様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　２５　日現在

機関番号：　３２６９２
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2017
課題番号：　２６３５０２８９

研究課題名（和文）大規模かつ多様な学習データを活用した知的協調スクリプト実行システムの開発と評価

研究課題名（英文）Development and evaluation of intelligent collaborative script　for large-scale and various learning data

研究代表者

稲葉　竹俊（INABA, Taketoshi）

東京工科大学・教養学環・教授

研究者番号：１０３８６７６６

交付決定額（研究期間全体）：（直接経費）　　3,700,000円

研究成果の概要（和文）：　コンピュータ支援協調学習研究において，相互作用の活性化のために学習者の活動の課題やプロセスをシナリオ化して構造化する協調スクリプトの研究が関心を集めている。本研究では大規模な人数の学習者を対象に複数の協調スクリプトの有効性を評価することを最終目標とした。
　しかし、その前提として、大規模な学習データの分析手法の確立のため、本研究では、会話データへのコーディングを深層学習技術によって自動化する手法を開発し，その精度を評価した。その結果，開発手法が機械学習のベースラインを凌駕する正解率を実現することが明らかになり、大規模なデータの定量的解析を簡便におこなうことが可能となった。

研究成果の概要（英文）：　In the area of Computer Supported Collaborative Learning (CSCL) research, scripting collaborative learning is a relatively new but promising approach to promote learning. The term scripting is used to describe ways of prescribing relevant elements for collaborative interaction.In this research, the goal was to evaluate the effectiveness of multiple collaborative scripts for a large number of learners.
　However, as a prerequisite, it was necessary to establish an analytical method capable of performing qualitative analysis of large-scale learning data in combination with quantitative methods. Therefore, we developed a method to automate coding to conversation using deep learning technology and evaluated its accuracy. As a result, our method realized the accuracy rate exceeding the baseline of machine learning, and it has become possible to perform qualitative analysis of large scale big data in a simple manner.

研究分野：　教育工学

キーワード：　スクリプト　コンピュータ支援協調学習　教育データ　深層学習　コーディング

１．研究開始当初の背景

コンピュータ支援協調学習研究において，相互作用の活性化のために学習者の活動の課題やプロセスをシナリオ化して構造化する協調スクリプトの研究が関心を集めている。しかし、大規模なデータに依拠したスクリプトの効果の研究は極めて少なく、大半の研究は小規模な事例研究の域を出ていない。

２．研究の目的

本研究の最終ゴールは協調スクリプトの実行システムを構築し、グループの特性や学習過程の状況に応じて動的にスクリプトを最適化する知的機能や柔軟なグループ生成機能を実装することで、その有効性を評価することにある。そのため以下のような目標を中心に研究を進める。1.　言語データ評価技法の探求：先行研究のコーディングを参考に、スクリプトに適応したコーディングスキームを確立し、チャット上での相互作用を質的かつ定量的に評価する。 2．分析手法の探求：実際の学習局面のモニタリングおよび適応的支援の実現のため、学習局面を分析する。

３．研究の方法

本研究組織が開発した協調学習支援システムを学内のクラウドから取得された大規模なチャットデータを対象に協調プロセスの質的データを定量的に分析するためコーディングスキームを開発し、深層学習技術を用いてコーディング作業を自動化し、その結果から協調学習プロセスを分析する。

４．研究成果

(1)　１．に述べた最終目標にむかう第一ステップとして，チャットデータのコーディングの自動化の技法を開発し，その精度の検証（検証１）と教育上の有効性の検証（検証２）の２つの検証を行った。

具体的には，相当量のチャットデータに手動でコーディングを行い，その一部をトレーニングデータとして機械学習の最新技術である深層学習に学習をさせ，その後，テストデータに自動コーディングを実施した。精度の評価にあたっては，機械学習による自動コーディングを実践した既存研究で用いられた機械学習アルゴリズムのベースラインとなるナイーブベイズや Support Vector Machines(SVM)との精度比較を行った。また，開発手法の教育的有効性の検証では，新たなチャットデータを対象に自動コーディングを行い，その結果からどのような知見を得ることができるかを検討した。

