The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

社会人向け教育における記述式問題回答自動分類

Automatic Classification of Description Problem Answers in Business Education

佐々木 健太[*1] Kenta Sasaki
鈴木 健一[*1] Kenichi Suzuki
乾 健太郎[*2] Kentaro Inui

[*1] グロービス経営大学院 Graduate School of Management, Globis University
[*2] 東北大学 Tohoku University

Recently, automatic text scoring research using machine learning is progressing in an education field. Most of the short description problems have specific model answers. However, about business education, most of them don't have specific model answers. Therefore, we investigated the possibility of classifying short description problem answers in a business education field, to develop an adaptive learning system. As a result, we confirmed that we could classify them with high accuracy if the number of answer characters is about 50 and answers can be clearly classified into less than three patterns. Then we could detect the words that contribute to the classification. Moreover, we found that the way of creating problems is one of the important factors to classify answers.

1. はじめに

近年，教育分野において，記述式問題回答の自動採点に関する研究が進んでいる [Sultan 16, Taghipour 16, 水本 18]. 実際，2020 年度から始まる新共通テスト（既存の大学入試センター試験の代替試験）において，採点支援システム，または，自動採点システムの導入が検討されている [石岡 16, 亀田 17]. 具体的には，それぞれの問題に対して複数の採点基準を用意しておき，回答がそれらの採点基準をどの程度満たしているのかによって採点するというものである．一方，このような手法を社会人向け教育にそのまま適用することは難しい．なぜなら，社会人向け教育では正解が一意に決まらない問題が多く，採点基準を明文化することが難しいからである．例えば，「リーダーに必要なスキルを述べよ．」という問題に対して，「人を動かす力」「ビジョンを描く力」などが模範回答として考えられるが，例えば「鈍感力」「体力」などの回答も間違いとは言い切れない．そこで，本研究では，社会人向け教育における記述式問題回答の自動分類の可能性を調査した．自動分類することができれば，回答のパターンによってフィードバックを換えるなど，学習者の回答に応じた適切な振り返りサービスを提供することができる．そして，例えば，学習者が回答した以外の回答パターンを知ることができれば，学習者の思考の幅を広げることができる可能性がある．また，分類タスクに落とし込むことで，数多くの既存研究がある文書分類の技術や知見を活かすことができると考えられる．

2. 先行研究

先行研究として，先に紹介した新共通テストの採点支援（自動採点）に関する研究と，Twitter (https://twitter.com/) に投稿されたツイートの分類を題材に分散表現を獲得する際に使用するデータセットに関する研究を紹介する．前者の研究として，[石岡 16, 亀田 17] らは採点基準を複数用意し，採点基準毎に回答がその基準を満たしているのかをチェックする手法を提案した．同義語，言い換え可能な表現等を同等に扱うことで，回答の表現の揺れを吸収している．そして，採点基準毎の点数を足し合わせたものを当該回答のチェックスコアとする．さらに，回答と模範回答や問題文とのコサイン類似度などを素性とした機械学習（ランダムフォレスト）による手法も提案し，チェックスコアを上回る精度が出ることを確認している．後者として，[Li 17] らはツイートの分類を題材にし，どのようなデータセットから分散表現を獲得すれば分類精度が高くなるのかについて述べている．まず，データセットとしてツイートデータ（スパム付きとスパム除去済みの 2 パターン）と一般データ（ニュース記事，Wikipedia 記事）の2種類を用意する．そして，それぞれを単独で使用する場合と組み合わせて使用する場合，さらには 2 単語以上から成るフレーズを考慮する場合と考慮しない場合で，ツイートの分類精度を比較した．なお，ツイートを構成する単語，もしくはフレーズの分散表現の最大値，最小値，平均値のいずれかを当該ツイートの分散表現としている．また，分類には LibLinear やランダムフォレストなどを使っている．結果，フレーズを考慮した上でツイートデータ（スパム除去済み）と一般データを組み合わせ，単語，もしくはフレーズの分散表現の平均値を当該ツイートの分散表現とした場合に最も分類精度が高くなったことを示した．

