K-018

協調学習における評価対象テキストの自動評定
Automatic scoring of texts in collaborative learning

福田 治輝† Haruki Fukuda　綱川 隆司† Takashi Tsunakawa　大島 純† Jun Oshima　大島 律子† Ritsuko Oshima
西田 昌史† Masafumi Nishida　西村 雅史† Masafumi Nishimura

1. はじめに

協調学習では，複数の学習者がグループ活動を通して協力し，意見を交わしながら課題に向き合う．協調学習内で行われる調整活動プロセスにおける調整能力の評価手法として CSSER (Collaboration Scenario-based Scale for Emotion Regulation)と呼ばれる手法が提案されている[1]．CSSER では協調学習で起こるあるシナリオにおいて登場する学習者の思考や発言を表す吹き出しに状況に応じた内容のテキストを記入させ，評価を行っている．評定作業を効率的に行うため，学習者の回答ならびに教授者側の回答への評定をオンラインで行えるシステム Computer-based Regulation Profiler (CRP) が開発されている[2]．CSSER を実施するには，グループ編成を行うまでの期間で評定作業を済ませる必要があるが，ルーブリックに基づいて人手でテキストを評定しているため，学習者数の多い場合は実用的でない．

本研究では，学習者の回答の評定を自動化する方法として．近年多くの自然言語処理タスクにおいて特筆すべき性能を示し注目を集めている BERT (Bidirectional Encoder Representations from Transformers)[3]を利用し，クラス分類を行うことにより評定結果を出力する方法を提案し，その効果を検証した．

2. CSSER

Oshima ら[1]により作成された CSSER シナリオは 4 種類のシナリオからなり，被験者自らが参加する協調学習場面が再現され，同じグループに属する登場人物同士の会話を描いたコマが続く形でストーリーが展開されている．シナリオの最後には，被験者自身を表現したキャラクターの上に空欄の吹き出しが設けられたコマが用意されており，その状況で考えている内容（以下，「心の声」とする）とその状況で発言する内容（以下，「発言」とする）を想定し，記述させる．それぞれのシナリオで記述された「心の声」と「発言」を，社会認識的側面と社会感情的側面の 2 側面から，既定のルーブリックに従って 1〜5 点の評定を与える．つまり，学習者の CSSER 評定は，4 シナリオ*2 つの記述項目（心の声，発言）* 2 つの評価側面（社会感情的側面，社会認識的側面）の計 16 の評定となる．

3. 提案手法

本章では，提案方法に用いる BERT[3]と，その適用方法を述べる．

3.1 BERT

BERT は自己注意機構のみを使用したモデルである Transformer[4]をベースとしている汎用言語表現モデルである．大規模なテキストコーパスで事前学習を行い，各タスクに対してファインチューニングすることにより，多くの自然言語処理タスクにおいて特筆すべき性能を示している．

3.2 適用方法

日本語 Wikipedia 記事（約 1800 万文）によってプレトレーニングされたモデル[5]を用いて，BERT への入力を CSSER の「心の声」または「発言」のテキスト，出力を社会認識的側面または社会感情的側面の 1〜5 点の評定とし，それぞれクラス分類を行うようにファインチューニングを行う．1 から 5 の各評定について求めた尤度が最大である評定を，自動評定の結果として採用する．

4. 評価

4.1 実験設定

2015 年度後期，2016 年度後期，2017 年度後期の国立大学情報系学部の初年時必修科目として行われた「学習マネジメント」の受講生に対して，初回講義の冒頭および最終回講義の冒頭の 2 回 CSSER を実施した．「心の声」と「発言」合わせて 20760 件（Collaboration シナリオ: 5250 件，Personal Priorities シナリオ: 5187 件，Team work シナリオ: 5152 件，Work and communication シナリオ: 5171 件）の回答（記入漏れ，解読できない悪筆等の回答データを除く）が得られた．2015 年度分の回答データに関しては，ルーブリックに基づき学習の調整活動の研究に精通する研究者の計 2 名で評定を行い，評定の差異があった箇所は合議の上で最終的な評定を決定している．2016，2017 年度分の回答データに関してはサンプルデータを用いた評定基準の調整をし，十分に合意形成を行った後に，2 名の評定者が作業分担をして評定を行っている．これらの得られた回答データおよび評定全てを用いた場合，およびシナリオ別の回答データおよび評定を用いた場合について評価実験を行った．

実験対象となるデータを 3:1:1 の割合でランダムにトレーニングセット，バリデーションセット，テストセットに分割し，実験を行う．ベースラインとして，「心の声」と「発言」のテキストを Doc2Vec[6]でベクトル化し，コサイン類似度が最も大きくなるものの評定を自動評定の値として採用する手法と比較する．評価指標として，Cohen の一致係数（Cohen’s coefficient of agreement：以下 κ 係数）および F 値のマイクロ平均を用いる．κ 係数は，2 人以上の評定者の評価の一致率を表す係数であり，評定者間の信頼性を求めることができる．BERT の主なパラメーターとして，語彙サイズ（サブワード含む）32,000，バッチサイズ 32，エポック数 3 を用いた．Doc2Vec についてはベクトルサイズ 300，エポック数 100 でモデルを作成し，ステップ数 10000 でテキストのベクトルを生成した．

