，N-gram 類似度よりも形態素の一致数の方が高い相関が出ることが分かっているからである [9]．

3.1 形態素類似度評価

形態素の一致数を得点として出力する評価関数を構築する．文章の形態素区切りを行うために本研究では，形態素解析器 CaboCha[2]を用いた．また辞書にはデフォルトのものでは小論文課題で取り扱われるような専門用語をカバーできていないため，専門用語を単語として評価することができる mecab-ipadic-NEologd[3]の 2017 年 6 月 27 日のものを利用した．その際，各形態素を名詞，動詞，形容詞で，自立語の内容語とそのほかの機能語に分類する．文章の意味をになわず，比較的どのような文章でも頻繁に出現する機能語を無視し，文章の内容を表す内容語のみの一致数で評価を行う．

文書 A, B に出現するそれぞれの内容語を a, b とすると内容語が一致しているかどうかを返す関数 $sim(a, b)$ は式 (1) で表される．それを利用した文書 A, B の内容語の一致数 $score\_match(A, B)$ は式 (2) となる．

$$sim(a,b) = \begin{cases} 1 & (a = b) \\ 0 & (a \neq b) \end{cases} \quad (1)$$

$$score\_match(A,B) = \sum_{a \in A, b \in B} P(a,b) \quad (2)$$

[2]https://taku910.github.io/cabocha/
[3]https://github.com/neologd/mecab-ipadic-neologd/blob/master/README.ja.md

3.2 *idf* 重みを用いた形態素類似度評価

Wikipedia 全文書 (2016/10/1 最新版) から *idf* 重みを計算した．その結果 1386126 単語の *idf* 重みを得ることができた．その重みで先ほどの手法の一致した各単語を重みづけし，その合計を得点とした．回答と講義内容で一致した形態素で *idf* 重みの値のないものに関しては，ノイズとして合計に加算していない．先ほどと同様に各文章の内容語を a, b とすると内容語が一致した場合，その形態素の *idf* 重みを返す関数 $sim_{idf}(a,b)$ は式 (3) で表される．それを利用した文書 A, B の内容語の一致数 $score\_match(A,B)$ は式 (4) となり，それを利用した文書 A, B の内容語の一致数 $score\_match\_idf(A,B)$ は式 (2) とする．

$$sim_{idf}(a,b) = \begin{cases} w_{idf} & (a=b) \\ 0 & (a \neq b) \end{cases} \quad (3)$$

$$score\_match\_idf(A,B) = \sum_{a \in A, b \in B} Q(a,b) \quad (4)$$

4 妥当性評価モジュール

妥当性評価モジュールでは，小論文内で記述されている説明の根拠，事例が正しく述べられているかどうかを確認するために，小論文に記述されている内容が世の中で言われていることとどの程度一致するかを評価する．本研究では Wikipedia を使用し，世の中でも言われているかどうかを比較する．妥当性評価モジュールでは，小論文と Wikipedia の文書を比較して一致度が高ければその文章は妥当だと評価するようにした．しかし，大量の Wikipedia の文書には，課題で取り上げるべき議題とはまったく関係ない文書も多く存在する．そのような文書と小論文との一致度は限りなく低く議題と関係ない文書と比較するだけ無駄であるため，それを避けるために小論文の課題に関係した文書のみを取得する必要がある．例えば，「多国籍企業とグローバリゼーション」に関する課題があった場合，これらに関連した文書を獲得することが望ましい．次節では関連文書の取得方法について記述する．

4.1 関連文書取得法

本研究ではいくつかの関連文書の取得法を開発しているが，本論文では本研究プロジェクトの以前の報告 [10] で最も良い成果を出した手法と新たに開発した手法を記述する．

4.1.1 講義内容などと Wikipedia 本文との単語ベクトル和のコサイン類似度を求める方法

本手法は先行研究 [10] で提案されている手法を利用する．その議題にのみ出現する独特な単語は，出現回数が少なくとも関連文書の取得に大きな役割を果たすと考えられる．そこで *idf* を考慮した講義内容などの本文と Wikipedia 本文との単語ベクトルのコサイン類似度を測って，類似した文書を取り出すという方法を提案する．関連文書を取得するまでの具体的な処理の流れを以下に示す．