3. 問題と回答

本研究では，問題 A，問題 B，問題 B'の 3 つの問題とそれぞれの回答を取り上げる．これらは人の思考パターンを掴むために開発された問題で，問題 A は回答が数パターンに集約される問題，問題 B は回答が数パターンに集約されずに発散する問題である．既存の文書分類の技術や知見を使えば，問題 A の回答はある程度の精度で分類でき，問題 B の回答は分類が難しいものの問題を細分化するなどの工夫をすれば分類が可能になると考えた．実際，問題 B を細分化したものが問題 B'である．問題を細分化することで回答が発散することを防ぐことができ，問題 A と同様に回答をある程度の精度で分類できると考えた．なお，他の問題と比較して回答数が多いという理由から，本研究ではこれらの問題とそれぞれの回答を取り上げる．

問題 A

夏休みも終盤，家族で外食に出かけることにしました.具体的にどんなお店を選ぶべきかについて，枠組みを置いて考えてみることにしました.以下の 2 つの枠組み，どのような状況であれば A が，どのような状況であれば B の枠組みの方が適切となるでしょうか.50 字程度で答えてください.

A いい雰囲気か，料理は満足できるか，手ごろな値段か
B 量はあるか，美味しいか，盛りつけはきれいか.

連絡先：kenta.sasaki@globis.ac.jp, suzuki-k@globis.ac.jp

問題 A の回答

グロービス経営大学院に通っている学生 111 名(20 代〜50 代の社会人)に回答して貰い, 回答データを集めた. そして, 各回答を 3 パターン(意味合いの回答, 具体的な回答, その他の回答)に分類した. 実施の回答を幾つか掲載する.

1. 意味合いの回答(42 件)
- A はレストランに興味がある場合. B は料理そのものに興味がある場合.

2. 具体的な回答(39 件)
- A は大人だけで外食に出かける場合, B は子供も含む家族全体で出かける場合.

3. その他の回答(判断が難しいものも含む)(42 件)
- A は, 価格と雰囲気を重視する場合に有効です.例えば, 大人数で個室を探したい時が考えられます. B は, ボリュームを重視する場合に有効です.例えば, 食べ盛りの子供が多い場合時が考えられます.

問題 B

「運動は, 毎日とは限らず, 実施のハードルもある.睡眠は, 毎日のことではあるが, 忙しい人にとって時間は減らせない.一方で, 食事は, 毎日のことで, 意識すれば, アクションにつなげられる」という根拠の「穴」をいくつか指摘してください.

※この問題の前段の説明で, 健康志向を意識した際に取り組める要素として,

1. 食事
2. 運動
3. 睡眠

の 3 つを挙げている. そして,「食事」は恒常性と実行容易性の観点からアクションに繋げられやすい一方で,「運動」「睡眠」は実施のハードルが高いことを述べている. その上で, その逆を押さえるという問題である.

問題 B の回答

問題 A と同様, グロービス経営大学院に通っている学生 15 名に回答して貰い, 回答データを集めた. そして, 各回答を 4 パターン(全ての要素に言及している回答, 一部の要素にしか言及していない回答, 根拠が弱い回答, イシューずれ回答(論点を外している回答))に分類した. 形式的には 4 パターンに分類できているものの, 根拠が弱い回答やイシューずれ回答はさらに幾つかのパターンに分けることができるので, 回答が発散していると言えるであろう. 実際の回答を幾つか掲載する.

1. 全ての要素に言及している回答(3 件)
- 運動は毎日ムリとされているが駅まで歩くことで運動は可能.睡眠時間は少なくともマクラ改善による質の良い睡眠が可能.食事は意識すればとあるが, 毎日の勤務が遅い時間であれば自炊やバランスの良い食事が取れない可能性もある.

2. 一部の要素にしか言及していない回答(5 件)
- 食事についても仕事による時間的制約や飲み会等があり得る.
- 食事はお金との関係性があるので, その費用内で収まるように健康を意識した食事が可能か検討する必要がある.

3. 根拠が弱い回答(4 件)
- 全ての人が定期的に食事をしているわけではない, そもそも食事をする時間がない, アクションすることの難しさ
- 食事は, 睡眠や運動と異なり一人でとるものとは限らないため, 自分の意思だけでは達成できないことがある.休みの日と出勤日で意識が変わってしまう.

4. イシューずれ回答(3 件)
- 睡眠を減らせないほど忙しいビジネスマンは, 食事への意識も回らない.
- 食べる物を選べない, 場所を選べない, 時間を選べない,
- 運動も意識すれば毎日(通勤時), 徒歩距離を伸ばすことでも実現可能.時間が取れないから睡眠が取れないなら食事も同様

問題 B'

1. 「運動は, 毎日できるとは限らず, 実施するハードルもある」という主張の「弱さ」はなんでしょうか.
2. 「睡眠は, 毎日のことではあるが, 忙しいヒトにとっては増やすことは難しい」という主張の「弱さ」はなんでしょうか.
3. 「食事は, 毎日のことで, アクションのハードルも高くない」という主張の「弱さ」はなんでしょうか.

