平成30年 7 月

ディープラーニング技術による教育ビッグデータの分析・可視化手法の
開発・評価　プロジェクト
共同プロジェクト（平成 27 年度～30 年度）

研究最終報告書

東京工科大学教養学環
（研究代表者　稲葉竹俊）

目次

第1章

全体報告

第1章 全体報告

1．1 プロジェクト活動のサマリー

1．1．1 プロジェクトの目的と成果

本プロジェクトの目的は、教養学環における取組課題であるアクティブラーニング授業の質向上のためのオンライン学習環境をクラウド上、とりわけ、本学の学習管理システムMoodle 上に構築し、八王子、蒲田両キャンパスにおけるオフライン（物理世界）での学習活動や教育活動を様々な形で支援し、その質と精度を高めることにある。そのため、本学クラウドサービスセンターとの協力体制のもと、Moodle から取得される多様な教育ビックデータを取得・蓄積し、これを最新の機械学習技術のディープラーニング技術を用いて解析し、学習状況を自動的にオンタイムで、多角的に評価する手法を開発しその有効性を評価することをめざした。また、その結果を授業担当の教員や学生に対して提示するに際しては、ウィークリーサマリーや図表などの表現形式で明示的に視覚化を行うことを目指した。これによって、教員と学生はリアルタイムでグループや個人の学習状況を掌握し、アクティブラーニングにおいて必須の教授活動であるタイムリーな学生への介入や振り返りによる学習者自身の学習活動の軌道修正が可能となると考えられる。図１がプロジェクトで構築を目指したシステムの全体像である。

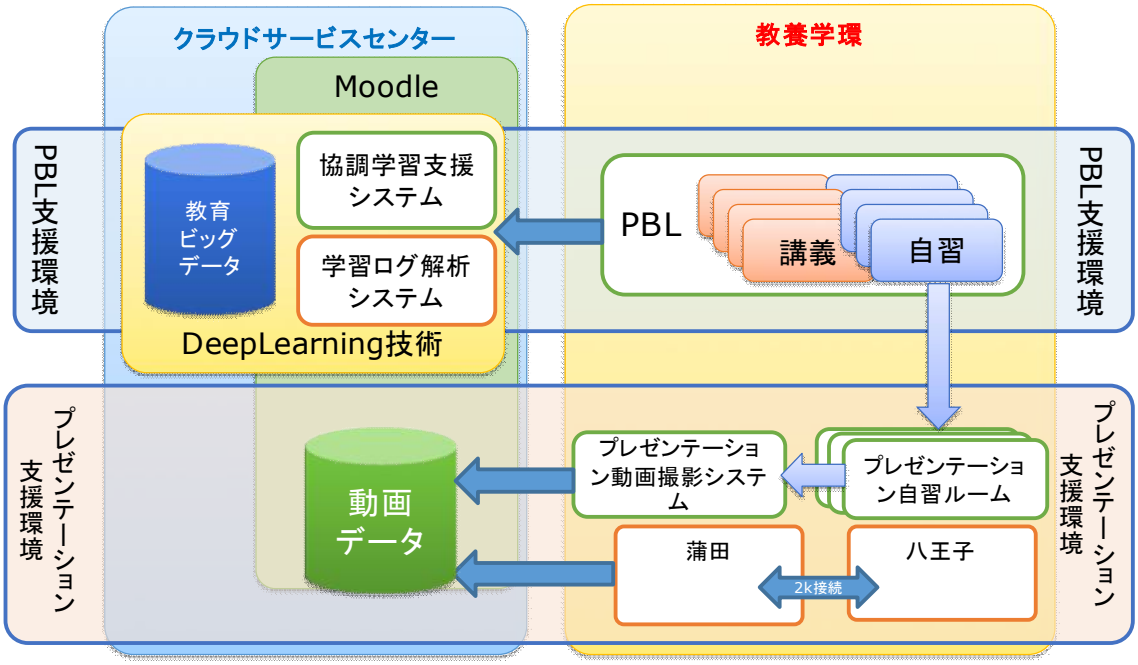


図１　プロジェクト構築システムの全体像

上の方針に従って、本プロジェクトで実際に対象としたのは以下のような教育データであり、それぞれのデータに依拠して、この報告書の 2 章以降でその詳細が記される研究成果を出すことができた。

1. 学習データ

キャリアデザイン等の PBL 授業や人文社会科目における反転授業の中核となるオンライン上でのグループ内での議論内容を示すチャットログや反転授業における予備学習のログなどである。チャットログについては、稲葉、安藤らが開発した協調学習支援システム（以下 CSCL と記す）を Moodle 上で運用しデータを取得する。研究成果については、2 章、3 章、4 章において詳述する。

2. 動画データ

PBL 等のプレゼンの撮影動画、Alice 等で行わる英語によるプレゼンや会話の撮影動画等である。プレゼン練習・指導を支援するため、学生達だけで操作可能なプレゼン録画アプリケーションを開発する。この録画データはシームレスに Moodle 上に自動的にアップされ、教員や学生が閲覧できるものとする。また、図 1 にもあるように、八王子キャンパスと蒲田キャンパス間を高品質画像の送受信可能な会議システムで結び、リアルタイムでの授業や議論を可能とした。研究成果については、2 章、5 章において詳述する。

1. 2 プロジェクトの実施概要

1. 2. 1 主要なメンバーと役割

プロジェクトに関連した教員と役割を表 1 に示す。

表 1　プロジェクト参加メンバー

		所属	役割
研究代表者	・稲葉竹俊	教養学環	全体統括、CSCL設計、ログ分析スキームの構築
共同研究者	・安藤公彦	クラウドサービスセンター	Moodleモジュールの開発・プレゼン支援環境構築
	・柴田千尋	CS学部	ディープラーニング技術の導入
	・高橋潔	教養学環	Moodleのログ抽出
	・豊田ひろ子	教養学環	ALICE学習環境整備
	・R.キャンベル	教養学環	ALICE学習環境整備
	・石塚美佳	教養学環	ALICE学習環境整備
	・村上康二郎	教養学環	反転授業実験と評価
	・松永信介	MS学部	反転授業実験と評価

1．2．2 活動期間

2015 年 9 月～2018 年 3 月

1．2．3 学術的成果

このプロジェクトに関係する成果の件数は以下のとおりである。

・学会誌学術論文（査読付き）2 件
・国際会議論文（査読付き） 2 件
・学会発表・予稿集採録 3 件
・国際会議・学会論文賞 3 件
・著作 1 件
・現在投稿中論文 2 件

なお、以上の成果の詳細リストについては本章の末尾に掲載してあるので、そちらを参照されたい。

1．2．4 産学上の成果

1．2．4．1 CEATEC2017 での成果発表

プロジェクトでの成果の公開と産学連携のチャンスを得るために本プロジェクトとクラウドサービスセンターとの合同で CEATEC2107 に以下のような要領で出展を行った。幸い連日、ブースには多くの来訪者があり、また、東京工科大学の OB、OG の訪問も予想以上に頻繁であった。これらの卒業生は CEATEC を見学中に偶然東京工科大学の名前を発見し、声をかけてくれたものである。

日程：2017 年 10 月 3 日（火）～10 月 6 日（金）
場所：幕張メッセ CEATEC JAPAN 2017 会場内
『ディープラーニング・対話・まなびプロジェクト」（教養学環学内共同プロジェクト）
展示エリア：「社会・街エリア」
ブース名：東京工科大学
小間番号：C029

また、この出展に合わせる形で、プロジェクトの活動の対外的なアピールのため Web ページを公開した。また、出展用のプロジェクト紹介パンフレット、プロジェクトの内容を要約したムービー等を作成した。

展示内容等については第 6 章で述べることとする。

1．2．4．2 産学共同研究

CEATEC の展示に関心を持ってくれて何らかのやり取りがあった企業は相当数あったが、そこから 3 社との産学共同研究が始まることとなった。2017 年 12 月から八王子に本社のあるムラウチドットコムとの共同研究が開始され、2018 年 5 月から約 1 年間の予定で FXCoin 株式会社との共同研究、6 月から約半年の予定でビズオーシャン株式会社との共同

研究が開始された。

これらの共同研究は CEATEC での出展がなければ成立しなかったものであり、今回の展示の具体的な成果であるといえる。

これらの産学共同研究については、第 7 章で具体的な取り組み課題等について紹介することにする。

1．2．5 教育上の成果

1．2．5．1 プロジェクトで構築したシステムの活用

本プロジェクトで構築した PBL 向け協調学習用 Moodle モジュールは教養学環が八王子の 3 学部（BS, CS、MS 学部）の 2 年生の学生に対して設計・運用をしているキャリアデザイン I および II での利用を想定して設計を行った。利用は全教員に義務化しているものではなく、現状では希望する教員のみが利用しているが、利用している教員からは PBL の効率的な授業運営に大いに役立つツールとして好評を得ている。また、今後はキャリアデザインのみならず、他の講義で行われる反転授業においても十分活用が可能なものである。このシステムについては 2 章で詳しく述べる。反転授業での協調学習システムの利用については、4 章でその具体例を示す。

1．2．5．2 2 つのキャンパスをリアルタイムで結ぶ会議システム

本プロジェクトの担当学部である教養学環は、蒲田キャンパスと八王子キャンパスのいずれかに本務をおく教員を成員としている唯一の教員組織であり、距離的に離れた 2 キャンパス間で教授会、アゴラ、様々な委員会や WG を開催する必要がある。これらの意思疎通が円滑にいく ICT インフラの整備が教育上も学務上も必須の課題であった。また、英会話の学習プログラムである ALICE も両キャンパスで展開するという企画もあり、本プロジェクトでは、北米の AVAYA 社の会議システム、SCOPIA XT4300 と大型スクリーンの 4 K テレビを両キャンパスの会議室に設置を行った。このシステムを活用した ALICE での実験授業については第 5 章で詳しく述べる。

1．3 残された課題

1.1.1 の図 1 で示した研究の全体像に依拠すれば、本プロジェクトで想定していた Moodle 上での環境構築のシステムの根幹部分については、PBL 支援環境、プレゼンテーション支援環境のいずれにおいても設計と構築は何とか無事に終えることができたといって過言ではない。ただ、このいずれの支援環境においても今後さらに取り組む必要のあるタスクが残っており、これらが残された課題である。

1．3．1 PBL 支援環境での課題

第 2 章、3 章で示すように、PBL で用いる協調学習支援システム本体およびそこから取得されるチャットデータの深層学習による解析システムはすでに Moodle 上で稼働している。しかし、深層学習の解析結果から、各グループや各学習者の活動の進捗状況を把握し、

これを教員や学生に例えば図 2 のようにダイアグラム等によって可視化するモジュールが未完成である。これについては、稲葉、柴田、安藤で現在取り組んでいる科研費の基盤研究等の枠内で完成を急ぐことにしたい。

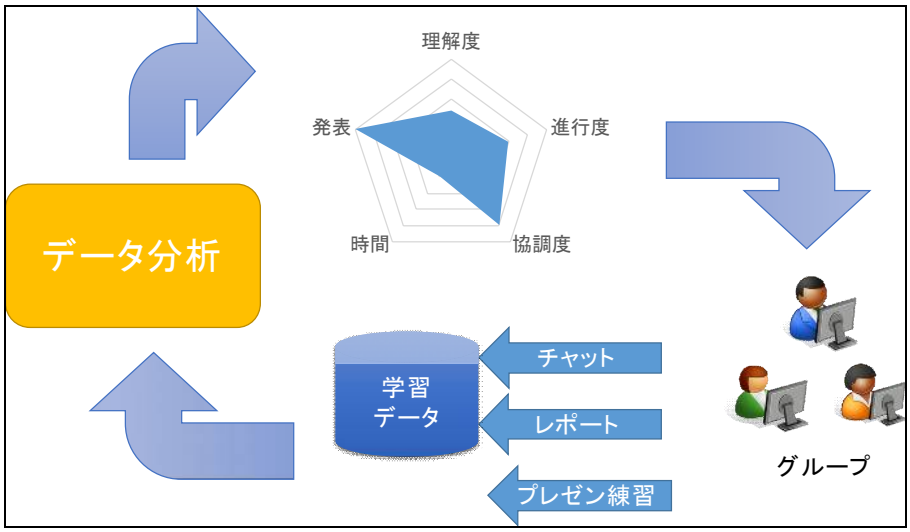


図 2 学習状況の可視化の流れの一例

また、反転学習の事前学習等で有効なデータとなる Moodle 上に残される各学習者の膨大な学習ログから教員が必要とするログだけを取り出して可視化するという学習ログ解析システムも未完成である。これについては、クラウドセンター主導で学内への導入が行われることを希望する。

1. 3. 2 プレゼンテーション支援環境での課題

2 章で示したように、Moodle 上ですでにこのシステムは稼働可能になっている。しかし、構築がようやく 2018 年 3 月に終了したことから、授業や演習での実運用にはまだ至っていない。今後は教員や学生に対して利用を積極的に呼び掛けていきたいと考えているが、まずは運用の試運転として、2018 年のキャリアデザインにおいて有志教員の利用から始めていく予定である。

１．４ 本報告書の執筆責任者

本報告書の各章の執筆者および協力者は以下のとおりである。

1 章：稲葉　竹俊

2 章：安藤　公彦

3 章：稲葉　竹俊、柴田　千尋

4 章：松永　信介　（協力者：村上　康二郎）

5 章：豊田　ひろ子　（協力者：R.キャンベル、石塚美佳）

6 章：稲葉　竹俊

7 章：稲葉　竹俊

１．５ 研究成果の詳細

１．５．１ 学会誌学術論文（査読付き）

・安藤公彦、柴田千尋、稲葉竹俊： 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、vol.43, pp.79-84、2017.

・Kimihiko Ando, JapanChihiro Shibata,Taketoshi Inaba, ”Coding Collaboration Process Automatically: Coding Methods Using Deep Learning Technology”,
The International Journal on Advances in Intelligent Systems, 10(3&4), pp.345-354, 2017.

１．５．２ 国際会議論文（査読付き）

・JapanChihiro Shibata, Kimihiko Ando, Taketoshi Inaba, “Towards Automatic Coding of Collaborative Learning Data with Deep Learning Technology”, in proceedings of eLmL 2017, The Ninth International Conference on Mobile, Hybirid and On-line Learning, in Nice, France, March 19-23, 2017.

・Takahiro Kanayama, Kimihiko Ando, Chihiro Shibata, Taketoshi Inaba, "Using Deep Learning Methods to Automate Collaborative Learning Process Coding Based on Multi-Dimensional Coding Scheme."in proceedings of eLmL 2018, The Tenth International Conference on Mobile, Hybrid, and On-line Learning, in Rome, Italy, March 25 - 29, 2018.

１．５．３ 学会発表・予稿集採録

・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 研究報告集、A1-1、pp.1-8, 2017.

・松村 佳記、村上 康二郎、安藤 公彦、稲葉 竹俊、松永 信介：法学の授業における反転学習とコンピュータ支援協調学習の事例研究、情報処理学会、第 80 回全国大会、2018.

・松永 信介・安藤 公彦・稲葉 竹俊：Moodle 環境を活用した反転学習用 CSCL システムの開発、

第 17 回科学技術フォーラム、2018.

１．５．４ 国際会議・学会論文賞

・Best Paper Award (The Ninth International Conference on Mobile, Hybrid, and On-line Learning), 2017.
受賞者：Kimihiko Ando, JapanChihiro Shibata,Taketoshi Inaba

・Best Paper Award (The Tenth International Conference on Mobile, Hybrid, and On-line Learning), 2018.
受賞者： Chihiro Shibata, Kimihiko Ando, Taketoshi Inaba
・2018 年度 CIEC 学会賞 論文賞（コンピュータ利用教育学会), 2018.
受賞者：安藤公彦、柴田千尋、稲葉竹俊

１．５．５ 著作

・稲葉竹俊、奥正廣、 工藤 昌宏 , 鈴木万希枝, 村上 康二郎：プロジェクト学習で始めるアクティブラーニング入門、コロナ社、2017.

第2章

プロジェクトで構築した学習支援システム概要

第 2 章　プロジェクトで構築した学習支援システム

2．1 はじめに

PBL（Project Based Learning）や反転学習では、その評価はグループとしての提出物や発表により行われるためグループ単位での評価となり、グループ内の各学生個別の評価が難しいという課題がある。学生間による相互評価を行うことである程度のグループ内評価は可能であるものの同一の基準に従った公平な評価は難しい。そこでこれらの協調学習にLMS（Learning Management System）を導入することで、従来教員から見えにくかったグループ内の活動状況の把握や適切な評価を可能とすることを目的とし、協調学習支援システムの構築とその実践を行う。

協調学習支援システムとしては、「グループ生成」「グループとしての課題提出」「グループに対する教員のフィードバック」などのグループ学習支援機能と、各グループ内の学習が問題なく進んでいるか教員が把握できるグループ学習評価機能が必要となる。

また、PBL や反転学習においてはグループ発表が必須となることが多いが、発表練習を学生自身が録画し課題として提出することで、円滑な講義進行と細やかな指導が可能となる。しかし、動画の課題提出は LMS の保存領域の圧迫やストリーミング再生のためのエンコードなどシステム面での負荷が大きい。そこで、GoogleDrive を活用した課題提出システムの開発も合わせて行う。

以降、本章では、2.2 項で開発した PBL 向け強調学習支援システムについて詳細に述べ、2.3 項では開発した GoogleDrive 連携課題提出システムについて詳しく述べる。

2．2 PBL 向け協調学習用 Moodle モジュール

2．2．1 目的

本学では LMS として Moodle を導入しているが、システムによる協調学習に必要となる「グループ生成」「グループとしての課題提出」「グループに対する教員のフィードバック」や「グループ学習評価機能」を包括的に利用できる機能は Moodle にはなく、新たに開発する必要がある。また「グループ学習評価機能」では、本プロジェクトの目的である教育ビックデータの可視化を可能とするための機能に重点を置き開発を行う。

2．2．2 モジュール概要

本モジュールは Moodle 内にインストールすることで、活動「PBL」として追加することができる。このモジュールでは Moodle コースに参加している学生を自動的に読み込み、参加している学生のみをグループに参加させることができる。またこのモジュール内に課題やチャット等の機能があり、もともと Moodle に備わっている「課題」や「チャット」機能とは独立している。これにより、本モジュールに特化した課題やチャットが可能となって

いる。また、このモジュール内にて、週の概念があり、本モジュールの活動を 1 つ追加するのみで、15 週の講義が可能となる。

2. 2. 3 グループ学習支援機能

本モジュールのグループ学習支援機能は下記のとおりである。

A) グループ編成機能

設定したグループ人数に合わせて自動的にグループの生成を行う

B) スケジュール機能

設定した日程に合わせて、ファイルや課題を配布する機能

C) 共有ファイル機能

学生がグループ内でファイルを共有する機能

D) 課題機能

スケジュールで設定できる課題であり、グループか個人かなど細かな設定ができる。

E) チャット機能

グループ内で学生がチャットをするための機能。教員が書き込むこともできる。

F) メッセージ機能

教員が学生単位やグループ単位でメッセージを送る機能

2. 2. 4 グループ学習評価機能

本モジュールでの学習評価機能は、大きく 2 つに別れ、1 つは課題提出状況の把握であり、もう 1 つはチャット状況の把握である。

課題提出については、各グループや個人ごとに提出の有無を一覧表示することができる。チャット状況の把握については、グループ間での発言数の差や、グループ内での発言数の差を偏差値を用いて一覧表示することができる。

また先進的なチャット評価機能として人工知能による質的評価を示すことができるがこれについては、次章「大規模なチャットデータの分析」で詳しく述べる。

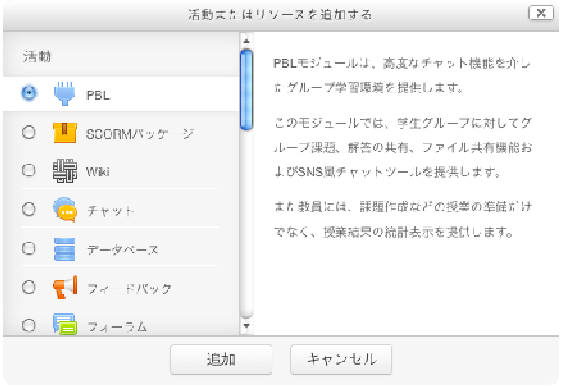
2．3 PBL 向け協調学習用 Moodle モジュール利用方法

2．3．1 活動の追加

Moodle の他のモジュール設定と同じように「編集モードの開始」ボタンをクリックして、編集モードに移行する。

「活動またはリソースを追加する」のリンクをクリックする。複数週の授業のために活動を設定しますので、「週」ではなく「トピック」などの日時に影響されない項目に活動を追加することをお勧めする。

リストの中から「PBL」を選択して、「追加」ボタンをクリックする。



2．3．2 PBL の初期設定

活動が追加されると設定画面が表示される。

「PBL の名称」の項目に『キャリアデザイン I』などの授業の名称を入力する。この名称が学生の画面に表示される。

「授業期間」の設定は 1 週から 16 週の間で指定できる。

「グループ人数」の設定は 2 人から 6 人の間で指定できる。この人数は、グループ自動編成時に利用するための値である。設定した値と異なる人数のグループを生成することもできる。

「キャリアデザイン向けに初期化する」のチェックを入れると、上記の「授業期間」と「グループ人数」の設定を無視して「15 週」「4 人グループ」として強制的に設定を行う。また、この後で説明する「スケジュールの設定」も一部自動で行う。

「チャット解析 API URL」の項目には、なにも入力しない。この機能は、特別な実験のためにのみ用いる。

ここで設定を行った内容に従って授業の雛形を準備する。なお、この設定は後から修正することが可能である。

2．3．3 画面の見方

(A)教員画面

設定が完了すると授業管理画面が表示さる。この画面には、教員権限を持つユーザのみ

がアクセスできる。学生が閲覧することはできない。

画面は、大きく分けて 5 つの要素から構成されている。

「スケジュール」

各週の名称や課題の設定などを行うことができる。ここで設定した内容が学生画面に表示される。

「Weekly Report」

週ごとの各グループの活動内容概略が表示される。チャット機能の利用頻度を偏差値で評価しているので、グループの活動具合を簡単に把握することができる。また、グループ名をクリックするとグループ内での個人の活動概要が表示される。

表示期間は、全期間または 1 週間ごととなる。ここでの週は、授業スケジュールの週とは一致しない。アクセス時間を起点とした 7 日分のデータが表示される。

「学生宛メッセージ」

受講者全員に対してメッセージを一斉送信することができる。このメッセージは、PBL モジュール内での通知であり、メールではない。

このメッセージには既読確認機能が搭載されおり、まだメッセージを閲覧していない学

生をリストアップすることができる。
　個人宛や個別グループ宛のメッセージは次項「グループ」から送信することができる。

「グループ」
　グループ編成機能と編成済みグループ管理機能が表示される。
　グループが編成されていない場合は、「グループ自動編成」と「グループ手動編成」のボタンが表示される。グループに編成されていない学生がいる場合は、人数と警告が表示される。
　すでにグループが編成されている場合、編成済みグループ一覧が表示される。ここで表示されたグループ名をクリックすることで、各グループの活動内容詳細を閲覧することができる。

「学生画面プレビュー」
　指定した学生の現在の学生画面をプレビューすることができる。ターゲットに「匿名ユーザ」を指定することで仮想の学生画面を表示することもできる。「匿名ユーザ」はグループ編成が終わっていない段階で学生画面を確認する際に使用する。

(B)学生画面
　学生が閲覧する画面であり、教員画面とほぼ同じ構成の画面が表示される。教員は「学生画面プレビュー」でのみ閲覧することができる。

「通知バッジ」

　未読のチャット、閲覧していない共有ファイル、未読のメッセージがあると「！」で通知する。

「スケジュール」
　教員画面での設定に従って、課題や配布物が表示される。

「チャット」

SNS ライクなチャット機能であり、必ず別ウインドウで起動される。チャットでは会話だけではなく、ファイル共有も行え、アップロードしたファイルは、自動的に共有ファイルに保存される。

「共有ファイル」

課題の下書きや参照すべきファイルを共有するためのファイル置き場であり、アップロードされたファイルは投稿者のみ削除することができる。

「教員からのメッセージ」

教員が送信したメッセージがすべて表示され、返信を行うことはできない。

「メンバーステータス」

グループに所属するメンバーのログイン履歴を表示する。

2. 3. 4 スケジュールの設定

画面の構成が確認できたら、まずはスケジュールの設定を行う必要がある。

設定したい週をクリックしてするとその週の設定項目が表示される。

「名称」は、その週のタイトルとして学生画面に表示され、フォームから抜けると自動的に保存される。

「概要」は、その週の説明となり、省略してもよく、フォームから抜けると自動的に保存

される。
「課題一覧」は、その週の課題がすべて表示される。課題を追加する際は、「＋課題の追加」をクリックする。また、登録済みの課題名をクリックすることで課題を編集・削除することができる。課題が存在しない週の場合は、設定する必要はない。
「配布物一覧」は、その週の配布物がすべて表示される。配布物を追加する際は、「＋配布物の追加」をクリックする。また、登録済みの配布物名をクリックすることで配布物を編集・削除することができる。配布物が存在しない週の場合は、設定する必要はない。
課題が設定された週には、赤いピンのアイコンが表示される。
配布物が設定された週には、クリップのアイコンが表示される。

(a) 課題の追加

課題の追加をクリックするとダイアログが表示される。
課題名、出題形式、締め切りを設定して追加ボタンをクリックする。
課題の提出は、回答ファイルをアップロードする形式になり、テキスト入力による回答はできない。
出題形式は、グループ課題と個人課題の 2 種類が選択でき、グループ課題は、グループの 1 名が提出すれば全員が提出した扱いになる。
締め切りは、設定した日付の 23 時 59 分 59 秒となる。締め切り後も課題提出を受け付けるが、期限後提出の印が表示される。

(b) 配布物の追加
配布物の追加をクリックするとダイアログが表示される。
配布ファイルを選択して追加ボタンをクリックし、タイトルを入力すると学生画面で表示される名称を変更することができる。タイトルを入力しても元のファイル名のまま配布される。

・課題の編集
課題名をクリックすると追加時と同様のダイアログが表示され、内容の編集、削除が行える。一度削除した課題を元に戻すことはできない。

・配布物の編集

配布物名をクリックすると追加時と同様のダイアログが表示され、タイトルの編集、配布物の削除が行える。学生の混乱を避けるため配布ファイルの差し替えはできない。一度削除した配布物を元に戻すことはできない。

2. 3. 5 学生の登録

学生が PBL モジュールを利用するためには、Moodle の受講登録を行う必要があるが、学生の登録は、いつでも行うことができ、受講者が増えた場合にも柔軟に対応することが可能である。

2. 3. 6 グループの生成

受講者の登録が済んだらグループを生成する必要がある。グループが編成されていない場合、2 つの編成方法が提示される。

(a)グループ自動編成

初めに設定したグループ人数を超えない範囲でグループを自動的に生成する。原則的に学籍番号順にグループメンバーを選択する。

(b)グループ手動編成

学生が所属するグループを手動で選択してくことができる。

グループの作成やメンバーの変更は、あとからでも行うことができます。

2. 3. 7 学生の画面を確認する

スケジュール設定とグループ編成が完了したら、学生画面を確認するために、「ターゲット」に表示したい学生を選択して「プレビュー」ボタンをクリックする。学生が見ることになる PBL モジュールの画面が表示され、教員用管理画面とほぼ同一の内容が表示されていることが確認できる。

学生画面プレビューは、あくまでプレビューであり、誤操作を防ぐため多くの機能が実行できないようになっている。

2. 3. 8 授業を進行する

実際に授業が始まると管理画面から、さまざまな活動報告にアクセスすることができる。

(a) 課題提出状況の確認

週の課題一覧にある「提出状況の確認」をクリックすると、出題形式に従って、グループまたは個人の課題提出状況が一覧で表示される。

(b) 活動報告の確認

「Weekly Report」には、各グループの活動内容の概要が一覧表示される。ここには、過去 7 日分の活動内容が表示されており、表の下部にある期間変更のリンクをクリックすることで全期間や前週の活動内容報告に切り替えることができる。

グループ名をクリックすることで、そのグループ内での各個人の活動内容報告が表示され、表示される期間は、現在表示中の Weely Report の期間と同じものになる。

発言数偏差値やファイルアップロード数などを確認して、活動活性が低いグループを簡単に見つけることが可能である。

(c)個別グループの確認

「グループ」に表示されているグループの名前をクリックすると各グループの活動詳細を確認することができる。

ここでは、課題の提出履歴、グループチャットの発言内容、共有ファイルの利用状況、学生向けメッセージの送信、ログイン履歴の確認が提供される。

課題の項目にある課題名をクリックすると提出履歴がすべて確認できる。

また、チャットと共有ファイルには、閲覧確認機能が搭載されており、チャットの発言横にあるチェックマークをクリックすると発言を確認した学生の氏名が表示される。ファイルのリストにある人のアイコンをクリックすると投稿者と閲覧者の氏名が表示される