まず，ある単語がどの程度珍しい単語であるかを調べる．そのために Wikipedia の各文書を MeCab[4] を用いて形態素解析し，数字を除く名詞，形容詞，動詞に分解する．その後，各単語の *idf* 値を計算することによって単語の珍しさを測る．

次に講義内容と質問文を同じく MeCab で単語に分解し，単語集合を作成しその単語集合の単語ベクトル和を求める．単語ベクトル和を求めるにあたり単語ベクトルには nwjc2vec[5] を使用する．nwjc2vec とは，国語研が 1 億程度のコーパスから word2vec で学習して作成した 300 次元の Skip-gram である．単語ベクトル和は式 5 によって求める．

$$WordVectorSum = \sum_{i=1}^{W} idf(t_i) \times V(t_i) \quad (5)$$

W は取り出した単語集合の総数である．$idf(t)$ はある単語 t の *idf* 値である．$V(t)$ はある単語 t の 300 次元の Skip-gram である．なお，Wikipedia または nwjc2vec に存在しない単語の $idf(t)$ と $V(t)$ は 0 とする．

同様の方法で Wikipedia の各文書の単語ベクトル和を求める．その後に講義内容などの単語ベクトル和と Wikipedia 各文書の単語ベクトル和のコサイン類似度を求める．そして，類似度の高かった上位 1000 件の文書を本節の方法で取得した文書とする．

4.1.2 LSI を用いた講義内容などと Wikipedia 本文とのコサイン類似度を求める方法

本節では新しく追加した LSI を用いた記事検索の手法について記述する．前節の手法では単語-文書間の概念を用いた類似性を測ることによって記事検索をしており，その手法は有用であった．よって，同じような単語-文書間の概念を用いた情報検索の手法として有名な LSI を使用した場合，どの程度の精度で文書を取得できるか確認するためにこの方法を提案する．文書を取得するまでの具体的な処理の流れを以下に示す．

まず，講義内容と質問文を形態素解析し，数字を除く名詞，形容詞，動詞に分解する．各単語の頻度と *idf* 値を掛け合わせたものを各ベクトルの要素として，Bag of Words による文書ベクトルを作成する．次に Wikipedia の 1 文書を形態素解析し，講義内容などから抽出した単語と同じ単語のみを抽出し，同様に文書ベクトルを作成する．全 Wikipedia の文書ベクトル作成した後，行を単語，列を文書とした Wiki 行列を作成する．出来た Wiki 行列を式 6 のように特異値分解する．

$$WikiMatrix = U \, \Sigma \, V^T \quad (6)$$

$m \times n$ 行列の $WikiMatrix$ に対して，U は $m \times m$ のユニタリー行列であり単語を表現する．V は $n \times n$ のユニタリー行列であり文脈を表現する．Σ は $m \times n$ の非対角成分は 0，対角成分は非負で大きさの順に並んだ行列である．この後，左特異値ベクトルである U を使用したいが，サイズが大きいため近似した行列を使用する．そのために U から k 列目以降の行列を削除した $m \times k$ 列の U' を作成するという次元圧縮を行う．この U' と文書ベクトルを掛け合わせることで単語文書行列が作成されるので，Wikipedia と講義内容などの文書間の類似度を測ることができる．文書間類似度は式 7 によって求まる．

$$DocSim = CosSim(U'^T d_i, U'^T q) \quad (7)$$

$CosSim$ はコサイン類似度を示す．d_i は $WikiMatrix$ の i 列目のベクトルであり 1 文書を示す．q は講義内容と質問文

[4] http://taku910.github.io/mecab/

[5] http://pj.ninjal.ac.jp/corpus_center/nwjc/subscription.html

の文書ベクトルである．全文書の文書間類似度を測った後，類似度の高かった上位 1000 件の文書を本節の方法で取得した文書とする．

5 評価実験

講義 1，2 の 2 つの小論文課題に各小問 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，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．今回は採点が完了している 161 人分のデータで実験を行った．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。