(2)　検証１：データとコーディングスキーム

①会話データセット

会話データセットは本研究組織が独自に開発した CSCL システムを大学の講義内で用いて，オンラインでの協調学習を行いシステム内のチャット機能から得られた学生間の会話である。本研究で利用する発言データ元の CSCL の利用状況を Table１に示す。１人の学生が複数の科目に参加しているため，グループ数×グループ人数よりも参加学生数が少なくなっている。

Table 1　発言データの概要

科目数	7 科目
グループ人数	3-4 人
時間	45 分～90 分
グループ数	202 グループ
参加学生数	426 人
データセット	11504 発言

②コーディングスキーム

本研究組織が作成したコード付与のためのマニュアルに従い，チャットの１発言に対し１つのラベルを付与する。ラベルは Table 2 に示す 16 種類となっており，このラベルのいずれかを付与する。

発言データは講義単位で分割されており，コーダー6 名が分担してコーディングを行った。その際に，各講義に対し 2 名のコーダーを割り当て，すべての発言についてその 2 名が，それぞれコーディングを行った。これらのコーディングの一致または不一致の結果を精査したところ，発言内容的に重複しているコードや，コーダーによりブレのあるコードがあることが判明したため，研究組織の合議によりコードの統合および一部コードの再コーディングを行った。この結果，2 名のコーダー間の一致率は 82.3%で，偶然によらない一致率を表す Kappa 係数は 0.800 という高い結果となり，深層学習のトレーニングデータとして十分実用に耐えうるものとなった。

Table 2　ラベルの種類

ラ　ベル	ラベルの意味	発言例
同意	肯定的な返答	いいと思います
提案	意見を伝えるまたは、YES/NO 質問	この五人で提出しませんか？
質問	YES/NO 以外の質問	タイトルどうしましょうかね
報告	自身の状況を報告する	複雑の方はなおしました
挨拶	他メンバーへの挨拶	よろしくお願いします
回答	質問や確認に対する返信	そうみたいです！
メタ	課題内容以外の発言 システムに対する意見など	はやくも自分の発言が消えるバグが
確認	課題内容や作業の進め方について確認	じゃあ提出していいですか？
感謝	他メンバーへの感謝	ありがとう！

愚痴	課題やシステムにたいする不満など	テーマがいまいちだよね；；
ノイズ	意味をなさない発言	?会?日??
依頼	誰かに作業を依頼する	どちらかが回答お願いします
訂正	過去の発言を訂正する	すいません児童の間違いです
不同意	否定的な返答	30 分は長すぎる気がします
転換	次の課題へ進めるなど、扱う事象を変える発言	とりあえずやりますか
ジョーク	他メンバーへのジョーク	そんなの体で覚えろ的な？(´・ω・`)

③深層学習を用いた自動コーディング手法

コーディングを自動的に行うために，本研究では，深層学習と呼ばれる技術を用いる。深層学習とは，近年劇的に発展した機械学習の一手法であり，数十から数百に及ぶ深いレイヤーと，しばしば数百万以上となる重みパラメータからなる巨大なニューラルネットワークを規模の大きなデータから学習させるものである。 学習に深層学習を利用するメリットとしては，予測精度の高さのほかに，以下の点があげられる。まず，既存の機械学習の手法では，人間が有効な特徴量を考え，それを抽出するためのプログラムを行う必要があったため，多大なコストと開発時間が必要となっていたが，深層学習では特徴の抽出までが内部で行われるため，そのコストが大幅に削減できる。また，モデルの学習には計算時間がかかるものの，一度学習が終われば，新しいデータへの適用は，極めて高速に行なうことができ，実用上，従来の機械学習の手法と遜色はない。

