Vol.2019-CLE-27 No.21
2019/3/22

オンライン短答式記述問題の解答に対する潜在的意味解析を用いた自動フィードバック手法の検討

生田 寛[1,a)]　中野 裕司[2,b)]　杉谷 賢一[2,c)]　久保田 真一郎[2,d)]

概要：LMS のオンライン記述式問題は解答が多様なため自動的に評価を行い，フィードバックを返すことが困難である．記述式問題には大きく分けてエッセイ式と短答式の問題形式があり，自動採点の先行研究には，解答の候補を有しないエッセイ式の問題（1000 字程度の小論文）を対象にしたものや，解答候補を有する短答式の問題（1 文または多くとも 2 文）を対象にしたものがある．しかし，解答候補を有しない短答式記述問題に対しての自動採点は解答文の自由度が上がり困難であると考えられる．本研究では，特定のテーマ設定がなされた解答候補を有しない短答式記述問題に対して，潜在的意味解析により評価対象の解答文を自動評価し，その結果をもとに LMS でフィードバックを返す仕組みを提案する．

キーワード：自動フィードバック，潜在的意味解析，LMS

A Study on Automatic Feedback Method Using Latent Semantic Analysis for Short Sentences in Response to Online Short-answer Questions

Kan Ikuta[1,a)]　Hiroshi Nakano[2,b)]　Kenichi Sugitani[2,c)]　Shin-ichiro Kubota[2,d)]

Abstract: In short answer question to impliment online on LMS, it is known for difficulties to evaluate and give feedbacks for students automatically because of various answers.Descriptive questions can be roughly divided into question formats of essay style and short-answer style, and prior studies of automatic grading are directed to essay style questions (no less than 1000 letters of essay) without answer candidates And short-answer questions having answer candidates (one sentence or at most two sentences). However, automatic grading for short sentences in response to short-answer questions without answer candidates is considered to be difficult because the degree of freedom of the answer sentence is increased. In this research, we grade short sentence answers automatically by latent semantic analysis for short-answer questions, which does not have answer candidates with specific theme settings, and based on the results, We propose a mechanism to give feedbacks for students on LMS.

Keywords: automatic feedback, latent semantic analysis, LMS

1. はじめに

LMS（Learning Management System）を用いたオンライン学習環境では，学習者の管理や教材の配信，ライブ授業，成績管理など多岐に渡って，教師と学生の学習を支えるシステムを使用することができ，代表的なものとして Moodle[1] や canvas[2] などがある．その中でもオンライン

[1] 熊本大学大学院自然科学教育部
Kumamoto University Graduate School of Science and Technology
[2] 熊本大学総合情報統括センター
Kumamoto University Center for Management of Information Technologies
[a)] kan@st.cs.kumamoto-u.ac.jp
[b)] nakano@cc.kumamoto-u.ac.jp
[c)] sugitani@cc.kumamoto-u.ac.jp
[d)] kubota@cc.kumamoto-u.ac.jp

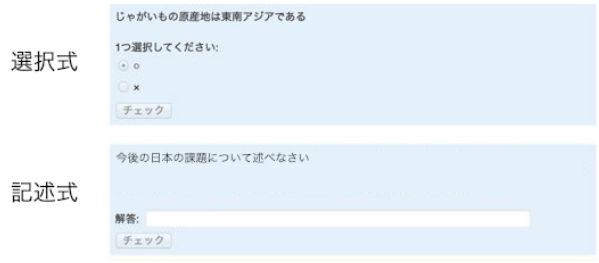

図 1 選択式と記述式の例

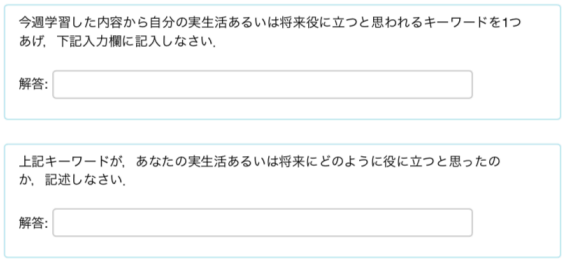

図 2 小レポートの問題文

テストは問題を作成し，テストすることで解答を一様に管理することができ，選択式問題や記述式問題などの問題形式がある．図 1 は選択式と記述式の例を示している．

選択式問題は正解が定まっており自動採点ができるため，解答者は即時にフィードバックを受けることができる．記述式問題では解答が多様で，自動採点や即時フィードバックが困難である．そのため，評価者は学生全員の答案に対して手動で採点を行わなければならず，多大な負担となる．この負担を軽減するために記述式問題の自動採点の研究が行われている．自動採点の研究は石岡 · 亀田らの「コンピュータによる小論文の自動採点システム Jess の試作」[3] や中島の「機械学習を利用した短答式記述答案の自動識別」[4]，石岡らの「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」[5] などが挙げられる．