また、グループ詳細ではグループ全体での活動と個人の活動を個別に表示することができる。

(d) メッセージの送信

PBL モジュール内で学生にメッセージを送ることができる。このメッセージは、モジュール利用時にのみ閲覧できる簡易的な通知機能である。メッセージに対する学生からの返信機能や学生から教員へのメッセージ機能はない。

学生の混乱を避けるため送信したメッセージの内容変更はできなくしている。送信内容に不備などがあった場合は、送信済みメッセージを削除してから再送信をする。。

送信範囲は全受講者、グループ、個人の 3 つが用意されている。

(e) グループの修正

追加の受講生が現れた場合やグループの変更があった場合、手動でグループの修正を行う必要がある。その場合は、「グループ」項目の下部にある「グループ手動修正」ボタンをクリックして行える。ただし、授業を開始してからグループを変更すると変更した学生のチャットや提出物は削除される。

2. 4 Google Drive 連携課題 Moodle モジュール

2. 4. 1 目的

Moodle にて課題として動画を提出させる場合、Moodle のストレージ容量や配信のためのエンコードなど、システムへの負担が非常に大きくなりその実現は難しい。しかし本学では学生全員に Google アカウントが付与されており、各学生は GoogleDrive として大きなストレージ容量が割り当てられている。また、GoogleDrive の場合、動画のエンコードも自動的に行われストリーミング再生も容易である。これらのことから、Moodle から各個人の GoogleDrive へ課題を提出および提出・閲覧履歴を管理する仕組みを構築することで、動画の提出に対応可能となる。

2. 4. 2 開発状況

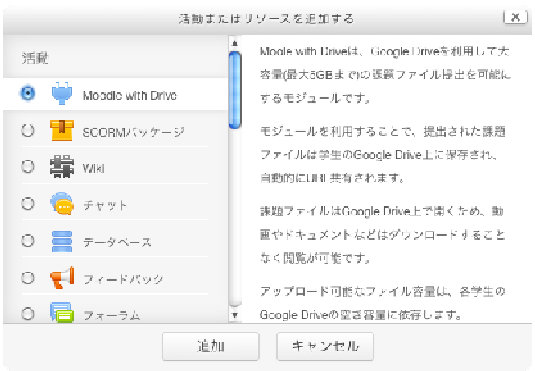
機能のほとんどは開発完了しており、すでに Moodle へとインストールが行われている。今後は実際の講義で運用し、細かな利用方法の調整をするのみである。ある程度のテスト運用を行った後全学に向けて、提供予定となっている。

また、本モジュールを使うことで、課題のアップロードおよびダウンロードは GoogleDrive で行うこととなり Moodle サーバーとの通信量を大幅に減らすことが可能となる。これによって Moodle サーバーを学外のクラウドに移行を視野にいれることができる。

2. 4. 3 利用方法

「小テスト」などの Moodle の標準モジュールと同じように編集モードに移行して活動を追加する。表示される名称は「Moodle with Drive」である。

活動を利用するには、「名称」と「提出期限」を入力する必要があり、表示モードは「提出」のまま変更する必要はない。



2．4．4 教員画面

授業ステータスには「提出」または「講評」のどちらのモードになっているかが表示される。

その下の表には学生の提出状況が一覧表示される。「学生用提出画面を確認する」から学生がどんな画面を見ているのか確認できる。教員もここから提出を試せるが、内容は学生に開示されない。

2．4．5 学生画面

コメント入力欄と提出ボタンから構成されています。初めて提出を行う際は、利用確認のため Google の認証画面が表示されます。ログインを行って連携を許可してください。

提出コメントに入力すること手学生は提出ファイルにコメントをつけて提出することができる

教員は、学生の提出ファイルに対して簡単なフィードバックコメントを送信することができ、フィードバックコメントは、対象の学生のみが閲覧することができる。

2．4．6 表示モードの切り替え

Moodle with Drive モジュールには、「提出」と「講評」の 2 つのモードが用意されており、授業の進行に合わせて任意にモードを切り替えることができる。提出モードは課題提出のためのモードで、提出されたファイルは、提出者本人と教員のみが確認できる。講評モードは講評向けのモードで、受講生全員がお互いの提出物を確認することができる。

第3章

大規模なチャットデータの分析

第3章　大規模なチャットデータの分析

3．1 はじめに

本章では前章でその仕様の詳細を述べた Moodle 上で稼働している協調学習支援システムの基幹をなすチャットシステムから取得されるチャットデータの分析手法について述べる。本研究ではこの手法の開発において、機械学習の先端技術である深層学習技術を用いている点とデータの分類ラベルにコーディング手法を用いている点に大きな特徴がある。

3．2 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の一つは，グループ内でのどのような知識や意味が共有され，どのような意見の対立や同調や調整があり，どのような議論によって知識構築が行われたのか，その社会的プロセスを社会構成主義的な観点から分析することである．また，その知見を活用することで，より有効な足場掛けの方法を提案したり，協調プロセスを活性化するような CSCL システムやツールの開発を行うことである．

CSCL の初期の研究においては，協調するグループ内の各個人に焦点をおいて，グループのどのような特性（グループサイズ，グループ構成，学習課題，コミュニケーションメディアなど）が個人の学習成果に有意に関与するかが主要な関心となっていた．しかし，これらの特性はお互いに複雑に関係し関連し合っているのであり，ある結果に関して因果関係を示すことはきわめて困難であることが次第に明らかになった．90 年代からは，CSCL 研究の関心は，グループ内の個人の学習がどのように成立するかという問題意識から離れ，ある学習がグループで生起している場合に，そのプロセスをグループの相互作用の仔細を明らかにすることで説明しようと試みるようになる[(1)]．

しかし，協調プロセスの分析を試みることは，単に研究の視点のシフトにとどまらず，その分析の方法の根本的な見直しを余儀なくされることになる．つまり，定量的な分析から定性的な分析へのシフトを伴うこととなる．もちろん CSCL システムに保存される定量的なデータにもグループ内の発言（contributions）数やグループメンバごとの発言数，また場合によってはシステムのインターフェース（ sentence opener）から取得される発言属性等の利用可能なデータがあるが，これらはきわめて表面的なデータにすぎない．分析のための最も重要なデータはチャットの発言，スカイプ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会話分析，ビデオ分析などのエスノメソドロジーが援用されてきた[(2)(3)]．

しかし，これらの研究はその性質上，限られた数のグループの協調活動を対象とした in-depth なケーススタディとなることが多く，ある程度一般性を持ち，他のコンテキストにおいても適用可能な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を言語学

的視点や協調学習活動の視点からコーディングを行って，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[4][5][6]．この手法の長所はかなり大規模な協調学習の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．既存研究においても，協調学習データのコーディング支援を試みたシステムは存在している．これらの研究では，コーディング入力自体は人力によって行われるものと[7][8]，機械学習の技術を用いて行ったものがある[9][10]．これに対して，本研究では深層学習技術によって，大規模な協調学習を対象にコーディングを自動化する手法を模索する．

3．2．1　教育データと Learning Analytics

教育機関で教育クラウドの導入が進展することで，LMS，e ラーニング，SNS,MOOC などにおいて生成されるデータが急速に増加しており，これらの教育ビッグデータを解析し学習活動や教育活動の支援につながるような知見を得ようとする Learning Analytics といわれる新しい研究アプローチが活性化している．学内の教育クラウドに統合された CSCL システムから取得される会話データや提出物，学習活動中の画像や音声も早晩，ビッグデータとして分析対象になることは確実であり，LA としての協調学習研究という新しい可能性を真摯に検討しなければならない時期になっていると思われる．このような背景から，本研究プロジェクトでは 5 年前から学内サーバで稼働していた CSCL システムを学内クラウド上の LMS である Moodle 内のモジュールとして再構築し，学内で運用し，協調学習データの収集，分析が可能な環境を構築した．

3．2．2 研究目的

本研究の最終目的は，上に述べたように大規模な協調学習データを LA の視点で解析を行い，今までのミクロレベルでのケーススタディでは得られなかった協調活動プロセスの活性化や非活性化のメカニズムを明らかにすること，さらに，その成果を踏まえて，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．本論文はその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の開発とその精度の検証を行うことにする．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する．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 Naïve Bayes や Support Vector Machines との精度比較を行うことで深層学習を用いたことの有効性を評価する．

3．3 データとコーディングスキーム

3．3．1 会話データ

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[11]%25 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 CSCL は非対面で用いるものであり，本データはすべて大学の大教室内で離れた面識のない学生同士間でグループを組んだ際のものとなる．またシステム上での学生の名前はニックネームとなっており，知人であったとしてもそれを知ることはできないようになっている．

本研究で利用する発言データ元の CSCL の利用状況を表 1 に示す．表 1 に示されているのはこの研究で利用する発言データの元となった科目のみで，実際にはさらに多くの講義で利用されている．科目数は 7 科目であり，どの科目でも 3-4 人のグループを組んでいる．時間は科目により異なり 45～90 分となっている．研究対象となったデータセットは合計で 11504 発言に及んでいる．すべての科目のグループの合計は 202 グループ，参加学生は 426 人となっているが，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 1 発言データの概要

科目数	7科目
グループ	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

表 2 に実際のチャットの会話例を示す．これは 3 名による会話例となる．

表 2　会話例

発言者	内容
D	どの辺を変えますか？
E	そこですよね…まず問題文は絶対かえなきゃだとは思うんですけど、推論式の方はどうしますかね
D	問題文の三行目だけを変えるのはどうですか？
D	推論式でいうと最後のつ以降です
E	それでいいと思います
F	良いと思います。どう変えます？

3．3．2 コーディングスキーマ

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，チャットの 1 発言に対し 1 つのコードを付与する．コードは表 3 に示す 16 種類となっており，このコード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著者

らで精査したところ，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，著者ら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 名のコーダ間の一致率は 82.3%で Kappa 係数は 0.800 という高い一致率となり，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．図 1 はデータセットのラベルの分布を示したものである．これを見ると，9 つのラベルが全体の 90 パーセント以上を占めており，それ以外のラベルがロングテールを形成していることがわかる．

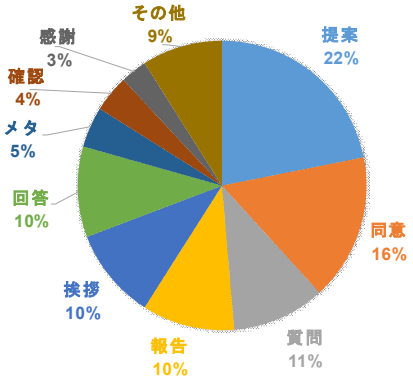


図 1 コーディングラベルの分布

表３　ラベルの分布

タグ	タグの意味	発言例
同意 了承	肯定的な返答	いいと思います
提案 意見	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
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！

タグ	タグの意味	発言例
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・`)
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,拒否	否定的な返答	30 分は長すぎる気がします
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???

3．4 深層学習を用いた自動コーディング手法

前節で述べたような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．主に，近年の

データ量の巨大化，および GPGPU に代表される並列計算技術の進展により，そのような大規模なニューラルネットを現実に訓練させることが可能になり，画像認識など限定的なタスクおいてはしばしば人間の認識率を上回ることが知られている

本節では，よく知られているネットワーク構造を持つ深層学習の手法をいくつか適用し，自動コーディングの正解率，F 値，および κ 係数を比較する．具体的には，(1)畳み込みニューラルネット(CNN)による分類モデル，(2)長短期記憶(LSTM)による分類モデル，(3)Sequence to Sequence (Seq2Seq) と呼ばれる手法に基づく分類モデル，の 3 つを比較する． また，古典的な機械学習の手法である SVM をベースラインとして用い，各種の深層学習の手法の適用により，個々のコーディングラベルについて，どの程度改善されるかについて述べる．

3．4．1 各手法における共通点

比較を行った深層学習を用いた 3 つの手法の詳細を述べる前に，各手法に共通する部分について記述する．一般的に，本研究における自動コーディングの学習は，チャットログデータの各発言を入力として，人手で付与されたコーディングラベルを出力するような，分類問題となる．そのため，各発言は単語の列として扱われ，ニューラルネットへと入力される．また，最終的に出力としては，各コーディングラベルの各々について，ニューラルネットが予測する確率値が出力される．

各単語は 50-100 次元程度のベクトル空間へ写像されて，ベクトルで表現される．この写像は，一般に単語の「埋め込み」と呼ばれ，埋め込まれた単語を表すベクトルは単語ベクトルと呼ばれる．まず，発言は単語ごとに区切られる．その後，得られた単語の列 w_1,..., w_T を，m 次元の単語ベクトルの列 v(w_1),..., v(w_T) へ変換した後，ニューラルネットへと入力する．したがって，ニューラルネットに対する入力は，単語ベクトルを並べたものになるため，T×m の行列と考えられる．

3．4．1．1 CNN による分類モデル

畳み込みニューラルネット(CNN)はもともと画像認識のために用いられたニューラルネットワークの構造であるが[12]，Kim ら[13]によって，極性判定などのテキスト分類の問題に適用しても，高い分類精度を得ることができることが知られている．Kim らの手法に基づいたコーディングの判別のための手法を図 2 に示す．まず，事前に word2vec[14]という手法により，Wikipedia などの大規模データから，そのなかに出現した各単語に対して，50-100 次元程度のベクトル表現を学習させる．その結果，意味の近い単語はベクトル空間内でも近いところへ写るように写像されるようになり，それらは単語ベクトルとよばれる．本研究では，単語ベクトルを， 日本語 Wikipedia 全文から作成する場合と，本研究に用いたチャットログデータから作成する場合の，両方のケースにおいて適用し，比較した．得られた単語ベクトルは，Kim らの手法に従い，畳み込み層により，T×ch の行列に写像された後(ch は CNN におけるチャネル数とよばれる定数)，max pooling により ch 次元のベクトルに写像される．さらに，そのベクトルを全結合層に入力し，Softmax 層を経て，各ラベ

ルに対する予測確率値を最終的に出力する．

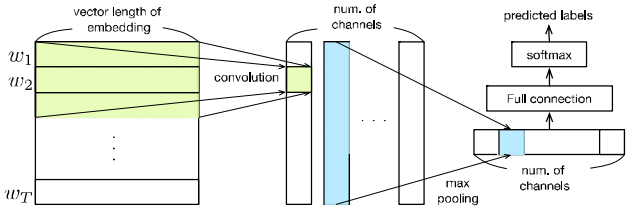


図 2 CNN による分類モデル

3．4．1．2 LSTM による分類モデル

長短期記憶(LSTM)[15]とは，リカレントニューラルネットワークの一種であり，主に音声や単語列などの時系列データを分類・認識する際に用いられるニューラルネットワークの構造である．時系列中に存在する長距離の依存関係を捉えることができることが知られている．本研究では，まず，v(w_1),..., v(w_T) を順次，LSTM に入力してゆき，最終的に v(w_T) を入力した後に得られる LSTM からの出力のベクトル o をえる．さらに o を，全結合層に入力し，Softmax 層を経て，各ラベルに対する予測確率値として出力する．なお，実際には，精度向上のために，2 層に重ねた LSTM を 2 セット用意し，単語列を正順と逆順の両方向で入力していき，それら 2 セットの出力のベクトルを連結したものを，全結合層の入力とした(図 2)．

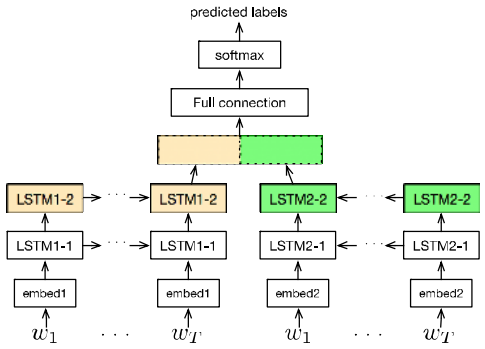


図 3 LSTM による分類モデル

3．4．1．3　Seq2Seq による分類モデル

各発言は会話文の一部であるから，コーディングラベルをより正確に予測するためには，会話文の文脈を考慮する必要がしばしばある．例えば，あるユーザによる発言 A「大化の改新って，いつ誰が起こしたか知っています？」を参照して，別のユーザーによる発言 B「大

化の改新は 645 年に中大兄皇子や中臣鎌足が中心となって起こしたらしい.」が存在した場合，A と B に対する正しいコーディングラベルはそれぞれ，「質問」と「回答」である．しかし，もし発言 A とは全く関係のない文脈で 発言 B があった場合，B の正しいコーディングラベルは「事実の提示」となる．このような発言間の関連性を捉えるため，本研究では，Seq2Seq とよばれるモデル(16)(17)を適用する．まず，「ソース」と「リプライ」と以降で呼ぶ，発言のペアを次のようにして作成する．(1) 発言 A が他の発言 B をシステムの機能を用いて明示的に参照している場合，B をソース，A をリプライとする．(2) 発言 A が明示的に参照している発言が存在しない場合，A の直前の発言をソース，A をリプライとする．(3)発言 A がスレッドの最初の発言の場合，空文をソース，A をリプライとする．

作成したソース・リプライのペアに対して，一つのコーディングラベルを出力するようなニューラルネットを構成し学習させる．ソースの単語列を u_1, ..., u_T'，リプライの単語列を w_1, ..., w_T とする．本研究では Seq2seq モデルを利用するが，このモデルは，encoder および decoder と呼ばれる 2 つの異なる LSTM を持ち，encoder の状態ベクトルを decoder の状態ベクトルへ代入することにより両者を接続する．本研究では，encoder にソースの単語ベクトルの列を入力した後，decoder にリプライの単語ベクトルの列を入力する．その後，最終的な decoder の出力ベクトルを，他の手法と同様，全結合層に入力し，Softmax 層を経て，各ラベルに対する予測確率値として出力する．この手法の場合も，実際には，精度向上のために，2 層に重ねた encoder と decoder のペアを 2 セット(すなわち全部で 8 個の LSTM)用意し，正順および逆順でソース及びリプライを入力して，それらの出力を連結する(Fig. 4).

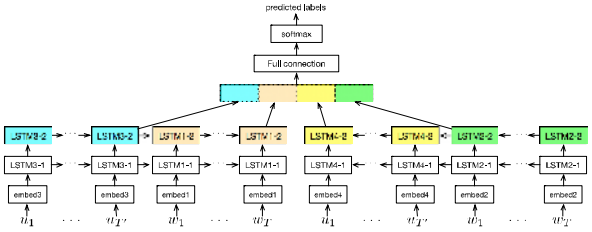


図 4 Seq2Seq による分類モデル

3. 5 実験と評価

3. 5. 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は, 文の形態素への分割を MeCab を用いておこなった．ま

た，頻度の低い単語を 「unknown」と置き換えた．人手によるコーディングによって一致をした，全部で 8,015 の発言のうち，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，総和が 1 になるように正規化したのち，上記分類器に入力した．

本研究で実装した各ニューラルネットの構造を決めるためのパラメータは，次のようにとった．すべてのモデルについて，単語ベクトルの次元数を 200，最後の全結合層の出力ベクトルの次元数を 200 とした．CNN に基づいたモデルにおいては，畳み込み層のパッチサイズを 4 ，チャネル数を 256 とした．LSTM および Seq2Seq を用いたモデルにおいては，全ての LSTM の出力ベクトルの次元数を 800 とした．

モデルの学習は，確率的勾配降下法 (SDG)の一種である Adam を用いた．また，すべての方法において，全結合層において，ドロップアウトを適用した．過適合をさけるため，CNN に基づいたモデルにおいては 30 世代，LSTM および Seq2Seq を用いたモデルにおいては 10 世代で，学習を終了させた． テストデータに対する正解率や F 値は，世代ごとに変動するが，その影響さけるため，最後の 5 世代によるニューラルネットの予測結果を平均した値を実験の結果として用いた．

3．5．2 実験結果

表 4 に前節で提案した DNN 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（正解率）を示す．正解率とは，モデルが出力した予測ラベルと，人手により付与された正解ラベルとが一致する割合である． 表 4 が示すように，全体として，DNN 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 DNN 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，正解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) だが正解率が高くなっている．

一方，全てのモデルの中で，Seq2Seq を用いたモデルが最も正解率が高くなっている (0.71)．SVM と比べて 5-7%，他の DNN モデルと比べても 3-4%高くなっている．

表 4 提案 DNN モデルおよびベースラインによるテストデータに対する予測精度(正解率)

Naïve Bayes		SVM(Linear)		SVM(RBF Kernel)	
unigram	uni+bigram	unigram	uni+bigram	unigram	uni+bigram
0.554	0.598	0.642	0.659	**0.664**	0.659

CNN		LSTM		Seq2Seq	
with wikipedia	w.o. wikipedia	single-direction	bidirection	bidirection	bidir. w. interm.
0.686	0.677	0.676	0.678	**0.718**	0.717

次に，ラベルの一致率の指標として一般的に用いられるカッパ係数で上記の結果を考察する．まず，LSTM を用いたモデルに対するカッパ係数は 0.63 となり，この場合でも，一致

度としては十分高い(Good to fair) 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，カッパ係数で 0.8 以上（Excellent）が好ましいとされており，より高い一致度が求められる．一方， Seq2Seq を用いたモデルに対するカッパ係数は 0.723 であり，0.8 には至らないものの，一致度の観点から見ても，大きく改善されていることがわかる．各発言をばらばらに捉えるのではなく，文脈の情報を何らかの形で考慮することの重要性がわかる．

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，各ラベルの適合率(precision)と再現率(recall) および F 値を 表 5 に示す．「挨拶(Greeting)」，「了承(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，「外部コメント(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かんがえられる．また，「返答(Reply)」でも F 値が低い (0.53)．Seq2Seq を用いたモデルにおいても，「返答(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 5) 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．「応答」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．より適切な発言のペアを抽出する手法が今後の研究では求められる．

表 5 Presition and Recall for each label (result of bi-directional LSTM)

	Presition	Recall	F1-value
Agreement	0.85	0.81	0.83
Proposal	0.73	0.74	0.73
Question	0.75	0.8	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

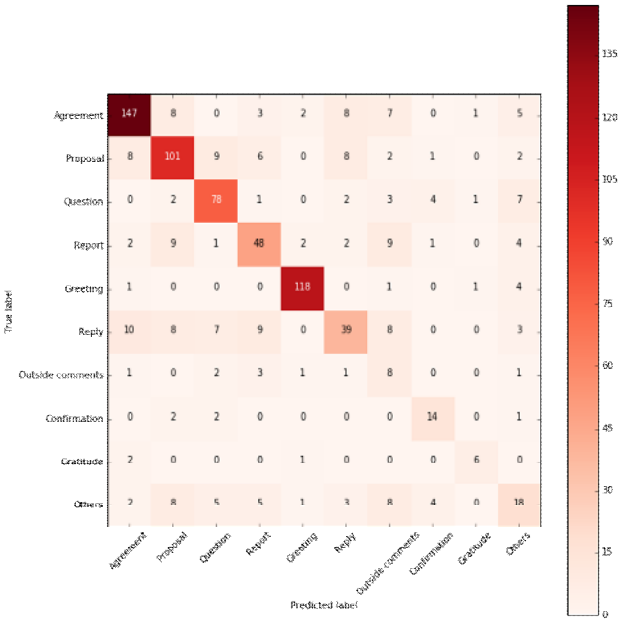


図 4 Confusion matrix for the Seq2S2q model.

3．6　開発手法の有効性の検証

3.3 および 3.4 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。

3．6．1 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。

表６　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

表７　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

３．６．２　自動コーディング結果

図５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。

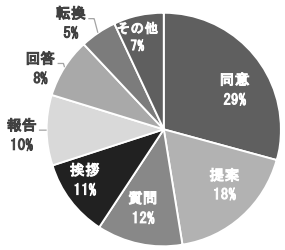


図５　自動コーディング結果の割合

３．６．３　提出物評価と発言内容

表８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。また表９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。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をしたかが重要であると考えられる。

表８ 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

表９ 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	全発言
全体	0.17	−0.16	0.04	0.09	**0.37**	0.05	−0.09	0.07
具体性	0.16	−0.08	0.08	0.00	**0.37**	0.10	−0.12	0.09
工夫	0.18	0.02	0.18	**0.20**	**0.38**	**0.26**	−0.08	**0.24**
適切性	−0.02	−0.04	−0.09	0.03	**0.32**	−0.01	−0.02	−0.01

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。変動係数が高いと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表 10 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。

表 11 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 1 人が担っていると考えられ，課題のほとんどをその 1 人が行うなどグループとしての機能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，1 方向的な会話になっていると考えられる。

表 10 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

表 11　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！

3. 6. 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。表 10 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。

3. 7 新しいコーディングスキーム

前節までの研究で用いていた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、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ことが研究の過程で明らかになった。

そこで、新たに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設計を行うことにした。

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[18]。表 12 に示すように、新コーディングは 4 つの次元からなり、チャットでの発言単位でコードが付与される。また、4 つの次元のコードは、それぞれ複数のラベルから 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 12 新コーディングスキーム

● 次元	● 内容
● Epistemic	● 課題解決への直接的な関わり方
● Argumentation	● 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

3．7．1 Epistemic 次元

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 13 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 13　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、Argumentation 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。

3．7．2 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 14 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。

表 14　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てい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。

3．7．3 Argumentation 次元

Argumentation 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。

Argumentation 次元のコード一覧を表 15 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 15　Argumentation 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)

Argumentation 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。

3．7．4 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 16 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。

表 16　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

Social 次元のコードは他者との関わりを表すコードなので、Social 次元のコードを分析することで、どれだけ活発な議論が行われたかや、グループの誰の意見が尊重されたかなどが推測できる。例えば「Quick consensus building」が多い議論は、ほとんど深い議論されることなく、出された意見をそのまま取り入れるだけの結果となっていることが予測される。

3．7．5 各コーディング次元間の関係とコードの付与

新コーディングスキームでは、「Participation」次元のコードは発言のログから自動的にシステムが生成するが、他のコードについては、コーダーによる人力での付与が必要となる。また、「Epistemic」のコードの結果によって、「Argumentation」「Social」「Coordination」のどの次元のコードを付与するかが決まる。

そのため、コーダーは発言内容および「Participation」次元のコードを分析し、「Epistemic」のコードを付与する。その後、「Epistemic」のコードが「On task」の場合、「Argumentation」および「Social」次元のコードを付与する。また「Social」次元のコードが「consensus building」の場合、必ず参照元の発言があるため「Refer」として発言番号を付与する。「Epistemic」のコードが「Off task」の場合は、「Coordination」次元のコードを付与する。図 6 に各次元の関係性を示す。

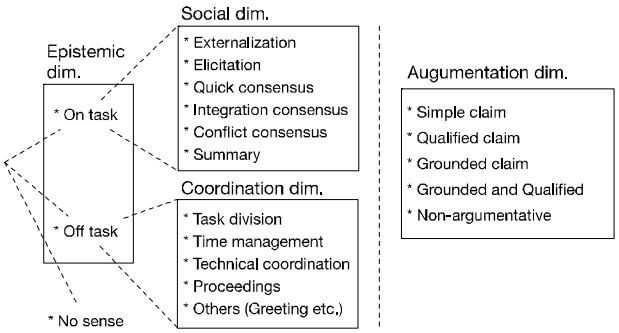


図６　各コーディング次元間の関係性

３．８ 実験と結果

先行研究として 3.3 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、Augmentation 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。

３．８．１ 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。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 3)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。

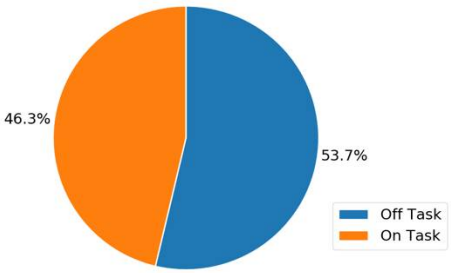


図７　　Epistemic 次元の各ラベルの割合

Argumentation 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 4)。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。

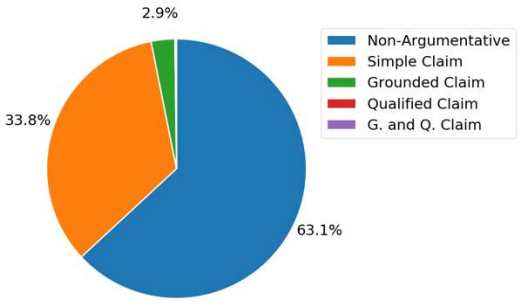


図 8　Argumentation 次元の各ラベルの割合

Coordination 次元についてもほぼ同様のことが観測できる(図 5)。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。

