6Q-07

Doc2Vec を用いた国語記述式答案の自動評価

鈴木　千尋†　　佐藤　直行‡

公立はこだて未来大学大学院システム情報科学研究科†‡

1. 研究の背景

大学入試センター試験に代わる新テストでは新たに記述式テストが導入されるが，採点にかかる時間や手間が増大することや多数の採点者が必要であること [1]，採点者間で採点結果の不一致が生じることなどが懸念されている．

この問題を受け，記述式答案の自動採点に関する研究が増えてきている．水本ら [2] は，国語記述式答案を対象として採点項目別に点数を出力するモデルを提案した．石岡ら [3] は，社会科の記述式答案を対象に採点結果と採点の根拠を併せて出力する自動採点システム JS^4 を開発した．ただし，上述のモデルや JS^4 の構築には大量の採点済み答案データが必要であり，一般の学習者が簡便に利用できるものではない．

そこで，自然言語処理技術の一つである Doc2Vec [4] に着目した．Doc2Vec は，単語や文章の意味を約数百次元のベクトルとして表す手法で，関連論文の検索や学会のトレンド解析などに利用されている．また，コサイン類似度推定法などと用いることで，2 つの文章の意味的類似度を計算できるとされている．Doc2Vec は採点済み答案データを必要としないため，従来の方法よりも簡便に記述式答案を評価できる可能性がある．

本研究では，Doc2Vec を用いて記述式答案の質的な良さを評価する手法を提案する．提案手法の有効性を評価するために，自由再生テスト，要約テスト，短答記述式テストの答案の自動評価を行い，提案手法の妥当性を検討した．

2. 提案手法

提案手法は，問題ごとに用意された 1 つの模範解答と答案の類似度により評価する方法である．記述式答案と模範解答のベクトルを得るために，Doc2Vec の実装である Python ライブラリ Gensim に「書き言葉均衡コーパス図書館サブコーパス」を学習データとして与え，100 次元のモデルを生成した．学習データの前処理は，不要な記号類の除去，文章の単語分割，表記の統一，ストップワードの除去を行った．文章の単語分割では，単語を基本形に直し，名詞，動詞，形容詞，副詞以外は除去した．生成したモデルにより記述式答案と模範解答を 100 次元のベクトルに変換し，コサイン類似度推定法を用いて記述式答案と模範解答のコサイン類似度（-1 から 1 までの値をとる）を算出する．答案の質的な良さはコサイン類似度の値で評価した．コサイン類似度の値が 1 に近ければ近いほど答案と模範解答は類似しており，質的に良いと解釈できる．

3. 自由再生テストの答案の評価

自由再生テストは文章内容をそのまま再生するテストで，答案と模範解答の類似を検出し易いと予想された．

2.1 評価方法

実験により自由再生テストの答案を収集し（約 3000 語の文章を 2 つ，$n = 10$），模範解答と答案の類似度を算出した．人手のスコアは，国語教育の専門家が模範解答を元に採点したもので，自由再生文の出来により 0 から 9 までのいずれかのスコアを与えた．

2.2 結果と考察

再生文（277 ± 38.93語）についての類似度と人手のスコアの無相関検定を行ったところ，文章 A ($\rho = 0.72,\ S = 46,\ p < .05$) と文章 B ($\rho = 0.92,\ S = 14,\ p < .0005$) の両方で有意な正の相関が得られた（図 1）．このことは，答案と模範解答に含まれる単語の一致度合いで答案の良さを評価できる場合に，提案手法が有効であることを示していると考えられる．

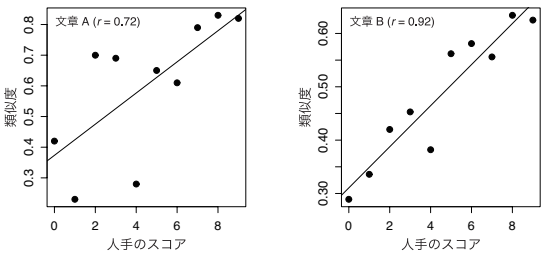

図 1　類似度と人手のスコアの相関

4. 要約テストの答案の評価

要約テストは，黙読した文章内容を 100 字程度

Evaluation of Descriptive Type Answers Using Doc2Vec
†Suzuki Chihiro・Future University Hakodate
‡Sato Naoyuki・Future University Hakodate