記述式問題には大きく分けてエッセイ式と短答式の問題形式があり，エッセイ式の問題は 1000 字程度の小論文の問題を扱い，短答式の問題は解答文に 1 文または多くとも 2 文の問題を扱う．本研究が自動判定の対象とする問題形式は短答式記述問題であり，問題形式の例として，2016 年，2017 年の 10 月から 12 月にかけて開講された授業を挙げる．この授業は大学の学部 2 年生約 800 人を対象として，IT 系国家資格のひとつ「IT パスポート試験」のテクノロジ系分野と IT 関連法務について知識習得を目標にしている授業であり，全 7 回のオンライン授業と期末試験から構成されている [6]．授業を受講後，毎回，出席確認を兼ねた小レポートが LMS の Moodle 上で実施されている．図 2 に小レポートで使用する問題文を示す．

問題文は最初に学習したキーワードを答える問題が設けられ，そのキーワードに従って短答式記述問題が設けられている．このような問題文では，得られる解答文は多様であり，評価者にとって多大な負担となる．また，解答候補を有しないため，問題に対して正解を設定し，自動採点やフィードバックを行うことが困難である．

日本語エッセイ式記述問題の自動採点の研究として，石岡 · 亀田らは，エッセイ式記述問題の 800∼1600 字程度の解答文に対して，修辞，論理構成，内容の 3 つの観点で自動採点を行うシステム Jess を開発している [3]．特に内容の採点では問題の質問文，新聞の社説やコラムを学習データとして潜在的意味解析（LSA：Latent Semantic Analysis）で要約した単語文書行列を作成し，解答文とのコサイン類似度から自動採点を行うが，学習データにない内容はうまく自動採点できないことが分かっている．本研究が対象とする問題形式であっても解答が多様であるため，学習データにない内容はうまく自動判定できない．

短答式の研究として，中島は，文字数制限があり解答に含めるべき文が定まる短答式記述問題において，複数の採点済みの解答を学習データとし，機械学習手法を用いて自動評価のための自動識別を試みている [4]．その結果，相応の識別力を確認したが，判定に失敗する解答文があることを確認している．短答式記述問題に関する自動採点の研究では，解答として含まれるべきキーワードや表現が存在し，解答例がある問題設定を前提としている．短答式で自由記述のような解答候補を有しない短答式記述問題は解答文の自由度が上がり，本研究が対象とする問題形式では自動採点が困難である．

本研究は，解答候補を有しない短答式記述問題であっても特定のテーマに絞ることで自動評価が有効であると考え，特定のテーマ設定がなされた解答候補を有しない短答式記述問題に対して，潜在的意味解析により評価対象の解答文を自動評価し，その結果をもとに LMS でフィードバックを返す仕組みを提案する．

2. 提案手法

本研究では特定のテーマ設定がなされた解答候補を有しない短答式記述問題を対象とし，問題に対する解答文に合わせて学習者にフィードバックを与えるシステムを提案する．図 3 に提案するシステムの全体像を示す．

システムの流れは，はじめに評価者が問題の作成と過去の学習者の解答文を評価者が決められた評価基準で判定し，その結果から単語文書行列の作成と類似度の閾値を決定する．学習者に問題の解答をしてもらい，提出した解答文を判定モジュールに入力し自動判別を行い，出力した判別結果をフィードバックモジュールに入力する．フィードバックモジュールでは学習者に星マークで類似度を通知し，悪い判定の学習者には強制的ではない解答の再提出を促す．