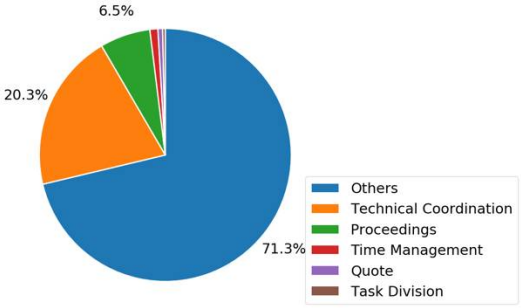


図 9　Coordination 次元の割合

また、Social 次元(図 6)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。

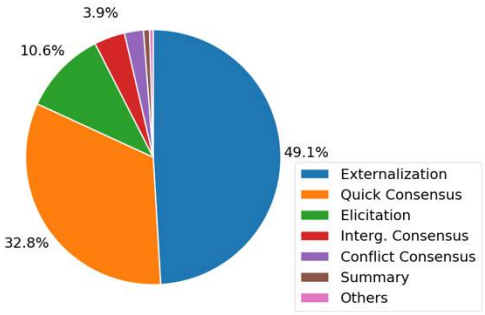


図 10　Social 次元の割合

3．8．2 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 17　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。評価らわかるように、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 18　Argumentation 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、Argumentation 次元に関しても高い精度が得られた。f1 値の micro 平均(micro-f1)は 87%であり、特に「Non-Argumentative」の f1 スコア（91%）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 19　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。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。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 20　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。

参 考 文 献

(1) Stahl, G., Koschmann, T. and Suthers D.: “Computer-supported collaborative learning”, In The Cambridge handbook of the learning science, K. Sawyer, Eds. Cambridge university press, pp.479-500 (2014)

(2) Koschmann T.: “Understanding　in action”, Journal of Pragmatics, 43, pp435-437 (2011)

(3) Koschmann T., Stahl G., and Zemel A.: “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)”, In Video reseach in the learning sciences, Goldman, R. , Pea,R., Barron B. and Derry S. Eds. Routledge, pp.133-144 (2007)
(4) Chi M.: “Quantifying qualitative analyses of verbal data: A practical guide ”, Journal of the Learning Science, 6(3), pp.271-315 (1997)
(5) Meier A., Spada H., and Rummel N.: “A rating scheme for assessing the quality of computer-supported collaboration processes”, International Journal of Computer Supported Collaborative Learning, 2, pp.63-86 (2007)
(6) Jeong H.: “Verbal data analysis for understanding interacitons”, In The International Handbook of Collaborative Learning, C. Hmelo-Silver, A. M. O’Donnell, C. Chan and C. Chin, Eds. Routledge, pp.168-183 (2013)
(7) Persico D., Pozzi, F. and Sarti L.: “Monitaring collaborative activities in computer supported learning”, Distance Education, 31(1), pp.5-22 (2010)
(8) Lipponen L., Rahikainen M., Lamillo J., and Hakkarainen K.: “Patterns of participation and discourse in elementary students’ computer-supported collaborative learning”, Learning and Instruction, 13, pp.487-509 (2003)
(9) Rosé,C. et al.: “Towards an interactive assessment framework for engineering design project based learning”, In Proceedings of DETC2007 (2007)
(10) Rosé, C. et al.: “Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning”, International Journal of Computer Supported Collaborative Learning, 3(3), pp.237-271 (2008)
(11) Inaba T. and Ando K.: “Development and Evaluation of CSCL System for Large Classrooms Using Question-Posing Script”, International Journal on Advances in Software, 7(3&4), pp.590-600 (2014)
(12) LeCun,Y. Bengio, Y. and Hinton G.: “Deep learning”, Nature, 521(7553), pp.436—444 (2015)
(13) Kim, Y.: “Convolutional neural networks for sentence classification”, arXiv preprint arXiv:1408.5882 (2014)
(14) Zhang,X., Zhao J. and Y.LeCun: “Character-level convolutional networks for text classification”, In Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS2015), pp.649-657 (2015)
(15) Hochreiter S. and Schmidhuber J.: “Long short-term memory”, Neural Computation, 9(8), pp.1735--1780 (1997)
(16) Bahdanau, D., Cho, K. and Bengio, Y.: “Neural machine translation by jointly learning to align and translate”, arXiv preprint arXiv, pp.1409.0473 (2014)
(17) Vinyals O. and Le, Q. V.: “ A Neural Conversational Mode”, arXiv preprint arXiv:1506.05869, (ICML Deep Learning Workshop 2015) (2015)
(18) Weinberger, A. and Fischer, F.: “A frame work to analyze argumentative knowledge construction in computer-supported learning”, Computer & Education, 46(1), pp.71-95 (2006)

第4章

反転学習

第 4 章　反転学習

4．1 概要

近年、反転学習とよばれる学びのスタイルが注目されている。授業を受けてから復習へ移行するという旧来の学習形態ではなく、事前に課せられた予習に取り組んだ上で当日の授業に臨むという新たな学習形態である。効果的・効率的なアクティブラーニングの展開に繋がるという意味で期待が大きく、21 世紀型の学習観の主流となりつつある。

本章では、本学で開講している一般教養人文社会系の法学と心理学の 2 科目を対象に、本学の学習管理システムである Moodle を活用して実施した反転学習の効果ならびにその課題について報告する。

4．2 アクティブラーニングと反転学習

4．2．1 アクティブラーニング

アクティブラーニングは、昨今の教育改革の一環として登場してきた概念であり、受け身の学びの姿勢を改め、主体的・能動的に取り組む学びのスタイルのことである。個における知識の定着や創造力の醸成はもちろんのことであるが、グループワークや討論などの他者との協働活動を通じて協調性・社会性を育むこともそのねらいとしている[1][2]。

本学では、初年次のフレッシャーズゼミ（I・II）や 2 年次のキャリアデザイン（I・II）などの授業科目で積極的に取り入れている。

4．2．2 反転学習

反転学習は、アクティブラーニングの思想と親和性の高い新たな学びの形態である。 従来の学びは、教室で講義を受け、その後に復習をしつつ、レポート課題に取り組むというスタイルであった。一方、反転学習では、事前に講義ビデオを視聴したり、クイズ問題に取り組むというような予習が課され、授業当日はその準備学習を前提に、より高度な内容の学びやグループディスカッションなどのアクティブラーニングを展開するというスタイルである。知識の定着やコミュニケーション力・協調力の醸成に有用であることから、近年普及している学習観である。

反転学習の利点は、従来学習者に任かされていた授業外学習の部分に対面授業と同様、教員の意図を大きく反映させられることにある。また授業外での学習（講義ビデオの視聴など）からの復習・応用などの新たな学習サイクルを生み出せるため、学習効果の向上が期待されている[3]。この反転学習は、2010 年頃からブームとなった大学講義の録画ビデオをオンラインで無償提供する MOOC（Massive Open Online Course）の動きに呼応している。MOOC の特徴は、然るべく学習した際には、大学からその科目に関する正式の修了証が得られるという点である。

4．3 システム環境と教材

事前学習と授業当日の学習はともに、本学が 2014 年度より運用している基盤学習管理システム Moodle のもとで展開することとした。本節では、事前学習用として Moodle に載せる教材(ミニ講義ビデオ＋クイズ問題)および授業当日の学習用として Moodle と連動して機能する CSCL について述べる。

4．3．1 事前学習用のミニ講義ビデオとクイズ問題

「法学」「心理学」ともに、事前学習として課されるのは、Moodle 上にアップされた 10 分程度のミニ講義ビデオ 3 本の視聴と、その講義内容に関連する数問の確認クイズへの解答である。

講義ビデオは、録画映像をそのまま流すのではなく、講師の解説と同期するように編集された講義資料を併用する仕様とした（図 4.1)。画面を 7：3 の比で縦に分割し、左側に講義資料が表示され、右側に講師映像がワイプのように映されるデザインとした。なお、ビデオの編集には Camtasia Studio を用いた。

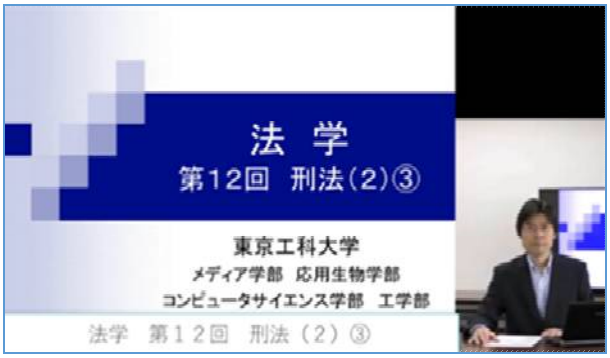


図 1　法学のミニ講義ビデオ

クイズ問題は、Moodle のクイズモジュールを活用して制作した。「法学」は刑法に関する 2 択式の 15 問、「心理学」は内発的動きづけに関する多肢選択式の 17 問を用意した（図 4.2)。

図 2　心理学の 4 択クイズ問題

4. 3. 2 CSCL

授業当日は、Moodle と連動して機能する CSCL を活用した。この CSCL 環境で各学生が事前に準備することは単純であり、ハンドルネームの設定と簡単なアンケートへの回答である。システムは、その登録情報をもとに、2〜3 人のグループを自動生成する。そして、ここから所定の課題テーマに関するクラウド上のグループワークが始まる。

各学生は教室内の誰と議論しているのかはわからないが、“挨拶・確認・質問・提案・報告”という 5 種の発言趣旨カテゴリーの中から適切と思われるものを一つ選択した上で　グループメンバーにメッセージを送る。システムはそれを受けて、図 4.3 のような発言　　ログを残す。

id	body	label
14893	(1)私は行為無価値論派です。理由は社会規範に反する行為を行ったのにもかかわらず、結果的に法律に触れなければ問題がないという判決は法に触れなければ何をしても良いという解釈が生まれるからです。 結果無価値論では、違法行為を働こうとした時点で罰することができるので)	提案

図 3　発言ログ

右端の label フィールドはどのカテゴリー趣旨で発言しているかを表しており、発言　内容そのものは中央の body フィールドに記録される。

4. 4 評価実験

4. 4. 1 概要

各科目の実施日および履修者数は次の通りである。なお、事前学習用のミニ講義ビデオとそれに付随するクイズ問題は、各実施日の一週間前に学生に公開した。学生の利用環境は、ネットワークに繋がるノート PC やスマートフォンである。

■ 科目：法学(前期)
- 実施日：2017 年 7 月 3 日
- 履修者数：154 名

■ 科目：心理学(後期)
- 実施日：2017 年 12 月 8 日
- 履修者数：54 名

■ 科目：法学(後期)
- 実施日：2017 年 12 月 14 日
- 履修者数：319 名

また、主な評価視点として、

- E1：事前学習の成績と授業当日の発言回数との相関（※ 法学(前期)は対象外）
- E2：授業当日のグループワークの量・質

を設けた。

4．4．2 実践･結果 ～ 法学（前期） ～

システムの試験運用であるとともに、E2 の評価基準を設定する事前調査の位置づけで実施した。この時点での暫定評価基準は次の通りである。

・A 評価（※ “挨拶”は除く）

label 選択が適切で、具体例なども挙げて議論の流れに沿った body である。

・B 評価（※ “挨拶”は除く）

label 選択は不適切と思われるが、body は具体例を伴い、議論の流れに沿っている。

・C 評価

具体例こそないが、課題に関係する body で、議論を促進する内容である。label 選択の妥当性は考慮しない。

・D 評価

課題に関係する要素が body にはないが、議論を遮断する発言ではない。label 選択の妥当性は考慮しない。

・E 評価（※ “挨拶”は除く）

課題から外れる、あるいは議論を妨げる body である。label 選択の妥当性は考慮しない。

表 4.1 は、上記基準のもとでシステムに記録された 566 発言のカテゴリーと評価を整理したものである。

表 1　発言のカテゴリーと評価（法学(前期)）

	A	B	C	D	E	小計
挨拶	−	−	36	25	−	61
確認	5	20	151	17	2	195
質問	−	6	57	10	1	74
提案	−	17	77	18	1	113
報告	5	13	83	21	1	123
小計	10	56	404	91	5	566

label は適度なバラツキがあり、最も多いのは“確認”であった。また、議論のまとめに至る“提案”や“報告”はともに全体の 4～5 分の 1 を占めていた。一方、議論の活性化のために欠かせない“質問”が、“挨拶”と同程度であるという結果となった。

評価基準に関しては、A と E は該当者が少なく、基準調整が必要であると考えられる。

4．4．3 実践･結果 ～ 心理学（後期） ～

4.4.2 の結果を受け、後期は 4.4.3 の法学（後期）とともに、評価基準を次のように簡素化した。変更の要点としては、先の A はそのまま残し、B と C をマージし，D と E をマージするということである。この方針のもとに新たに設定した評価基準 A・B・C は次の通りである。なお、“挨拶”に該当する発言は、冒頭あるいは最後でのやりとりに限られ、議論の本質とは関係ないので、一律 C 評価とした。

・A 評価（※ “挨拶”は除く）

label 選択が適切で、具体例なども挙げて議論の流れに沿った body である。

・B 評価（※ “挨拶”は除く）

具体例こそないが、課題に関係する body で、議論を促進する内容である。label 選択の妥当性は考慮しない。

・C 評価

課題に関係する要素が body にはないが、議論を遮断する発言ではない。label 選択の妥当性は考慮しない。

表 4.2 は、事前調査のデータをもとに作成した発言の種別とその分布割合である。事前調査で確認されなかった A 評価の“質問”と“提案”については、発言ログを精査した結果、それに該当すると思われるものがいくつかあったので、確率としては 0 にはせず、“確認”や“報告”の半分（2.5／566）とした。なお、この増分の調整は最も人数の多い“確認”の B 評価（旧 C 評価）のところで行っている。下表は、そのような調整を行った上での確率分布である。四捨五入処理があるものの、全確率は 1.001 となっている。

表 2　発言のカテゴリーと評価の確率分布

	A	B	C
挨拶	−	−	0.108
確認	0.009	0.293	0.034
質問	0.004	0.111	0.020
提案	0.004	0.166	0.034
報告	0.009	0.170	0.039

（※ 各数値は小数点以下第 4 位を四捨五入）

また併せて、各発言のカテゴリーとその評価に応じた素点換算表を用意した（表 4.3）。積極的な発言カテゴリーである“提案”の中で A 評価に値するものを基準値 1 とし、その他を基本的に 0.2 刻みで定めた。なお、“挨拶”については C 評価しかないので、このルールとは関係なく、素点を 0.1 に設定した。

表 3　発言の素点換算表

	A	B	C
挨拶	–	–	0.1
確認	0.6	0.4	0.2
質問	0.8	0.6	0.4
提案	1	0.8	0.6
報告	1.2	1	0.8

まず、評価視点 E1 に関してであるが、図 4.4 のような散布図を得た。横軸が各学生の事前学習のクイズ問題の点数（満点 17）で、縦軸が授業当日の発言回数である。

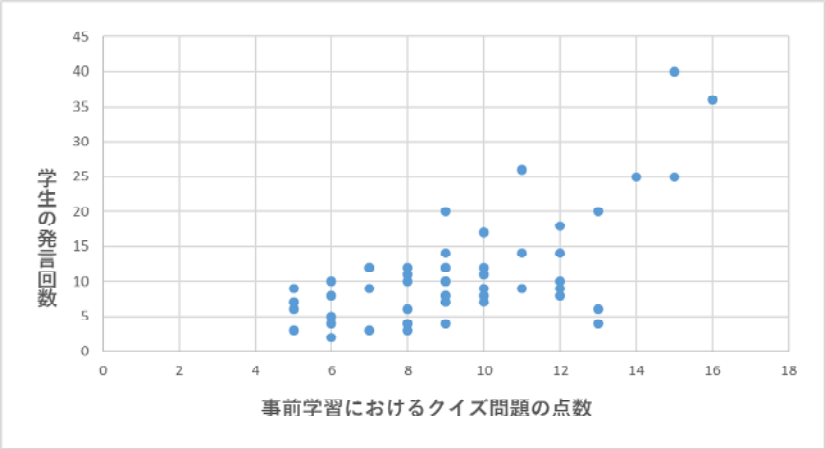


図 4　心理学（後期）の散布図

やや弱いものの正の相関（相関係数≒0.6）を確認した。ただ、発言ログを精査すると、事前学習のクイズの正答数が多い学生の一部が授業当日にあまり発言しておらず、必ずしも実態を反映していないようにも思える。CSCL において必要不可欠なコミュニケーション力という別のスキル要素が、優秀な学生の当日の行動に出たとも考えられる。

次に、評価視点 E2 に関してであるが、素点換算表をもとに全 605 発言を分類・累積加算したところ、表 4.4 のような結果となった。

表 4　心理学（後期）の素点合計

	A	B	C	小計
挨拶	—	—	12.9	12.9
確認	25.2	55.2	24.2	104.6
質問	4.8	21.0	2.8	28.6
提案	26.0	40.0	7.8	73.8
報告	9.6	10.0	16.0	35.6
小計	65.6	126.2	63.7	255.5

表 4.2-4.3 に基づく期待値は約 0.587 であるが、表 4.4 に基づく全発言の素点平均は 0.42（≒255.5／605）となり、数値上で期待値を下回った。

4．4．4 実践･結果　～ 法学（後期） ～

4.4.3 の心理学と同様の手続き・基準で実践した。なお、クイズ問題は前期と同じ 2 択 15 問とした。

まず、評価視点 E1 に関してであるが、図 4.5 のような散布図を得た。横軸と縦軸のラベルは心理学のときと同様である。相関係数は約 0.4 となり、心理学を下回ったものの、分布を見る限り、相関が窺える。

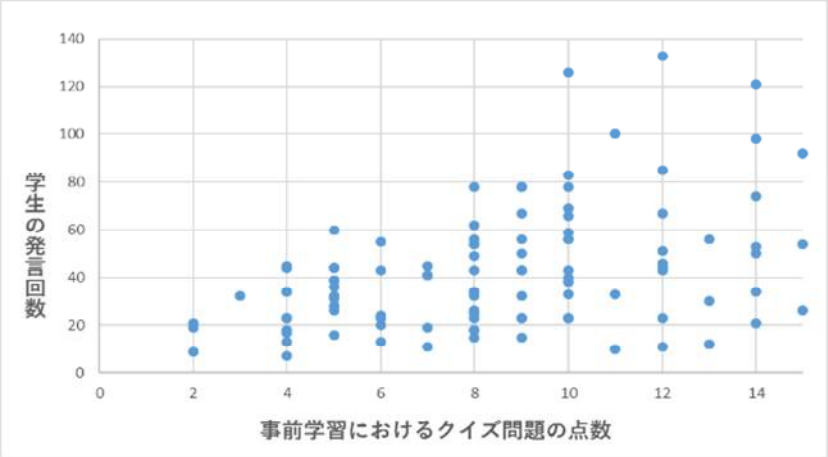


図 4　法学（後期）の散布図

次に、評価視点 E2 に関してであるが、素点換算表をもとに全 3780 発言を分類・累積加算したところ、表 4.5 のようになった。

表 5　法学（後期）の素点合計

	A	B	C	小計
挨拶	–	–	75.6	75.6
確認	136.2	315.6	139	590.8
質問	24.8	127.8	18	170.6
提案	151	250.4	45	446.4
報告	56.4	289	119.2	464.6
小計	368.4	982.8	396.8	1748

全発言の素点平均は 0.462（≒1748／3780）となり、期待値 0.587 を下回った。履修者数が多いことに加え、遅刻者が多かったことが一因と考えられる。本システムは、授業開始時に教員管理のもとで一斉にグループを組む仕様となっているが、遅刻者へのグループ割当てはその都度手動で行っている。

4．5 まとめ

本章では、本学で開講している人文社会系科目の法学と心理学を対象とした反転学習、およびその支援環境である CSCL について論じた。事前調査も含め、学習効果は数値上にも表れており、法学（前期）では、期末試験（60 点満点）の点数が、前年度の 24.1 点から 25.7 点へと増えた。また、科目担当教員からも「従来の授業に比べて事前準備が大変であったものの、当日は密の濃い授業が展開できた」という評価を得た。

一方、課題も山積である。当日学生が取組む課題テーマは共通であったため、本来議論するグループメンバーではなく、隣の友人と議論する様子が散見された。その都度注意をすることでこの問題は解消されていったが、遠隔利用においても本質的な問題は変わらない。本来議論するメンバーとは別の友人に SNS を通じて助言を求めるということは可能であり、監視が難しい潜在的な課題といえる。また、遅刻者への対応も大きな課題である。教室内に限ったことではないが、システム（管理者）は一定のタイミングで、その時点でのシステムへの参加者をグループに振り分ける。しかし、遅刻者への自動対応はまだ実装されていない。

今回の試みは人文社会系科目が対象あったが、今後は数学や物理などの数理・自然科学系の科目で実践してみて、その効果の差異を見るのは興味深い。また、AI を活用して隣の友人とのテーマの共有がないシステム環境にできればと考えている。

第5章

遠隔会議システムを活用した英会話授業

第５章　遠隔会議システムを活用した英会話レッスン

５．１　はじめに

Active Learning Center for English、通称 ALICE は、学生の英会話力、コミュニケーション能力、異文化理解力の向上のために設置されている。2016 年度末まで八王子キャンパスの図書館棟３階にあった。ネイティブ講師が、少人数制の対面型の英会話レッスンを行っていた。環境ウィークやクリスマスのイベントには大勢の学生が参加した。その一方で、蒲田キャンパスでは、英会話スクールのネイティブ講師が、ランチタイムに対面型の英会話レッスンを行っていた。２つのキャンパスの学生同士の交流と、ALICE の運営費節減のために、遠隔会議システムを活用して、八王子キャンパスの学生たちに蒲田キャンパスのランチタイムの英会話レッスンを受けさせてはどうか、という案が浮上した。本学では、そのような英会話レッスンの方法も効果も未知のものだった。そこで、ネイティブ講師 A と学生８名、ネイティブ講師 B と学生９名の２つのグループに、対面型と遠隔会議型の英会話レッスンを行ってもらい、それぞれの特徴を観察し、違いを分析し考察した。本編はそのレポートである。

５．２　調査概要

５．２．１　講師プロフィール

調査協力をしてもらった講師はいずれも外国人男性だった。講師 A はベネゼエラ出身で大学院生だった。英語は第二言語だったが流暢だった。大学での正規の英語の授業の指導経験はなかった。講師 B はイギリス出身で、英語は母語であった。日本の大学で長年にわたり正規の英語の授業の指導経験があるベテランだった。それぞれ同じ学生のグループを、2016 年 11 月第 4 水曜日と 12 月第 1 水曜日の 2 回に渡って、45 分間指導してもらった。1 回目のセッションでは、八王子キャンパスの図書館棟内の ALICE で対面型の英会話レッスンを、2 回目のセッションでは、講師に蒲田キャンパスに出向いてもらい、八王子キャンパスにいる学生たちに対して、遠隔会議型の英会話レッスンをしてもらった。2 回目は、両キャンパスの遠隔会議システムが置かれた会議室が使われた。モニターは大型で見やすく、音響スピーカーとマイクは高性能だった。レッスンはビデオ録画された。

５．２．２　受講者プロフィールおよび学習スタイル

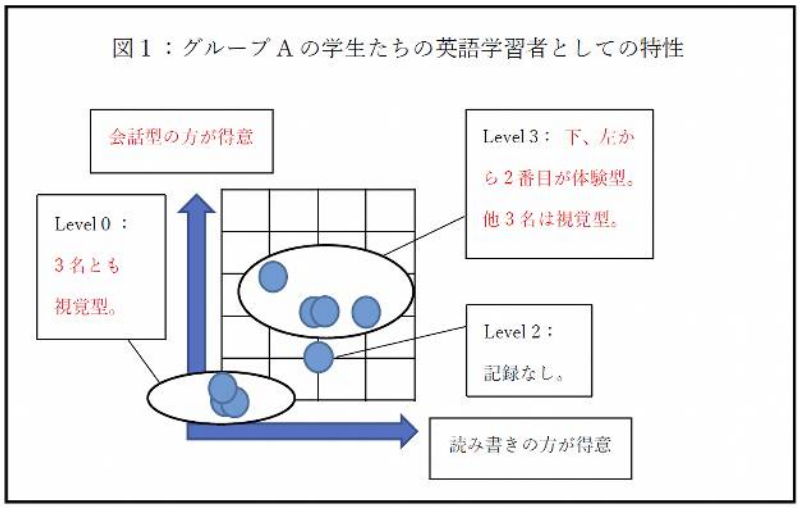
講師 A のグループ（＝グループ A）には８名の学生が、講師 B のグループ（＝グループ B）には９名の学生（うち１名は１回目に参加せず、2 回目にのみ参加した）がいた。講師たちには、2 回とも、普段と同様の指導スタイルでレッスンをしてもらった。学生たちには、同じグループでレッスンを 2 回受けること以外は参加条件を出さず、学生の英語力のレベルや英語学習動機、講師の好み等に基づいて、２つのグループの特徴を均質化することは

なかった。

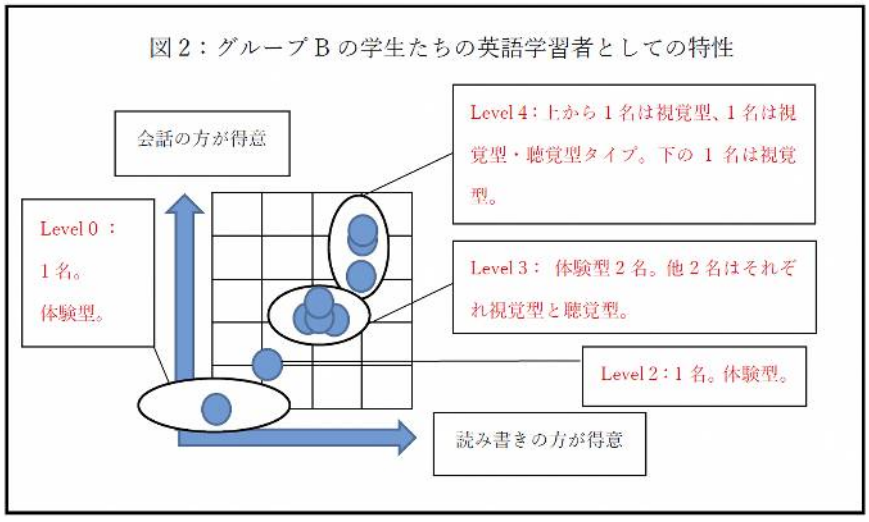
学生には、学習スタイル志向、英語力自己査定、英語学習動機に関するアンケート調査、2 回分のセッションへの評価と感想の記述をしてもらった。そのデータは、彼らのレッスンへの取り組み方、対面型と遠隔会議型指導への反応の違いを説明するために使った。学習スタイル志向のデータによって、絵や図等を見るのを好む視覚型、歌や話等を聞くのを好む聴覚型、手作業等を好む体験型であるかが確認された。英語力自己査定のデータによって、英語力に関する自信の程度、会話か読み書きのどちらを比較的得意とするか確認された。

概ね、A グループでは、英語があまり得意ではない学生たちが、親しみやすい若手で明るい外国人の講師と、異文化交流をメインに英会話活動を行っていた。学生は講師の英語を熱心に聞くが、グループワークになると日本語でおしゃべりをしながら活動し、比較的英語が得意な学生がグループ発表をしていた。講師も、学生がわかるまで徹底して説明してはいなかった。B グループでは、英語が比較的得意な学生たちが、熱心に英語を教えてくれる外国人の講師と、英語学習を行い、積極的に英会話を行っている様子が観察された。グループワークで、学生たちは日本語を使用することがあったが、講師が英語を使うように注意した。学生がわからなそうにしていると、講師は学生に近づき言葉や身振り、プリントの絵を指さす等して、熱心に説明していた。

A グループの 8 名の学生たちのうち、約半数の 3 名は英語力に自信がなく（レベル０）、4 名は英語力に少し自信があった（レベル３）。学習スタイル志向は、ひとりが体験型で、その他は視覚型だった。講師の指導スタイルは、絵などをあまり使わず、常に自然なスピードの英語を話し続けて授業を行うというものだった。学生が英語を理解できない表情をしていると、ホワイトボードにヒントを書いて説明を試みた。オールイングリッシュでレッスンを行い、日本語の使用はまったくなかった。