問題 B'の回答

この問題もグロービス経営大学院に通っている学生 23 名に回答して貰い, 回答データを集めた. 問題 B の回答のように発散することはなく, 各回答を2~3パターンに分類することができた. 例えば, 1 つ目の「運動」問題に対する回答は 2 パターン(反証の回答, その他の回答)に分類できた.

1. 反証の回答(13 件)
- 階段を上る, 一駅歩くなど, 日常生活の中でできる運動もあること

2. その他の回答(10 件)
- 個人の生活環境によりハードルに差が生まれるので, 一般論としての根拠が弱い.

4. 手法

本研究の手法を述べる. まずは回答の分散表現獲得方法, 次に回答の分類方法について説明する. そして, 最後に回答の分類に寄与している単語の特定方法を説明する.

4.1 回答の分散表現獲得

回答の分散表現獲得には, 先行研究で紹介した [Li 17] らの手法を参考にした. まず, コーパスとして, Wikipedia のダンプデータに加え, グロービス経営大学院が出版している書籍のテキストデータを利用した. そして, word2vec で学習したものから, 単語の分散表現を得た. なお, 単語分割には MeCab を利用し, 単語の分散表現は 200 次元, ウインドウサイズは 5 とした. そして, 回答を構成する単語分散表現の平均を当該回答の分散表現とした(図 1). 但し, 助詞や句読点などはノイズになり得ると考え, 名詞, 動詞, 形容詞の単語に限定した.

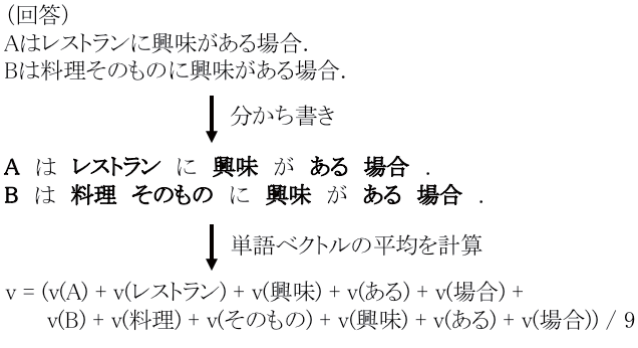

図 1 回答の分散表現化

4.2 回答の分類

回答の分類アルゴリズムには SVM (Support Vector Machine) を用い，無償で利用可能な LIBSVM (https://www.csie.ntu.edu.tw/~cjlin/libsvm/) を利用した．カーネルには RBF カーネルを採用し，ハイパーパラメータはグリッドサーチによって最も精度が高くなるものに決めた．なお，精度は leave-one-out 交差検証で評価した．

4.3 回答の分類に寄与している単語の特定

本研究では，回答の分類に寄与している単語（分類において影響力の大きい単語）の特定も行った．分類に寄与している単語を特定できれば，学習者へのフィードバックを回答のパターンによって換える場合，そのフィードバックが返ってくる理由を学習者に提示することができるかもしれない．その結果，フィードバックへの納得感が高まり，学習者の理解の向上に繋がる可能性がある．具体的には，例えば「A は新婚の場合なら適切であり，B は小さい子供がいる家族なら適切である.」という回答の場合，

1. A は~~新婚~~の場合なら適切であり，B は小さい子供がいる家族なら適切である．
2. A は新婚の~~場合~~なら適切であり，B は小さい子供がいる家族なら適切である．

（以下省略）

の 8 パターンの回答を用意する．今回，「A」「B」も除く単語の対象となるが，これらは全ての回答に出現するので対象外とした．そして，それぞれの回答をモデルに掛け，各回答の予測スコア（分類超平面からの距離）を算出した．予測スコアが悪化している（絶対値が小さくなっている）回答ほど，その回答で除いた単語の影響力が大きいと考えられる．前記の例で，1 番目の回答で予測スコアが最も悪化している場合，「新婚」という単語が分類に最も影響していると言える．