図 3 の判定モジュールでは，解答候補を有しない短答式記述問題に対して，学習者の解答文の良し悪しをを判定す

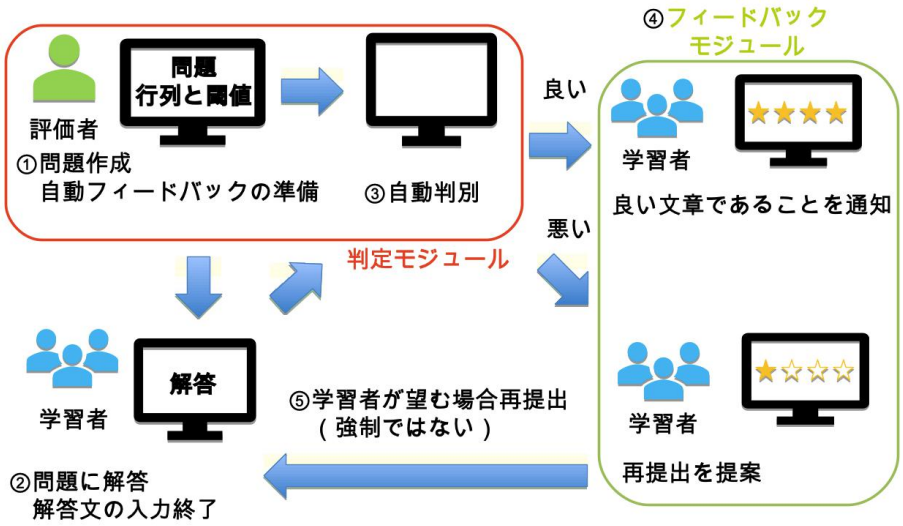

図 3　システムの全体像

る．その判定には過去の学習者が解答した解答文を利用する．過去の解答文に対して，評価者が決められた評価基準で，正解とする文（良い文）とフィードバックの必要がある文（悪い文）を判定する．その良い文と判定された文との意味的な類似度をもとに，良い文であるか，悪い文であるか，学習者の解答文を判定する．判定モジュールは，学習プロセスと動作プロセスで構成する（図 4）．学習プロセスでは，良い文と悪い文を判定するための類似度の閾値を定める．次に学習者が提出した解答文を動作プロセスに入力し，過去の解答文との類似度を算出し，学習プロセスで定めた閾値から解答文を良い文と悪い文に自動判別する．

2.1　学習者が提出する解答文が妥当であるか判定するモジュール（判定モジュール）

本モジュールでは，評価者が決められた評価基準をもとに過去の解答文を評価し，評価の結果，妥当であると判定された文と，学習者が提出する解答文とが，意味的に似ているかを類似度で算出し，閾値により提出された解答文が妥当か否かを判定する．判定は類似度の大きさから 4 つのグループに分類する．

本モジュールは過去の解答文とその判定結果をもとに，LSA による単語文書行列を構成する処理と，閾値を機械的に決定する処理（学習プロセス），LSA による単語文書行列によって判定する処理（動作プロセス）の 3 つの処理で構成される．

2.1.1　単語文書行列の構成

文と文の類似度は，文を単語に分割し，文を単語の出現頻度で表した単語文書ベクトルをもとに 2 つの文の単語文書ベクトルのコサイン類似度で表される．本研究では，良い文と判定された複数の文により，1 つの解答文の単語文書ベクトルを行とする単語文書行列を構成し，LSA により意味が要約された行列を構成する [7]．LSA 要約行列もまた，その行数は構成する文章の数になるため，提出された解答文とのコサイン類似度は，各行の文章ごとに算出され，行の数だけ類似度が存在する．この複数の類似度のうち，中央値を代表値として閾値により判定する．

単語文書行列の作成のために，文を形態素解析エンジン Mecab[8] を使用して分割し，単語群を作成する．単語文書行列の作成時に使用する単語群は，文章に含まれる全ての単語から自然言語処理において悪影響を及ぼすとされるストップワード [9] を除いた単語を使用し，全ての単語を原型で使用する．石岡 · 亀田が開発した Jess では，評価対象の解答文を単語文書行列に含めて LSA により意味が要約された行列が構成される．石岡らのように判定の対象となる解答文を含めて LSA により単語文書行列を構成すると，悪い文が高い類似度を持ち，良い文との区別が困難になるため，本研究では解答文を単語文書行列に含めずに，LSA を行い比較する．