一方、B グループの 9 名の学生たちは、1 名だけが英語力に自信がなく（レベル０）、4 名は英語力に少し自信があり（レベル３）、3 名は英語に自信があった（レベル４）。トップレベルの学生たちは、会話よりも読み書きに自信があった。学習スタイル志向は、5 名が体験型で圧倒的に多く、その他は視覚型と聴覚型だった。講師の指導スタイルには、最初に絵や文字が書かれたプリントを複数枚使用して学生に机上ワークをさせ、レッスンの最後には手を使った活動を行うという流れがあった。自然なスピードの英語で指示を出し、ときどきホワイトボードも活用して説明していた。授業の最初に配るプリントには、その日の活動に必要となる語彙や表現が書かれており、それを絵とマッチングさせる問題や、クロスワードパズルを完成させる問題があった。また、授業中、講師は学生たちの間を行き来し、英語で声掛けを行い、進捗を確認していた。



５．３　対面型授業（1 回目のセッション）の記録

５．３．１　グループＡ

1 回目のセッションでは、まず「卒業後大金があったら何をしたいか」というトピックで、講師が自分の例を述べ、学生たちに問いかけた。最初、全体的に学生たちは黙ってお

り、ほとんど英語を話さなかった。やがて、英語に比較的自信のある学生たちが、旅行に行く、会社を作る等と答えた。次に、講師はプリントを配り「マンガでストーリーを作ろう」というトピックで、学生にグループワークをさせた。学生たちは、日本語でアイディアを出し合い、紙に絵を描く等した後、代表者がそれを見せながら、英語でストーリーを発表した。講師は「グループのメンバー全員のアイディアがすべて入っていますか」「ストーリーを市場に売り込もうとするとき、セールスポイントを３つ述べてください」等と質問や指示を出すが、学生は理解することができず、答えられなかった。講師が、ストーリーの内容について、例えば「パイロットになるために勉強しましたか、それとも大金でライセンスを買いましたか」等、具体的な質問をすると答えられた。最後には、お互いに作品を回覧し、人気投票をした。

英語に自信のある学生たちは、グループの代表として発表をする等、英語を比較的話していたが、英語に自信のない学生たちは、もっぱら聞き手になっていた。学習者はほとんどが視覚型で、講師もそのような学生のニーズを察知していたのか、ホワイトボードを使って説明しようとすることがあった。しかし、質問の内容がより深い内容になると、講師が言っている英語を学生が理解できなくなり、英会話のやり取りが不可能になった。そのようなとき、講師は話題を変えて別の質問や違う学生と会話を始めた。

写真１：A グループの対面型レッスン

写真２：B グループの対面型レッスン

５．３．２ グループ B

１回目のセッションのテーマは「ホテル」。講師はプリントを配り、学生に語彙問題に取り組ませ、グループごとに完了までの時間を競わせていた。講師はグループの間に座り、間近で活動を観察し､友だちの答えをコピーしようとする学生がいると優しく注意をした。答え合わせをすると、まず､「どちらのホテルに滞在したいか」というトピックで英会話の活動が始まった。プリントにある日本の２つのホテル（ひとつは伝統的な旅館、もうひとつは直島にある個性的なホテル）のうち、どちらに宿泊したいかを、講師がひとりひとりの学生に近づき、質問して答えさせていた。ある学生が､「プールがあるから直島のホテ

ルに泊まりたい」と答えると、講師は「そのプールでは泳げないのですが」とコメントを返す等していた。次に、プリントで、世界各地の変わったホテル（氷や水で作られている等）の写真を見ながら、「どのホテルを体験したいか」というトピックで、英会話の活動を行った。学生たちは興奮して、日本語でおしゃべりをしていた。最後に「折り紙で民宿を作ろう」という活動を行った。講師が学生の間を回りながら、手元で折り方を見せながら英語で説明し、学生の作業を確認していた。学生たちは作っている最中も、完成させた作品を見せ合っているときも笑顔だった。事後のアンケートでは、ある学生は「民宿、大切にします」と記述していた。

B グループには、比較的英語力に自信があり、旅行や将来の仕事に英語を使いたいと思う等、英語学習の動機が高い学生が集まっていた。学習スタイル志向が、体験型である学生が多数いた。講師には、最初に学生たちに語彙学習をさせ、次にプリントを配って、多様な写真等を見せながら英会話活動を行い、最後に、英語でクラフト作品を作らせるという、ルーティーンのある指導スタイルがあった。学生たちは、ときどき日本語のおしゃべりをしたが、頭を働かせ手を使って、積極的に英語の発声をしていたことから、講師の指導スタイルと学生たちの学習者としての特性が適合しているように思われた。

５．４　遠隔会議型レッスン（2 回目のセッション）の記録

５．４．１　グループ A

2 回目のセッションでは、講師は学生たちと同じ空間にはおらず、モニターの向こう側に立っていた。学生たちはコの字に座っていた。最初に、講師が「国内外で旅行したい場所」というテーマを提示し、グループを作るように指示した。すると、学生たちは一斉に日本語で話し出した。学生たちはその後も日本語を話し続けた。講師が、国内で旅行したい場所とその理由を英語で質問すると、学生たちは、‘Fukuoka. I'm from Fukuoka.’ ‘Okinawa.　It’s hot.’ のように、英語で答えた。講師が、次の学生に対して、How about you? と言うが、指名がなかったので、学生は誰の番なのかがわからず、黙っていた。長野－生誕地、北海道－自然、京都－金閣寺、と英語で答えが続いた。ひとりの学生が、神社を英語で言えなかったため、他の学生たちが ‘shrine’ だと教えてサポートした。新潟―花火、群馬－温泉、宮城―牛タン、福島―自然、と学生がそれぞれ答えた。次は、海外で行きたい国。学生は日本語で話していたが、自分の番になって質問されると、英語で答えた。エジプト―ピラミッドを見たいから。ハワイ―海とダイビング。フランス―フランス料理。行きたい国に関しては、‘I want to go to + 場所.’ の構文で答えられる学生たちが多数いた。モニター越しだが、講師の声はよく聞こえた。講師が話し合ってくださいと言うと、学生たちは日本語で積極的に話し合っていた。講師が、真ん中のグループから発表するように指示し、場所の特徴や理由を具体的に言うように促した。英語に比較的自信のある学生が代表となり発表した。

次に、教室に置かれていた袋が配られ、ペアワークとなった。講師はモニター越しに、ホワイドボードを使って指示を明確に伝えようとするが、ホワイトボードが奥に置かれており、学生は読みにくい様子だった。袋には、問題と解決法が書かれた紙が入っており、これを使って「問題と解決法」というトピックで英会話をすることになった。学生は紙に

書かれた問題を読み上げ、互いに解決法を考えて、英語で発表した。一組ずつ組み合う紙があり、それを順番に、学生たちが読み上げていった。講師は、自分なりの解決法を言っても構わないと言った。講師は、まだ発言していない学生に発言するように促した。ひとりの学生が、袋を届け、受け取った学生が読み上げた。講師は、このようなやり方でペアワークをするように指示したが、学生たちはペアワークを始めなかった。

そこで、講師は学生たちの注意を引き寄せるために、自分自身の問題を板書した。それは、'I broke up with my girlfriend.' だった。ひとりの学生が 'Don't be so sad.'、別の学生が 'That's too bad.' と反応した。講師が別の側からも発言するようにと促すと、ひとりの学生が手を挙げて、'You could get a new girlfriend.' と言った。講師は笑ってうれしそうに反応した。しかし、講師がペアワークをするように改めて指示したが、学生たちは英語で活発にペアワークをすることはなかった。

その後も、講師はモニターの向こう側で板書を行い、ペアワークさせようと試みたが、うまくゆかなかった。結局、クラス全体で行うことになった。ひとりの学生が問題を読み上げ、クラスの誰か他の学生が解決法を英語で読み上げるやり方で活動が続いた。講師が問題を読み上げる番になり、'I don't wake up easily.' と言ったら、ある学生が 'You should set more alarm.' と答えた。講師は聞き取れず、学生は同じ答えを繰り返していったが、コミュニケーションが成立しないまま終わってしまった。別の問題 'I hate working.' に対して、英語の得意な学生がひとつ解決法を言った。講師はもっと解決法を言うように、他の学生たちにも促すが発言はなかった。講師が 'Anything is fine. Don't worry.' と優しく声掛けしていた。'My friend won't see me.' という問題に対して、他の学生が解決法を言わずに、'I'm always late.' と問題の方を読み上げてしまった。講師はホワイトボードを使って説明しようとした。講師は、それぞれの問題に対して、学生から解決法を言わせた。講師が、こちらから側から発言してくださいと言うと、前の問題に対して、ようやくひとりの学生が、 'You should ask him or her why.' と答えた。終了時間となり活動が終わった。

写真３－１：A グループの遠隔会議型レッスン（学生側）

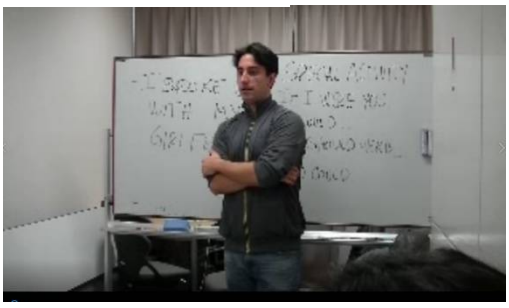


写真３－２：A グループの遠隔会議型レッスン（講師側）

５．４．２　グループＢ

２回目のセッションでは、テーマは「余暇」。講師が学生のグループを３つ編成した。学生たちはコの字型に座っていた。講師が、ひとりの学生に名前を聞き、その学生に教室に置かれたプリントを配るように頼んだ。学生たちは、1 回目同様、まず、プリントで、テーマ関連の語彙のマッチング問題とクロスワードパズルを行った。学生たちは真剣に取り組んでいた。講師はモニターの向こう側に立っていたが、やがて座る。学生たちが解き終わると答え合わせをした。学生たちが答えを言い、講師が確認した。女子グループが最初に終わったので勝者となり、皆で拍手し笑顔が見られた。

モニター越しに、講師が板書をし、学生たちに文字が読めるかどうか確認した。学生たちは大丈夫であると答えた。教室に置かれた別プリントが配られた。講師は、音楽の種類に関して、‘active’ なものか ‘sedate’ なものか、A か S を書くように指示をした。学生は日本語を使いながら問題を解いていた。講師はモニターの向こうに立って、学生の様子を見ていた。講師が、また別の新しいプリントを配るように指示した。‘Extreme Activity’ という言葉を導入し、 ‘Extreme Sports’ に関して説明した。講師は、イギリスで行われている伝統的なスポーツを紹介した。

講師は、学生たちが自由時間にしたいと思う活動について考えるように言う。講師は、日本語ではなく、英語で話すように、学生たちの注意を促す。学生たちは、日本語で話すのをやめ、お互いに英語で話し始める。講師は、プリントにあった活動でやってみたいものを話してもよいと言った。講師は座って、モニターを通して学生を観察していた。‘karaoke’ と言う学生の声に、講師が反応して英語で質問すると、その質問がマイクを通して教室全体に聞こえた。学生たちは、英語を使って話し続ける中、新たな活動として、娯楽番組を制作すると仮定して、紹介されている３つの ‘Extreme Sports’ の中からひとつを選び、選んだ理由を発表するように、講師が学生に指示した。学生がひとりずつ、講師に対して答えた。講師は学生の名前を知らないので、Next person. と指示した。

最後に、青いひもとプリントが配られ、ペアで‘Cat’s cradle’ を作る「あやとり」をした。モニター越しに、講師がデモンストレーションしようとするが、学生は講師の方を見ずに取り組もうとした。講師が学生たちを制止すると、’Sorry, sorry.’と、ある学生が謝った。講師が手元を見せながら説明するが、モニターのすぐ傍に座って講師の説明を見聞きしていても、理解できない学生がいた。近くの学生が「こうするの」と見せながら教えて助けた。いよいよ完成の段階で、学生たちが戸惑っていると、講師が、再度 ‘Everybody, look at the camera.’ と言って、アシスタントと「あやとり」をして見せた。学生たちはそれを見て試した。講師が ‘In Japanese you call this *River*?’ と聞くが、学生は「あやとり」に夢中で、誰も質問に答えず、日本語でも話していた。しかし、質問を聞いていて「リバーまでいかないわ。Very difficult.」とつぶやく学生がいた。さらに、別のもの

を作ろうとした。‘Everybody, look at the screen?’ と講師が言って手元を見せた。しかし、講師は、うまくできないペアに気づき、成功したペアが助けに行くように指示した。講師は、全員ができたことを確認してレッスンが終了した。

写真４－１：B グループの遠隔会議型レッスン（学生側）

写真４－２：B グループの遠隔会議型レッスン（講師側）

５．５　対面型レッスンと遠隔会議型レッスンへの学生の反応の比較分析と考察

５．５．１　比較分析

アンケート調査の結果から、学習スタイル志向、英語力に関わらず、両方の学生たちが、対面型活動の方が遠隔会議型活動よりも好きだったことがわかった（表１）。A グループでは 1 人の学生が、B グループでは 2 人の学生が、「テレビで会話するのが新鮮で面白かった」と自由欄に記述したが、3 人とも、通常は対面型の方が好だと回答した。B グループは全員が対面型を好んだ。学習スタイル志向別には、両方のグループの体験型の学習スタイルを好む学生たちが全員、遠隔会議型活動の方に「違和感がある」とコメントを記述していた。A グループの体験型の 2 人は「英語が難しかった」「活動は楽しかったが私自身リスニングの力が弱いと思った」と書いた。B グループの体験型の 1 人は「いつもと雰囲気が違ったので、違和感があった」と記述した。学生たちが対面型を好んだ主な理由は、「講師と直接会話できた方が、コミュニケーションがとりやすい」、「誰が指名されているかがわかりやすい」、「映像よりも顔の表情がわかりやすいし、声も聞き取りやすい」、というものだった（表１）。

興味深いことに、英語が苦手な学生が多かった A グループと、英語が得意な学生が多かった B グループの反応の違いは、「英語活動への積極的な取り組み」にあった。A グループは遠隔会議型レッスンに積極的に取り組み、B グループは対面型レッスンに積極的に取り組んだ（表２）。レッスンの観察によると、A グループは、遠隔会議型のレッスンで、モニター越しの講師側を英語圏、自分たちの側を日本語圏と分ける傾向があった。学生同士で日本語を気兼ねなく使っており、ペアワークでも英語を使おうとしなかった。しかし、モニター越しに講師がひとりひとりの学生に質問をすると、積極的に答えようとした。一方、B グループは、対面型活動の方が積極的に取り組めており、講師が直接、同じ部屋にいてくれるので、コミュニケーションがスムーズになる、というコメントがあった。英語力が比較的高いからこそ、話す内容が難しくなったとき、英会話を持続させるために、講師との直接的なインタラクションがあった方が頑張れるのかもしれない。また、B グループの講師は、レッスンの最後に手作業を伴う活動（折り紙、あやとり）をしていたため、間近で講師の手元が見える対面型の方が、作業に取り組みやすかったと考えられる。

表 1：形態別活動に関する好み	A グループ（8 人）	B グループ（8 人）
どちらのセッションが好きですか？	対面型：87.5%	対面型：100%
	遠隔会議型：0%	遠隔会議型：0%
	その他：12.5%	その他：0%
またそれはなぜですか？		
【A グループ】		
■対面型が良い：		
先生と直接会話できたほうがより分かりやすいから。		
先生との距離が近い方が好き。コミュニケーションがとりやすい。映像だと表情がわかりづらい。		
目の前に先生がいた方が分かりやすいと感じました。		
画面越しでも大丈夫だが、やはり対面でやった方が、コミュニケーションが取りやすかったので、どちらかと言えば対面型が好きです。		
前と似た感じにはできたが、話す時に誰に話すとか実際にみんながその場にいるときと比べて少し戸惑ったから。		
身近に先生がいてくれた方が内容を理解しやすいため。		
■その他：		
クロスワードが前回より難しかったから。		
【B グループ】		
■対面型が良い：		
テレビ越しだと少しやりづらかったので。		
誰に言っているのかがわかりやすいから。		
直接話した方がコミュニケーションを取りやすいから。		
話す人と同じ部屋にいた方がスムーズに進められるから。また、テレビからの音声が少しだけ聞き取りづらかった。		

表 2：英語活動に関する質問へ回答	A グループ （8 人）		B グループ（8 人）	
セッション	1 回目	2 回目	1 回目	2 回目
英語活動は楽しかったですか？	4.00	3.60	3.20	2.80
英語活動の内容は興味深かったですか？	3.55	3.75	3.20	3.10
英語活動に積極的に取り組めましたか？	2.66	3.25	3.25	2.50
ALICE に今後も参加したいと思いましたか？	3.55	3.50	3.22	3.10

*最低点 1、最高点 4 のスケールを使用

5．5．2 遠隔会議型英会話レッスン実施に関する提案

(1) 英語が苦手な学生のグループには、身近で軽いトピックで英会話をする。

英語が苦手な学生たちは、身近で軽いトピックであれば英会話を楽しめる。最初は身近

で軽いトピックを中心に英会話活動を行い、徐々に社会的で抽象的なトピックを導入するべきである。事前の語彙学習も助けになる。反転レッスンで予習をさせる、あるいは授業の開始時に、プリントで語彙学習をさせる等、学生たちを英語に浸してから会話練習を行うとよいだろう。

(2) 英語が苦手な学生たちには、講師がひとりずつ英会話をする。

会議型であると、全体に向かって話すことになるので、講師が「こちら」「あちら」と指示しても、学生たちは誰が指名されているのかがわからず、沈黙してしまう。また、モニターが境界線となり、講師の側は英語圏で、学生側は日本語圏となり、学生たちが日本語で躊躇なくおしゃべりをしてしまう。誰の順番なのかを明らかにして、ひとりひとり講師と英会話をさせるとよい。そうすることで、日本語でのおしゃべりを抑えることができる。英語が苦手でも、新しいシステムによる学習に好奇心がある学生はいるので、うまく活用すると効果が期待できる。

(3) 英語が得意な学生たちには、英会話に必要な語彙等の事前準備をさせておく。

英語が得意な学生たちには、モニター越しのオンライン・レッスンによる英語学習が成功する可能性がある。それを確かなものにするためには、あらかじめ語彙学習をさせ、予想される質問に関して自分なりの答えを英語で用意させるなど､事前準備をさせるとよい。質問によっては、メッセージを作る時間がかかるので、考えを整理させておく。英語力がないと、学生側は日本語に染まってしまうので、各自の英語力が用意されている状態で英会話活動に取り組ませ、会議型レッスンの効果を確保したい。

(4) システム機器を使いこなすこと。

学生の英語の得意不得意に関わらず、システム機器を適切に操作するスキルは必要不可欠だ。講師あるいはサポート・スタッフにその能力がなければならない。特に、体験型の学習スタイル志向を持つ学生たちは、講師と直接インタラクションをしたいと欲する傾向があるので、臨場感が必要になる。ひとりひとりの学生と英会話ができるように、カメラの位置を変え、ズーム機能を使う。特定のグループにコメントを伝えるために、マイクの位置を変え、ボリュームを調整する。プリントの特定の箇所の説明をするために、そこをズーム機能で拡大する。英語を使いながら手作業を行うとき、講師や学生の手元が見えるように、ズーム機能を使う。あるいは個人のオンラインレッスンと会議型レッスンのハイブリッドシステムにする。つまり、大型モニターだけでなく、ノート PC で講師と学生個人あるいは学生数名が直接つながり、講師と同じ画面上に現れている状態で、英会話の練習をする等、様々な工夫が考えられる。

5．6 結語

本研究では､対面型と遠隔会議型の英会話レッスンの事例調査を行った｡英語力の違い、学修志向スタイルの違いによって反応が異なることがわかった。レッスンの指導を有効にするために、システム機器を適切に操作できなければならないこともわかった。印象的だったのは、遠隔会議型の英会話レッスンを面白いとコメントした学生たちでさえも、通常の対面型のレッスンの方が好きだと答えたことである。本調査では、遠隔会議型レッスンを行う際、システムのカメラの位置、ズーム機能などを特別に工夫して操作することがな

かったので、もしそのようにしていれば、別の回答が得られたかもしれない。また理系大学１校の小規模な事例調査であった。オンラインレッスンは個人型ものが主流だが、グループを対象にした会議型レッスンの良さは、参加者同士が互いのメッセージを共有できること、協働作業ができることである。対面型レッスンを行うことが難しくなった場合、それに代わる有効な手段として、今後も遠隔会議型のレッスンの研究を続けてゆく必要があるだろう。

第6章

CEATEC JAPAN 2017 展示報告

第 6 章 CEATEC JAPAN 2017 展示報告

6．1 展示動機

対外的なプロジェクトでの研究成果の発表は学会や国際会議等ではプロジェクトの始動時から積極的に行ってきた。しかし、アカデミーの領域を超えて実業界や実社会への成果のアピールを行うことも本プロジェクトのミッションの一部であることを看過してはならない。そのため本プロジェクトでは電子技術や IT 技術の展示会としては国内最大級の規模で開催される CEATEC を展示の場所として選択した。選択理由としては、この展示会が来訪者の数と来訪者の来訪目的の多様性という点で抜きんでていることであった。また、事前のプレスリリースや主催者側で事前に行われる展示内容のインタビュー等、主催者側から提供される広報サービスの充実も考慮したポイントである。

6．2 展示概要

日程：2017 年 10 月 3 日（火）～10 月 6 日（金）
場所：幕張メッセ CEATEC JAPAN 2017 会場内
『ディープラーニング・対話・まなびプロジェクト」（教養学環学内共同プロジェクト）
展示エリア：「社会・街エリア」
ブース名：東京工科大学
小間番号：C029
配布物：プロジェクト紹介パンフレット、日本語論文

写真1　展示ブース

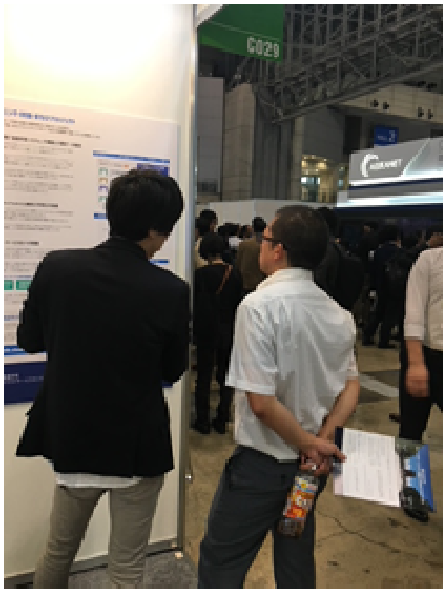


写真 2　接客

6．3 展示内容

展示内容としては、2章、3章で取り上げたPBL協調学習支援システムのチャット機能とその会話内容を深層学習の技術によって自動的にリアルタイムで分類・コーディング（タグ付け）を行う機能のデモを中核に据えることとし、それを来訪者にアピールするため、展示名は「ディープラーニング・対話・まなびプロジェクト」とした。また、本プロジェ

クトを進めるにあたって協力をしてくれたクラウドセンターが開発した学生・教員ポータルサイトの出欠管理システムについても公開を行った。以下は展示パネルの一部である。

図１　展示パネル

また、実社会での応用を想定し、本プロジェクトで開発した手法が、SNS の分析やコールセンターでの会話内容の分析、さらには社内でのチャット内容の分析などに幅広く応用が可能である点もパンフレット、パネル、来客者への説明において特に強調したポイントである。

ブース内にはパネルが 3 枚設置され、さらにチャットシステムを説明したムービーを常時流した。また、今回の展示用に作成したパンフレットと日本語論文をセットで来訪者に配布を行った。デモでは、ユーザにデモ用ノート PC 上のチャット画面から会話を入力してもらい、それに対してリアルタイムでその会話内容を示す分類タグが表示されるという仕組みを実体験してもらった。

6.4 展示来訪状況と成果

幸いにもブースへの来訪者は CEATEC 開催期間を通じて相当な数に達した。正確な数を把握しなかったのは残念ではあるが、説明や質疑が必要だった来訪者も一日 100 名近くはいたと思われる。特にうれしかったのは、会場を来訪していた本学の OB,OG が東京工科大学という名前を会場で偶然発見し、来訪してくれたことである。また、在学生の訪問も少なからずあった。

当然、来訪者の大部分は企業関係、自治体関係の来訪者であり、チャットの分析については関心を示してくれる方も少なくはなかった。すでに開催中に、直近でのアポのリクエストをしてくれた企業が 4 企業あった。その 4 企業とは、CEATEC 直後からプレゼンやミーティング等での交渉があり、最終的には 2017 年度中に 2 企業と共同研究を進めることが決定した。また 2018 年度になって、CEATEC 開催期間中はコンタクトがなく、その後にパンフレットや論文をみてプロジェクトに関心をもった 1 企業との間で研究を行う運びとなった。これらの企業との共同研究の概要については次の 7 章で示す。

第7章

企業との共同研究の概要

7 章　企業との共同研究

7．1

6 章でも述べたように、CEATEC 後に複数の企業と共同研究について交渉があり、その結果現在 3 社の企業とプロジェクトの間で研究契約が締結され、実際の研究が進行中である。これらの企業等は秘密保持契約を同時に締結している関係で研究の仔細をすべて明らかにすることはできないが、契約の範囲内でその概要を示す。

7．2 株式会社ムラウチドットコムとの共同研究

ムラウチドットコムとの共同研究のタイトル、研究期間、研究費は以下の通りである。

・タイトル：muragon における AI（「人工知能」）的アプローチ
・研究期間：平成 29 年 12 月 1 日～ 平成 30 年 5 月 1 日
・研究費：　1,874,000 円

ムラウチドットコムは八王子に本社をおく地元企業である。２つのブログサイトの運営が中核的事業の一つとなっている。この２つのブログとは、「にほんブログ村」と『muragon』の２つのサイトである。前者は、いわゆるブログポータルサイトであり、ブロガーが他社のサイトで執筆したブログへのアクセス数を増やすために、ブログ村に自身のブログを登録することで、このブログ村から自身のブログへとリンクが生成され、アクセス数のランキングなどがページ上に公開される。これに対して、muragon はムラウチドットコムが運営をするブログサイトである。

ブログ村に登録する際、ブロガーは自身のブログの内容をあらわすようなタグ（カテゴリーとサブカテゴリー）を選択することが必要である。このタグによって、ブログ村のブログは分類されている。一方、murgon ではこのタグはつけないことも可能となっていて、分類不可能なブログが大量に発生している。また、これらのブログと他の同タイプのブログとの関連付け等もできず現在まで未着手の課題となっていた。

今回の共同研究ではムラゴンのブログに対して、にほんブログ村のカテゴリーのタグを自動付与する技法の開発がテーマとなっている。これによって、ブログ間の関連付けが飛躍的に増えることになる。また、それによって、ムラゴンに来訪するユーザがより長時間、より多くのブログを閲覧する機会が生まれることが期待される。

7．3 FXcoin 株式会社との共同研究

FXcoin との共同研究のタイトル、研究期間、研究費は以下の通りである。

・タイトル：シナリオベースの AI チャットシステムを実現するための技術・技法の開発と評価
・研究期間：平成 30 年 5 月 10 日から平成 31 年 5 月 9 日まで
・研究費：2,100,000 円

Fxcoin はドイツ銀行で外国為替営業部長であり、東京外国為替市場委員会の副議長を務

めていた大西知生氏が新たに設立した仮想通貨交換業務に特化した企業である。

昨年あたりから顧客からの質問やクレームに対してコールセンター等での友人による電話対応に替わって、一部の業務を AI チャットボットに移行する企業が増えつつある。しかし、これら既存のチャットボットの精度はまだまだ発展途上にあると言わざるをえない。今回の研究では、本プロジェクトで開発したコーディングに依拠した深層学習による分類技法を用いて顧客対応チャットボットを開発し、その精度を検証する。

7．4 株式会社ビズオーシャンとの共同研究

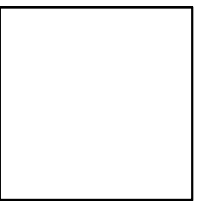
ビズオーシャンとの共同研究のタイトル、研究期間、研究費は以下の通りである。

・タイトル：多様なスタイルの業務報告文から帳票を生成するためのモデルの構築
・研究期間：平成 30 年 6 月 1 日から平成 30 年 12 月 1 日まで
・研究費：2,000,000 円

ビズオーシャンは日本最大級のビジネステンプレートサイト「書式の王様」等を運営するデジタル帳票の提供を主たる業務にしている企業である。2017 年、この帳票を LINE 上から音声入力を行い、Excel や Word の文書として生成するサービス「SPALO」をリリースした。現在までのところ音声入力は帳票の項目ことに一対一で行われており、自然な会話入力で複数分を一度に入力することはできない。今回の研究では、とりあえず営業担当の複数文からなる商談報告書全体を一度に入力したことを想定し、それが帳票フォーマットに変換を可能にするようなモデルの設計を目指している。

附録 A

外部研究発表論文等



Towards Automatic Coding of Collaborative Learning Data with Deep Learning Technology

Chihiro Shibata
School of Computer Sciences
Tokyo University of Technology
Tokyo, Japan
email:shibatachh@stf.teu.ac.jp

Kimihiko Ando
Cloud Service Center
Tokyo University of Technology
Tokyo, Japan
email:ando@stf.teu.ac.jp

Taketoshi Inaba
Graduate School of Bionics, Computer and Media Sciences
Tokyo University of Technology
Tokyo, Japan
email:inaba@stf.teu.ac.jp

***Abstract*— In Computer Supported Collaborative Learning (CSCL) research, gaining a guideline to carry out appropriate scaffolding by analyzing mechanism of successful collaborative interaction and extracting indicators to identify groups where collaborative process is not going well, can be considered as the most important preoccupation, both for research and for educational implementation. And to study this collaborative learning process, different approaches have been tried. In this paper, we opt for the verbal data analysis; its advantage of this method is that it enables quantitative processing while maintaining qualitative perspective, with collaborative learning data of considerable size. However, coding large scale educational data is extremely time consuming and sometimes goes beyond men's capacity. So, in recent years, there have also been attempts to automate complex coding by using machine learning technology. In this background, with large scale data generated in our CSCL system, we have tried to implement automation of high precision coding utilizing deep learning methods, which are derived from the leading edge technology of machine learning. The results indicate that our approach with deep learning methods is promising, outperforming the machine learning baselines, and that the prediction accuracy could be improved by constructing models more sensitive to the context of conversation.**

Keywords-CSCL; leaning analytics; coding scheme; deep learning methods.

