多次元コーディングスキームに依拠した協調学習プロセスの自動コーディングの精度検証

靳展*†、安藤公彦*‡、柴田千尋*‡、稲葉竹俊*‡

＊ 東京工科大学

Evaluation of Automatic Coding for Collaborative Learning Process Based on Multi-Dimensional Coding Scheme

Zhan Jin*†、 Ando Kimihiko*‡、 Chihiro Shibata*‡、 Taketoshi Inaba*‡

* Tokyo University of Technology

† d2113002a2@edu.teu.ac.jp
‡{ ando、 shibatachh、 inaba }@stf.teu.ac.jp

概要: 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メカニズムを分析し、協調プロセスがうまく進行していないグループを識別する指標を抽出し、適切な足場掛けを行う指針を得ることは、きわめて重要な課題といえる。協調プロセス分析のため、会話データへのコーディングと統計的分析が研究方法としてしばしば採用されるが、本研究プロジェクトでは、深層学習技術による高精度のコーディングの自動化の手法を開発し、その精度と有効性を評価してきた。我々の行った先行研究ではスピーチアクトに基づく 16 のラベルで構成されるコーディングスキームに依拠して、教師付データを作成し、深層学習の対象とした。本論では、より包括的に協調プロセスを掌握することをめざして、5 つの次元をもつ多層的なコーディングスキームを新たに構築し、これに基づいて深層学習技術による自動コーディングを行い、その精度を検証することにした。さらに精度検証で使用したデータとは異なるデータセットに対して自動コーディングを行い、その結果の分析を行った。

Abstract: In computer-supported collaborative learning research、 it may be a significantly important task to figure out guidelines for carrying out an appropriate scaffolding by extracting indicators for distinguishing groups with poor progress in collaborative process upon analyzing the mechanism of interactive activation. And for this collaborative process analysis、 coding and statistical analysis are often adopted as a method. But as far as our project is concerned、 we are trying to automate this huge laborious coding work with deep learning technology. In our previous research、 supervised data was prepared for deep learning based on a coding scheme consisting of 16 labels according to speech acts. In this paper、 with a multi-dimensional coding scheme with five dimensions newly designed aiming at analyzing collaborative learning process more comprehensively and multilaterally、 an automatic coding is performed by deep learning methods and its accuracy is verified. In addition、 we apply our methods to predict another dataset for verification and investigate the correlation between the multidimensional coding labels and the assessments given by professionals manually.

キーワード: コンピュータ支援協調学習、協調プロセス、コーディングスキーム、深層学習

Keywords: Computer Supported Collaborative Learning、 Collaborative Process、 Coding Scheme、 Deep Learning

1. はじめに

1.1 協調プロセスの分析

コンピュータ支援協調学習（以下 CSCL）研究の目下の最大の研究課題は、グループ内でどのような知識や意味が共有され、どのような議論によって知識構築が行われたのかを明らかにするため、その社会的プロセス、認知プロセスを仔細に分析することである。また、その知見を活用することで、協調プロセスを活性化したりするような足場掛け機能を有する CSCL システムやツールの開発を行うことである。

しかし、協調プロセスの分析を行うには、単に定量的な分析では全く不十分であり、定性的な分析へのシフトを伴うこととなる。その主な理由としては、分析対象の主たるデータはチャットの発言、Skype 等のツール上での映像と音声、協調学習の過程で作成される様々なアウトプットなどであり、これらの分析のためには単に定量的な分析では全く不十分であることが挙げられる [1]-[4]。

しかし、これらの研究は in-depth なケーススタディとなることが多く、他のコンテキストにおいても適用可能な一般性を有した指針を導出することは、決して容易ではないという弱点を持っている。そのため、一定量のヴォリュームをもった協調学習で生成される言語データの各発言に、言語学的視点や協調学習活動の視点から、その特性を適切に表すラベル付け(以後、コーディングと呼ぶ)を行って、分析を行う verbal analysis の手法を用いる研究が近年行われるようになってきている[5]。この手法の長所はかなり大規模なデータを対象に定性的な視点を維持しつつ、定量的な処理を行える点である。しかし、コーディングを人力で行う事はきわめて時間と労力を要する作業であり、さらにデータがビッグデータになった場合は、人力では不可能になることが予想される。

我々の研究プロジェクトにおいても、2017 年に発表した一連の先行研究において、大量の協調学習データのコーディングの自動化のため、深層学習技術を援用し、一定の成果を上げてきた[6]-[8]。本論文においても昨年度同様に深層学習技術を用いた自動コーディングの精度を検証するが、その際に、新たに、より多面的的かつ包括的に協調学習プロセスを掌握するため、多次元のコーディングスキームを採用し、それに依拠して人力でコーディングを行い、教師付データを構築した。このデータを対象に深層学習を行い、その精度を検証する。

1.2 研究目的

本研究の最終目標は、上に述べたように大規模な協調学習データの解析を行い、リアルタイムでの協調プロセスのモニタリングや活性化していないグループへの足場掛け等の実際の学習、教育の場での支援を実装することである。本論文では、その最終目標実現に向けて、より協調学習プロセスを包括的に表現できる多次元コーディングスキームに依拠した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、その精度の検証を行う。