5. 結果と考察

まずは，問題 A による分類結果と分類に寄与している単語について述べる．次に，問題 B'による分類結果と分類に寄与している単語について述べる．

5.1 分類結果（問題 A）

SVM による分類結果を表1に示す．まずは，「1. 意味合いの回答」と「2. 具体的な回答」の 81 件のデータセットだけを使って分類（二値分類）した．結果，分類誤りは 81 件中 3 件だけであり，96.3%(=78/81)という非常に高い精度であった（表1）．また，分類誤りとなった 3 件も予測スコアは非常に小さく，つまり，誤り方として惜しいものであることが分かった（表 2）．ここで，予測スコアが正であれば「1. 意味合いの回答」，負であれば「2. 具体的な回答」と予測したことになる．

表1 問題 A の回答分類結果（概要）

		予測	
		1. 意味合い	2. 具体的
実際	1. 意味合い	39	2
	2. 具体的	1	38

表 2 問題 A の回答分類結果（詳細）※予測スコアで昇順ソート

予測スコア	正解ラベル	回答
1.45	1	A は料理以外の要素も大事な場合，B は料理自体が大切な場合
（略）	（略）	（略）
0.15	**2**	**旅行中などはじめての土地で食事をする場合には A, 平日のランチなどすでに何度も行ったことのあるお店から選ぶ場合には B**
0.07	1	A は雰囲気重視の場合，B は食欲重視の場合
-0.01	**2**	**A.レストランに興味がある場合.B.料理そのものに興味がある場合.**
-0.05	2	A は異性と時間を楽しみたい状況，B は部活の仲間と楽しみたい状況
-0.08	2	小さい子ども連れ，バリアフリー環境が必要など，制約がある場合は A，制約はなく料理のみで判断できる場合には B
-0.08	**1**	**食事だけでなく，外食の体験そのものの満足度を最大化したい場合は A，食事に関する満足度を最大化したい場合は B**
（略）	（略）	（略）
-1.61	2	A は核家族(親と子供)，B は祖父母な親戚も含めた大勢の状況

次に，「1. 意味合いの回答」と「2. 具体的な回答」のデータからモデルを作成し，「3. その他の回答」のデータ 42 件のラベルを予測した．ここで，予測スコアの絶対値が大きい 2 件の回答と，絶対値が小さい 2 件の回答を表 3 に示す．絶対値が大きい回答については，「3. その他の回答」とラベル付けしたものの，「1. 意味合いの回答」と「2. 具体的な回答」にラベル付けしても問題なかった回答だと考えられる．一方，絶対値が小さい回答については，特に予測スコアが-0.04 の回答は「1. 意味合いの回答」と「2. 具体的な回答」の双方に言及している．この場合はどちらとも言えないというのが正解であり，予測スコアの絶対値が小さくなっていることから，予測結果は妥当であったと言える．

表 3 予測スコアの絶対値が大きい回答と小さい回答

予測スコア	回答
1.22	料理そのものに関心がある場合は A，料理だけではなくコストパフォーマンスが大切な場合は B
0.00	外食の目的に応じて変わる，コミュニケーションが目的なら A，食べることが目的なら B
-0.04	A は，価格と雰囲気を重視する場合に有効です.例えば，大人数で個室を探したい時が考えられます．B は，ボリュームを重視する場合に有効です.例えば，食べ盛りの子供が多い場合時が考えられます．
-0.67	参加者が雰囲気の良いお店が好きな場合は A，参加者に美食家が多い場合は B

5.2 回答の分類に寄与している単語（問題 A）

今回対象とした回答は，「お店そのものを重要視する場合は A，料理そのものを重要視する場合は B（回答 1）」「A は新婚の場合なら適切であり，B は小さい子供がいる家族なら適切である.（回答 2）」の 2 つである．結果を表 4，表 5 に示す．回答 1 では，「料理」「店」の影響力が大きく，「そのもの」「場合」の影響力が小さいことが分かる．「料理」「店」は回答を意味付ける重要な単語の一つであるが，「そのもの」「場合」はその単語が無くても文の解釈はほとんど変わらないという点で，分類に寄与している単語をうまく抽出できたと言えるであろう．回答 2 では，「子供」「家

族」の影響力が大きく,「小さい」「適切」の影響力が小さいことが分かる. これも回答 1 と同様,「子供」「家族」は回答を意味づける重要な単語の一つであるが,「小さい」「適切」はその単語が無くても文の解釈はほとんど変わらないという点で, 分類に寄与している単語をうまく抽出できたと考えられる.