I. INTRODUCTION

A. Analysis of collaborative process

One of the greatest research interests in the actual Computer Supported Collaborative Learning (CSCL) research is to analyze its social process from a social constructionist viewpoint, and key research questions are as follows: how knowledge and meanings are shared within a group, what types of conflict, synchronization and adjustment of opinions occur, and how knowledge is constructed from discussions. And answering to these questions enables to develop more effective scaffolding methods and CSCL system and tools.

In earlier researches at initial stage of CSCL, the focus was on each individual within a collaborating group, and the main point of interest had been how significantly a personal learning outcome was affected by characteristic types of a group (such as group size, group composition, learning tasks, and communication media) [1]. However, it gradually became clear that those characteristics are complexly connected and intertwined with each other, and showing causal relation to a specific result was extremely difficult. From the 1990s, the interest in CSCL research had moved away from awareness of the issue on how a personal learning is established within a group, to attempting to explain the process by clarifying the details of group interactions when learning is taking place within a group [2].

However, attempting to analyze collaborative process goes beyond merely shifting a research perspective; it also leads to fundamental re-examination of its analytical methodology. In other words, this involves a shift from quantitative analysis to qualitative analysis. Naturally, there are useful data among quantitative data saved within CSCL system, such as the number of contributions within a group, the number of contributions by each group member, and in some cases contribution attributes obtained from system interface (sentence opener), but those are very much a mere surface data. The most important data for analysis are contributions in chats, images/sounds within tools such as Skype, and various outputs generated in the process of collaborative learning; for analysis of those, ethnomethodologies such as conversation analysis and video analysis have been invoked [3] [4].

However, those researches by their very nature tend to be in-depth case studies of collaborative activities with a limited number of groups and have the disadvantage of not at all being easy to derive a guideline that has a certain level of universality and can be applicable in other contexts.

 ISBN: 978-1-61208-541-8

Therefore, researches have been carried out using verbal data analysis method that carry out coding from a perspective of linguistic or collaborative learning activities on a certain volume of language data generated in collaborative learning and analyzing them [5][6][7]. The advantage of this method is that it enables quantitative processing while maintaining qualitative perspective, with collaborative learning data of considerable size as the subject, while coding them manually is an extremely time consuming task which goes sometimes beyond men's capacity. For example, Persico et al. developed a technological tool which helps the tutors to code the contributions in chats and displays quantitative information about the qualitative information and coding data [8]. However, given that the coding procedure itself remains manual in most existing studies [9][10], there is an insurmountable limit in front of big data. Hence, we seek an automatic coding technique for a large scale collaborative learning data with deep learning methods.

B. Educational data and Learning Analytics

With the progress of educational cloud implementation in educational institutions, data generated in Learning Management System (LMS), e-learning, Social Network Service (SNS), Massive Open Online Course (MOOC) and others are increasing rapidly, and a new research approach called Learning Analytics (LA) that tries to gain knowledge that would lead to support of learning and educational activities by analyzing those educational big data is becoming more active [11][12]. Big educational data obtained from CSCL system integrated in educational cloud at a campus, such as conversation data, submitted documents and images/sounds of learning activities, will certainly become a subject for analysis in the near future: therefore, it is believed that we are coming into a time when it is necessary to seriously examine a new possibility of collaborative learning research as LA. Due to such background, in this research we have reconstructed CSCL system that has been operating in a campus server for the last five years as a module within Moodle, which is a LMS within the campus cloud, and have already structured an environment that can be operated within the campus and collect/analyze collaborative learning data.

C. The goal and purpose of this study

The goal of our research is to analyze large-scale collaborative data from LA perspective as described above and discover the mechanism of activation and deactivation of collaborative activity process which could not be gained from micro level case studies up to now. Furthermore, this research, based on its results, aims to implement supports in authentic learning/educational contexts, such as real-time monitoring of collaborative process and scaffolding to groups that are not becoming activated.

In this paper, as the first step towards this goal, we present work in progress, which attempts to develop an automation technique for coding of chat data and verifies its accuracy. To be more specific, a substantial volume of chat data is coded manually, and has a part of that learnt as training data in deep learning methods, which are derived from the leading edge technologies for machine learning; afterwards, automatic coding of the raw data is carried out. For validation of accuracy, the effectiveness of using deep learning methods is assessed by comparing accuracy against Naive Bayes and Support Vector Machines, which are baselines of machine learning algorithm used in existing studies that carried out automatic coding by machine learning.

D. Structure of this paper

This paper is structured as follows. In Section II, we present the related work. The Section III describes our datasets and coding scheme. The approach with deep learning methods for automatic coding is discussed in Section IV. Then, our experiment and results from our evaluation are described in Section V. Section VI concludes the paper.

II. RELATED WORK

Since deep learning can often outperform existing machine learning methods, such as SVMs, it has been applied in various research areas, such as image recognition and natural language processing [13]. Text classification is an important task in natural learning processing, for which various deep learning methods have been exploited extensively in recent studies. A structure called a CNN has been applied for text classification using word- or character-level modeling [14][15]. LSTM [16] and gated recurrent units (GRUs) [17] are popular structures for RNNs. Both structures are known to outperform existing models, such as n-grams, and are thus widely available as learning models for sequential data like text. RNNs are also applied to text classification in various ways [18][19]. For instance, Yang et al. used a bidirectional GRU with attention modeling by setting two hierarchical layers that consist of the word and sentence encoders [18].

In the field of CSCL, some researchers have tried to apply text classification technology to chat logs. The most representative studies would be Rosé and her colleagues'works [20][21][22]. For example, they applied text classification technology to a relatively large CSCL corpus that had been coded by human coders using the coding scheme with 7 dimensions, developed by Weinbeger and Fisher [21][23]. McLaren's Argunaut project took a similar approach: he used online discussions coded manually to train machine-learning classifiers in order to predict the appearance of these discussions characteristics in the new e-discussion[24]. However, it should be pointed out that all these prior studies rely on the machine learning techniques before deep learning studies emerge.

III. DATA AND CODING SCHEME

In this section, we explain how we collected our dataset and what coding scheme we adopted to categorize the dataset.

 ISBN: 978-1-61208-541-8

A. *Data Description*

Our dataset obtained through chat function within the system, comes from conversations among students while carrying out online collaborative learning in university lectures using CSCL, which had been previously developed by the researchers of this study [25].

This CSCL is used without face to face contact; therefore, these data are all from occasions when unacquainted and separated students formed groups within lecture halls at the campus. And within the system all names of students are shown in nicknames, so that even if students knew each other they would not recognize each other.

The overview of CSCL contributions data used in this research is shown in Table 1. The number of lectures is seven and all classes of these lectures form groups of three to four; in fact, there are a lot of data that we could not process by coding them in this research. Learning times vary depending on the class, from 45 to 90 minutes. In total, the dataset contains 11504 contributions; there are 202 groups from all the classes, with 426 participating students; since students attend multiple classes, the number of participating students are smaller than the product of number of groups and number of students in a group.

Table 2 shows a conversation example of chat. This is a conversation example of three students.

TABLE I. CONTRIBUTIONS DATA USED IN THIS STUDY

Number of Lectures	7 Lectures
Member of Groups	3-4 people
Learning Time	45-90 mintutes
Number of Groups	202 groups
Number of Students	426 students

TABLE II. CONVERSATION EXAMPLE (TRANLATION FROM JAPANESE)

Talker	Contents
D	Where do you want to change?
E	That's right … I guess, first of all, we definitely need to change the question, and then, what about the well-formed formula?
D	How is it that changes only the third line of the question?
D	Regarding the well-formed formula, it's the final part after ⊃.
E	That's good idea.
F	I agree. How do we want to change that?

B. *Coding scheme*

In accordance with our manual for code assignment, one code label is assigned to one contribution in a chat. There are 16 types of code labels as shown in Table 3, and one of those labels is assigned for all cases.

All labels in our dataset are coded by two people; the coincidence rate between the labels assigned was 67%. However, when we reviewed the resultant coding data, it was discovered that there were duplicated labels for some contributions, and some labels had variances depending on the coder; therefore, after conferring among us, we unified labels and re-coded the contributions. The resultant number of labels assigned is shown in Table 3. Concordance rate is 82.3% and this is a high concordance rate with 0.800 Kappa coefficient, and we consider this to be sufficiently practical for use as an educational dataset in deep learning methods. Fig. 1 shows the frequencies of the labels in the dataset. Nine labels describe more than 90% of occurrences; label occurrences appear to have a long-tail distribution. The main purpose of this study is to learn and infer these labels from posted contributions.

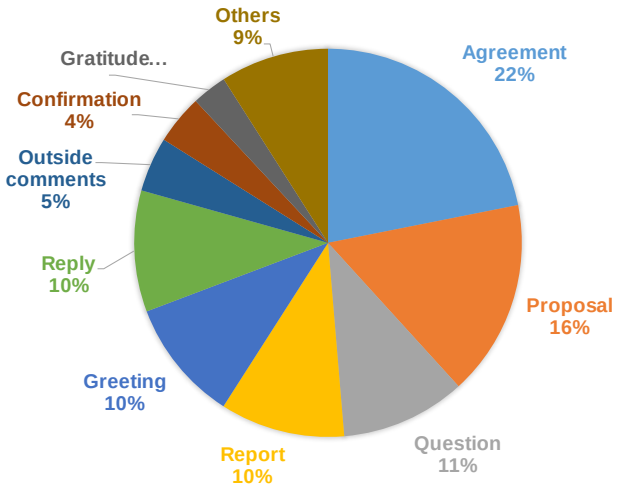


Figure 1. Ratio of each conversational coding labels

IV. APPROACH -- DEEP LEARNING

In recent years, deep learning technology has led to dramatic developments in the field of artificial intelligence. Deep learning is a general framework of learning methods that use neural networks with millions of weight parameters. The weights in neural networks are optimized so that their output coincides with labels in the given data. With the recent development of parallel computing using Graphics Processing Units (GPUs) and optimization algorithms, machines are able to learn large numbers of parameters from large datasets at realistic costs.

To try automatic coding, we adapt three types of deep neural network (DNN) structures: a convolutional neural network (CNN) -based model and two bidirectional Long short-term memory (LSTM) -based models, LSTM and Sequence-to-Sequence (Seq2Seq). The first and second models take only a single contribution as input and cannot refer to context information in the conversation. Conversely, the Seq2Seq model can capture context information by using a pair of sentences as its input, which represent source and replay contributions.

A. *CNN-based model*

The CNN-based model uses the network architecture proposed by Kim et al. (Fig. 2). Before training, all words in

TABLE III. List of labels

Tag	Meaning of tag	Contribution example	Number of times used
Agreement	Affirmative reply	I think that's good	5033
Proposal	Conveying opinion, or yes/no question	How about five of us here make the submission?	3762
Question	Other than yes/no question	What shall we do with the title?	2399
Report	Reporting own status	I corrected the complicated one	2394
Greeting	Greeting to other members	I'm looking forward to working with you	2342
Reply	Other replies	It looks that way!	2324
Outside comments	Contribution on matters other than assignment contents Opinions on systems and such	My contribution is disappearing already; so fast! A bug	1049
Confirmation	Confirm the assignment and how to proceed	Would you like to submit it now?	949
Gratitude	Gratitude to other members	Thanks!	671
Switchover	A contribution to change event being handled, such as moving on to the next assignment	Shall we give it a try?	625
Joke	Joke to other members	You should, like, learn it physically? :)	433
Request	Requesting somebody to do some task	Can either of you reply?	354
Correction	Correcting past contribution	Sorry, I meant children	204
Disagreement	Negative reply	I think 30 minute is too long	160
Complaint	Dissatisfactions towards assignments or systems	I must say the theme isn't great	155
Noise	Contribution that does not make sense	?meet? day???	143

the data are converted to word vectors. Word vectors are often obtained by pre-training using another external dataset. In this study, we implemented two types of word vectors: 1) vectors obtained by applying word2vec (the skipped gram model with negative sampling) to all Japanese text in Wikipedia, and 2) randomly initialized vectors that are tuned simultaneously with the CNN.

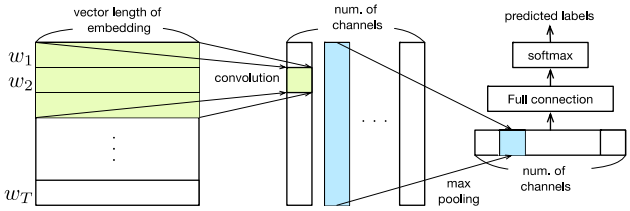


Figure 2. CNN-based model

B. *Bidirectional LSTM-based model*

An LSTM is a recurrent neural networks (RNNs) that is carefully constructed so that it can capture long-distance dependencies in sequential data. Generally speaking, an RNN consists of input vector x_t and output vector y_t for each time t. To obtain the output $y_{\{t\}}$, the previous output vector $y_{\{t-1\}}$ is fed to the neural network along with the current input vector x_t. The LSTM has another hidden vector, c_t, called the *state vector* in addition to the input and output vectors. While the state vector is also output from the neural network, it is computed to track long-distance relations through a function called a *forget gate*, which is designed to decide whether the state vector should be changed. We feed word vectors into the two-layer LSTM network sequentially in both the forward and reverse directions. After all words in a contribution are input, both output vectors are concatenated and fed into the two-layer fully-connected network and the softmax layer to obtain classification results. Fig. 3 illustrates this architecture.

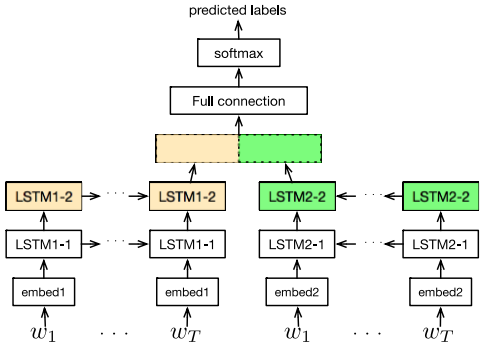


Figure 3. Bidirectional LSTM-based

C. *Bidirectional Seq2Seq-based model*

Each contribution is a part of a conversation; therefore, to classify labels more accurately, we must account for conversational contexts. To do this, we convert all contributions in conversations into pairs of *source* and *reply* contributions. Even if a user posts a contribution that does not explicitly cite another, we assume that it cites a previous contribution. We also suppose that the first contribution of each conversation cites the empty string. To construct a model that regards the source contribution as a conversational context and the reply as a representation of the user's intention, we use the Seq2seq framework. Seq2seq

 ISBN: 978-1-61208-541-8

[26] was originally proposed as a neural model using RNNs for machine translation, and later applied to other tasks, such as conversational generation [27]. It consists of two separate LSTM networks, called the encoder and decoder. We use two-layer LSTM networks for both the encoder and decoder. Words are sequentially fed in both the forward and reverse directions. Output vectors from decoders are concatenated and fed into the two-layer fully-connected network and the softmax layer (Fig. 4).

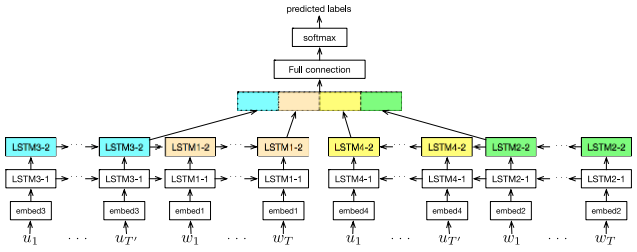


Figure 4. Bidirectional Seq2Seq-based model

V. EVALUATION

For each contribution, we trimmed sentences beginning with the symbol ">," which were automatically generated by the system. Since all the data consist of Japanese text, morphological analysis was needed. We split texts into words using a tool called MeCab. Replacing low-frequency words with "unknown," the vocabulary size was decreased to approximately 4,000. Each contribution was given two labels annotated by different people; we removed contributions that were assigned two different labels. We used 90% of the remaining 8,015 contributions as training data and 10% as test data. The accuracy of the learning result for each model is measured with the test data.

A. Baseline Methods

For comparison, we used three classifiers; Naive Bayes, a linear support vector machine (SVM), and an SVM with a radial basis function (RBF) kernel. We also used two types of feature sets: unigrams only and unigrams and bigrams. For the SVM classifiers, in order to improve the classification accuracy, input vectors were obtained by normalizing zero-one vectors whose elements represent occurrences of unigrams or bigrams.

B. Model Parameters and Learning

Model parameters, such as the vector sizes of layers, are determined as follows. Both the size of word embedding and the size of the last fully connected layer are 200 for all models. We set the patch size of the convolutional layer in the vertical direction to 4 and the number of channels to 256 for the CNN-based models. We set the size of both LSTM layers to 800 for the LSTM and Seq2Seq models.

Models are learned by stochastic descent gradient (SDG) using an optimization method called Adam. To avoid overfitting, iteration was stopped at 10 epochs for the LSTM-based methods and 30 epochs for the CNN-based methods. Due to the fluctuation in accuracy results between epochs, we took the average of the last 5 epochs to measure the accuracy of each model. To prevent overfitting, dropout was applied to the last and second-last fully connected layers.

C. Experimental Results

Table 4 shows the accuracies of the three DNN models and baseline methods. Overall, the DNN models outperform the baselines, even as the SVMs maintain their high performance. Among baseline methods, the SVM with the RBF kernel achieved the highest accuracy. For the CNN-based models, using word vectors trained using the Wikipedia data slightly enhanced accuracy. For LSTM-based models, bidirectional processing yielded slightly higher accuracy than single-directional processing.

There was no significant difference in the accuracies of the CNN model using Wikipedia and the bidirectional LSTM model. Both of these methods outperformed the best of SVMs by 1–2%.

Seq2Seq model outperformed other methods clearly; the best of SVMs by 5-6% and other DNN models by 3-4%.

TABLE IV. PREDICTIVE ACCURACIES FOR BASELINES AND DEEP-NEURAL-NETWORK MODELS

Naïve Bayes		SVM(Linear)		SVM(RBF Kernel)	
unigram	uni+bigram	unigram	uni+bigram	unigram	uni+bigram
0.554	0.598	0.642	0.659	**0.664**	0.659

CNN		LSTM		Seq2Seq	
with wikipedia	w.o. wikipedia	single-direction	bidirection	bidirection	bidir. w. interm.
0.686	0.677	0.676	0.678	**0.718**	0.717

The kappa coefficient for the bidirectional LSTM model was 0.63, which is sufficiently high. However, to automatically comprehend and judge the activities of users from only the labels inferred by machines, the kappa coefficient must be improved. By using the Seq2Seq model, which is able to capture the contextual information from the source or the adjacent contribution, the kappa coefficient was improved to 0.723.

Hereafter, we analyze the misclassification of each label individually. The precision and recall for each label are shown in Table 5. Of the ten most frequent labels, the precision of "Greeting" predictions were highest (F1: 0.94) and that of "Agreement" was the second highest (F1: 0.83). "Question" was also predicted with high accuracy (F1: 0.77). These results are consistent with our intuition, as both seem to be easy to infer from the contributions themselves, without knowing their context. In contrast, as Table 5 shows, the label "Reply" was hard for our model to predict. That performed worst with respect to the recall, tending to be misclassified as an "Agreement", "Proposal" or "Report," as shown in the confusion matrix (Fig. 5). This can be solved if richer context in neighboring contributions is used as input to classifiers in addition to the source contribution.

VI. CONCLUSION AND FUTURE WORK

As the first step to analyze collaborative process of big educational data, we tried to automate time-consuming

 ISBN: 978-1-61208-541-8

TABLE V. PRESITION AND RECALL FOR EACH LABEL (RESULT OF BI-DIRECTIONAL LSTM)

	Presition	Recall	F1-value
Agreement	0.85	0.81	0.83
Proposal	0.73	0.74	0.73
Question	0.75	0.8	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

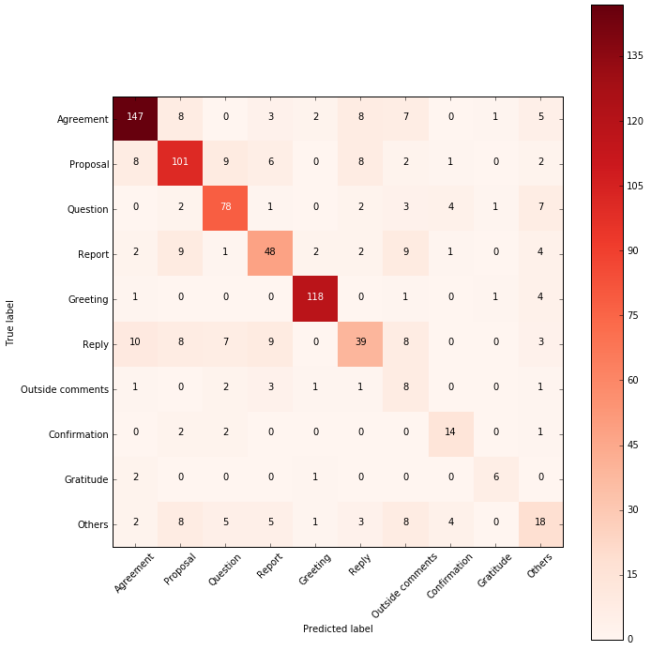


Figure-5. Confusion matrix for the Seq2S2q model.

coding task by using deep learning methods. The result was promising; our approach, particularly, Seq2Seq model outperformed other methods clearly; the best of SVMs by 5-6% and other DNN models by 3-4%. It seems that this model could obtain almost the same predictive accuracy with other coding schemes than ours, for the reason that our coding scheme is sufficiently complex with 16 labels, based not on the surface information, but on the contextual significance of each contribution.

As for the future research directions, we may have two approaches to pursue. The first approach concerns coding scheme. Our scheme, based on speech acts, was sufficiently complex, but not global. To capture the collaborative process more precisely, it will be necessary to construct a coding scheme which is more sensitive to details of interaciton and social cognitive process of learning. The second approach is about DNN models. To improve prediction accuracy, it may be effective to introduce an attention model to our DNN models. In addition, the context of conversation should be considered. To capture context more precisely, it may be necessary to construct more complex models that take multiple preceding contributions as input vectors.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number 26350289 and 16K01134.

REFERENCES

[1] G. Stahl, T. Koschmann, and D. Suthers, “Computer-supported collaborative learning,” In The Cambridge handbook of the learning science, K. Sawyer, Eds. Cambridge university press, pp.479-500, 2014.

[2] P. Dillenbourg, P. Baker, A. Blaye, and C. O’Malley, “The evolution of research on collaborative learning,” In Learning in humans and machines: Towards an interdisciplinary learning science, P. Reimann and H. Spada, Eds. Oxford: Elservier, pp. 189-211, 1996.

[3] T. Koschmann, “Understanding understanding in action,” Journal of Pragmatics, 43, pp435-437, 2011.

[4] T. Koschmann, G. Stahl, and A.Zemel, “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),” In Video reseach in the learning sciences, R. Goldman, R. Pea, B. Barron and S. Derry, Eds. Routledge, pp.133-144, 2007.

[5] M. Chi, “Quantifying qualitative analyses of verbal data : A pratical gauide ,” Journal of the Leanrnig Science, 6(3), pp.271-315, 1997.

[6] A. Meier, H. Spada, and N. Rummel, “A rating scheme for assesseing the quality of coputer-supported collaboration processes,” International Jounala of Computer Suppported Collaborative Learning, 2, pp.63-86, 2007.

[7] H. Jeong, “Verbal data analysis for understanding interacitons,” In The International Handbook of Collaborative Learning, C. Hmelo-Silver, A. M. O’Donnell, C. Chan and C. Chin, Eds. Routledge, pp.168-183, 2013.

[8] D. Persico, F. Pozzi, and L. Sarti, “Monitaring collaborative activities in computer supported learning,” Distance Education, 31(1), pp.5-22, 2010.

[9] L. Lipponen, M. Rahikainen, J. Lamillo, and K. Hakkarainen, “Patterns of participation and discourse in elementary students’computer-supported collaborative learning,”, Learning and Instruction, 13, pp.487-509, 2003.

[10] S. Schrire, “Knowledge building in asynchronous discussion groups: Going beyond quantitative analysis,” Computer & Education 46, pp.49-70, 2006.

[11] 1st Internationa Conference on Learning Analytics and Knowledge. [Online]. Avaiable from: https://tekri.athabascau.ca/analytics/ 2016.11.29

[12] B. R. Schaun and P. S. Inventado, “Educational data mining and learning analytics,” In Learning Analytics, J. A. Larusoon and B. White, Eds. Springer, pp.61-75, 2014.

[13] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, 521(7553), pp.436--444, 2015.

[14] Y. Kim, “Convolutional neural networks for sentence classification,” arXiv preprint arXiv:1408.5882, 2014.

[15] X. Zhang, J. Zhao, and Y.LeCun. “Character-level convolutional networks for text classification,” In Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS2015), pp.649-657, 2015.

 ISBN: 978-1-61208-541-8

[16] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Computation, 9(8), pp.1735--1780, 1997.

[17] J. Chung, C. Gulcehre, K. Hyun Cho and Y. Bengio, “Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling,” arXiv preprint arXiv:1412.3555, 2014.

[18] Z. Yang, et al., “Hierarchical Attention Networks for Document Classification,” In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics(NAACL2016), Human Language Technologies, 2016.

[19] D. Tang, B. Qin, and T. Liu, “Document modeling with gated recurrent neural network for sentiment classification,” In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing(EMNLP2016), pp.1422–1432, 2015.

[20] C. Rosé, et al., “Towards an interactive assessment framework for engineering design project based learning,” In Proceedings of DETC2007, 2007.

[21] C. Rosé, et al., “Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning,” International Journal of Computer Supported Collaborative Learning, 3(3), pp.237-271, 2008.

[22] G. Gweon, S. Soojin, J. Lee, S. Finger and C.Rosé, “A framework for assessment of student project groups on-line and off-line,” In Analyzing Interactions in CSCL: Methods, Approaches and Issues, S. Putambekar, G.Erkens and C. Hmelo-Silver Eds. Springer, pp.293-317, 2011.

[23] A. Weinberger and F. Fischer, “A frame work to analyze arugmetative knowledge construciton in computer-supported learning,” Computer & Education, 46(1), pp.71-95, 2006.

[24] B. McLaren, O. Scheuer, M. De Laat, H. Hever and R. De Groot, “Using machine learning techniques to analysze and support mediation of student e-discussions,” In Proceedings of artificial intelligence in education, 2007.

[25] T. Inaba and K. Ando, “Development and Evaluation of CSCL System for Large Classrooms Using Question-Posing Script,” International Journal on Advances in Software, 7(3&4),pp.590-600, 2014.

[26] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” arXiv preprint arXiv, pp.1409.0473, 2014.

[27] O. Vinyals and Q. V. Le, “ A Neural Conversational Mode,” arXiv preprint arXiv:1506.05869, (ICML Deep Learning Workshop 2015), 2015.