2.1.2　本研究における潜在的意味解析

潜在的意味解析は単語文書行列の作成，特異値分解，次元削減の大きく 3 つの手順に分けて行う．本研究における次元削減について説明する．

潜在的意味解析の過程で特異値分解を行うことで対角要素が特異値となる特異値行列を得ることができる．この特異値の削減を行うことを次元削減と呼び，特異値を任意の k 番目の値まで選択することで削減を行う．私たちは行列の大きさを表すフロベニウスノルムにより，次元を削減した場合の行列の大きさが変化するが割合を考察し，削減する次元を決定する．特異値の値が小さい方から順に削減していき，ランク $k+1$ の行列のフロベニウスノルム N_{k+1} とランク k の行列のフロベニウスノルム N_k との差により，削減された次元の影響を考察できる．削減された次元による影響が大きいとその差は大きく，削減された次元による影響が小さいとその差は小さくなる．今回は，特異値の値

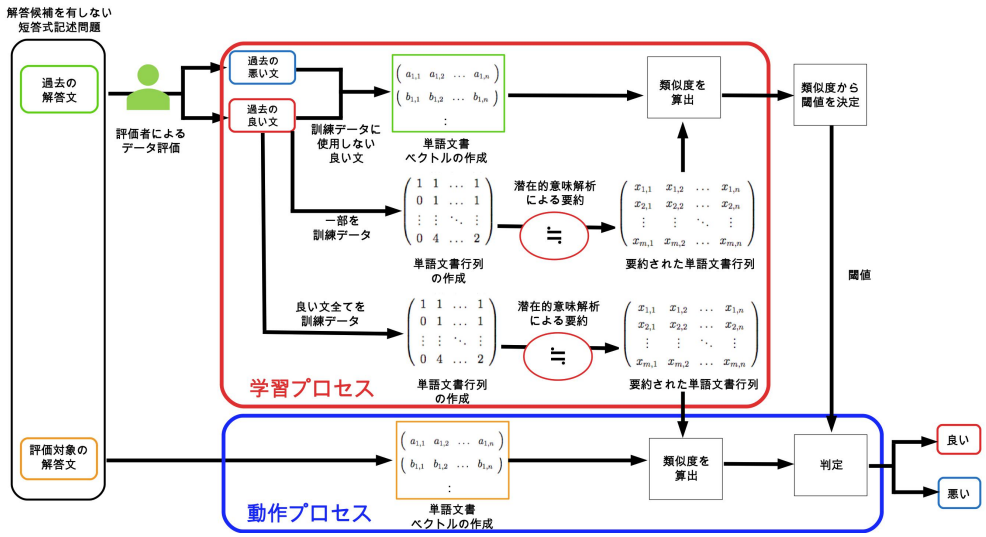

図 4　判定モジュールの学習プロセスと動作プロセス

が小さい方から順に次元を削減した場合のフロベニウスノルムの差 $N_{k+1} - N_k$ をプロットし，その変化が大きく変化した場合のランク k までを有効な次元とし，ランク $k+1$ 以降の特異値を 0 として次元削減を行った

2.2　学習プロセス

学習プロセスは評価者が過去の解答文を一定の評価基準に沿って評価し判定した良い文と悪い文が入力される．以下に学習プロセスの流れを記述する．

(1)　評価者が評価し判定した，良い文と悪い文が入力される．

(2)　評価者が良い文と判定した文書の数を 5 分割するように無作為に組分けする．

(3)　組分けした良い文のうち，2 組を訓練データとして単語文書行列を作成し，潜在的意味解析により要約された単語文書行列 X を作成する．

(4)　訓練データに使用しない評価者が評価した良い文のうち，残り 3 組の良い文とそれと同数の悪い文をテストデータとして単語文書ベクトル V_1~V_n を作成する．

(5)　要約された単語文書行列 X の各行と解答文の単語文書ベクトル V_1~V_n とでそれぞれコサイン類似度を求める．

(6)　ベクトルごとに算出された複数の類似度の中央値を計算して，良い文と悪い文を判定するための閾値を 0.05 刻みで変更したそれぞれの場合で評価者による評価とシステムによる評価における F 値を算出する．

(7)　(3) ~ (6) を良い文 5 組から 2 組を選出する組み合わせ 10 パターン全てで繰り返して行う．

(8)　0.05 刻みで変更した各閾値における F 値の 10 パターンの平均を計算し，動作プロセスでの判定で使用する F 値がもっとも大きくなる閾値を決定する．