本研究では，深層学習手法として， 畳み込みニューラルネット(CNN)による分類モデル，長短期記憶(LSTM)による分類モデル，Sequence to Sequence (Seq2Seq) による分類モデルの 3 つを適用する。このうち，Seq2Seq モデルは，エンコーダー及びデコーダーとよばれる 2 つの LSTM のユニットから構成された深層ニューラルネットワークであり，それぞれのパートに，ペアをなす単語列を入れて分類問題や文生成の学習を行うものである。例えば，翻訳システムであれば，ある言語の文とその対訳文が，質疑応答システムであれば，質問文と応答文がそのペアにあたる。

さらに古典的な機械学習の手法である SVM を用いたモデルをベースラインとして用いる。各モデルの精度の検証は，自動コーディングの一致率，および Kappa 係数を比較する。各分類モデルの技術的詳細および詳細な実験結果については，本研究組織の既存論文を参照されたい[1]。

④実験概要

前述のような，収集した発言および人手によるコーディングラベルをデータセットとして学習を行い，各モデルにおいて，どの程度コーディングが正しく予測できたかを，比較・検証する。

まず，データの前処理として，MeCab を用いて文の形態素への分割をおこない，頻度の低い単語を 「unknown」と置き換えた。そして，人手によるコーディングによって一致をした 8,015 の発言のみを抽出し，90% を訓練データ，10% をテストデータとした。
ベースラインの手法としては，ナイーブベイズ，線形 SVM，RBF カーネルを用いた SVM を適用した。また，それらの手法に使用する特徴量として，ユニグラムの出現の有無，およびバイグラムの出現有無を{0,1}で表した 2 値ベクトルを用いた。また，SVM における分類精度をあげるために，2 値ベクトルを，ベクトルの L2 ノルムが 1 になるように正規化したのち，上記分類器に入力した。

⑤実験結果

Table 3 に我々が提案したモデルと，ベースラインとなるモデルのテストデータに対する予測精度（一致率）を示す。ここでの一致率は，人手により付与されたラベルとモデルが出力した予測ラベルとが一致する割合である。Table.3 が示すように，全体として，提案モデルの結果はベースラインモデルの結果よりも精度が高くなっていることがわかる。前述の 3 つのモデルのうち，CNN を用いた手法と LSTM を用いた手法の間には，一致率にほとんど差異がないことがわかる(0.67-0.68)。これらの手法は，ベースラインである SVM(0.64-0.66) に比べて僅か(2-3%程度) だが一致率が高くなっている。

Table 3 提案モデルおよびベースラインによる予測精度(一致率)

ナイーブベイズ	線形 SVM	RBF カーネルを用いた SVM	CNN	LSTM	Seq2Seq
0.598	0.659	0.664	0.686	0.678	0.718

一方，全てのモデルの中で，Seq2Seq を用いたモデルが最も一致率が高くなっている(0.718)。SVM と比べて 5-7%，他のモデルと比べても 3-4%高くなっている。

次に，偶然によらない一致率を意味する Kappa 係数を用いて上記の結果を考察する。まず，LSTM を用いたモデルに対する Kappa 係数は 0.63 となり，十分高い結果を得ているといえる。しかし，一般的に，機械による自動コーディングの判別結果を信用に足りる形で利用するためには，Kappa 係数で 0.8 以上が好ましいとされており，より高い一致率が求められる。一方， Seq2Seq を用いたモデルに対する Kappa 係数は 0.723 であり，0.8 には至らないものの，大きく改善されていることがわかる。Seq2Seq は返信元も入力した

モデルであり，各発言をばらばらに捉えるのではなく，文脈の情報を考慮することが精度向上の一因となったと考えられる。

⑥考察

上の実験結果は，Seq2Seq モデルが，文脈情報を考慮したことで他の方法を上回ることを示している。また，今回用いたコーディングスキームが，各発言の文脈上の意味を表現した 16 のラベルからなるものであり，十分に複雑性を有していたことを考慮すると，今回とは異なるスキームにおいても，このモデルを用いる事で，今回と同程度の予測精度を得ることができると思われる。また，解析に要する時間は，十分に短く，リアルタイム処理に耐えられると考えられる。