 ISBN: 978-1-61208-541-8

論　文

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

安藤公彦・柴田千尋・稲葉竹俊

抄録

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た。その結果，開発手法が既存の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った。また，大規模な協調学習データを対象にした，リアルタイムでの協調プロセスの解析と教育的介入の実現可能性が示唆された。

◎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

Analysis of Collaborative Learning Processes by Automatic Coding Using Deep Learning Technology

Kimihiko Ando, Chihiro Shibata, Taketoshi Inaba

Abstract

In Computer Supported Collaborative Learning research, gaining a guideline to carry out appropriate scaffolding by analyzing mechanism of successful collaborative interaction and extracting indicators to identify groups where collaborative process is not going well, can be considered as the most important preoccupation, both for research and for educational implementation. For the process analysis, coding and statistical analysis to chat data are often adopted as a research method. In this research, we developed a method for automating highly accurate coding by deep learning technology, and its accuracy and effectiveness was evaluated. As a result, it became clear that our method realizes the high accuracies outperforming the machine learning baselines. In addition, the feasibility of analysis of collaborative processes and instructional intervention in real time, was suggested.

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

1　序論

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の一つは，グループ内での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，その社会的プロセスを社会構成主義的な観点から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。もちろん各グループやメンバーごとの発言数，また場合によってはシステムのインターフェース（sentence opener など）から取得される発言属性等の利用可能な定量的データがあるが，これらはきわめて表面的なデータにすぎない。最も重要なデータはチャットの発言，スカイプ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会話分析，ビデオ分析などのエスノメソドロジーが援用されてきた[1][2]。

しかし，これらの研究はその性質上，限られた数のグループの協調活動を対象とした 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[3]。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。既存研究においても，協調学習データのコーディング支援を試みたシステムは存在している。これらの研究では，コーディング自体は人力によって行わ

連絡先：東京工科大学クラウドサービスセンター
Contact：ando@stf.teu.ac.jp

れるものと[4]，機械学習の技術を用いて行ったものがある[5][6]。本研究では，機械学習による既存の自動コーディングではまだ実用に耐えるだけの精度には達していない点に注目し，深層学習技術によって，大規模な協調学習を対象にコーディングを自動化する手法を模索し，既存の研究によって示された精度を凌駕することをめざす。

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つの検証を行う。

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する。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）との精度比較を行う。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する。

2　検証1：データとコーディングスキーム

2.1　会話データセット

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。本研究で利用する発言データ元のCSCLの利用状況をTable 1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

2.2　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，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著者らで精査したところ，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，著者ら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。Fig. 1にデータセットのラベルの割合を示す。

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？ (ˆ・ω・ˆ)

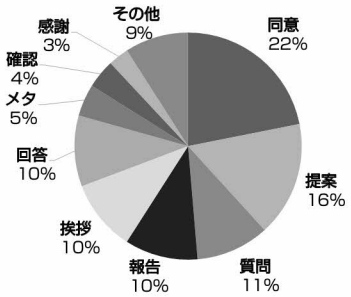


Fig. 1　コーディングラベルの分布

2.3　深層学習を用いた自動コーディング手法

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，(1) 畳み込みニューラルネット（CNN）による分類モデル，(2) 長短期記憶（LSTM）による分類モデル，(3) 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[7]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，著者達の既存論文を参照されたい[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。

2.4.2　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。

モデルの学習後，発言の解析，即ちコーディングラベルを得るに必要な時間を計測すると，GPU を用いた場合は，1 発言あたりの平均で 77msec，CPU のみを用いた場合は，143msec であった。

2.4.3　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 章で提示した 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。

3.1　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

3.2　自動コーディング結果

Fig. 2 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。

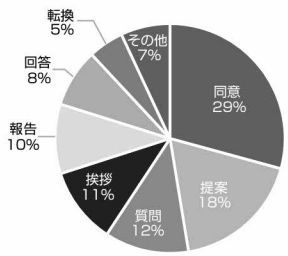


Fig. 2　自動コーディング結果の割合

3.3　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。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

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	全発言
全体	0.17	-0.16	0.04	0.09	**0.37**	0.05	-0.09	0.07
具体性	0.16	-0.08	0.08	0.00	**0.37**	0.10	-0.12	0.09
工夫	0.18	0.02	0.18	**0.20**	**0.38**	**0.26**	-0.08	**0.24**
適切性	-0.02	-0.04	-0.09	0.03	**0.32**	-0.01	-0.02	-0.01

た。変動係数が高いと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」の発言を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1人が担っていると考えられ，課題のほとんどをその1人が行うなどグループとしての機能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，1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！

3.4　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2つの課題を追求していく。まず，コーディングスキームの再検討が急務となる。本研究のスキームはスピーチアクトに基づく，十分に複雑性を有するものではあったが，協調プロセス全体をとらえるような包括的なものではなかった。社会的認知プロセスの詳細を表現できるスキームの構築が求められる。次に，深層学習手法の再検討を行う必要がある。予測の精度を向上させるため，会話の文脈をさらなる考慮の対象とすべきである。そのため，複数の先行する発言を入力ベクトルとする，より複雑なモデルを構築する必要があると思われる。

謝辞

本研究の一部は，科研費（26350298及び16K01134）の助成を受けたものである。

参考文献

[1] Koschmann, T, “Understanding in action”, Journal of Pragmatics, 43, 2011, pp435-437.

[2] Koschmann, T., Stahl, G., and Zemel, A., “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, Routledge, 2007, pp.133-144.

[3] Chi, M., “Quantifying qualitative analyses of verbal data: A practical guide ”, Journal of the Learning Science, 6 (3), 1997, pp.271-315.

[4] Persico, D., Pozzi, F. and Sarti, L., “Monitoring collaborative activities in computer supported collaborative learning”, Distance Education, 31 (1), 2010, pp.5-22.

[5] Rosé, C. P., Gweon, G., Arguello, J., et al., “Towards an interactive assessment framework for engineering design project based learning”, Proceedings of DETC2007, 2007.

[6] Rosé, C. P., Wang, Y., Cui, Y., et al., “Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning”, International Journal of Computer Supported Collaborative Learning, 3 (3), 2008, pp.237-271.

[7] Sutskever, I., Vinyals, O., Le, Q. V, “Sequence to Sequence Learning with Neural Networks”, Advances in Neural Information Processing Systems 27, 2014, pp. 3104—3112.

[8] 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.

2017. 9. 1 受理　2017. 10. 13 掲載決定

著者略歴

安藤公彦（あんどう　きみひこ）
◎現在の所属：東京工科大学クラウドサービスセンター
◎専門分野：学習支援システム，学習管理システム

柴田千尋（しばた　ちひろ）
◎現在の所属：東京工科大学コンピュータサイエンス学部
◎専門分野：機械学習，並列分散処理

稲葉竹俊（いなば　たけとし）
◎現在の所属：東京工科大学教養学環
◎専門分野：コンピュータ支援協調学習，e ラーニング

Coding Collaboration Process Automatically: Coding Methods Using Deep Learning Technology

Kimihiko Ando
Cloud Service Center
Tokyo University of Technology
Tokyo, Japan
email:ando@stf.teu.ac.jp

Chihiro Shibata
School of Computer Sciences
Tokyo University of Technology
Tokyo, Japan
email:shibatachh@stf.teu.ac.jp

Taketoshi Inaba
Graduate School of Bionics, Computer and Media Sciences
Tokyo University of Technology
Tokyo, Japan
email:inaba@stf.teu.ac.jp

Abstract— **In Computer Supported Collaborative Learning (CSCL) research, gaining a guideline to carry out appropriate scaffolding by analyzing mechanism of successful collaborative interaction and extracting indicators to identify groups where collaborative process is not going well, can be considered as the most important preoccupation, both for research and for educational implementation. And to study this collaborative learning process, different approaches have been tried. In this paper, we opt for the verbal data analysis; the advantage of this method is that it enables quantitative processing while maintaining qualitative perspective, with collaborative learning data of considerable size. However, coding large scale educational data is extremely time consuming and sometimes goes beyond men's capacity. So, in recent years, there have also been attempts to automate complex coding by using machine learning technology. In this background, with large scale data generated in our CSCL system, we have tried to implement automation of high precision coding utilizing deep learning methods, which are derived from the leading edge technology of machine learning. The results indicate that our approach with deep learning methods is promising, outperforming the machine learning baseline. But the prediction accuracy could be improved by constructing coding schemes and models more sensitive to the context of collaboration and conversation. Therefore, we propose a new coding scheme that can represent the context of learning more comprehensively and accurately at the end of this paper for the next research.**

Keywords-CSCL; leaning analytics; coding scheme; deep learning methods.

I. INTRODUCTION

This article is an extended version of a conference paper presented at eLmL 2017, the Ninth International Conference on Mobile, Hybrid and On-line Learning [1]. It introduces more information on the theoretical background of this study and especially a new coding scheme, based on the experiment results.

A. Analysis of collaborative process

One of the greatest research interests in the actual Computer Supported Collaborative Learning (CSCL) research is to analyze its social process from a social constructionist viewpoint, and key research questions are as follows: how knowledge and meanings are shared within a group, what types of conflict, synchronization and adjustment of opinions occur, and how knowledge is constructed from discussions. And answering to these questions enables to develop more effective scaffolding methods and CSCL system and tools.

In earlier researches at initial stage of CSCL, the focus was on each individual within a collaborating group, and the main point of interest had been how significantly a personal learning outcome was affected by characteristic types of a group (such as group size, group composition, learning tasks, and communication media) [2]. However, it gradually became clear that those characteristics are complexly connected and intertwined with each other, and showing causal relation to a specific result was extremely difficult. From the 1990s, the interest in CSCL research had moved away from awareness of the issue on how a personal learning is established within a group, to attempting to explain the process by clarifying the details of group interactions when learning is taking place within a group [3].

However, attempting to analyze collaborative process goes beyond merely shifting a research perspective; it also leads to fundamental re-examination of its analytical methodology. In other words, this involves a shift from quantitative analysis to qualitative analysis. Naturally, there are useful data among quantitative data saved within CSCL system, such as the number of contributions within a group, the number of contributions by each group member, and in some cases contribution attributes obtained from system interface (sentence opener), but those are very much a mere surface data. The most important data for analysis are contributions in chats, images/sounds within tools such as Skype, and various outputs generated in the process of

collaborative learning; for analysis of those, ethnomethodologies such as conversation analysis and video analysis have been invoked [4][5].

However, those researches by their very nature tend to be in-depth case studies of collaborative activities with a limited number of groups and have the disadvantage of not at all being easy to derive a guideline that has a certain level of universality and can be applicable in other contexts. Therefore, researches have been carried out using verbal data analysis method that carry out coding from a perspective of linguistic or collaborative learning activities on a certain volume of language data generated in collaborative learning and analyzing them [6][7][8]. The advantage of this method is that it enables quantitative processing while maintaining qualitative perspective, with collaborative learning data of considerable size as the subject, while coding them manually is an extremely time consuming task, which goes sometimes beyond men's capacity. For example, Persico et al. developed a technological tool which helps the tutors to code the contributions in chats and displays quantitative information about the qualitative information and coding data [9]. However, given that the coding procedure itself remains manual in most existing studies [10][11], there is an insurmountable limit in front of big data. Hence, we seek an automatic coding technique for a large scale collaborative learning data with deep learning methods.

B. Educational data and Learning Analytics

With the progress of educational cloud implementation in educational institutions, data generated in Learning Management System (LMS), e-learning, Social Network Service (SNS), Massive Open Online Course (MOOC) and others are increasing rapidly, and a new research approach called Learning Analytics (LA) that tries to gain knowledge that would lead to support of learning and educational activities by analyzing those educational big data is becoming more active [12][13]. Big educational data obtained from CSCL system integrated in educational cloud at a campus, such as conversation data, submitted documents and images/sounds of learning activities, will certainly become a subject for analysis in the near future: therefore, it is believed that we are coming into a time when it is necessary to seriously examine a new possibility of collaborative learning research as LA. Due to such background, in this research we have reconstructed CSCL system that has been operating in a campus server for the last five years as a module within Moodle, which is a LMS within the campus cloud, and have already structured an environment that can be operated within the campus and collect/analyze collaborative learning data.

C. The goal and purpose of this study

The goal of our research is to analyze large-scale collaborative data from the perspective of LA as described above and discover the mechanism of activation and deactivation of collaborative activity process which could not be gained from micro level case studies up to now. Furthermore, this research, based on its results, aims to implement supports in authentic learning/educational contexts, such as real-time monitoring of collaborative process and scaffolding to groups that are not becoming activated.

In this paper, as the first step towards this goal, we present work in progress, which attempts to develop an automation technique for coding of chat data and verifies its accuracy. To be more specific, a substantial volume of chat data is coded manually, and has a part of that learnt as training data in deep learning methods, which are derived from the leading edge technologies for machine learning; afterwards, automatic coding of the raw data is carried out. For validation of accuracy, the effectiveness of using deep learning methods is assessed by comparing accuracy against Naive Bayes and Support Vector Machines, which are baselines of machine learning algorithm used in existing studies that carried out automatic coding by machine learning.

D. Structure of this paper

This paper is structured as follows. In Section II, we present the related work. The Section III describes our datasets and coding scheme. The approach with deep learning methods for automatic coding is discussed in Section IV. Then, our experiment and results from our evaluation are described in Section V. In Section VI, taking account of experimental results, we propose a new coding scheme. Section VI concludes the paper.

II. RELATED WORK

Since deep learning can often outperform existing machine learning methods, such as SVMs, it has been applied in various research areas, such as image recognition and natural language processing [14]. Text classification is an important task in natural learning processing, for which various deep learning methods have been exploited extensively in recent studies. A structure called a CNN has been applied for text classification using word- or character-level modeling [15][16]. LSTM [17] and gated recurrent units (GRUs) [18] are popular structures for RNNs. Both structures are known to outperform existing models, such as n-grams, and thus are widely available as learning models for sequential data like text. RNNs are also applied to text classification in various ways [19][20]. For instance, Yang et al. used a bidirectional GRU with attention modeling by setting two hierarchical layers that consist of the word and sentence encoders [19].

In the field of CSCL, some researchers have tried to apply text classification technology to chat logs. The most representative studies would be Rosé and her colleagues' works [21][22][23]. For example, they applied text classification technology to a relatively large CSCL corpus that had been coded by human coders using the coding scheme with multiple dimensions, developed by Weinbeger and Fisher [22][24]. McLaren's Argunaut project took a similar approach: he used online discussions coded manually to train machine-learning classifiers in order to predict the appearance of these discussions characteristics in the new e-discussion [25]. However, it should be pointed

out that all these prior studies rely on the machine learning techniques before deep learning studies emerge.

III. DATA AND CODING SCHEME

In this section, we explain how we collected our dataset and what coding scheme we adopted to categorize the dataset.

A. Data Description

Our dataset obtained through chat function within the system, comes from conversations among students while carrying out online collaborative learning in university lectures using CSCL, which had been previously developed by the researchers of this study [26].

This CSCL is used without face to face contact; therefore, these data are all from occasions when unacquainted and separated students formed groups within lecture halls at the campus. And within the system all names of students are shown in nicknames, so that even if students knew each other they would not recognize each other.

The overview of CSCL contributions data used in this research is shown in Table I. The number of lectures is seven and all classes of these lectures form groups of three to four; in fact, there are a lot of data that we could not process by coding them in this research. Learning times vary depending on the class, from 45 to 90 minutes. In total, the dataset contains 11504 contributions; there are 202 groups from all the classes, with 426 participating students; since students attend multiple classes, the number of participating students are smaller than the product of number of groups and number of students in a group.

Table II shows a conversation example of chat. This is a conversation example of three students.

TABLE I. CONTRIBUTIONS DATA USED IN THIS STUDY

Number of Lectures	7 Lectures
Member of Groups	3-4 people
Learning Time	45-90 mintutes
Number of Groups	202 groups
Number of Students	426 students

TABLE II. CONVERSATION EXAMPLE (TRANLATION FROM JAPANESE)

Talker	Contents
D	Where do you want to change?
E	That's right … I guess, first of all, we definitely need to change the question, and then, what about the well-formed formula?
D	How is it that changes only the third line of the question?
D	Regarding the well-formed formula, it's the final part after ⊃.
E	That's good idea.
F	I agree. How do we want to change that?

B. Coding scheme

In accordance with our manual for code assignment, one code label is assigned to one contribution in a chat. There are 16 types of code labels as shown in Table III, and one of those labels is assigned for all cases.

All labels in our dataset are coded by two people; the coincidence rate between the labels assigned was 67%. However, when we reviewed the resultant coding data, it was discovered that there were duplicated labels for some contributions, and some labels had variances depending on the coder; therefore, after conferring among us, we unified labels and re-coded the contributions. The resultant number of labels assigned is shown in Table III. Concordance rate is 82.3% and this is a high concordance rate with 0.800 Kappa coefficient, and we consider this to be sufficiently practical for use as an educational dataset in deep learning methods. Fig. 1 shows the frequencies of the labels in the dataset. Nine labels describe more than 90% of occurrences; label occurrences appear to have a long-tail distribution. The main purpose of this study is to learn and infer these labels from posted contributions.

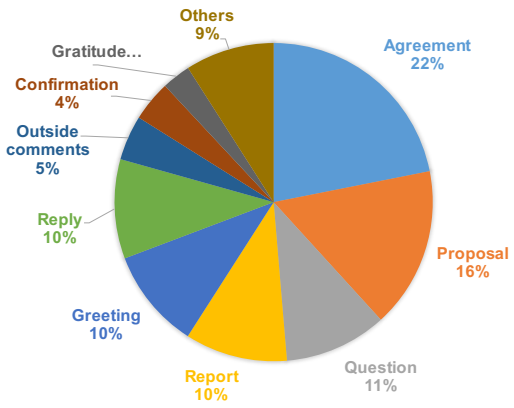


Figure 1. Ratio of each conversational coding labels

IV. APPROACH –DEEP LEARNING

In recent years, deep learning technology has led to dramatic developments in the field of artificial intelligence. Deep learning is a general framework of learning methods that use neural networks with millions of weight parameters. The weights in neural networks are optimized so that their output coincides with labels in the given data. With the recent development of parallel computing using Graphics Processing Units (GPUs) and optimization algorithms, machines are able to learn large numbers of parameters from large datasets at realistic costs.

To try automatic coding, we adapt three types of deep neural network (DNN) structures: a convolutional neural network (CNN) based model and two bidirectional Long short-term memory (LSTM) based models, LSTM and Sequence-to-Sequence (Seq2Seq). The first and second models take only a single contribution as input and cannot refer to context information in the conversation. Conversely, the Seq2Seq model can capture context information by using

TABLE III. List of labels

Label	Meaning of label	Contribution example	Number of times used
Agreement	Affirmative reply	I think that's good	5033
Proposal	Conveying opinion, or yes/no question	How about five of us here make the submission?	3762
Question	Other than yes/no question	What shall we do with the title?	2399
Report	Reporting own status	I corrected the complicated one	2394
Greeting	Greeting to other members	I'm looking forward to working with you	2342
Reply	Other replies	It looks that way!	2324
Outside comments	Contribution on matters other than assignment contents	My contribution is disappearing already; so fast!	1049
	Opinions on systems and such	A bug	
Confirmation	Confirm the assignment and how to proceed	Would you like to submit it now?	949
Gratitude	Gratitude to other members	Thanks!	671
Switchover	A contribution to change event being handled, such as moving on to the next assignment	Shall we give it a try?	625
Joke	Joke to other members	You should, like, learn it physically? :)	433
Request	Requesting somebody to do some task	Can either of you reply?	354
Correction	Correcting past contribution	Sorry, I meant children	204
Disagreement	Negative reply	I think 30 minute is too long	160
Complaint	Dissatisfactions towards assignments or systems	I must say the theme isn't great	155
Noise	Contribution that does not make sense	?meet? day???	143

a pair of sentences as its input, which represent source and replay contributions.

A. CNN-based model

The CNN-based model uses the network architecture proposed by Kim et al. (Fig. 2). Before training, all words in the data are converted to word vectors. Word vectors are often obtained by pre-training using another external dataset. In this study, we implemented two types of word vectors: 1) vectors obtained by applying word2vec (the skipped gram model with negative sampling) to all Japanese text in Wikipedia, and 2) randomly initialized vectors that are tuned simultaneously with the CNN.

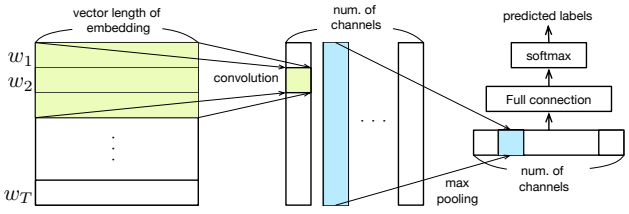


Figure 2. CNN-based model

B. Bidirectional LSTM-based model

An LSTM is a recurrent neural networks (RNNs) that is carefully constructed so that it can capture long-distance dependencies in sequential data. Generally speaking, an RNN consists of input vector x_t and output vector y_t for each time t. To obtain the output $y_{\{t\}}$, the previous output vector $y_{\{t-1\}}$ is fed to the neural network along with the current input vector x_t. The LSTM has another hidden vector, c_t, called the *state vector* in addition to the input and output vectors. While the state vector is also output from the neural network, it is computed to track long-distance relations through a function called a *forget gate,* which is designed to decide whether the state vector should be changed. We feed word vectors into the two-layer LSTM network sequentially in both the forward and reverse directions. After all words in a contribution are input, both output vectors are concatenated and fed into the two-layer fully-connected network and the softmax layer to obtain classification results. Fig. 3 illustrates this architecture.

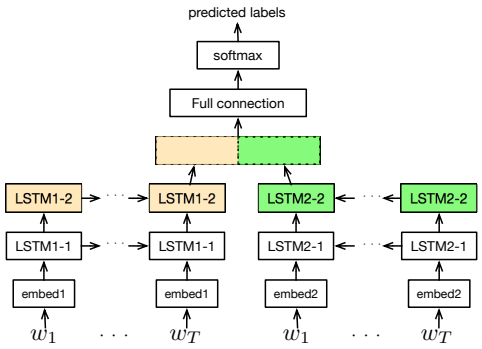


Figure 3. Bidirectional LSTM-based

C. Bidirectional Seq2Seq-based model

Each contribution is a part of a conversation; therefore, to classify labels more accurately, we must account for conversational contexts. To do this, we convert all

contributions in conversations into pairs of *source* and *reply* contributions (Table IV). Even if a user posts a contribution that does not explicitly cite another, we assume that it cites a previous contribution. We also suppose that the first contribution of each conversation cites the empty string. To construct a model that regards the source contribution as a conversational context and the reply as a representation of the user's intention, we use the Seq2seq framework. Seq2seq [27] was originally proposed as a neural model using RNNs for machine translation, and later applied to other tasks, such as conversational generation [28]. It consists of two separate LSTM networks, called the encoder and decoder. We use two-layer LSTM networks for both the encoder and decoder. Words are sequentially fed in both the forward and reverse directions. Output vectors from decoders are concatenated and fed into the two-layer fully-connected network and the softmax layer (Fig. 4).

TABLE IV. Examples for source and replay contributions

Source (u)	Replay (w)	Label
(None)	How about five of us here make the submission?	**Proposal**
(None)	I must say the theme isn't great.	**Com-plaint**
How about five of us here make the submission?	It sounds great!	**Reply**
I must say the theme isn't great.	If we had another hour, we could change it…	**Agree-ment**
It sounds great!	Thanks!	**Gratitude**

V. EVALUATION

A. Data Preprocessing

For each contribution, we trimmed sentences beginning with the symbol ">," which were automatically generated by the system. Since all the data consist of Japanese text, morphological analysis was needed. We split texts into words using a tool called MeCab [29]. Replacing low-frequency words with "unknown," the vocabulary size was decreased to approximately 4,000. Each contribution was given two labels annotated by different people; we removed contributions that were assigned two different labels. We used 90% of the remaining 8,015 contributions as training data and 10% as test data. The accuracy of the learning result for each model is measured with the test data.

B. Baseline Methods

For comparison, we used three classifiers; Naive Bayes, a linear support vector machine (SVM), and an SVM with a radial basis function (RBF) kernel. We also used two types of feature sets: unigrams only and unigrams and bigrams. For the SVM classifiers, in order to improve the classification accuracy, input vectors were obtained by normalizing zero-one vectors whose elements represent occurrences of unigrams or bigrams.

C. Model Parameters and Learning

Model parameters, such as the vector sizes of layers, are determined as follows. Both the size of word embedding and the size of the last fully connected layer are 200 for all models. We set the patch size of the convolutional layer in the vertical direction to 4 and the number of channels to 256 for the CNN-based models. We set the size of both LSTM layers to 800 for the LSTM and Seq2Seq models. The set of parameters were needed to be chosen so that their prediction accuracy of the model will not be reduced, and at the same time, the computational cost of learning is in the range of reasonable time. Generally, the vector size of LSTM layers is needed to be increased for better prediction accuracy when it is inappropriately small. On the other hand, if it is sufficiently large, increasing their size is almost in vain for better accuracy. For instance, if we set it larger than that of our setting, say 1000 or 2000, we will get almost the same value of accuracy as the result of the experiment. Thus, we empirically decided it so as to achieve the nearly optimal accuracy and to minimize computational cost. Meanwhile, we need to carefully choose the vector size of the last fully connected layer. Our model easily suffers from over fitting if we set it too large. On the other hand, if we set it too small, our model is suffered from the lack of the expression capability. Thus, we should set it moderately; not so small to

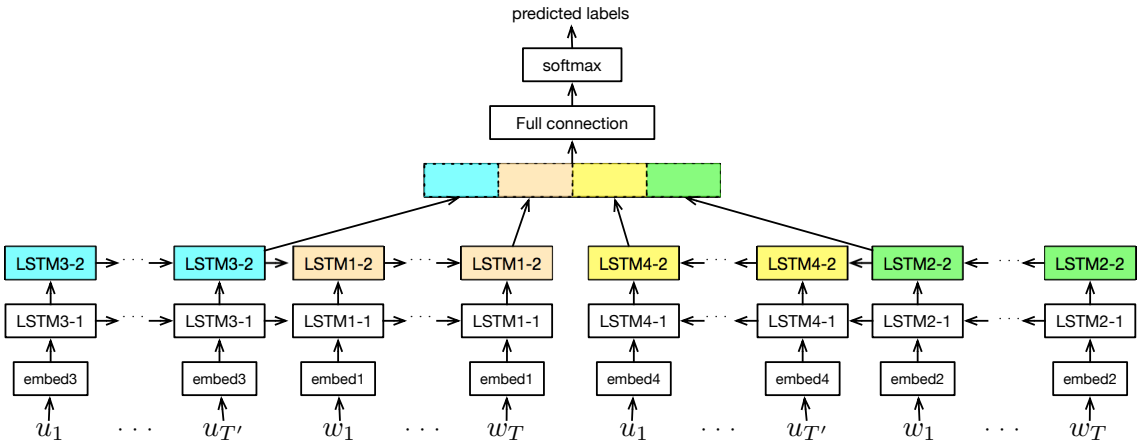


Figure 4. Bidirectional Seq2Seq-based model

have the sufficient capability to learn accurately, and not so large to avoid the over fitting problem. We obtained 200 as an appropriate value for the vector size of the last layer through several experiments.

Models are learned by stochastic descent gradient (SDG) using an optimization method called Adam. To avoid overfitting, iteration was stopped at 10 epochs for the LSTM-based methods and 30 epochs for the CNN-based methods. Due to the fluctuation in accuracy results between epochs, we took the average of the last 5 epochs to measure the accuracy of each model. To prevent overfitting, dropout was applied to the last and second-last fully connected layers. Figure 5 shows the learning curves of the CNN-based model with Wikipedia and the bi-directional Seq2Seq-based model. The y-axis shows the accuracy on the test data. As the figure shows, the accuracy converges approximately after around 10 epochs for the Seq2Seq-based model. On the other hand, it converges after around 30 epochs. The numbers of epochs that are needed for convergence largely depend on the models.

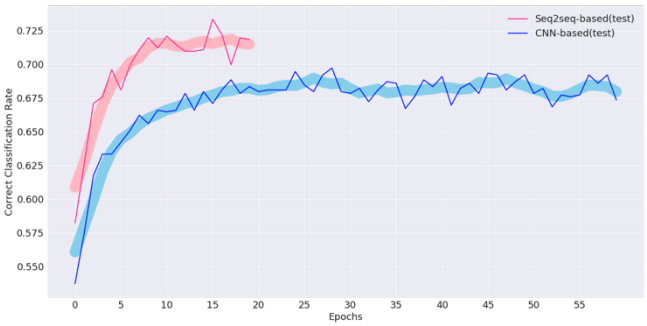


Figure 5. Learning curves of Seq2Seq-based and CNN-based models

D. Experimental Results

Table V shows the accuracies of the three DNN models and baseline methods. Overall, the DNN models outperform the baselines, even as the SVMs maintain their high performance. Among baseline methods, the SVM with the RBF kernel achieved the highest accuracy. For the CNN-based models, using word vectors trained using the Wikipedia data slightly enhanced accuracy. For the LSTM-based models, bidirectional processing yielded slightly higher accuracy than single-directional processing.