(9)　動作プロセスで使用するため，評価者が判定した良い文全てを訓練データとし，潜在的意味解析を行い要約された単語文書行列 X' を作成する．

このようにして良い文と悪い文を判定するための閾値と良い文全てを訓練データとして潜在的意味解析で要約された単語文書行列 X' が動作プロセスに出力される．

2.3　動作プロセス

動作プロセスは現在の学習者から得られる評価対象の解答文と学習プロセスで決定した閾値と良い文全てを訓練データとして潜在的意味解析で要約された単語文書行列 X' が入力される．以下に動作プロセスの流れを記述する．

(1)　評価対象の解答文から単語文書ベクトル V_1'~V_n' を作成する．

(2)　学習プロセスで作成した単語文書行列 X' の各行と評価対象となる解答文の単語文書ベクトル V_1'~V_n' とでコサイン類似度を求める．

(3)　ベクトルごとに複数算出された類似度の中央値を計算し，学習プロセスで設定した閾値に従って，類似度の中央値が閾値より大きい場合良い文，小さい場合悪い文に判定する．

このようにして評価対象となる解答文の判定結果が出力される．

提案手法での各プロセス内で作用するモジュールと入出力をまとめた図を図 5 に示す．

これらのプロセスにより，評価対象となる解答文の判定を行う．

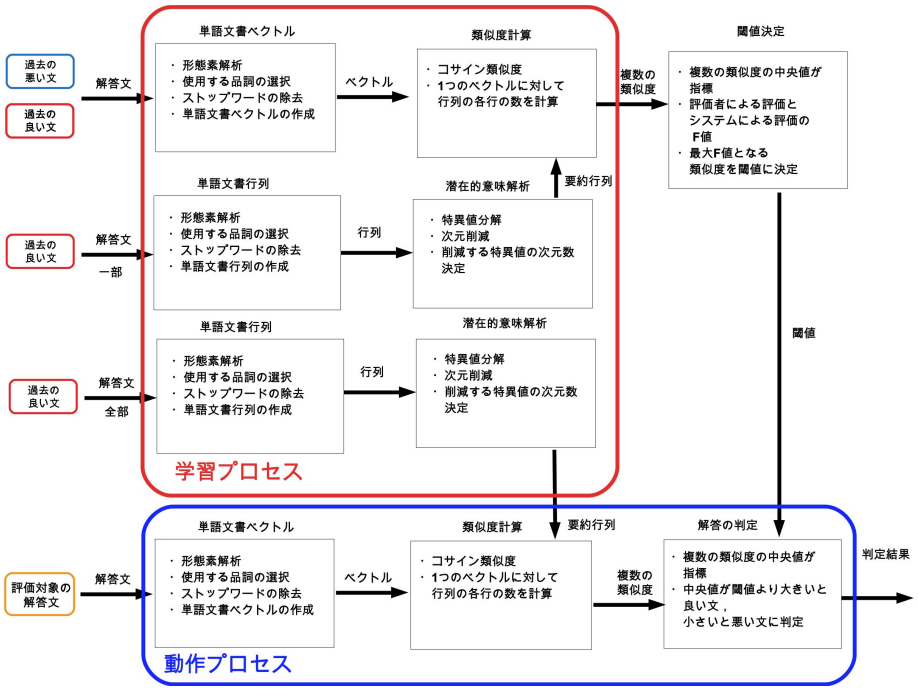

図 5　システムのモジュール図

表 1　評価者が評価した結果

	2016		2017	
	良い文	悪い文	良い文	悪い文
キーワード W_1	23	30	25	31
キーワード W_2	33	66	15	74
キーワード W_3	32	49	24	35

2.4　判定結果によりフィードバックを与えるモジュール（フィードバックモジュール）

本モジュールでは，判定モジュールにより，学習者が提出した解答文が妥当ではないと判定された場合に学習者に再提出を促すようなフィードバックを与える．判定モジュールは，類似度により，4 つのグループに分類されるので，類似度がもっとも小さいグループを星マーク 1 つで表し，次に小さいグループを星マーク 2 つ，中程度の類似度を星マーク 3 つもっとも大きなグループを星マーク 4 つで表現する．

フィードバックには提出した解答文とその類似度に合わせた星マークを表示して，星マークが 2 つ以下の場合には再提出を促す文章を表示する．

3.　判定モジュールの実験

特定のテーマ設定がなされた解答候補を有しない短答式記述問題の解答文として，第 1 節で説明した授業の小レポート問題の解答文を対象とする．2016 年と 2017 年で共通した 3 つのキーワードに対する解答文を評価者が評価した結果を表 1 に示す．