(3)　検証 2：開発手法の有効性の検証

(2)で提示した Seq2Seq に依拠した手法を用いて，実際のチャットデータを自動コーディングさせ，どのような分析が可能になるのかを考察する。

①チャットデータ

Table 4 に本検証で自動コーディングの対象となるチャットデータの詳細を示す。講義の最終課題はグループ単位で提出する課題であり，「新しい教育テレビ番組を提案せよ」というものだが，「タイトル」「学習課題」「対象者」「番組内容」「工夫点や特徴」を含むこととなっている。また，各グループの提出物は教員により「具体性」「工夫」「適切性」で各 3 段階（良い，普通，悪い）に評価され，その合計から「総合」評価が付けられている。具体性とは，提案内容から番組内容が現実性をもって想像できるかどうか，工夫は手法やコンセプトに独自性があるかどうか，適切性は番組内容と番組対象者との適合性がどの程度あるかを評価した。各評価がつけられたグループ数を Table 5 に示す。

Table 4　チャットデータ

日時	2017 年 7 月 17 日および 24 日
講義名	教育メディア論
課題内容	教育番組の提案
学習時間	合計 2 時間
学生数	138 人
グループ人数	3 人
グループ数	46 グループ
全発言数	2743 発言

Table 5　各評価がつけられたグループ数

	良い	普通	悪い
総合	7	20	19
具体性	10	18	18
工夫	13	19	14
適切性	12	25	9

②　自動コーディング結果

Fig.1 に全 2743 発言を自動コーディングした結果の各タグの割合を示す。学習で利用したラベルの割合と比べると，同意と転換が増えたことがわかる。また，提案，回答は減っている。

Fig.1　自動コーディング結果の割合

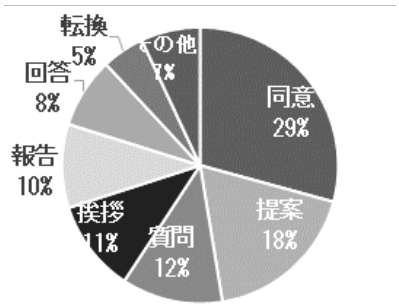

Table 6 提出部評価と平均発言数（ラベル別）

(a) 総合

評価	同意	提案	質問	挨拶	報告	回答	転換	その他	計
良い	20.1	8.7	6.7	6.4	8.0	4.6	3.0	5.4	62.6
普通	16.9	10.2	7.2	6.4	5.8	5.3	2.9	5.2	59.8
悪い	15.8	11.5	6.6	6.0	4.7	4.5	3.3	6.4	58.4

(b) 具体性

評価	同意	提案	質問	挨拶	報告	回答	転換	その他	計
良い	19.6	9.9	7.5	5.7	7.8	5.6	2.6	5.6	64.0
普通	16.6	10.2	6.7	6.8	5.4	4.6	3.2	5.1	58.5
悪い	15.8	11.2	6.7	5.9	4.7	4.7	3.2	6.4	58.3

(c) 工夫

評価	同意	提案	質問	挨拶	報告	回答	転換	その他	計
良い	18.8	9.4	7.7	6.8	7.2	5.4	3.1	6.0	64.1
普通	17.2	12.3	6.8	6.3	5.6	5.5	2.8	6.2	62.5
悪い	14.9	9.1	6.1	5.6	4.4	3.4	3.4	4.9	51.6

(d) 適切性

評価	同意	提案	質問	挨拶	報告	回答	転換	その他	計
良い	17.8	10.6	6.3	5.9	7.8	4.8	2.9	4.9	60.8
普通	15.9	10.2	7.0	6.6	4.8	4.9	3.2	5.8	58.2
悪い	18.7	11.4	7.2	5.6	5.2	4.9	3.0	6.7	62.1