具体的には、先行研究でも活用した相当量のチャットデータに手動で、新たに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、その精度の評価を行う。さらに、これらの実験データとは別の新たな協調学習時のチャットデータを対象に自動コーディングを行うことで、どのような知見を得ることができるかを検証する。

1.3 本論文の構成

本論文は、まず第2章では、我々の先行研究の概要と結果を示す。第3章では、今回新たに考案したコーディングスキームの概要について述べ、続く第4章において本研究の実験とその結果を示すことにする。5章では、新たな教育データを対象に、3章、4章で示した新コーディングスキームの深層学習を用いた自動コーディングを行い、その結果をもとに学習プロセスの分析を試みた。最後に第 6 章において、結論と今後の展望を提示して、本論文の末尾とする。

表 1 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	はやくも自身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	課題やシステムに対する不満など	テーマがいまいちだよわ‥
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・`)

2. 先行研究

先行研究において、柴田ら[6]は、協調学習分析のためのスキームとして 16 ラベル(表 1)からなるコーディングラベルを提案している。深層学習を用いて学習を行い、比較的高精度に予測精度を達成している。その概要を以下に示す。

2.1 会話データセット

研究対象となった会話データセットは著者らが独自に開発した 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[9]。ちなみに、今回の研究でもこの同じデータセットを用いることにする。この発言データ元の CSCL の利用状況を表2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。

表 2 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

2.2 コーディングスキーム

著者らが作成したコード付与のためのマニュアルに従い、すべての発言について 2 名のコーダーがそれぞれコーディングを行い、チャットの 1 発言に対し 1 つのラベルを付与した。これらのコードの一致または不一致の結果を著者らで精査したところ、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、一部コードの再コーディングを行った。ラベルは前述のように表 1 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した。

図 1 にデータセットのラベルの割合を示す。

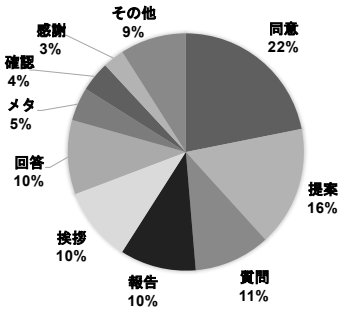

図 1 コーディングラベルの分布

2.3 深層学習を用いた自動コーディング手法

先行研究では、深層学習手法として、(1)畳み込みニューラルネット(CNN)に基づくもの、(2)長短期記憶(LSTM)に基づくもの、(3)Sequence to Sequence (Seq2Seq) に基づくものの合計 3 つのアーキテクチャを適用し、比較を行っている。このうち、Seq2Seq ベースのアーキテクチャ 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[10][11]。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。

なお、以降では、混乱を避けるため、用語として、手法としての深層学習のネットワークの構造のことを「アーキテクチャ」、実際にデータから学習させた結果得られたネットワークやパラメータを「学習済みモデル」あるいは単に「モデル」と呼ぶことにする。実験では、古典的な機械学習の手法である SVM による学習済みモデルをベースラインとして用いた。各モデルの精度の検証は、自動コーディングの一致率、および偶然によらない一致率を意味する Kappa 係数を比較する。各アーキテクチャの技術的詳細および学習パラメータなどの詳細や実験結果については、著者達の既存論文を参照されたい[6]-[8]。

2.4 実験と評価

2.4.1 実験の概要

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・検証した。

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。また、深層学習以外の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、比較を行った。

2.4.2 実験結果

表 3 に先行研究にて提案したモデルと、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。表 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。

表 3 提案モデルおよびベースラインによる予測精度

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、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。

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。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。

最後に、どのような場合に誤分類が起きるかを、各コーディングラベルごとに分析する。LSTM を用いたモデルに対する、各ラベルの精度(precision)と再現率(recall) および F 値を 表 4 に示す。「挨拶(Greeting)」、「同意 (Agreement)」および「質問(Question)」に対する F 値が最も高いことがわかる(それぞれ 0.94、 0.83、0.77)。これらの結果は、発言の外形から文意を深く捉えなくても容易に判断できるケースが多いため、人間の感覚にも一致しているといえる。それに対して、「メタ(Outside comments)」が最も F 値が低い(0.25)。これは、意見交換すべき内容とは全く関係のない、冗談などを意図した発言が該当するが、それを判断するためには文意を深く捉える必要があるためと考えられる。また、「回答(Reply)」でも F 値が低い (0.53)。Seq2Seq を用いたモデルにおいても、「回答(Reply)」でも F 値は若干改善するものの、依然として低いことがわかっており、混同行列(図 2) を見ても、「同意(Agreement)」や「提案(Proposal)」、「報告(Report)」などへ誤分類されていることがわかる。「回答」は「質問」に対応するものであることがほとんどであること、および「質問」の F 値は高いことから、今回用いた手法では、「ソース」と「リプライ」の発言ペアの抽出が、不十分となっていると結論できる。

表 4 各ラベルの精度と再現率(LSTM)

	Precision	Recall	F1-value
Agreement	0.85	0.81	0.83
Proposal	0.73	0.74	0.73
Question	0.75	0.80	0.77
Report	0.64	0.62	0.63
Greeting	0.94	0.94	0.94
Reply	0.62	0.46	0.53
Outside comments	0.17	0.47	0.25
Confirmation	0.58	0.74	0.65
Gratitude	0.67	0.67	0.67