TABLE V. PREDICTIVE ACCURACIES FOR BASELINES AND DEEP-NEURAL-NETWORK MODELS

Naïve Bayes		SVM(Linear)		SVM(RBF Kernel)	
unigram	*uni+bigram*	*unigram*	*uni+bigram*	*unigram*	*uni+bigram*
0.554	0.598	0.642	0.659	**0.664**	0.659

CNN		LSTM		Seq2Seq	
with wikipedia	*w.o. wikipedia*	*single-direction*	*bidirection*	*bidirection*	*bidir. w. interm.*
0.686	0.677	0.676	0.678	**0.718**	0.717

There was no significant difference in the accuracies of the CNN model using Wikipedia and the bidirectional LSTM model. Both of these methods outperformed the best of SVMs by 1-2%.

The Seq2Seq model outperformed other methods clearly; the best of SVMs by 5-6% and other DNN models by 3-4%.

The kappa coefficient for the bidirectional LSTM model was 0.63, which is sufficiently high. However, to automatically comprehend and judge the activities of users from only the labels inferred by machines, the kappa coefficient must be improved. By using the Seq2Seq model, which is able to capture the contextual information from the source or the adjacent contribution, the kappa coefficient was improved to 0.723.

Hereafter, we analyze the misclassification of each label individually. The precision and recall for each label are shown in Table VI. Of the ten most frequent labels, the precision of "Greeting" predictions were highest (F1: 0.94) and that of "Agreement" was the second highest (F1: 0.83).

TABLE VI. PRESITION AND RECALL FOR EACH LABEL (RESULT OF BI-DIRECTIONAL LSTM)

Label	Presition	Recall	F1-Value
Agreement	0.85	0.81	0.83
Proposal	0.73	0.74	0.73
Question	0.75	0.8	0.77
Report	0.64	0.62	0.63
Greeting	0.94	0.94	0.94
Reply	0.62	0.46	0.53
Outside Commnets	0.17	0.47	0.25
Confirmation	0.58	0.74	0.65
Gratitude	0.67	0.67	0.67

"Question" was also predicted with high accuracy (F1: 0.77). These results are consistent with our intuition, as both seem to be easy to infer from the contributions themselves, without knowing their context. In contrast, as Table VI shows, the label "Reply" was hard for our model to predict. That performed worst with respect to the recall, tending to be misclassified as an "Agreement", "Proposal" or "Report," as shown in the confusion matrix (Fig. 6). This can be solved if richer context in neighboring contributions is used as input to classifiers in addition to the source contribution.

VI. NEW CODING SCHEME

As indicated in some case that Replay may include a meaning of Agree in the coding scheme based on speech acts used in the current study, the fact that the definition of one label may sometimes overlap the definition of another label has become a factor making it difficult to assign a label always with accuracy and reliability just in artificial intelligence coding but also in manual coding as well. In addition to these technical problems, more importantly, labels based on speech acts which express the linguistic characteristics of the conversation are insufficient for the analysis of the learning process. With this single linguistic scheme, one can not clearly realize whether members of a group engage in activities to solve the task, how members

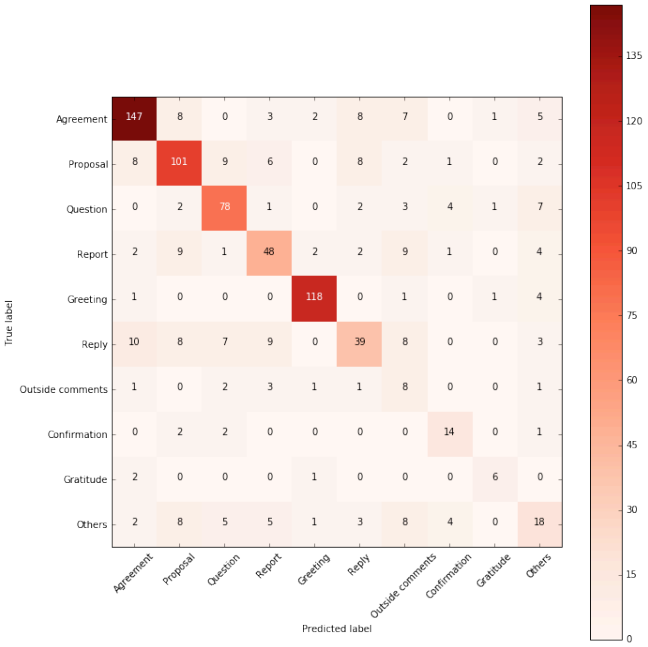


Figure 6. Confusion matrix for the Seq2S2q model.

coordinate each other in terms of task division, time management, etc. during their collaboration, how each member constructs his argument, how members discuss and negotiate each other. From those described above, we propose a new coding scheme so that the automated coding accuracy will improve and that we may understand more accurately and globally collaborative process.

Our new coding scheme is constructed based on the multi-dimensional coding scheme proposed by Weinberger et Fischer who try to analyze whole samples of discourse corpora on multiple process dimensions and "better understand how specific processes of computer-supported collaborative learning contribute to and improve individual acquisition of knowledge" [24]. As shown in Table VII, our scheme consists of five dimensions, while Winberger and Fischer's one has four dimensions without Coordination dimension. We provide labels basically regarding a statement in a chatting as a unit similarly to way we used in the study. In addition, while such values as number of statements are provided as Participation dimension labels, those in other four dimensions are provided by selecting one label from among multiple labels. In other words, since one label is given for each dimension for one statement, a plurality of labels will be assigned to one statement. Therefore, the coding work with this scheme is extremely complicated and takes a lot of time, but the merit of automated coding is even greater. Each dimension is described in detail below.

A. *Participation dimension*

As shown in Table VIII, Participation dimension is for measuring participation frequency in argumentation. Since this dimension is defined as quantitative data mainly including number of statements, number of letters of statements, time for and interval of statements, there is no need for neither manual nor artificial intelligence coding, requiring a coding just by statistical processing on a database.

Even though Participation dimension labels are capable of analyzing quantitatively different aspects of participation in conversations since they work on specific number of statements or the like, they are incapable of qualitatively analyzing such as whether the contribution has contributed to problem solving.

TABLE VII. NEW CODING SCHEME

Dimension	Description
Participation	Frequency of participation in argumentation
Epistemic	How to be directly involved in problem solving
Argumentation	Ideal assertion in argumentation
Social	How to cope with others' statements
Coordination	How to coordinate to advance discussion smoothly

B. *Epistemic dimension*

This dimension represents whether each statement is directly related to problem solving as a task and the labels are classified as shown in the table below depending on contents of statements. Labels of this dimension are provided to all statements.

Weinberger and Fischer's scheme has 6 categories to code epistemic activities which consist in applying the theoretical concepts to case information. But, as shown in Table IX, we set only two categories here, because we want to give generality that we can handle as many problem solving types as possible.

TABLE VIII. PARTICIPATION DIMENSION

Category	Description
Number of statements	Number of statements of each member during sessions
Number of letters of a statement	Number of letters during a single speech
Time for statement	Time used for a statement
Interval of statements	Time elapsed since last statement
Statements distribution	Standard deviation of each member within a group

TABLE IX. LABELS IN EPISTEMIC DIMENSION

Label	Description
On Task	Statements directly related to problems
Off Task	Statements without any relationship with problems

"On Task" here indicates such statements which are directly related with assigned problem solving and statements with any of contents described below are regarded as "Off Task."

- Statements asking meaning of problems and how to advance them
- Statements to allocate tasks
- Statements regarding the system

Labels in Epistemic dimension are regarded to be the most basic ones for qualitative analysis since they represent whether they are directly involved in problem solving. For example, it is understood that almost no effort has been made on a problem if there is less "On task" labels.

Besides, Argumentation and Social dimension labels as referred to in the next section and beyond are provided only if Epistemic dimension is "On Task" and those in coordination dimension are provided only if Epistemic dimension is "Off Task."

C. Coordination dimension

Labels of Coordination dimension are provided only if Epistemic dimension labels are "Off Task" and the statements are not directly but indirectly involved in problems. While a list of Coordination dimension labels is shown in Table X, labels are provided not to all of statements of "Off task" but only one label is provided to any statement which falls under the label. For responses to statements to which Coordination dimension labels are provided, those in the same Coordination dimension are provided.

"Task division" here refers to a statement to decide who to work on which task requiring division of tasks for advancing problem solving. "Time management" is a statement to coordinate degree of progress in problem solving, and for example, such statements fall under the definition that "let's check it until 13 o'clock," and "how has it been in progress?" "Meta statement" refers to a statement for clarifying what the problem is when intention and meaning of the problem is not understood. "Technical coordination" refers to questions and opinions about how to use the CSCL System.

TABLE X. LABELS OF COORDINATION DIMENSION

Label	Description
Task division	Allotment of tasks
Time management	Check of temporal and degree of progress
Meta statement	Questions to ask meaning of problems
Technical coordination	How to use the system, etc.

Since Coordination dimension labels are provided to statements for executing problem solving smoothly, it is believed to be possible to predict progress in arguments by analyzing the timing that the labels were provided. In case of less Coordination dimension labels recognized, it is also predicted that smooth relationships have not been built up within the groups.

In a case that a lot of these labels have been provided in many groups, on the other hand, it is assumed that there is some sort of defect in contents of the problems or systems.

In addition, it should be noted that this dimension is not set in Weinberger and Fischer's scheme.

D. Argument dimension

Labels of Argument dimension are provided to all statements when Epistemic labels are "On Task", indicating attributes such as whether each statement includes the speaker's opinion and whether the opinion is based on any ground. Labels of this dimension are provided to just one statement content without considering whether any ground was described in other statement.

A list of Argument dimension labels is shown in Table XI. Here, presence/absence of grounds is determined whether any ground to support the opinion is presented or not but it does not matter whether the presented ground is reliable or not. A qualified claim represents whether it is asserted that presented opinion is applied to all or part of situations to be worked on as a task. "Euphemism" indicates such statements with low confidence rating that presented opinion is just a prediction or shows only possibility. "Non-Argumentative moves" refer to statements without including any opinion and simple questions are also included in this tag.

Labels in Argument dimension are capable of analyzing the logical consistency of statement contents. For example, if a statement is filled just with "Simple Claim" it is assumed as a superficial argument.

In comparison with Weinberger and Fischer's scheme, we introduce a new label "Euphemism". But we do not set for now the categories of macro-level dimension in which single arguments are arranged in a line of argumentation such as arguments, counterarguments, reply, for the reason that it seems difficult that the automatic coding by deep learning methods for this macro dimension works correctly.

TABLE XI. LABELS IN ARGUMENT DIMENSION

Label	Description
Simple Claim	Simple opinion without any ground
Qualified Claim	Opinion based on a limiting condition without any ground
Grounded Claim	Opinion based on grounds
Grounded and Qualified claim	Opinion with limitation based on grounds
Euphemism	Unconfident and ambiguous opinion
Non-argumentative moves	Statement without containing opinion (including questions)

E. Social dimension

Labels in Social dimension are provided when Epistemic code is "On task" but they are provided not to all statements "On task" but to a statement which conforms to Epistemic code. This dimension represents how each statement is related to those of other members within the group. Therefore, it is required to understand not only a statement but also the previous context. A list of this dimension labels is shown in Table XII.

TABLE XII. CODE OF SOCIAL DIMENSION

Label	Description
Externalization	Externalization: No reference to other's opinion
Elicitation	Questionning the learning parner or proviking a reacion from the learning partner
Quick consensus building	Prompt consensus formation
Integration-oriented consensus building	Consensus formation in an integrated manner
Conflict-oriented consensus building	Consensus forming based on a confrontational stance

"Externalization" here refers to a statement without reference to those of others and it is provided mainly to statements as a point of argument origin such as in the beginning of argument on certain topic. "Elicitation" is provided to such statements which require others to extract information such as questions.

From its property as a statement to be made in response to other's opinion, "Consensus building" is classified into the following three labels. "Quick consensus building" is provided to a statement aiming at achieving prompt agreement with other's opinion. In particular, it is provided to a case to agree without delivering any specific opinion. "Integration-oriented consensus building" is provided to statements with an intention to achieve agreement with other's opinion while adding its own opinion. "Conflict-oriented consensus building" is provided to statements which adopt a confrontational stance or request revision against other's opinion.

A sub-dimension called as "Refer" in Social dimension represents which statement is referred to in the statement coded as "Consensus building". Labels in "Refer" dimension are provided without exception only if Social dimension labels belong to "Consensus building."

Since Social dimension labels represent relationship with others, it is possible to estimate how lively discussions were conducted or whose opinion in the group was respected by analyzing Social dimension labels. For example, arguments including a lot of "Quick consensus building" are assumed to be a result obtained just by taking a delivered opinion directly with almost no profound discussion.

F. Each coding and Learning toward artificial intelligence

In the new coding scheme, "Participation" dimension labels are automatically generated from statement logs, whereas other labels require manual coding by a coder in order to build up training data for deep learning and test data. Further, labels to be provided are decided by selecting from any of the dimensions of "Argumentation", "Social" and "Coordination" depending on a result of "Epistemic" labels. Therefore, coder provides "Epistemic" labels based on analysis of "Participation" dimension labels. Subsequently, "Argumentation" and "Social" dimension labels are provided if the "Epistemic labels are "On task." In addition, in a case that "Social" dimension labels belong to "Consensus building", statement number is provided as "Refer" since there exists reference source statement without exception. In a case that "Epistemic" labels are "Off task", those in "Coordination" dimension are provided.

VII. SUMMARY AND FUTURE WORK

This section recapitulates the findings of this study and suggests briefly some future issues.

A. Summary

As the first step to analyze collaborative process of big educational data from the perspective of LA, we tried to automate time-consuming coding task by using deep learning methods.

First, we developed a coding scheme based on the speech acts, coded manually for the remarks, and created training data and test data for deep learning. Next, three DNN models, that is, CNN-based model, LSTM-based model, Seq2Seq-based model were constructed for automatic coding, and their accuracy of automatic coding was verified. In addition, we also compared accuracy with SVMs, which are the baselines of classical machine learning. The result was promising; our approach, particularly, Seq2Seq model outperformed other methods clearly; the best of SVMs by 5-6% and other DNN models by 3-4%. It seems that this model could obtain almost the same predictive accuracy with other coding schemes than ours, for the reason that our coding scheme is sufficiently complex with 16 labels, based not on the surface information, but on the contextual significance of each contribution.

B. Future work

As for the future research directions, we may have two approaches to pursue.

The first approach is about DNN models. To improve prediction accuracy, it may be effective to introduce other network structures such as memory networks [30] instead of DNNs that consist of RNNs and CNNs. Memory networks make a vector from conversation by taking weighted mean of vectors of all sentences. Those weights play a role of attention since they correspond to importance of each sentence. In addition, the context of conversation should be considered. To capture context more precisely, it may be necessary to construct more complex models that take multiple preceding contributions as input vectors.

The second and most important approach concerns coding scheme. Our scheme, based on speech acts, was sufficiently complex, but not global. In order to more accurately and comprehensively grasp various collaborative learning activities such as individual cognitive process, social cognitive process, coordination among members, it will be necessary to construct a coding scheme which is more sensitive to details of interaction and social cognitive process of learning. Therefore, we proposed a new coding scheme with five dimensions, namely the participation dimension, the epistemic dimension, the coordination dimension, the argument dimension, the social dimension. With this new scheme, we are coding all the datasets again to constitute training data and test data for deep learning, in order to verify if this scheme contributes to a more precise understanding of the collaborative process and to improve the accuracy of automatic coding by our DNN models.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number 26350289 and 16K01134. We also thank to Ryo Yoshinaka for his thoughtful comments.

REFERENCES

[1] C. Shibata, K. Ando, and T. Inaba, "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.

[2] G. Stahl, T. Koschmann, and D. Suthers, "Computer-supported collaborative learning," In The Cambridge handbook of the learning science, K. Sawyer, Eds. Cambridge university press, pp. 479-500, 2014.

[3] P. Dillenbourg, P. Baker, A. Blaye, and C. O'Malley, "The evolution of research on collaborative learning," In Learning in humans and machines: Towards an interdisciplinary learning science, P. Reimann and H. Spada, Eds. Oxford: Elservier, pp. 189-211, 1996.

[4] T. Koschmann, "Understanding understanding in action," Journal of Pragmatics, 43, pp. 435-437, 2011.

[5] T. Koschmann, G. Stahl, and A. Zemel, "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)," In Video reseach in the learning sciences, R. Goldman, R. Pea, B. Barron and S. Derry, Eds. Routledge, pp. 133-144, 2007.

[6] M. Chi, "Quantifying qualitative analyses of verbal data : A pratical guide ," Journal of the Learning Science, 6(3), pp. 271-315, 1997.

[7] A. Meier, H. Spada, and N. Rummel, "A rating scheme for assesseing the quality of coputer-supported collaboration processes," International Jounal of Computer Suppported Collaborative Learning, 2, pp. 63-86, 2007.

[8] H. Jeong, "Verbal data analysis for understanding interacitons," In The International Handbook of Collaborative Learning, C. Hmelo-Silver, A. M. O'Donnell, C. Chan and C. Chin, Eds. Routledge, pp. 168-183, 2013.

[9] D. Persico, F. Pozzi, and L. Sarti, "Monitoring collaborative activities in computer supported learning," Distance Education, 31(1), pp. 5-22, 2010.

[10] L. Lipponen, M. Rahikainen, J. Lamillo, and K. Hakkarainen, "Patterns of participation and discourse in elementary students'computer-supported collaborative learning," Learning and Instruction, 13, pp. 487-509, 2003.

[11] S. Schrire, "Knowledge building in asynchronous discussion groups: Going beyond quantitative analysis," Computer & Education 46, pp. 49-70, 2006.

[12] 1st Internationa Conference on Learning Analytics and Knowledge. [Online]. Avaiable from: https://tekri.athabascau.ca/analytics/, Nov. 29, 2017.

[13] B. R. Schaun and P. S. Inventado, "Educational data mining and learning analytics," In Learning Analytics, J. A. Larusoon and B. White, Eds. Springer, pp. 61-75, 2014.

[14] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, 521(7553), pp. 436-444, 2015.

[15] Y. Kim, "Convolutional neural networks for sentence classification," arXiv preprint arXiv:1408.5882, 2014.

[16] X. Zhang, J. Zhao, and Y. LeCun. "Character-level convolutional networks for text classification," In Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS2015), pp. 649-657, 2015.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, 9(8), pp.1735-1780, 1997.

[18] J. Chung, C. Gulcehre, K. Hyun Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," arXiv preprint arXiv:1412.3555, 2014.

[19] Z. Yang et al., "Hierarchical Attention Networks for Document Classification," In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL2016), Human Language Technologies, 2016.

[20] D. Tang, B. Qin, and T. Liu, "Document modeling with gated recurrent neural network for sentiment classification," In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP2016), pp. 1422–1432, 2015.

[21] C. Rosé et al., "Towards an interactive assessment framework for engineering design project based learning," In Proceedings of DETC2007, 2007.

[22] C. Rosé et al., "Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning," International Journal of Computer Supported Collaborative Learning, 3(3), pp. 237-271, 2008.

[23] G. Gweon, S. Soojin, J. Lee, S. Finger and C.Rosé, "A framework for assessment of student project groups on-line and off-line," In Analyzing Interactions in CSCL: Methods, Approaches and Issues, S. Putambekar, G.Erkens and C. Hmelo-Silver Eds. Springer, pp. 293-317, 2011.

[24] A. Weinberger and F. Fischer, "A frame work to analyze arugmetative knowledge construciton in computer-supported learning," Computer & Education, 46(1), pp. 71-95, 2006.

[25] B. McLaren, O. Scheuer, M. De Laat, H. Hever and R. De Groot, "Using machine learning techniques to analysze and support mediation of student e-discussions," In Proceedings of artificial intelligence in education, 2007.

[26] T. Inaba and K. Ando, "Development and Evaluation of CSCL System for Large Classrooms Using Question-Posing Script," International Journal on Advances in Software, 7(3&4), pp. 590-600, 2014.

[27] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv, pp.1409.0473, 2014.

[28] O. Vinyals and Q. V. Le, " A Neural Conversational Mode," arXiv preprint arXiv:1506.05869, (ICML Deep Learning Workshop 2015), 2015.

[29] T. Kudo, "MeCab: Yet Another Part-of-Speech and Morphological Analyzer". http://mecab.sourceforge.net/, Nov 29, 2017.

[30] S. Sukhbaatar, A. Szlam, J. Weston and R. Fergus, "End-to-end Memory Networks," Proceedings of the 28th International Conference on Neural Information Processing Systems, pp. 2440-2448, 2015.

Using Deep Learning Methods to Automate Collaborative Learning Process Coding Based on Multi-Dimensional Coding Scheme

Takahiro Kanayama
Graduate School of Bionics, Computer and Media Sciences
Tokyo University of Technology
Tokyo, Japan
email:g211701051@edu.teu.ac.jp

Kimihiko Ando
Cloud Service Center
Tokyo University of Technology
Tokyo, Japan
email:ando@stf.teu.ac.jp

Chihiro Shibata
Graduate School of Bionics, Computer and Media Sciences
Tokyo University of Technology
Tokyo, Japan
email:shibatachh@stf.teu.ac.jp

Taketoshi Inaba
Graduate School of Bionics, Computer and Media Sciences
Tokyo University of Technology
Tokyo, Japan
email:inaba@stf.teu.ac.jp

***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.**

Keywords-CSCL; coding scheme; deep learning methods, automatic coding

I. INTRODUCTION

A. Analysis on Collaborative Process

One of the greatest research topics in the actual Computer Supported Collaborative Learning (CSCL) research is to analyze its social and cognitive processes in detail in order to clarify what kinds of knowledge and meanings were shared within a group as well as how and by what arguments knowledge construction was performed. In addition, it is also required to develop CSCL system and tools with scaffolding function which may activate collaborative process by utilizing such knowledge.

However, because main data for the collaborative process analysis include contributions over chatting, images and voices on tools such as Skype, and various outputs prepared in the course of collaborative learning, it is totally inadequate to perform just quantitative analysis in order to analyze such data. Therefore, CSCL research changed direction more or less to qualitative research [1] -[4].

As these qualitative studies often result in in-depth case study, however, they have a downside that it is not easy at all to derive guidelines with generality, which are applicable also to other contexts. Therefore, studies have been conducted in recent years based on an approach of verbal analysis in which labeling for appropriately representing properties (hereinafter referred to as coding) is performed to each contribution in linguistic data of certain volume generated over the collaborative learning from perspectives of linguistics and collaborative learning activities [5]. On the other hand, an advantage of the approach is its capability of quantitative processing for significantly large scale data while keeping qualitative perspective. However, it is a task requiring significant time and labor to perform coding manually and it is expected to become impossible to perform coding manually in a case that data becomes further bigger in size.

In our research project, we have achieved certain results in a series of previous studies reported last year in eLmL 2017 and the like, using deep learning technique for automatic coding of vast amount of collaborative learning data [6]-[8]. In this paper, while verification is performed for accuracy of the automatic coding based on deep learning technique similarly to last year, supervised data has been constructed by conducting coding manually depending on adopted multi-dimensional coding scheme in order to newly analyze collaborative learning process in a more multilateral and comprehensive manner.

B. Purpose of This Study

The final goal of our research project is to implement support at authentic learning and educational settings such as real time monitoring of collaborative process and scaffolding

 ISBN: 978-1-61208-619-4

for inactive groups based on analyses of large scale collaborative learning data as mentioned above.

As a further development of our previous research, a technique for automatizing coding of chat data is developed based on a multi-dimensional coding scheme capable of expressing collaborative learning process more comprehensively and its accuracy is verified in this paper.

Specifically, after newly performing coding manually for substantial amount of the same chat data, which was used in the previous studies, a part of it is learned as training data by deep learning methods and then automatic coding is conducted for the test data. For accuracy verification, we try to verify the accuracy of automatic coding by calculating precision and recall of automatic coding of test data in each dimension. We also evaluate what type of misclassification occurred frequently in each dimension.

C. Structure of This Paper

In this paper, the outline and results of our previous work are shown in Section II. Our coding scheme newly developed this time is described in Section III. Section IV presents the dataset with the statistics of the new coding labels assigned by the human coders. Our experiments and results of the study are shown in subsequent Section V. Finally, in Section VI, we present the conclusion and future work to complete the paper.

II. PREVIOUS WORK

Outline of our previous work [6] is shown below.

A. Conversation Dataset

Conversation dataset for the study conducted last year is based on conversations among students obtained from chat function within the system performing online collaborative learning by using CSCL originally developed by the authors for lectures in the university [9]. By the way, we will add that this data is also used in the research of this paper. Usage situation of CSCL as the source of the dataset is shown in Table I. Since students participated in multiple classes, number of participant students is less than the number obtained by multiplying number of groups and that of group members.

B. Coding Scheme

According to a manual for coding prepared by the authors, a label was assigned to each contribution of chat. Any of the 16 types of labels as shown in Table II was assigned. The ratio of each label is shown in Figure 1.

C. Automatic Coding Approach Based on Deep Learning

In the previous study, we adopted three types of Deep Neural Network (DNN) structures: 1) Convolutional Neural Networks (CNN), 2) Long-Short Term Memory (LSTM) and 3) Sequence to Sequence (Seq2Seq). Of the three models, Seq2Seq model is a deep neural network consisting of two LSTM units called encoder and decoder, and learning of classification problem and sentence generation is performed by entering pairs of strings of words to each part [10][11]. For example, the pair corresponds to a sentence in certain language and its translated sentence in case of translation system as well as to question sentence and response sentence in case of question and answer system, respectively.

In addition, a model based on Support Vector Machine (SVM), which is a traditional machine learning approach is used as a baseline. Accuracy of each model is verified by comparing automatic coding concordance rate and Kappa coefficient. About technology and experiment results in detail for each classification model, please refer to existing literatures of the authors [6]-[8].

TABLE I. CONTRIBUTIONS DATA USED IN THIS STUDY

Number of Lectures	7 Lectures
Member of Groups	3-4 people
Learning Time	45-90 mintutes
Number of Groups	202 groups
Number of Students	426 students
Dataset	11504 contributions

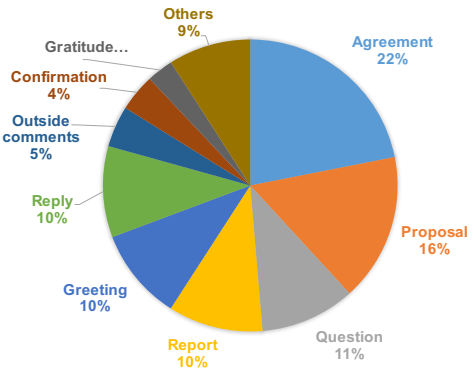


Figure 1. Ratio of each conversational coding labels

TABLE II. LIST OF LABELS

Label	Meaning of label	Contribution example
Agreement	Affirmative reply	I think that's good
Proposal	Conveying opinion, or yes/no question	How about five of us here make the submission?
Question	Other than yes/no question	What shall we do with the title?
Report	Reporting own status	I corrected the complicated one
Greeting	Greeting to other members	I'm looking forward to working with you
Reply	Other replies	It looks that way!
Outside comments	Contribution on matters other than assignment contents / Opinions on systems and such	My contribution is disappearing already; so fast! / A bug
Confirmation	Confirm the assignment and how to proceed	Would you like to submit it now?
Gratitude	Gratitude to other members	Thanks!
Complaint	Dissatisfactions towards assignments or systems	I must say the theme isn't great
Noise	Contribution that does not make sense	?meet? day???
Request	Requesting somebody to do some task	Can either of you reply?
Correction	Correcting past contribution	Sorry, I meant children
Disagreement	Negative reply	I think 30 minute is too long
Switchover	A contribution to change event being handled, such as moving on to the next assignment	Shall we give it a try?
Joke	Joke to other members	You should, like, learn it physically? :)

D. Experiment and Assessment

1) Outline of experiment

For the data set with manually prepared coding labels as described above, we compared the prediction accuracy of automatic coding for each model.

With separation of sentences into morpheme using MeCab conducted at first as a preprocessing of data, words with low use frequency were substituted by "unknown". Subsequently, just 8,015 contributions were extracted and 90% and 10% of them were sorted into data for training and test, respectively.

Naive Bayes, Linear SVM, and SVM based on RBF Kernel were applied as baseline approaches.