これらのデータを提案手法に当てはめ，2016 年の解答文を過去の解答文，2017 年の解答文を評価対象の解答文として学習プロセスの検証実験および動作プロセスの評価実験を行った．2016 年の解答文を学習プロセスにかけて，閾値と行列を算出し，2017 年の解答文を動作プロセスで算出される複数の類似度によって判定する．

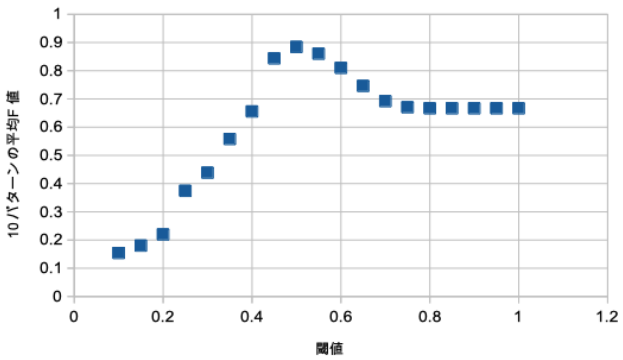

図 6　各閾値ごとの平均 F 値 (キーワード W_1)

3.1　学習プロセス

2016 年の解答文をもとに検証実験を行った．提案手法の学習プロセスによって得られた各閾値ごとの 10 パターンの平均 F 値をキーワード W_1 の場合のみ図 6 に示す．

図 6 より平均 F 値が最大の 0.88 となる 0.5 を閾値と決定する．他のキーワードでも同様に閾値決定した結果を表 2 にまとめる．

次に評価実験に使用するため，2016 年の良い文全てを用いて潜在的意味解析で要約した単語文書行列を作成しておく．

表 2　キーワードごとに決定した閾値と最大の平均 F 値

	閾値	F 値
キーワード W_1	0.5	0.88
キーワード W_2	0.55	0.68
キーワード W_3	0.6	0.85

表 3　判定結果：解答文の本数の内訳 (キーワード W_1)

判定		評価者 悪い	評価者 良い
類似度	悪い	13	1
	良い	18	24

表 4　判定結果：解答文の本数の内訳 (キーワード W_2)

判定		評価者 悪い	評価者 良い
類似度	悪い	52	0
	良い	22	15

表 5　判定結果：解答文の本数の内訳 (キーワード W_3)

判定		評価者 悪い	評価者 良い
類似度	悪い	26	8
	良い	9	16

表 6　判定した F 値と正確度

	閾値	F 値	正確度
キーワード W_1	0.5	0.58	0.70
キーワード W_2	0.55	0.83	0.75
キーワード W_3	0.6	0.75	0.71

3.2　動作プロセス

2017 年の解答文，検証実験で決定した閾値，作成した単語文書行列をもとに評価実験を行った．提案手法の動作プロセスによって判定した結果，得られた解答文の本数の内訳をキーワードごとに表 3～ 表 5 に示す．また，判定結果から得られた F 値と正確度を表 6 に示す．

表 3～ 表 6 より，学習プロセスで決定した閾値による判定でキーワード W_1 は悪い文の判定がうまくいかずに，F 値が 0.58 と低い値を示し，他のキーワードでは相応の F 値を示している．正確度はどのキーワードでも 0.70 以上の値を示した．

4.　考察

実験結果より提案手法の学習プロセスにより決定した閾値で動作プロセスの判定を行った結果，正確度がどのキーワードでも 0.70 以上であるが高い値とは言えない．しかし，7 割の学習者には正しい判断でフィードバックを与え，適切な学びを促すことができると考えられる．また，動作プロセスでも各閾値ごとの F 値と正確度を確認したところ F 値が最大となるときの閾値と正確度は表 7 のような結果が得られた．

表 7　F 値が最大となるときの閾値と正確度

	閾値	最大 F 値	最大正確度
キーワード W_1	0.7	0.81	0.75
キーワード W_2	0.7	0.91	0.84
キーワード W_3	0.65	0.83	0.78