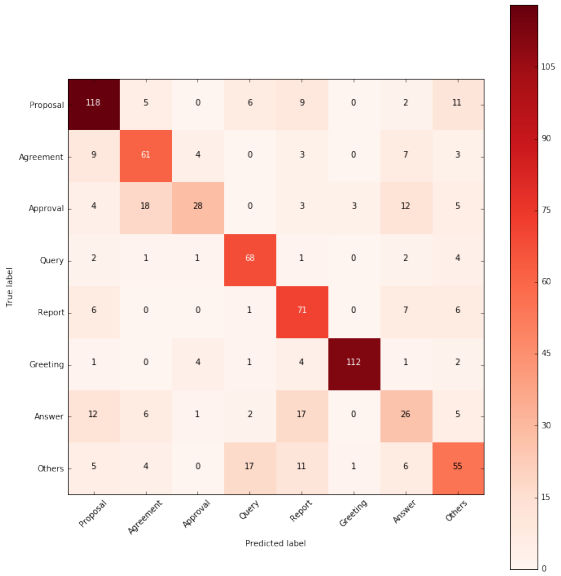

図 2 Seq2Seq モデルの混同行列

3. 新コーディングスキーム

先行研究で用いたスピーチアクトに基づくコードは、Replay が Agree の意味を含むことがあるように、1 つのコードが他のコードを内包することから、人工知能のみならず、人力によるコーディングにおいても、判断が難しくなる要因となっている。また、本コードは、すべてのコードが 1 つの次元で定義されているため、異なる次元にあるべきコードが同じ次元として定義されており協調学習を多元的に分析することが難しい。

これら技術的問題に加えてさらに重要なのは、スピーチアクトに依拠した言語学的特徴にのみ着目したスキームでは、協調学習のプロセスを包括的に表現するには不十分な点である。この一次元のスキームでは、グループの各人が問題解決にどの程度関与しているのか、どのような分業や時間管理が行われたのか、どのような議論の展開があったのか、メンバー間でどのような意見交換や意見のすり合わせがあったのかといった協調プロセスの本質に関わる問題に答えることは極めて困難である。

以上のことから、本研究では自動コーディング精度のさらなる向上と、CSCL 分析に有効な新たなコードを定義し、その付与方法を含め新コーディングスキームとして提案する。

提案する新コードは、Weinberger らが示した多次元のコードを用いるフレームワークを参考にし、本システムに適応させたものである[12]。表 5 に示すように、新コーディングは 5 つの次元からなり、基本的に先行研究と同様にチャットでの発言単位でコードが付与される。また、Participation 次元は発言数などの数値等がコードとして付与されるが、それ以外の 4 つの次元のコードは複数のラベルから 1 つが選択され付与される。以下、各次元について詳細に述べる。

表 5 新コーディングスキーム

次元	内容
Participation	議論への参加度合い
Epistemic	課題解決への直接的な関わり方
Argument	議論における主張のあり方
Coordination	他者の発言との関わり方
Social	議論を円滑に進めるための調整の仕方

3.1 Participation 次元

Participation 次元は、議論への参加度合いを測る次元である。この次元は、主に発言数や発言文字数、発言時間、発言間隔など、定量的なデータとして定義されるため、人力や人工知能によるコーディングは必要なく、データベース上の統計処理でのコーディングとなる。表 6 にその一覧を示す。

表 6 Participation 次元

要素	意味	例
発言数	セッション中の各メンバーの発言数	59 回
発言文字数	1 会話中の文字数	15 文字
発言時間	発言した時間	2017/8/21 15:15:01
発言間隔	前に発言したときからの時間	3:01.05

Participation 次元のコードは具体的な発言数などを扱うため、どれだけ会話に参加したかなど積極性を分析できるが、その会話が課題の解決に寄与したかなどの質的な分析はできない。

3.2 Epistemic 次元

Epistemic 次元のコードは、各発言がタスクである課題の解決に直接関係しているかを表し、発言内容により表 7 のように分類される。この次元のコードはすべての発言に付与される。

表 7 Epistemic 次元のコード

要素名(ラベル)	意味
On Task	課題に直接関係のある発言
Off Task	課題に関係のない発言
No Sense	内容が意味不明の発言

ここで、「On Task」は、課された課題の解決に直接関係のある発言であり、下記に示す内容の発言は「Off Task」となる。

- 課題の意味や進め方を問う発言
- タスクを割り振る発言
- システムに対する発言

Epistemic 次元のコードは課題の解決に直接かかわっているかどうかを表すため、質的な分析の最も基本的なコードとなる。例えば「On task」のコードが少ない場合、課題への取り組みはほとんどなされていないか、なされていたとしても課題に対する質的な深い議論は行なわれていないと考えられる。

なお、Argument 次元および Social 次元のコードは Epistemic 次元が「On Task」のときのみ付与され、Coordination 次元のコードは Epistemic 次元が「Off Task」のときのみ付与される。