でまとめるテストである．文章中の表現を抽出するだけでなく，自分なりの表現で言い換える必要があり，自由再生テストに比べて答案の評価が難しいと予想された．

3.1 評価方法

実験により要約テストの答案を収集し（約4000語の文章を2つ，$n = 12$），模範解答と答案の類似度を算出した．人手のスコアは，国語教育の専門家が模範解答を元に採点したもので，要約文の出来により0から3までのいずれかのスコアを与えた．

3.2 結果と考察

類似度と人手のスコアの無相関検定を行った結果，文章Aで有意な正の相関の傾向が得られたが $(r = 0.10,\ \mathrm{t}(10) = 0.31,\ p = .76)$，文章Bでは有意な相関は得られなかった $(r = 0.53,\ \mathrm{t}(10) = 1.99,\ p = .07)$（図2）．その理由としては，提案手法では模範解答の言い換えが含まれるタイプの答案は評価が困難であることが考えられる．

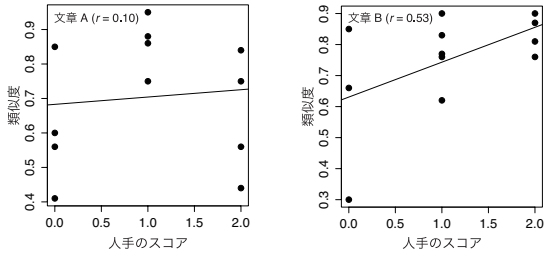

図 2　類似度と人手のスコアの相関

4. 短答記述式テストの答案の評価

短答記述式テストは，文章内容に関する問いに数十字から百字程度で解答するもので，2次試験など大学入試レベルのテストである．

4.1 評価方法

実験により短答記述式テストの答案を収集し，模範解答と答案の類似度を算出した．問題は，“抽出型”が3題，“表現工夫型”が4題の計7題から成る．“抽出型”は，傍線部に対応する内容を文中から抽出しまとめるタイプの問題である．“表現工夫型”は，文中の表現だけでは不十分で自分なりの表現を工夫して解答しなければならないタイプの問題である．人手のスコアは，国語の専門家が採点基準に従い採点した．

4.2 結果と考察

類似度と人手のスコアの無相関検定を行ったところ，抽出型の問題で有意な正の相関または有意な正の相関の傾向が得られた．このことは，実験1と同様，記述式答案と模範解答に含まれる単語の一致度合いで記述式答案の良さを評価できる場合に，提案手法が有効であることを示していると考えられる．一方，表現工夫型の問題では有意な相関は得られなかった．その理由としては，表記は全く異なるが特定の文章中でのみ同じ内容を表す単語が含まれるタイプの答案は評価が困難であることが考えられる．

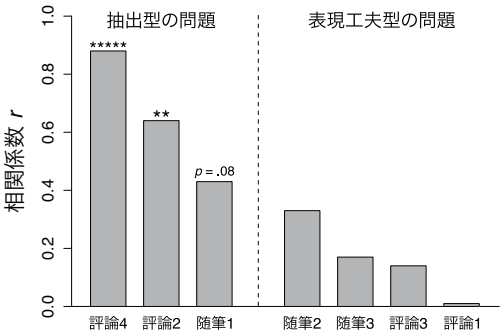

図 3　類似度と人手のスコアの相関係数

5. まとめ

本研究では，Doc2Vecを用いて記述式答案を評価する手法を提案し，その有効性を評価した．その結果，自由再生テストと抽出型の問題で類似度と人手のスコアの間に有意な正の相関が得られた．このことは，答案と模範解答に含まれる単語の一致度合いで答案の良さを評価できる場合に，提案手法が有効であることを示している．一方，要約テストと表現工夫型の問題では有意な相関は得られなかった．その理由としては，提案手法では模範解答の言い換えが含まれるタイプの答案や，表記は全く異なるが特定の文章中でのみ同じ内容を表す単語が含まれるタイプの答案は評価が困難であることが考えられる．これは複数の模範解答を用いることで改善できる可能性があるが，今後の検討を要する．

参考文献

[1]独立行政法人 大学入試センター：大学入学共通テストの導入に向けた試行調査（プレテスト）（平成29年11月実施分）の結果報告の概要，(2018).
[2]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).
[3]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).
[4] Le, Q. and Mikolov, T.: Distributed representations of sentences and documents, *Proc. 31st International Conference on Machine Learning* (*ICML 2014*), pp.1188-1196 (2014).