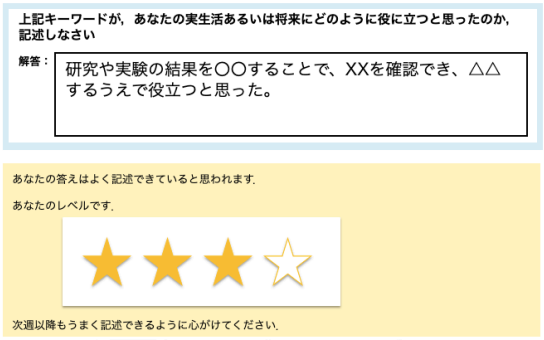

図 7　フィードバック例 (良い場合)

表 6 と表 7 を比較するとどのキーワードでも閾値を上げた方が良い判定を示しており，現在の閾値決定手法では最適な閾値を設定できていないことが確認できる．今後，閾値決定手法を改善することで更なる判定精度の向上を目指す．

4.1　LMS へのフィードバックモジュール実装

現時点の判定精度での LMS のオンライン短答式記述問題へのフィードバックモジュール実装を考える．フィードバックモジュールでは判定モジュールで提出された解答に対する評価を行った結果，良い文である場合は評価とその通知をし，悪い文である場合は評価と再提出を促すような通知をする．再提出は強制ではなく，あくまでも解答者が望む場合に再提出を行えるようにし，解答者の自主性を尊重することを想定している．学習意欲の高い学習者に対して有効なフィードバックである．次に実際に提出された解答へのフィードバック例を良い場合と悪い場合でそれぞれ図 7，図 8 に示す．

図 7，図 8 では解答文に対して黄色の領域でフィードバックを与えている．どちらの図にも表示されている星は解答文の類似度の大小による 4 つのグループ分けを示している．閾値で良い文と悪い文を完全に二分することは困難であるため，システムによる推定評価を学習者に視覚的に与え，学習者は自分の解答文が高い評価を得られるかを今一度検討する機会を得ることができる．また，図 7 では良い文であることの通知をしているが，図 8 では「キーワードがどういうものだから，どのようなことができるように

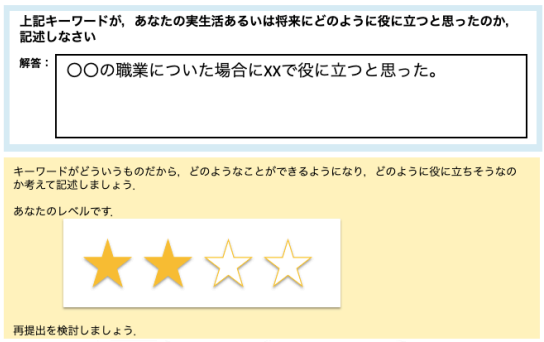

図 8　フィードバック例 (悪い場合)