3.3 Coordination 次元

Coordination 次元のコードは、Epistemic のコードが「Off Task」のときにのみ付与され、課題の解決に直接は関わらないが、間接的に関わる発言の場合付与される。表 8 に Coordination 次元のコードの一覧を示すが、「Off task」の発言全てにコードが付与されるのではなく、これらコードに当てはまるときにのみ 1 つが付与される。また、Coordination 次元のコードが付与された発言に対する応答は、同じ Coordination 次元のコードが付与される。

ここで、「Task division」は、課題を進めるうえで分業が必要であり、誰がどのタスクを行うかを決める発言である。「Time management」は課題の進行具合を調整する発言であり、「〇〇時までに調べましょう」や「進捗はどうですか？」などの場合があてはまる。「Technical coordination」は CSCL のシステムの使い方に対する質問や実行した操作の報告などである。「Proceedings」は議事進行であり、どの様に課題を進めるかなどの発言である。

表 8 Coordination 次元のコード

要素名(ラベル)	意味
Task Division	タスクの分配
Time Management	時間進行、進行具合の確認
Technical Coordination	システムの使い方等
Proceedings	議事進行

Coordination 次元のコードは、課題の解決をスムーズに行うための発言に対して付与されるため、どのようなタイミングで付与されいてるかを分析することによって、議論の進行具合が予測できると考えれる。また、Coordination 次元のコードが少ない場合は、グループ内での円滑な人間関係の構築ができていないとも予想できる。

一方、これらのコードが多くのグループで多数付与された場合、課題内容やシステム等に何らかの不具合があると推測できる。

3.4 Argument 次元

Argument 次元のコードは、Epistemic のコードが「On Task」のときの発言全てに付与され、各発言に発言者の意見があるかどうか、そしてその意見に根拠があるかなどの属性を示す。この次元のコードは 1 つの発言内容のみを対象とし、他の発言で根拠を述べたかどうかは考慮しない。

Argument 次元のコード一覧を表 9 に示す。ここで、根拠の有無は、意見の元となる根拠が示されているかどうかであり、提示された根拠の信頼性は問わない。また、限定条件とは、提示された意見がタスクとして扱うすべての状態にあてはまると主張しているのか、それとも一部にのみあてはまると主張しているのかを表す。例えば「～の場合は」や「～と比べて」などの文節が含まれている場合が当てはまる。「Non-Argumentative moves」は、意見を含まない発言であり、単純な質問の場合もこのラベルに含まれる。

表 9 Argument 次元のコード

要素名(ラベル)	意味
Simple Claim	根拠のない単なる意見
Qualified Claim	根拠なく、限定条件のある意見
Grounded Claim	根拠をもった意見
Grounded and Qualified Claim	限定付きかつ根拠をもった意見
Non-Argumentative Moves	意見のない発言。(質問も含む)

Argument 次元のコードは発言内容の高度さを分析できる。例えば、「Simple Claim」ばかりであればそれは表層的な議論だと推測できる。

3.5 Social 次元

Social 次元のコードは、Epistemic のコードが「On task」のときに付与されるが、「On task」の発言全てではなく、Epistemic のコードに一致したときのみ付与される。この次元は各発言がグループ内の他メンバーの発言にどのように関わっているかを表す。よって、1 つの発言だけでなく、それまでの文脈も読み取る必要がある。表 10 にこの次元のコードの一覧を示す。

ここで、「Externalization」は他者への発言の参照がない発言であり、主に議論のトピックの開始時など議論の起点となるべき発言に付与される。「Elicitation」は質問など他者へ情報の引き出し要求をする発言に付与される。

「consensus building」は他者の発言を受けて何らかの意見を述べる発言であり、その方向性から下記の 3 つのコードに分類される。「Quick consensus building」は他者の意見などに早急な合意を目指すための発言に付与される。特に意見などなく賛成する場合に付与される。「Integration-oriented consensus building」は自身の意見も追加しながら、他者の意見への合意を目指す発言に付与される。「Conflict-oriented consensus building」は、他者の意見への対立や改変を求める発言に付与される。

Social 次元には「Refer」というサブ次元があり、どの発言を参照しているかを表し、通常参照した発言の発言番号等がコードとして付与される。「Refer」次元のコードは Social 次元のコードが「consensus building」のときのみ必ず付与される。

表 10 Social 次元のコード

要素名(ラベル)	意味
Externalization	外化:他者の発言への参照がない
Elicitation	情報の引き出し
Quick consensus building	早急な合意形成
Integration-oriented consensus building	統合を目指す合意形成
Conflict-oriented consensus building	対立を目指す合意形成
Summary	他の発言をまとめた発言

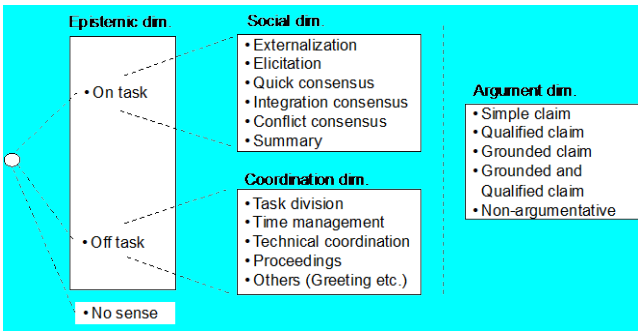

図 3 各コーディング次元間の関係性

4. 実験と結果

先行研究として 2.3 節で述べた手法のうち、最も精度の高かった Seq2Seq ベースのアーキテクチャを用いて、新しい次元の学習を行った。各次元ごとに、別々のデータを用意し、合計で 4 回独立した学習を行い、4つの別々の学習済みモデルを作成した。データの大きさとしては、それぞれ、Epistemic 次元に対して 8、460 個、Augment 次元に対して 7、795 個、Coordination