Table 7　提出物評価と発言数との相関係数

	同意	提案	質問	挨拶	報告	回答	転換	全
全体	0.17	-0.16	0.04	0.09	**0.37**	0.05	-0.09	0
具体性	0.16	-0.08	0.08	0.00	**0.37**	0.10	-0.12	0

③　提出物評価と発言内容

Table 6 に各項目の評価ごとに，付与され

たラベルの平均数を示す。また，Table.7 に総合，具体性，工夫，適切性の各評価を良い=3，普通=2，悪い=1 として，各タグの発言数との相関係数を示す。太字の項目が相関係数 0.2 以上の弱い相関のある項目である。この結果から，各評価とも発言数の多さよりも「報告」の数に対し正の相関があり，「報告」が多いほど評価が高いことがわかる。工夫に関しては，グループ内でどれだけ多く会話をしたかが重要であると考えられる。

一方，グループ内での各メンバーの発言数の差が提出物の評価に関係するかどうか比較するために，グループ内の各メンバーのタグごとの発言数の変動係数を求めた。変動係数が高い場合は、そのタグの発言が一人だけが多く発言しているなどグループ内での会話数の差が大きいことを表している。各タグの変動係数と各項目の評価（良い=3，普通=2，悪い=1 として計算）との相関係数を示したものが Table 8 である。太字の項目が相関係数の絶対値が 0.2 以上の弱い相関のある項目である。相関係数の高い項目はすべて負の相関であり，グループ内での会話数の差が大きいと，評価が悪くなることを表している。ここでも「報告」の発言数の偏りと評価には相関があり，「報告」の発言数が偏ると評価が悪くなる傾向があることを示している。また，「適切性」に限って言えば「報告」の偏りは無相関であり，「同意」「提案」に偏りがあると評価が悪くなる傾向があることがわかる。「同意」については，「具体性」にも弱い相関があり，メンバー間で偏りなく「同意」の発言することが良い評価になる傾向があることがわかる。

以上のことから，「報告」と「同意」の発言数や発言数の偏りが，各項目の評価に関係しているといえる。これらのコードが付与された発言が議論にどのように影響しているかを考察する。

Table 9 に「報告」と「同意」のラベルが付与された実際の発言を抜粋する。「報告」の発言は課題の内容自体ではなく，作業の進め方や進行状況の報告など，議論のコーディネーションの成立に寄与している。つまり，報告の発言の多さは，進行状況を相互に把握しながら課題を進めており，非対面で起こりがちなそれぞれが自分のタスクにのみ集中してしまうなどのコミュニケーション不足が回避されていることを示しているといえるだろう。また，「報告」の発言数の偏りは，課題の提出等の課題進行を特定の一人のメンバーが担っていると考えられ，グループ内でのコミュニケーション力が低いことが予測される。

「同意」は他の発言を必ず参照しつつ，肯定する役割を担っている。当初，「提案」や「質問」の数が評価に高い相関を持つと仮定していたが，実際にはグループ内での「同意」の偏りに対し相関が高い。これは「同意」が必ず「提案」や「質問」との対になっているのに対し，「提案」・「質問」は必ずしもそれに対する返答があるとは限らないためと考えられる。つまり，会話が成立しているときに「同意」というタグが付与されたと推測され，それが偏るということはグループ内で，一方向的な会話になっていると考えられる。

Table 8　提出物評価と発言数の偏りとの相関係数

	同意	提案	質問	挨拶	報告	回答	転換	全発言
全体	-0.14	0.02	-0.06	-0.09	**-0.25**	0.12	-0.12	-0.03
具体性	**-0.22**	-0.03	0.07	0.08	**-0.27**	0.14	-0.11	-0.07
工夫	-0.11	-0.05	-0.14	-0.20	**-0.24**	-0.08	-0.17	-0.07
適切性	**-0.29**	**-0.35**	0.14	**-0.22**	0.01	-0.07	-0.09	-0.11

Table 9　報告と同意の内容

報告の例 1	提出しました。 一応確認お願いします。
報告の例 2	いえ、まだ書いてないです。
報告の例 3	僕が今作りますね
同意の例 1	了解です
同意の例 2	よさそうですね。自分はこれでいいと思います
同意の例 3	大丈夫だと思います！