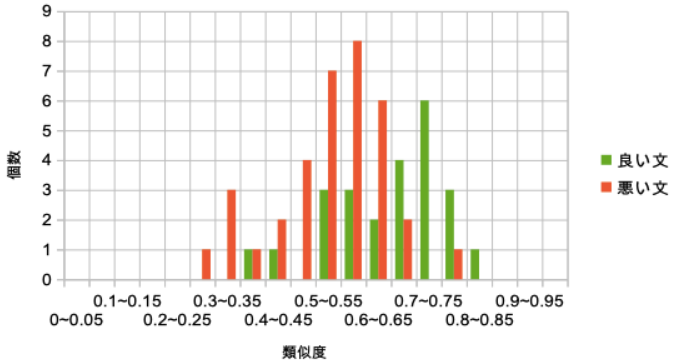

図 9　閾値の刻み幅 0.05 の類似度の度数分布

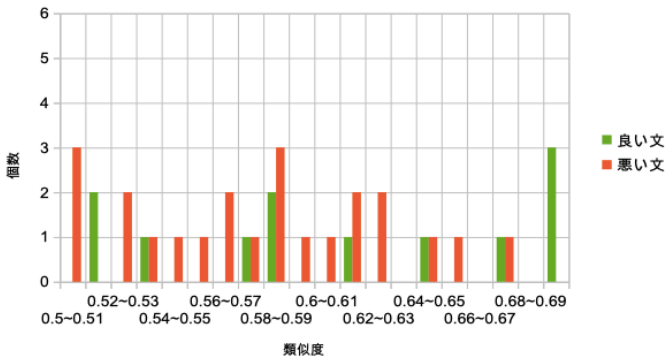

図 10　閾値の刻み幅 0.01 の類似度の度数分布

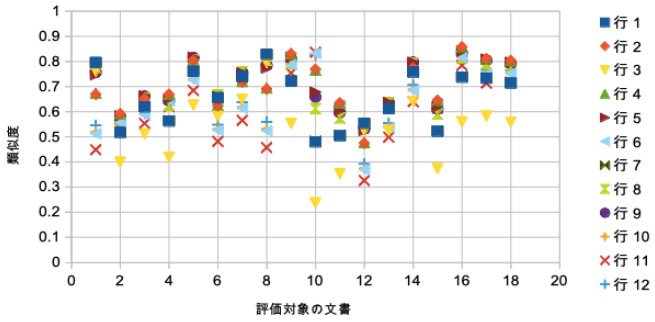

図 11　行列の各行に対する評価対象の文書の類似度の分布

なり，どのように役に立ちそうなのか考えて記述しましょう．」といった再提出を促すようなフィードバックを与えている．

5. まとめ

本研究では，特定のテーマ設定がなされた解答候補を有しない短答式記述問題に対して，判定モジュールの実験では，LSA により要約した単語文書行列と閾値決定手法により，評価対象の解答文を良い文か悪い文かに自動識別し，3 つのどのキーワードに対する解答文であっても正確度が 0.70 以上の値を示したが，決定した閾値が最適な閾値でないことが確認できた．この結果をもとにフィードバックモジュールで学習者に星マークでシステムによる推定評価を掲示し，星 2 つ以下の場合に再提出を促すフィードバックを与える仕組みを提案した．

閾値決定方法の改善として，類似度の扱い方，閾値の刻み幅の変更，類似度の分布傾向調査が挙げられる．提案手法では類似度のパラメータとして中央値を用いている．中央値は複数の解答文が良い文の多くと類似する場合の判定に有効であると考えていたが，他のパラメータでの検証も行うことで更なる精度向上に繋がると考えられる．また，閾値の刻み幅は 0.05 と設定している．閾値の刻み幅 0.05 の類似度の度数分布を図 9 に示す．

図 9 から多くの解答文が類似度 0.5~0.7 の範囲において多く分布していることが分かった．この 0.5~0.7 の範囲において閾値の刻み幅を 0.01 に変更した場合の類似度の分布を図 10 に示す．

図 10 から分かるように閾値を細かく設定することで，精度の向上は可能であるが，良い文と悪い文を完全に二分することは困難であることを確認した．

次に良い文で作成した単語文書行列の各行に対する評価対象の文書の複数の類似度の分布を図 11 に示す．

図 11 では行 1~ 行 12 は良い文で作成した単語文書行列の各行を示している．類似度の分布傾向を見ると，全体的に高い類似度を示すデータや高い類似度と低い類似度が入り混じっているデータを確認できた．これらのベクトルデータを用いて LinearSVC や k-means 法などの機械学習手法で傾向をパターン化し，閾値決定手法に取り入れることで判定精度の向上に繋がると考えられる．

参考文献

[1] Moodle, https://moodle.org/, (確認日:2019 年 2 月 23 日)
[2] canvas, https://www.canvaslms.com/, (確認日:2019 年 2 月 23 日)
[3] 石岡 恒憲, 亀田 雅之, コンピュータによる小論文の自動採点システム Jess の試作, 計算機統計学 第 16 巻 第 1 号, pp3 − 19, 2003
[4] 中島 功滋, 機械学習を利用した短答式記述答案の自動識別, 日本教育工学会 第 26 回全国大会, pp639-640, 2010 年
[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, 信学技報 NLC, 87- 92, 2016

[6] 情報処理概論ガイダンス https://mahara.kumamoto-u.ac.jp/view/view.php?t=VM652svlH13bkjpcxfwS,（確認日:2019 年 2 月 23 日）
[7] DEERWESTER, S., DUMAIS, S. T., FURNAS, G. W., LANDAUER, T. K. & HARSHMAN, R. (1990), Indexing by latent semantic analysis, Journal of the American Society for Information Science, 41 (6), pp. 391407.
[8] MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://taku910.github.io/mecab/ （確認日:2019 年 2 月 23 日）
[9] SlothLib ストップワードリスト - OSDN.net http://svn.sourceforge.jp/svnroot/slothlib/CSharp/Version1/SlothLib/NLP/Filter/StopWord/word/Japanese.txt,（確認日:2019 年 2 月 23 日）