課題 3： 文化のグローバリゼーションは、私たちの生活にどう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。

講義 2 の内容： 自然科学の構成と科学教育

課題 1： 「科学的」とはどのような条件をみたす必要があるのか 100 字以内で答えよ。

課題 2： 講義で解説した自然科学の二つの側面を参考に、自然科学が果たす役割について 400 字以内で論ぜよ。

課題 3： 「Scientific and Technological Literacy for All」の狙いを考慮し、これからの科学教育はどうあるべきか 500 字以上 800 字以内で論ぜよ

5.1 理解力評価

まず表 1 に理解力モジュールについて評価した結果を示す．

表 1: 内容語の一致数と理解力の点数との相関

設問	内容語の一致数	*idf* 重みづけ有	有効件数
1-問 1	0.372	**0.383**	154
1-問 2	0.427	0.345	152
1-問 3	0.627	0.550	153
2-問 1	0.682	**0.719**	154
2-問 2	0.507	**0.543**	152
2-問 3	0.617	**0.636**	152

ここで有効件数が設問によって異なるのは無回答の回答を取り除いて実験を行ったからである．
単純な内容語の一致数に比べ，内容語に *idf* による重みづけを行った場合の方が多くの場合で相関が向上した．これは通常の形態素の一致数では，設問特有の単語も「する」のような比較的どの文章でも使われているような単語を同様に 1 形態素として計算していた．一方この手法では単語の重要度を設定することができるため，相関が上がったと考えられる．また使用した辞書に関しては mecab-ipadic-NEologd を用いることで「する」のような比較的どのような文章にも表れるような単語は *idf* の値が低く，今回の講義 1 における重要単語であると考えられる「ジニ係数」のような単語の *idf* の値が高くなっていたことが見て取れた．

5.2 取得した文書のタイトル一覧

本節では，4.1 節の各関連文書取得法で取得した文書を評価する．評価方法は，取得した文書の上位のタイトルを見て人手で判断する．

表 2: 関連文書取得法で取得した文書の上位 10 件のタイトル

	単語ベクトル和 (講義 1)	LSI(講義 1)	単語ベクトル和 (講義 2)	LSI(講義 2)
タイトル	開発経済学	ムハンマド・ビン・ラーシド・アール・マクトゥーム	問題解決	自然の斉一性
	経済的不平等	インド	操作主義	科学におけるロマン主義
	グローバル資本主義	メキシコ	デザイン思考	自然観
	自由貿易	ブラジル	再現性	ネイチャーライティング
	空洞化	ポーランド	二重相続理論	自然写真
	貧困	超大国	情報	ガイストクラッシャー
	自由貿易協定	スペイン	第二言語習得の理論	鳳来寺山自然科学博物館
	進歩的活用理論	フランス	科学的方法	兵庫県立六甲山自然保護センター
	東アジア共同体	世界都市	ヒューマンファクター	野村圭佑_(ナチュラリスト)
	マスツーリズム	アメリカ合衆国	一般システム理論	長野県の観光地

各関連文書取得法で取得した文書のタイトルの上位 10 件を表 2 に示す．取得してきた文書の良し悪しに関してだが，講義 1 はグローバリゼーションや経済格差について記述している文書を取得することが望ましい．表 2 を見てみると，単語ベクトル和の手法は講義に即した文章が取れていると言える．対して LSI の方は世界の国に関する記事ばかりで経済格差に関した記事は取れていなかった．原因として，今回の LSI ではクエリである講義内容の文書に出てくる「グローバリゼーション，世界，文化，企業」の 4 つの単語がかなりの高頻度で出てきており，この単語たちのみが重要視され，ほかの単語の影響をほとんど受けなかったため経済に関する記事が取れなかったと思われる．対策としては単語の頻度による重みを軽減し幅広い単語から，単語-文書間の概念を構築できるようにすることがあげられる．