次元に対して 3、510 個、 Social 次元に対して 2、619 個の発言が、モデルの学習のためのデータとして用いられた。

4.1 各次元の人手によるコーディング結果

まず、前述のように、深層学習を用いて学習させるため、各発言に対して人手によりコーディングを行った。すべての発言に対して、2 名のコーダーがそれぞれにラベルを付与し、また2つのラベルが一致したデータが正解データとして利用した。以下に全発言を各次元の各ラベルの割合を示す。これらのグラフは、機械学習の観点から見ると、正解データにしめるラベルの割合を示していると言える。

Epistemic 次元(図 4)においては、「On Task」と「Off Task」 の割合はほぼ拮抗しており、一般的にいって、二値分類の典型的なタスクであり、機械学習を用いて比較的予測しやすいと考えられる。

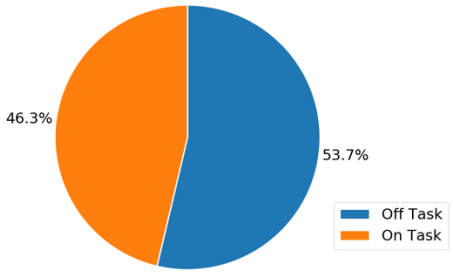

図 4　Epistemic 次元の各ラベルの割合

Argument 次元では、何らかの主張が含まれている発言以外のもの「Non-Argumentative」と「Simple Claim」 が合計で 95% 以上を占めていることがわかる(図 5)。したがって、機械学習の観点から、一般的に言って、上記2つについては、比較的分類しやすいと考えられるが、残りの Claim については、データ数の問題から、十分に学習できないと考えられる。

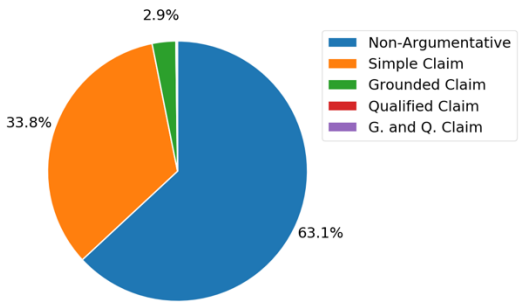

図 5　Argument 次元の各ラベルの割合

Coordination 次元についてもほぼ同様のことが観測できる(図 6)。議論のコーディネートとは関係のない発言 「Other」が全体の約 3/4 を占めており、「Technical Coordination」と「Proceedings」 がそれに続いている。

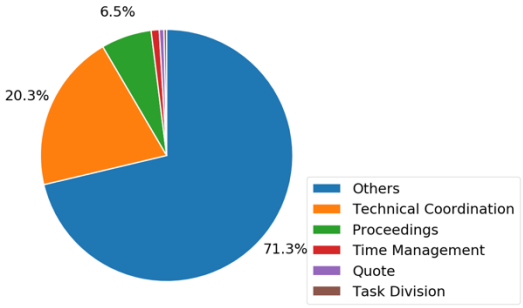

図 6　Coordination 次元の割合

また、Social 次元(図 7)については、上記 2 つの次元と比較して、やや割合のバランスが良いと言え、主なラベルについては、一般的に言って、機械学習により一定の精度を持って予測することが可能であると考えられる。

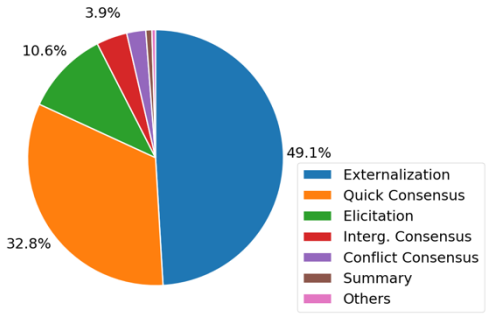

図 7　Social 次元の割合

4.2 各次元の深層学習による予測精度

次に、学習させた DNN モデルを用いて、テストデータに適用し、各次元のラベルを予測させた。

表 11 Epistemic 次元の精度と再現率

	Precision	Recall	F1-value	Support
On Task	0.90	0.91	0.90	390
Off Task	0.92	0.91	0.91	456
Average(Micro) / Total	0.91	0.91	0.91	846

表 11 に、Epistemic 次元に対する実験結果を示す。表からわかるように、On Task と Off Task の両方において、十分に高い精度を得ることができている。精度と再現率が全て 90%を超えている。一方で、正解データより、2名の人間コーダー間の一致率を計算したところ、91%であった。したがって、Seq2seq モデルによる自動コーディングは。人間に匹敵する精度を得ることができていると言える。