† 静岡大学 情報学部　Faculty of Informatics, Shizuoka University

表 1　実験結果

		社会認識的側面				社会感情的側面			
		発言		心の声		発言		心の声	
手法	シナリオ	κ係数	F 値	κ係数	F 値	κ係数	F 値	κ係数	F 値
BERT	全シナリオ	0.511	0.653	0.398	0.603	0.439	0.611	0.385	0.580
	Collaboration	0.500	0.659	0.366	0.711	0.440	0.587	0.396	0.558
	Personal Priorities	0.456	0.619	0.369	0.545	0.587	0.739	0.421	0.634
	Team work	0.507	0.640	0.320	0.603	0.514	0.660	0.416	0.599
	Work and communication	0.611	0.778	0.325	0.599	0.445	0.664	0.318	0.591
Doc2Vec	全シナリオ	0.160	0.387	0.089	0.395	0.141	0.458	0.303	0.615
	Collaboration	0.135	0.370	0.127	0.485	0.137	0.368	0.088	0.340
	Personal Priorities	0.143	0.376	0.102	0.334	0.091	0.283	0.099	0.303
	Team work	0.173	0.384	0.145	0.463	0.132	0.417	0.150	0.402
	Work and communication	0.239	0.448	0.069	0.421	0.289	0.480	0.115	0.380

4.2 結果

実験結果を表 1 に示す．全ての結果において，BERT による推定は Doc2Vec による推定よりも高い精度となった．全体的な傾向として，心の声より発言の方が高い精度で推定できている．

BERT によって出力される 1 から 5 の各評定について個別で推定される尤度は回答データ毎にばらつきがあり，一つの評定に対して非常に高い尤度がつくものから，上位の複数の評定が近い値の尤度になる場合も見られる．そこで，尤度が高いものほど推定精度が高いのではないかと考え，テストセットの中で出力した評定の尤度で降順にソートを行い，上位の回答データのみでκ係数の計算を行った．縦軸がκ係数，横軸がκ係数を算出するのに利用したデータの割合として，全てのシナリオの発言データの社会認識的側面のデータを用いて計算を行った結果が以下である（図 1）．尤度が高いテキストほど，高い推定精度が得られるといえる．

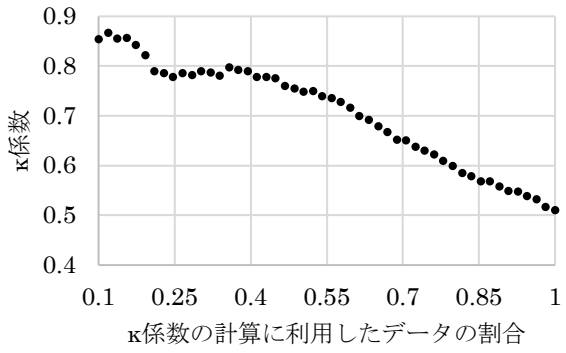

図 1　評価対象としたデータ量（高尤度順）と推定精度の関係

5. おわりに

CSSER では学習者の思考や発言を表す吹き出しに状況に応じた内容のテキストを記入させ、これをルーブリックに基づいて人手で評定することにより評価を行っているが，人手による作業は精度が高いものの時間がかかりすぎ，即時的なフィードバックを返せないという大きな欠点がある．この評定作業を自動化する手法として，BERT の適用方法を検討し，評価した．実験の結果，Doc2Vec を利用した手法と比較して高い精度を得られていることを示した．シナリオや評定項目によっては，2 名の人手での評定精度はκ係数 0.72 程度であり，現状の推定精度であっても BERT が出力する尤度が高いテキストほど高い精度での評定ができることが確認できたため，尤度が高く自動評定の精度が高いテキストは自動で評定し，尤度が低いものに関しては人手での評定を行うことで評定作業の効率化を見込めると考えられる．人手での評定と併用する場合，類似テキストの提示など評定支援の機能に関しても今後検討していく．

謝辞

本研究は JSPS 科研費 JP19H01714 の助成を受けたものである．

参考文献

[1] R. Oshima and J. Oshima, “Collaboration Scenario-based Scale for Emotion Regulation: Measuring Learners’ Agency to Regulate Own, Others’ and Group Emotions,” in Proceedings of EdMedia + Innovate Learning 2015, 2015, pp. 796–801.
[2] 神戸優, 大島律子, 大島純, “学習調整能力評価サイトの Web デザイン検討,” 日本教育工学会第34回全国大会講演論文集, pp. 391–392, 2018.
[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” arXiv:1810.04805 2018.
[4] A. Vaswani et al., “Attention Is All You Need,” in Proc. of the 31st International Conference on Neural Information Processing Systems, pp. 6000-6010 2017.
[5] 柴田知秀, 河原大輔, 黒橋禎夫, “BERTによる日本語構文解析の精度向上,” 言語処理学会 第25回年次大会 発表論文集, pp. 205-208 2019.
[6] Q. V. Le and T. Mikolov, “Distributed Representations of Sentences and Documents,” in Proc. of the 31st International Conference on Machine Learning, pp. 1188-1196 May 2014.