④　考察

開発手法によって，新規の大規模チャットデータに対しても自動コーディングが可能となることが明らかとなった。また，実際の授業実践に向けて，1．リアルタイムな状況把握と教育的介入や 2．学習評価の精緻化の可能性が示唆されたと考えられる。

前者については，議論が停滞しているグループや，グループの中で孤立しているメンバーを検知し，適宜なんらかの支援を行うことが可能となると思われる。例えば，本検証で示されたように「報告」が少なくコーディネーションが不十分なグループに対して，システムから作業分担や作業の現状報告を促す指示を配信し，共同作業を支援するなどが想定される。

後者については，グループ学習終了後に，各グループの議論全体のプロセスを評価したり，グループ内でのラベルの偏りから，一人の意見のみで成り立っているグループや議論には参加していない学生を発見したりすることができる。たとえば，本検証において，メンバー間の発言数が均等で，課題の評価が「良い」であったグループにおいて，ラベル別の発言を見ると，1 名のメンバーに「報告」が偏っているグループが存在した。Table.8 から「報告」が偏っているというこ

とは評価が低い傾向があるとわかる。この場合，グループ内に問題を抱えている可能性が高いといえる。チャット内容を精査すると，報告を多くしていたメンバーが課題を進め，提出物もほとんどその当人が作成していた。このように，提出物や発言数などからではわからない暗箱状態のプロセス評価が，比較的簡易に実施できる可能性が示唆されたと思われる。

(4) まとめ

本研究では，大規模データから協調プロセスを分析するため，深層学習技術を活用することで，きわめて煩雑で非常な時間を要するコーディングの自動化を行った。その結果，本研究で提案した Seq2Seq モデルは，他の方法を上回る結果となった。また，この手法を用いて，現実の授業においてリアルタイムの状況把握と介入および学習評価の精緻化の実現可能性が示唆された。

今後は、複数のデザインの異なった協調スクリプトを用いることで、協調プロセスにどのような差異が生じるのかを実験・分析してくことが課題となる。

参考文献

[1] Shibata, C., Ando, K., Inaba, T., “Towards Automatic Coding of Collaborative Learning Data with Deep Learning Technology”, The Ninth International Conference on Mobile, Hybrid, and On-line Learning, 2017, pp.65-71.

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計３件）

① Chihiro Shibata, Kimihiko Ando, Taketoshi Inaba, ” Coding Collaboration Process Automatically: Coding Methods using deep learning technology”, International Journal on Advances in Intelligent Systems, 査読あり, vol.10, 2017, 345-354
② 安藤公彦，柴田千尋，稲葉竹俊、深層学技術を用いた自動コーディングによる協調学習のプロセスの分析、コンピュータ＆エデュケーション、査読あり、43巻、2017、79-84
③ Taketoshi Inaba, Kimihiko Ando, “Development and Evaluation of CSCL System for Large Classrooms Using Question-Posing Script”, International Journal on Advances of Software, 査読あり, vol.7, 2014, 590-600

〔学会発表〕（計 1 件）

① 安藤公彦，柴田千尋，宮坂秋津，稲葉竹俊、深層学習による協調学習データの自動コーディングに向けて、教育システム情報学会研究会、2017

〔図書〕（計１件）

① 稲葉竹俊、奥正廣、工藤昌宏、鈴木万希枝、村上康二、コロナ社、プロジェクト学習で始めるアクティブラーニング入門、2017、94

６．研究組織

(1)研究代表者

稲葉 竹俊 (INABA, Taketoshi)
東京工科大学・教養学環・教授
研究者番号：10386766

(2)研究分担者

安藤 公彦（ANDO, Kimihiko）
東京工科大学・片柳研究所・助教
研究者番号：00551863

松永 信介 (MATSUNAGA, Shinsuke)
東京工科大学・メディア学部・准教授
研究者番号：60318871