表 12 Argument 次元の精度と再現率

	Precision	Recall	F1-value	Support
Non-Argumentative	0.87	0.97	0.92	491
Simple Claim	0.89	0.72	0.80	264
Grounded claim	0.58	0.52	0.55	21
Qualified Claim	0.00	0.00	0.00	1
Average(Micro) / Total	0.87	0.87	0.87	777

表 12 に示すように、Argument 次元に関しても高い精度が得られた。各ラベルのデータ数は異なるため、モデルの分類能力は、全発言数を母数とした平均値(micro 平均)で評価する必要がある。F 値の micro 平均(micro-f1)は 87%であり、特に「Non-Argumentative」の F 値(92%)は十分高かった。概して、この次元に対しても提案した DNN モデルを用いると正しく分類することができると言える。しかし、「Simple Claim」の分類精度は高い(89%)が、再現率(72%)が非常に低い。分類結果の詳細を調べると、「Simple Claim」データの 1/4 が「Non-Argumentative」に誤分類されていた。これは、小さい意見を含むデータと意見なしデータの区別が難しいためである。

表 13 Coordination 次元の精度と再現率

	Precision	Recall	F1-value	Support
Others	0.91	0.91	0.91	242
Technical Coordination	0.81	0.80	0.81	82
Proceedings	0.58	0.70	0.64	20
Time Management	0.33	0.25	0.29	4
Quote	0.00	0.00	0.00	1
Task Division	0.00	0.00	0.00	2
Average(Micro) / Total	0.85	0.86	0.85	351

提案手法によって Coordination 次元が高い精度を達成した。表 13 に示すように、micro-f1 は 85%であった。「Others」と「Technical Coordination」の精度は高かったが、「Time Management」と「Proceedings」の精度が非常に低い。これはデータが少ないため、正確に学習することが非常に困難である。スパースラベルに対応方法は今後の課題の一つとして検討する必要がある。

表 14 Social 次元の精度と再現率

	Precision	Recall	F1-value	Support
Externalization	0.86	0.61	0.72	127
Quick	0.71	0.93	0.81	88
Elicitation	0.56	0.97	0.71	29
Interg. Consensus	0.17	0.14	0.15	7
Conflict Consensus	0.00	0.00	0.00	6
Summary	0.00	0.00	0.00	3
Others	0.00	0.00	0.00	2
Average(Micro) / Total	0.75	0.72	0.70	262

他の次元と比べて、Social 次元の精度が比較的に低かった。Social 次元のラベルを分類するとき、会話の背景と深い意味を理解する必要があるため、少ないデータで学習することが困難と考えられる。「Externalization」の精度は高い(86%)が、再現率(61%)が非常に低い。詳しい分類データによると、一部「Externalization」データは「Elicitation」と「Quick」に誤分類されてしまった。今後結果を改善するために、この原因を追求する必要がある。

5. 手法の検証

3 章で提案した新コーディングスキームに対して、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。

表 15 チャットデータ

日時	2017 年 7 月 17 日及び 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

表 16 各評価が付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

5.1 チャットデータ

表 15 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実現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表 16 に示す。

5.2 自動コーディング結果

4 章の実験で得られた 4 つの学習済みモデルを利用し、全 2743 発言に対して自動コーディングの処理を行った。表 17 から表 20 までに 4 つの次元の各ラベルの発言数を示す。

表 17 Epistemic 次元の自動コーディングの結果

ラベル	発言数
On Task	1633
Off Task	1110

表 18 Argument 次元の自動コーディングの結果

ラベル	発言数
Simple Claim	1082
Non-argumentative moves	1638
Grounded Claim	23
Grounded and Qualified claim	0
Qualified Claim	0

表 19 Coordination 次元の自動コーディングの結果

ラベル	発言数
Others	2368
Technical coordination	360
Proceedings	15
Time management	0
Task division	0

表 20 Social 次元の自動コーディングの結果

ラベル	発言数
Externalization	2170
Elicitation	152
Quick consensus building	421
Integration-oriented consensus building	0
Conflict-oriented consensus building	0
Summary	0
Others	0

5.3 提出物評価と発言内容

表 21 から表 24 までに各次元の評価ごとに、付与されたラベルの平均数を示す。また、表 25 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、各ラベルの発言数との相関係数を示す。太文字の項目が相関係数の絶対値 0.2 以上の弱い相関のある項目である。この結果から、Epistemic 次元の「On Task」、 Argument 次元の「Non-argumentative」、Coordination 次元の「Others」と「Technical Coordination」、 Social 次元の「Externalization」に関して、発言数の多さより正の相関があり、この5つラベルが多いほど工夫の評価が高いことがわかる。また、工夫の評価に関しては、全体的に発言が多いほうが良い評価となる傾向がある。工夫に関しては、グループ内でどれだけ多く会話がしたかが重要であると考えられる。総合、具体性と適切性に関しては、Social 次元の「Elicitation」と負の相関があり、発言数が多いほど評価が低いことがわかる。

表 21 提出物評価と平均発言数(Epistemic)