表 4 回答の分類に寄与している単語(回答 1)

単語	予測スコア
料理	0.73
店	1.00
重要	1.00
する	1.16
視	1.19
そのもの	1.19
場合	1.22

表 5 回答の分類に寄与している単語(回答 2)

単語	予測スコア
子供	-0.95
家族	-0.97
いる	-1.04
新婚	-1.10
場合	-1.10
小さい	-1.13
適切	-1.39

5.3 分類結果(問題 B')

SVM による分類結果を表 6 に示す. 結果, 分類誤りは 23 件中 5 件となり, 78.3%(=18/23)という高い精度であった(表 6). 問題 A の時よりも精度が落ちた要因として学習データの数が少なかったことが挙げられ, 学習データの数が増えれば問題 A の時と同様に非常に高い精度で分類できると考えられる.

表 6 問題 B'の回答分類結果(概要)

		予測	
		1. 反証	2. その他
実際	1. 反証	12	1
	2. その他	4	6

5.4 回答の分類に寄与している単語(問題 B')

今回対象とした回答は,「毎日の通勤で歩くなど, 運動の見方によっては毎日実行可能.(回答 3)」「実施するハードルが具体性に欠けている点(回答 4)」の 2 つである. 結果を表 7 と表 8 に示す. 予測スコアが正であれば「1. 反証の回答」, 負であれば「2. その他の回答」と予測したことになる. 回答 3 では,「歩く」「通勤」の影響力が大きく,「可能」「見方」の影響力が小さいことが分かる. また, 回答 4 では,「性」「具体」の影響力が大きく,「する」「いる」の影響力が小さいことが分かる. 影響力が大きい単語は回答を意味付ける重要な単語の一つであるが, 影響力が小さい単語はその単語が無くても文の解釈はほとんど変わらない. これより, 問題 A の時と同様に, 分類に寄与している単語をうまく抽出できたと言えるであろう.

表 7 回答の分類に寄与している単語(回答 3)

単語	予測スコア
歩く	0.57
通勤	0.69
毎日	0.80
運動	0.86
実行	0.99
可能	1.00
見方	1.14

表 8 回答の分類に寄与している単語(回答 4)

単語	予測スコア
性	-0.75
具体	-0.79
欠け	-0.81
実施	-0.85
点	-0.85
ハードル	-0.89
する	-1.01
いる	-1.13

6. 結論と今後

本研究では, 社会人向け教育における記述式問題回答自動分類の可能性について述べた. 回答が 50 文字程度, 且つ, 3 パターン程度に明確に分かれる問題においては, シンプルな手法でも高精度に回答を分類できることが分かった. さらに, 分類に寄与している単語を特定することもできた. また, 回答が発散する問題においては, 問題を細分化することで, 高精度で回答を分類でき, 且つ, 分類に寄与している単語を特定することができた. これにより, 社会人向け教育における記述式問題において, 回答のパターンによってフィードバックを換えるなど学習者の回答に応じた適切な振り返りサービスを提供することができる可能性があることが確認できた.

今後は, 学習者の回答に応じた適切な振り返りサービスが, 本当に学習者の理解の向上に繋がるのか検証したい. また, 問題の細分化などの工夫をしてもあらゆる問題の回答をパターン化できるとは限らないので, そのような場合にどのようなアプローチを取るのが良いのかについて考えていきたい. さらに, 50 文字を超えるような長い回答の場合でも, 本研究のような回答の自動分類が可能であるのかについても調査したい.

参考文献

[Li 17] Quanzhi Li, Sameena Shah, Xiaomo Liu and Armineh Nourbakhsh: Data Sets: Word Embeddings Learned from Tweets and General Data, In Proc. of ICWSM, pp.428-436 (2017)

[Sultan 16] Md Arafat Sultan, Cristobal Salazar and Tamara Sumner: Fast and Easy Short Answer Grading with High Accuracy, In Proc. of NAACL, pp.1070-1075 (2016)

[Taghipour 16] Kaveh Taghipour and Hwee Tou Ng: A Neural Approach to Automated Essay Scoring, In Proc. of EMNLP, pp.1882-1891 (2016)

[水本 18] 水本智也, 磯部順子, 関根聡, 乾健太郎: 採点項目に基づく国語記述式答案の自動採点, 言語処理学会 第 24 回年次大会 発表論文集, pp.552-555(2018)

[石岡 16] 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, 信学技報, vol.116, no. 379, pp.87-92(2016)

[亀田 17] 亀田雅之, 石岡恒憲, 劉東岳: 短答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)