次に講義 2 に関してだがこちらは，自然科学や科学教育について記述している文書を取得することが望ましい．表 2 を見てみると，講義 2 は講義 1 よりも限定的な内容であり記事も少ないためかどちらの手法でも講義に即した記事を得ることができなかった．単語ベクトル和が記事を取得できなかった原因として，講義 2 の方がより抽象的な話であったこと，重要な単語が英単語のため形態素解析ができなかったため単語の共起関係がうまく取れなかったことが挙

げられる．LSI が記事を取得できていない原因は講義 1 の時と同じであった．

5.3 妥当性評価

本節では妥当性の評価に関して記述する．*idf* による単語マッチで採点を行う．取得した文書は M 文，小論文は N 文から構成されているので，総当たりで「名詞」,「形容詞」,「動詞」の単語マッチを行いマッチした単語の数を数えたものに *idf* の重みを加えたものを式 8 で定義するスコア S として出力する．

$$S = \sum_{m \in M} \sum_{n \in N} IDFWordMatch(\text{Wikipedia}_m, Essay_n) \qquad (8)$$

求めたスコア S と人手で採点したスコアとの相関を取ることによって取得した文書を評価する．その結果を表 3 に示す．

表 3: 各方法で取得した文書を用いて採点したスコアと人手のスコアとの相関係数

設問	単語ベクトル和	LSI
1-問 1	0.0205	**0.0299**
1-問 2	**0.455**	0.391
1-問 3	0.301	**0.344**
2-問 1	**0.0758**	0.00438
2-問 2	**0.412**	0.261
2-問 3	**0.360**	0.233

各小論文の有効件数は 5.1 節で示したものと同じである．講義 1 に関して単語ベクトル和の方が良質な記事が取れていたにも関わらず単語マッチによる採点では LSI の手法の方が相関係数が高かった．おそらく LSI は講義でかなり高い頻度で出てきている数種類の単語が多く出ている記事を取得しており，それとのみ大量に単語マッチしたため相関が上がったと考えられる．単語マッチでも単語頻度の重みを軽減する必要がある．講義 2 に関しては，全ての設問において単語ベクトル和の方が相関係数が良かった．これは LSI で取得してきた記事が文字数の少ない記事ばかりであったため，あまり単語マッチがなされず相関が低い結果となったと考えられる．

LSI で取得してきた記事でも問 1 よりも問 2 と問 3 の方が相関係数が高いため，この手法もまたエッセイタイプの問題の方が有効であることがわかる．

6 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．

7 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] 竹内孔一, 大野雅幸, 泉仁宏太, 田口雅弘, 稲田佳彦, 飯塚誠也, 阿保達彦, 上田均. 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. 言語処理学会第 23 回年次大会発表論文集, pp. 839–842, 2017.

[2] E.V. Steendam, M. Tillema, G. Rijlaarsdam, and H. van den Bergh. Measuring Writing: Recent Insights into Theory, Methodology and Practices. Brill Academic Pub, 2012.

[3] 石川巧. 「いい文章」ってなんだ？—入試作文・小論文の歴史. 筑摩書房, 2010.

[4] 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.

[5] 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.

[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] 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.

[8] Tshuneori Ishioka, Kohei Yamaguchi, and Thuneori Mine. Rubric-based Automated Japanese Short-answer Scoring and Support System Applied to QALab-3. In *Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies*, pp. 152–158, 2017.

[9] Masayuki Ohno, Koichi Takeuchi, Kota Motojin, Masahiro Taguchi, Yoshihiko Inada, Masaya Iizuka, Tatsuhiko Abo, and Hitoshi Ueda. *Construction of Open Basic Data for Automatic Scoring of Essay and Evaluation of Automatic Scoring Method at Current Stage*. PACLING-2017, 2017.

[10] 泉仁宏太, 竹内孔一, 大野雅幸, 田口雅弘, 稲田佳彦, 飯塚誠也, 阿保達彦, 上田均. 小論文採点支援のための関連文書取得法の考察. 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. 47–51, 2017.