(a)総合			
評価	On Task	Off Task	合計
良い	38.7	22.7	61.4
普通	35.4	23.2	58.6
悪い	33.7	24.7	58.4
(b)具体性			
評価	On Task	Off Task	合計
良い	40.6	22.6	63.2
普通	33.5	23.7	57.2
悪い	33.9	24.4	58.3
(c)工夫			
評価	On Task	Off Task	合計
良い	39.6	23.8	63.5
普通	35.9	25.3	61.2
悪い	30.1	21.5	51.6
(d)適切性			
評価	On Task	Off Task	合計
良い	36.9	21.2	58.2
普通	33.2	24.9	58.2
悪い	38.3	23.8	62.1

表 22 提出物評価と平均発言数(Argument)

(a)総合				
評価	Non-argumentative	Simple Claim	Grounded Claim	合計
良い	33.7	27.6	0.1	61.4
普通	35	22.9	0.6	58.6
悪い	35.8	22.2	0.4	58.4
(b)具体性				
評価	Non-argumentative	Simple Claim	Grounded Claim	合計
良い	34.4	28.3	0.5	63.2
普通	34.9	21.8	0.5	57.2
悪い	35.8	22.1	0.4	58.3
(c)工夫				
評価	Non-	Simple	Grounded	合計
	argumentative	Claim	Claim	
良い	38.1	24.9	0.5	63.5
普通	35.9	24.7	0.6	61.2
悪い	31.4	19.9	0.4	51.6
(d)適切性				
評価	Non-argumentative	Simple Claim	Grounded Claim	合計
良い	31.4	26.4	0.3	58.2
普通	36.5	21.1	0.6	58.2
悪い	36.3	25.4	0.3	62.1

表 23 提出物評価と平均発言数(Coordination)

(a)総合				
評価	Others	Technical Coordination	Proceedings	合計
良い	53.7	7.4	0.3	61.4
普通	50.4	7.8	0.3	58.6
悪い	50.5	7.5	0.4	58.4
(b)具体性				
評価	Others	Technical Coordination	Proceedings	合計
良い	54.1	8.8	0.3	63.2
普通	49.9	6.9	0.3	57.2
悪い	50.3	7.7	0.3	58.3
(c)工夫				
評価	Others	Technical Coordination	Proceedings	合計
良い	54.3	8.8	0.3	63.5
普通	52.7	8.2	0.3	61.2
悪い	45.5	5.7	0.4	51.6
(d)適切性				
評価	Others	Technical Coordination	Proceedings	合計
良い	50.2	7.8	0.2	58.2
普通	49.8	7.9	0.5	58.2
悪い	55.2	6.8	0.1	62.1

表 24 提出物評価と平均発言数(Social)

(a)総合				
評価	Quick Consensus	Externalization	Elicitation	合計
良い	8	51.4	2	61.4
普通	9.4	46.4	2.8	58.6
悪い	9.3	44.9	4.3	58.4
(b)具体性				
評価	Quick Consensus	Externalization	Elicitation	合計
良い	8.4	52.4	2.4	63.2
普通	9.2	45.1	2.9	57.2
悪い	9.4	44.7	4.2	58.3
(c)工夫				
評価	Quick Consensus	Externalization	Elicitation	合計
良い	8.8	51.8	2.8	63.5
普通	9.2	48.8	3.2	61.2
悪い	9.3	38.6	3.8	51.6
(d)適切性				
評価	Quick Consensus	Externalization	Elicitation	合計
良い	7.5	48.7	2	58.2
普通	10.1	44.3	3.8	58.2
悪い	8.6	49.9	3.7	62.1

表 25 提出物評価と発言数との相関係数

次元	ラベル	総合	具体性	工夫	適切性
Epistemic	On Task	0.11	0.15	**0.24**	−0.02
	Off Task	−0.09	−0.08	0.11	−0.11
Argument	Non−argumentative	−0.05	−0.04	**0.20**	−0.14
	Simple Claim	0.14	0.18	0.17	0.05
	Grounded Claim	−0.05	0.03	0.05	−0.02
Coordination	Others	0.05	0.07	**0.20**	−0.09
	Technical Coordination	0	0.06	**0.26**	0.06
	Proceedings	−0.06	−0.02	−0.09	0
Social	Quick Consensus	−0.08	−0.1	−0.04	−0.12
	Externalization	0.11	0.14	**0.27**	−0.01
	Elicitation	**−0.37**	**−0.31**	−0.16	**−0.26**
全発言		0.04	0.07	**0.22**	−0.06

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ラベルごとの発言数の変動係数を求めた。変動係数が高いとそのラベルの発言が一人だけ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。各ラベルの変動係数と各項目の評価との相関係数を示したものが表 26 である。太文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が全て負の相関であり、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、「Quick Consensus」の発言数の偏りと評価には相関があり、「Quick Consensus」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。

表 26 提出物評価と発言数の偏りとの相関係数

次元	ラベル	総合	具体性	工夫	適切性
Epistemic	On Task	−0.06	−0.07	−0.07	−0.01
	Off Task	0.05	0.05	−0.12	0.09
Argument	Non−argumentative	0	0.07	−0.10	0.10
	Simple Claim	0.14	0.09	0.06	0.14
	Grounded Claim	0.01	0.02	0.10	0.03
Coordination	Others	0.18	0.12	0.06	0.18
	Technical Coordination	0.10	0.07	−0.04	0.3
	Proceedings	−0.05	0.01	−0.05	0.30
Social	Quick Consensus	**−0.31**	**−0.28**	**−0.28**	**−0.26**
	Externalization	0.18	0.13	0.19	0.10
	Elicitation	0.05	0.01	−0.07	0.17
全発言		0.14	0.08	0.05	0.14

表 27 に Social 次元の「Elicitation」と「Quick Consensus」が付与された実際のが発言を抜粋する。「Elicitation」は質問によって他者の発言を求める発言であり、発言の内容によると、学生が課題の内容や進め方などに関して混乱してしまうと考えられる。「Quick Consensus」は他者の意見などに早急な合意を目指す

ための発言であり、発言の内容によると、グループ内の会話が盛り上がらないと予測される。

表 27 「Elicitation」と「Quick Consensus」の内容

「Elicitation」の例 1	結局どれですか？
「Elicitation」の例 2	内容はどうしますか？
「Elicitation」の例 3	提出しますか？
「Elicitation」の例 4	どうしますか？
「Quick Consensus」の例 1	それはいいですね
「Quick Consensus」の例 2	そうですね
「Quick Consensus」の例 3	タイトル難しいですね。。。
「Quick Consensus」の例 4	じゃあこれで決定ですね。

5.4 考察

提案手法によって、新たなチャットデータに対しても多次元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、教員により各グループの評価に基づいて、グループ内の問題点を早めに発見し、解決策を探索することが可能となった。

6. 結論と展望

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、5 つの次元をもつ多層的なコーディングスキームを提案した。また、各次元について自動コーディングの精度を評価した。この手法を用いて、現実の授業において学習評価の精緻化の実現可能性を検証した。

今後の研究方向性として、2つのアプローチを追求していく。一つは Social 次元に対して各ラベルの分類精度の向上が重要な課題となる。例えば、MemNN など Seq2seq 以外の DNN モデルを導入し、その分類結果を本手法と比較する価値がある[13]。二つ目は、提案したコーディングスキームの構造について改善することが必要と考えられる。現状には、提案した新たしい次元が各自のスキームを持っている。複数のスキームを統一し、異なる次元とその次元に属するラベルを階層的な構造で管理することにより、各次元間の関係情報を上手く利用できる新たなスキームを構築することを目指す[14]。

参考文献

[1] Stahl、 G.、 Koschmann、 T.、 & Suthers、 D. (2014). Computer-supported collaborative learning. In R. K. Sawyer (Ed.)、 *The Cambridge Handbook of the Learning Sciences*、 Cambridge: Cambridge University Press、 479-500.

[2] Dillenbourg、 P.、 Baker、 M.、 Blaye、 A.、 & O'Malley、 C. (1996). The evolution of research on collaborative learning. *Learning in Humans and Machine: Towards an interdisciplinary learning science*、 P. Reimann and H. Spada、 Eds. Oxford: Elservier、 189-211.

[3] Koschmann、 T. (2011). Understanding understanding in action. *Journal of Pragmatics*、 43(2)、 435-437.

[4] Koschmann、 T.、 Stahl、 G.、 & Zemel、 A. (2007). The video analyst's manifesto (or The implications of Garfinkel's policies for the development of a program of video analysis research within the learning science). *Video Research in the Learning Sciences*、 R. Goldman、 R. Pea、 B. Barron and S. Derry、 Eds. Routledge、 133-144.

[5] Chi、 M. T. H. (1997). Quantifying qualitative analyses of verbal data : a practical guide. *Journal of the Learning Science*、 6(3)、 271-315.

[6] Shibata、 C.、 Ando、 K.、 & Inaba、 T. (2017). Towards automatic coding of collaborative learning data with deep learning technology. *The Ninth International Conference on Mobile*、 *Hybrid*、 *and On-line Learning*、 65-71.

[7] 安藤公彦、 柴田千尋、 稲葉竹俊. (2017).「深層学習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*、 43、 79-84.

[8] 安藤公彦、 柴田千尋、 宮坂秋津、 稲葉竹俊. (2017).「深層学習技術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 2017 年度第 2 回研究会*.

[9] Inaba、 T.、 & Ando、 K. (2014). Development and evaluation of CSCL system for large classrooms using question-posing script. *International Journal on Advances in Software*. 7(3&4)、 590-600.

[10] Bahdanau、 D.、 Cho、 K.、 & Bengio、 Y. (2014). Neural machine translation by jointly learning to align and translate. *CoRR*、 arXiv preprint arXiv、 1409.0473.

[11] Vinyals、 O.、 & Le、 Q.V. (2015). A Neural Conversational Mode. *CoRR*、 arXiv preprint arXiv、 1506.05869.

[12] Weinberger、 A.、 & Fischer、 F. (2006). A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computer & Education*、 46(1)、 71-95.

[13] Sukhbaatar、 S.、 Szlam、 A.、 Weston、 J.、 & Fergus、 R. (2015). End-to-end memory networks. Proceedings of the 28th International Conference on *Neural Information Processing Systems*、 2440-2448.

[14] Scafino、 F.、 Pio、 G.、 Ceci、 M.、 & Moro、 D. (2015). Hierarchical multidimensional classification of web documents with multiwebclass. International Conference on *Discovery Science*、 236-250.