2) Experiment Results

Table III shows prediction accuracy (concordance rate) of models proposed in the previous study and those adopted as baseline for test data. The concordance rate here refers to a proportion that manually assigned label conforms with predicted label output by a model. It is proved, as Table III shows, that accuracy of the proposed model's result is higher than that of baseline model. Among the three models as described above, it is found that there is almost no difference in concordance rate between the approaches based on CNN and LSTM (0.67-0.68). These approaches show concordance rates a little bit higher (around 2 to 3%) compared with SMV as a baseline approach (0.64-0.66).

On the other hand, a model based on Seq2Seq showed the highest concordance rate among all of the models (0.718), higher by 5 to 7% and 3 to 4% compared with SVM and other models, respectively.

TABLE III. PREDICTIVE ACCURACIES FOR BASELINES AND DEEP-NEURAL-NETWORK MODELS

Naïve Bayes		SVM(Linear)		SVM(RBF Kernel)	
unigram	uni+bigram	unigram	uni+bigram	unigram	uni+bigram
0.554	0.598	0.642	0.659	**0.664**	0.659

CNN		LSTM		Seq2Seq	
with wikipedia	w.o. wikipedia	single-direction	bidirection	bidirection	bidir. w. interm.
0.686	0.677	0.676	0.678	**0.718**	0.717

Then, results as described above are discussed using Kappa coefficient, which means concordance rate excluding accidental ones. At first, it may be said that LSTM model has achieved sufficiently higher result as the Kappa coefficient for the model shows 0.63. In general, Kappa coefficient of 0.8 or higher is believed to be preferable for utilizing automatic coding discrimination result by a machine in a reliable manner, however, further higher concordance rate is required. In case of Seq2Seq model, on the other hand, Kappa coefficient is 0.723 with great improvement, if not reaching 0.8.

The experiment results above have suggested that Seq2Seq model is superior to other approaches due to consideration for context information. Since Seq2Seq is a model with reply sources entered, it is believed that the improvement in the accuracy has been partly caused by not separate capturing of each contribution but consideration of the context information.

III. NEW CODING SCHEME

As our previous studies mentioned some cases that Replay may include a meaning of Agree in the coding scheme, the fact that the definition of one label may sometimes overlap the definition of another label has become a factor making it difficult to assign a label always with accuracy and reliability. In addition to these technical problems, more importantly, labels based on speech acts, which express the linguistic characteristics of the conversation are insufficient for the analysis of the learning process. With this single linguistic scheme, it is almost impossible to realize whether members of a group engage in activities to solve the task, how members coordinate each other in terms of task division, time management, etc. during their collaboration, how each member constructs his argument, how members discuss and negotiate each other. From those described above, we propose a new coding scheme so that the automated coding accuracy will improve and that we may understand more accurately and globally collaborative process.

Our new coding scheme is constructed based on the multi-dimensional coding scheme proposed by Weinberger et Fischer [12]. As shown in Table V, our scheme consists of five dimensions, while Weinberger and Fischer's one has four dimensions without Coordination dimension. We provide labels basically regarding a contribution as a unit similarly to way we used in the previous studies. In addition, while such values as number of contributions are provided as Participation dimension labels, those in other four dimensions are provided by selecting one label from among multiple labels. In other words, since one label is given for each dimension for one contribution, a plurality of labels will be assigned to one contribution. Therefore, the coding work with this scheme is extremely complicated and takes a lot of time, but the merit of automated coding is even greater. Each dimension is described in detail below.

TABLE V. NEW CODING SCHEME

Dimension	Description
Participation	Frequency of participation in argumentation
Epistemic	How to be directly involved in problem solving
Argumentation	Ideal assertion in argumentation
Social	How to cope with others' statements
Coordination	How to coordinate to advance discussion smoothly

A. Participation Dimension

Participation dimension is for measuring degree of participation in arguments. As this dimension is defined as quantitative data including mainly number of contributions and its letters, time of contributions, and interval of contributions, coding is performed by statistical processing on the database while requiring neither manual nor artificial intelligent coding. The list is shown in Table VI.

Since Participation dimension labels handle number of specific contributions, it is possible to analyze quantitatively different aspects of participation in conversations but

 ISBN: 978-1-61208-619-4

impossible to perform qualitative analysis such as whether the conversation contributed to problem solving.

TABLE VI. PARTICIPATION DIMENSION

Category	Description
Number of contributions	Number of contributions of each member during sessions
Number of letters of a contribution	Number of letters during a single speech
Time for contribution	Time used for a contribution
Interval of contributions	Time elapsed since last contribution
contributions distribution	Standard deviation of each member within a group

B. Epistemic Dimension

This dimension shows whether each contribution is directly associated with problem solving as a task and the labels are classified depending on contents of the contributions as shown in Table VII. This dimension's labels are assigned to all contributions.

Weinberger and Fischer's scheme has 6 categories to code epistemic activities, which consist in applying the theoretical concepts to case information. But, as shown in Table VII, we set only two categories here, because we want to give generality by which we can handle as many problem-solving types as possible. "0n Task" here refers to contributions directly related to and such contributions with contents as shown below belong to "Off Trask".

- Contributions to ask meaning of problems and how to proceed with them
- Contributions to allocate different tasks to members
- Contributions regarding the system

Since Epistemic dimension represents whether directly related to problem solving, it works as the most basic code for qualitative analysis. In case of less "On Task" labels, for example, it is believed that almost no effort has been made for the task.

Besides, labels of Argument and Social dimensions are assigned when Epistemic dimension is "On Task", whereas those of Coordination dimension are assigned only when it is "Off Task".

TABLE VII. LABELS IN EPISTEMIC DIMENSION

Label	Description
On Task	contributions directly related to problm solving
Off Task	contributions without any relationship with problem solving
No Sense	contributions with nonsensical contents

C. Coordination Dimension

Coordination dimension code is assigned only when Epistemic code is "Off Task" and it is also assigned to such contributions that relate to problem solving not directly but indirectly. A list of Coordination dimension labels is shown in Table VIII but the labels are assigned not to all contributions of "Off Task" but just one label is assigned to such contributions that correspond to these labels. In addition, in case of replies to contributions with Coordination dimension labels assigned, labels of the same Coordination dimension are assigned.

"Task division" here refers to a contribution to decide who to work on which task requiring division of tasks for advancing problem solving. "Time management" is a contribution to coordinate degree of progress in problem solving, and for example, such contributions fall under the definition that "let's check it until 13 o'clock," and "how has it been in progress?" "Meta contribution" refers to a contribution for clarifying what the problem is when intention and meaning of the problem is not understood. "Technical coordination" refers to questions and opinions about how to use the CSCL System. "Proceedings" refer to contributions for coordinating the progress of the discussion.

Since Coordination dimension labels are assigned to such contributions that intend to problems smoothly, it is believed to be possible to predict progress in arguments by analyzing timing when the code was assigned. Further, in case of less labels of Coordination dimension, it may be predicted that smooth relationship has not been created within the group.

On the other hand, if a large number of these labels were assigned in many groups, it may be understood that there exists any defect in contents of the task or system.

TABLE VIII. LABELS OF COORDINATION DIMENSION

Label	Description
Task division	Splitting work among members
Time management	Check of temporal and degree of progress
Technical coordination	How to use the system, etc.
Proceedings	Coordinating the progress of the discussion.

D. Argument Dimension

Labels of Argument dimension are provided to all contributions, indicating attributes such as whether each contribution includes the speaker's opinion and whether the opinion is based on any ground. Labels of this dimension are provided to just one contribution content without considering whether any ground was described in other contribution.

A list of Argument dimension labels is shown in Table IX. Here, presence/absence of grounds is determined whether any ground to support the opinion is presented or not but it does not matter whether the presented ground is reliable or not. A qualified claim represents whether it is asserted that presented opinion is applied to all or part of situations to be worked on as a task. "Non-Argumentative moves" refer to contributions without including any opinion and simple questions are also included in this tag. Also, as a logical consequence, this label is assigned to all off-task contribution in the Epistemic dimension.

Labels in Argument dimension are capable of analyzing the logical consistency of contribution contents. For example, if a contribution is filled just with "Simple Claim" it is assumed as a superficial argument.

In comparison with Weinberger and Fischer's scheme, we do not set for now the categories of macro-level dimension in which single arguments are arranged in a line of argumentation such as arguments, counterarguments, reply, for the reason that it seems difficult that the automatic coding by deep learning methods for this macro dimension works correctly.

 ISBN: 978-1-61208-619-4

TABLE IX. LABELS IN ARGUMENT DIMENSION

Label	Description
Simple Claim	Simple opinion without any ground
Qualified Claim	Opinion based on a limiting condition without any ground
Grounded Claim	Opinion based on grounds
Grounded and Qualified claim	Opinion with limitation based on grounds
Non-argumentative moves	contribution without containing opinion (including questions)

E. Social Dimension

Labels in Social dimension are provided when Epistemic code is "On task" but they are provided not to all contributions "On task" but to a contribution which conforms to Epistemic code. This dimension represents how each contribution is related to those of other members within the group. Therefore, it is required to understand not only a contribution but also the previous context. Table X shows a list of labels of the dimension.

"Externalization" refers to contributions without reference to other's contributions and it is assigned to contributions to be an origin of arguments mainly at the start of argument on a topic. "Elicitation" is assigned to such contributions that request others for extracting information including question. "Consensus building" refers to contributions that express certain opinion in response to other's contribution and they are classified into the three labels below. "Quick consensus building" is assigned to such contributions that aim to form prompt consensus with other's opinion. It is assigned to a case to give consent without any specific opinion. "Integration-oriented consensus building" is assigned to such contributions that intend to form consensus with other's opinion while adding one's own opinion. "Conflict-oriented consensus building" is assigned to such contributions that confront with other's opinion or request revision of the opinion. "Summary" is assigned to contributions that list or quote contributions that have been posted.

Since Social dimension code represents involvement with others, it may be understood how actively the argument was developed or whose opinion within the group was respected by analyzing Social dimension labels. For example, it may be assumed that arguments with frequent "Quick consensus building" result in accepting all opinions provided with almost no deep discussion.

F. Learning for each code granting and artificial intelligence

In the new coding scheme, "Participation" dimension labels are automatically generated from contribution logs, whereas other labels require manual coding by human coders in order to build up training data for deep learning and test data. Further, labels to be provided are decided by selecting from any of the dimensions of "Argument", "Social" and "Coordination" depending on a result of "Epistemic" labels. "Argument" and "Social" dimension labels are provided if the "Epistemic labels are "On task." In a case that "Epistemic" labels are "Off task", those in "Coordination" dimension are provided.

TABLE X. CODE OF SOCIAL DIMENSION

Label	Description
Externalization	No reference to other's opinion
Elicitation	Questionning the learning parner or proviking a reacion from the learning partner
Quick consensus building	Prompt consensus formation
Integration-oriented consensus building	Consensus formation in an integrated manner
Conflict-oriented consensus building	Consensus forming based on a confrontational stance
Summary	Statment listing or quoting contributions

IV. DATASET AND STATISTICS

A. Target Dataset

The raw dataset is taken from the real conversation log of the CSCL system, which is the same one as that of previous study (Table I). On this dataset, the coding labels were newly annotated based on the new coding scheme. Labeling was manually carried out by several people in parallel. The human coders were lectured about the new coding scheme by a professional in advance in order to code labels as accurately as possible. To evaluate the accuracy of the manual coding, we had each contribution annotated by two annotators and measured the coincidence rate for each dimension of the new coding scheme.

B. Manual Coding and Preprocessing

While 9,962 contributions were manually coded in all, some contributions do not make sense as a text of CLSL. For instance, the duplicated posts, the blank posts, and the contributions that consist of only ASCII art can be mentioned. Such kinds of contributions were marked as "non-sense" when the annotators labeled, and removed or simplify ignored when the computer read them. After that, 9,197 contributions were remained as the useful data, on which the substantial jobs such as learning and classification are feasible.

The coincidence rates of the coding labels given by two human coders are significant for understanding the difficulty of the prediction, as well as to see the correctness of the manually coded labels. Table XI shows the coincidence rate, the number of the valid contributions, and that of the coincidence contributions for each dimension. For the Epistemic dimension, since the coincidence rate is high for human coders, we can expect that it is also easy for machines to classify them. On the other hand, for the Social dimension, since the coincidence rate is low for human coders and the valid samples are sparse, the opposite result is expected.

TABLE XI. THE VALID CONTRIBUTIONS AND THE COINCIDENT CONTRIBUTIONS

	# of valid contributions	# of coincidence contributions	Rate
Epistemic	9,197	8,460	0.92
Argumentation	9,083	7,765	0.85
Coodination	4,543	3,510	0.77
Social	3,917	2,619	0.67

 ISBN: 978-1-61208-619-4

C. Statistics of the New Coding Lables

In this subsection, we describe the statistics of the new coding labels assigned by the human coders with respect to each dimension. As we have multiple coders classify them, the statistics depend on the coders. When making a dataset for machines, we limit the contributions so as to have the same label assigned by the human coders. Thus, we describe the statistics of such contributions.

The ratios of “On task” and “Off task” in the Epistemic dimension are shown in Figure 2. In our dataset, the ‘On task’ contributions were a bit fewer than the ‘Off task.’ This implies that, at least from the view point of the conversation log, the cost of the communication was more than the cost of discussion in group work. Although this result is just an instance obtained by applying our CLCS system to the actual group works for limited lectures, we can at least conclude that the communication cost is not small in a group work.

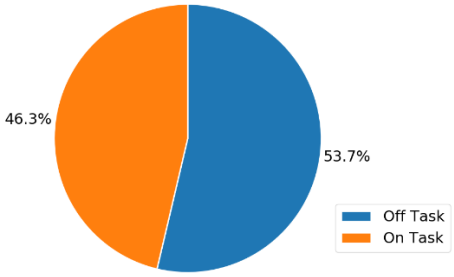


Figure 2. Ratio in the Epistemic dimension

Figure 3 shows the ratios of the labels in the Social dimension. Recall that its domain is On-task contributions. The label “Externalization” accounted half of the On-task contributions. The “Quick consensus building” followed it. Meanwhile, the ratios of the “Summary” and the “Consensus Buildings” except for the “Quick” one were small. These statistics show that the actual discussion mainly consisted of expressions of their opinions. Although we found that the contributions building consensus rarely come up in a real group work, we believe that they are the important keys for the discussion. Thus, we may can weight them when we assess the contribution to the discussion by students.

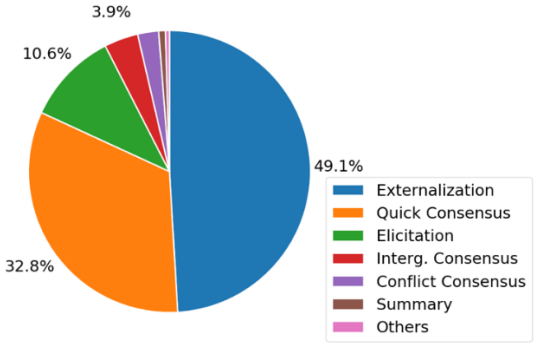


Figure 3. Ratio in the Social dimension

With respect to the "Coordination" dimension, the domain of which is the Off-task contributions, the most of them are assigned to "Other" as Figure 4 shows. The contributions labeled "Other" consist of short sentences that are not significant for neither discussion nor coordination of the group work. The representative examples are greetings and kidding. Meanwhile, the statistics show that the contributions except for "Other" also occupies more than a quarter. Since these kinds of contributions are related to coordinating tasks in the group work, they can be thought as important contributions for the assessment.

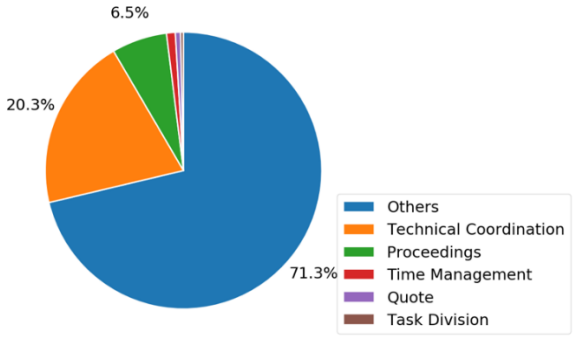


Figure 4. Ratio in the Coodination dimension

The labels in the "Argument" dimension are assigned independently of other dimensions. Thus, its domain spans both the On-task and the Off-task contributions. As shown in Figure 5, the label "Non-Argumentative moves" occupied more than 60 % of all. The label "Simple Claim" occupied the second percentage. To assess the discussion of the group work, at least it is necessary to remove the "Non-Argumentative" contributions and pay attention to which kind of claim is presented, even if almost every claim can be classified into the "Simple Claim". Therefore, the automatic coding for this dimension is as valuable as for the other three dimensions.

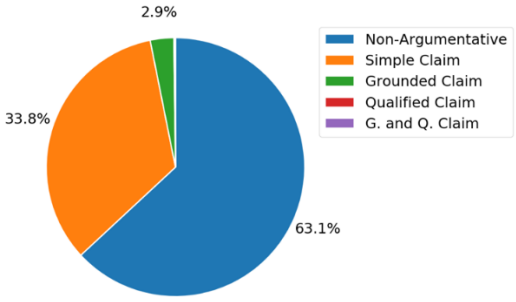


Figure 5. Ratio in the Argument dimension

V. EXPERIMENTS

A. Approach to Learing and Classifcation

As described in Section II, deep neural networks (DNNs) outperform other machine learning methods significantly at

least for the coding labels proposed by our previous studies [6]-[8]. Their results of the experiments show that the Seq2Seq-based model achieves the highest accuracy among several DNN structures. Thus, we apply the Seq2seq-based model to classify our new coding labels in this paper.

The new coding scheme has four axes to be labeled as discussed in Section III; the Epistemic, the Coordination, the Argument, and the Social dimension. In the following experiments, the labels in each axis, or the dimensions, are learned and classified. There are solid dependencies among the Epistemic, the Coordination and the Social dimensions, while the Augment dimension is independent of the other dimensions. As shown in Figure 6, there is a dependency tree among the former three dimensions. For instance, the label of the Social dimension is assigned only if that of the Epistemic is "On task." Therefore, the number of available contributions for learning is different for each classification task. In the following experiments, since we use the samples that have the coincidence labels only, the number of the available contribution was 8,460 for the Epistemic, 7,795 for the Augmentation, 3,510 for the Coordination, and 2,619 for the Social.

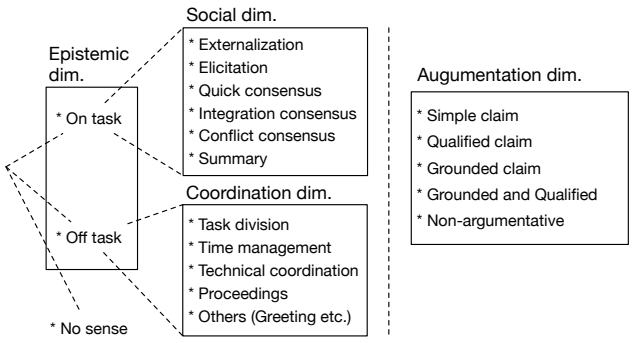


Figure 6. Dependency of Dimentions

B. Parameter Settings

We set the parameters for learning to the same values as in our previous study. They include the various kinds of the parameters such as the number of layers, the vector sizes of layers, the option of the optimization algorithms, learning rate, etc. The details can be referred to our previous studies [6]-[8].

C. Results for the Epistemic Dimension

The results of the experiments show that the On and Off tasks can be classified correctly with sufficiently high accuracy (Figure 7). The Seq2seq-based model achieves more than 90 % in both precision and recall (Table XII). Since the coincidence ratio by two human coders is 91%, we can say that the accuracy of automatic coding, which is comparable to human beings was obtained for the Epistemic dimension.

TABLE XII. PRECISION AND RECALL FOR THE EPSTEMIC DIMENTION

	Precision	Recall	F1-Score	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

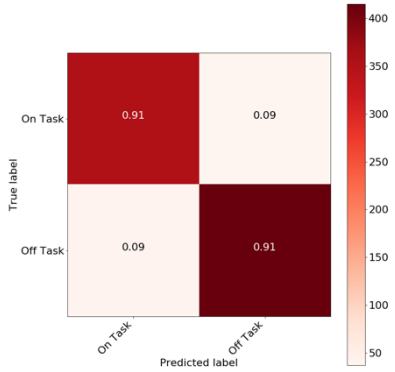


Figure 7. Confusion matrix for the Epstemic dimension

D. Results for the Argument Dimension

The classification accuracy is also high for the Argument dimension. The micro-averaged F1 score is 87 % (Table XIII). Especially, the F1 score for the label "Non-argumentative Moves" is high sufficiently (92 %), which means that our model can surely recognize whether the contribution has any substantial meaning as a claim or not. On the other hand, while the precision for the "Simple Claim" is high (89 %), the recall for it is low (72 %). According to the confusion matrix shown in Figure 8, a quarter of the Simple Claim is misclassified into the Non-argumentative. This is because it is difficult to distinguish contributions that have a very small opinion from that have no opinions.

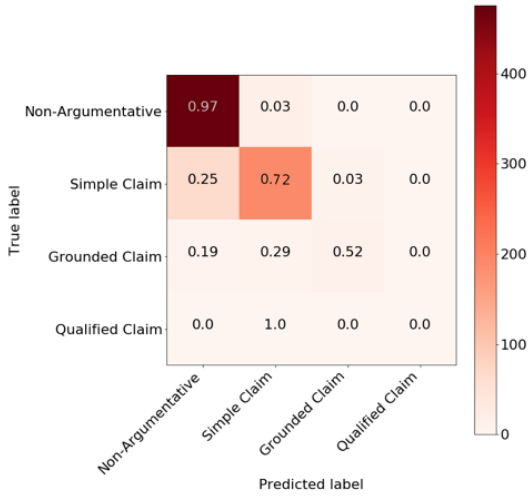


Figure 8. Confusion matrix for the Argument dimension

 ISBN: 978-1-61208-619-4

TABLE XIII. PRECISION AND RECALL FOR THE ARGUMENT DEMENTIONFFIGU

	Precision	Recall	F1-Score	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

E. *Results for the Coordination Dimension*

Regarding the Coordination dimension, our model also achieved high classification accuracy. Seeing that the number of supports varies greatly among the labels, we should evaluate the classification ability of the model by the micro-averaged accuracies over all coding labels. As Table XIV shows, the micro-averaged F1 score was 85 %.

According to the results for each label (Figure 9), the following is observed. The major labels such as "Other" and "Technical coordination" are classified correctly with high precisions, while the minor labels such as "Time Management", "Quote" and "Task Division" are not. Because the data for those miner labels are very limited, which have less than 50 contributions, it is quite difficult to learn them accurately. One of our future issues is to find some way to deal with those sparse labels.

TABLE XIV. PRECISION AND RECALL FOR THE COORDINATION DIMENTION

	Precision	Recall	F1-Score	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

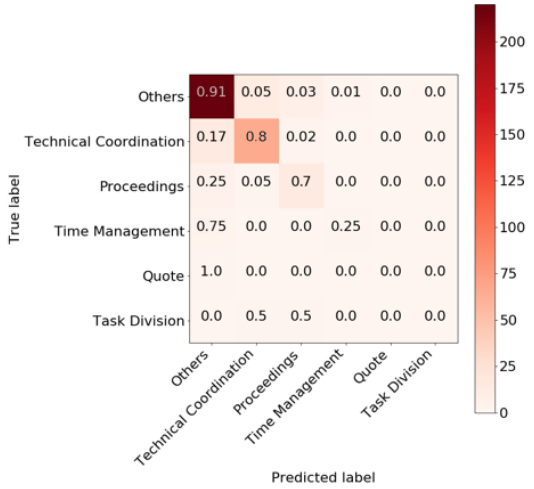


Figure 9. Confusion matrix for the Coodination dimension Results for the Social Dimension

F. *Results for the Social Dimension*

Comparing to the other dimensions, the accuracy was relatively low for the Social dimension. The F1 score was 70 % (Table XV). Since labeling the Social sometimes needs understanding the deep meaning of the contribution and the background story of the discussion, it seems to be difficult for machines to learn them correctly with limited data.

According to Figure 10, the recall of the label "Externalization" is especially low (61 %), while those of "Quick Consensus" and "Elicitation" are high sufficiently (93 % and 97 %, respectively). According to the confusion matrix in Figure 10, there is a major reason that worsen the accuracy; the Externalization labels are easily misclassified to the Quick Consensus and to the Elicitation, but not vice versa. This fact also explains the reason why the precisions for the Quick Consensus and the Elicitation are low though the recalls for them are high. To improve the result, it is necessary to pursue the causes of these two types.

TABLE XV. PRESCISION AND RECALL FOR THE SOCIAL DIMENTION

	Precision	Recall	F1-Score	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	0.72	0.72	0.70	262

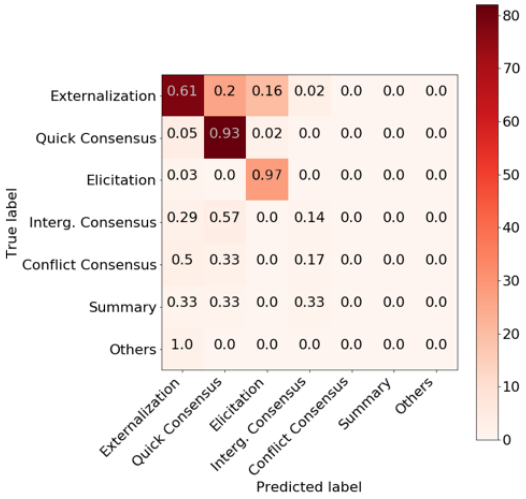


Figure 10. Confusion matrix for the Social dimension

VI. SUMMARY AND FUTRE WORK

A. *Summary*

In this study, we proposed a newly designed coding scheme with which we tried to automate time-consuming coding task by using deep learning technology.

 ISBN: 978-1-61208-619-4

We have constructed a new coding scheme with five dimensions to analyze different aspects of the collaboration process. After manually coding a large volume dataset, we proceeded to the machine learning of this dataset using Seq2seq model. Then, we evaluated the accuracy of this automatic coding in each dimension. Except some typical types of the misclassifications, the results were overall very good. These results indicate with certainty that we can introduce this model to authentic educational settings and that even for large classes that have many students, we can perform real-time monitoring of learning process or ex-post analysis of big educational data.

B. Future Work

As for the future research directions, we may have two approaches to pursue.

The first approach is about some typical misclassifications in the Social Dimension. To improve prediction accuracy, one could make more explicit and comprehensible the referential relation between a contribution and others even for the machines, if one indicates contributions to which a contribution refers. For example, with regard to the typical misclassification mentioned above between “Externalization” and “Quick Consensus” or “Elicitation”, since contributions labeled “Externalization” have no reference to other contributions, we can hope to effectively reduce these misclassifications with this kind of indicator. In addition, as the next step of this paper, it seems to be worth trying to compare the accuracy using DNN models other than Seq2seq and other network structures such as memory networks [13].

The second approach concerns the intrinsic structure of our coding scheme. Since the scheme contains different dimensions and under each dimension different labels are hierarchically organized, it is very interesting to discover not only correlations among dimensions, but also among labels belonging to different dimensions [14]. If we can input the information about the correlation between such labels in some form at the time of automatic classification, the accuracy of automatic coding can be further improved.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number 26350289, 17H02004 and 16K01134.

REFERENCES

[1] G. Stahl, T. Koschmann, and D. Suthers, "Computer-supported collaborative learning," In The Cambridge handbook of the learning science, K. Sawyer, Eds. Cambridge university press, pp. 479-500, 2014.

[2] P. Dillenbourg, P. Baker, A. Blaye, and C. O’Malley, "The evolution of research on collaborative learning," In Learning in humans and machines: Towards an interdisciplinary learning science, P. Reimann and H. Spada, Eds. Oxford: Elservier, pp. 189-211, 1996.

[3] T. Koschmann, "Understanding understanding in action," Journal of Pragmatics, 43, pp. 435-437, 2011.

[4] T. Koschmann, G. Stahl, and A. Zemel, "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)," In Video reseach in the learning sciences, R. Goldman, R. Pea, B. Barron and S. Derry, Eds. Routledge, pp. 133-144, 2007.

[5] M. Chi, "Quantifying qualitative analyses of verbal data : A pratical gauide ," Journal of the Learning Science, 6(3), pp. 271-315, 1997.

[6] C. Shibata, K. Ando, and T. Inaba, "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.

[7] K. Ando, C. Shibata, and T. Inaba, "Analysis of collaborative learning processes by automatic coding using deep learning technology", Computer & Education, 43, pp.79-84, 2017.

[8] K. Ando, C. Shibata, and T. Inaba, "Coding collaborative learning data automatically with deep learning methods", JSi SE Research Report, 32, 2017.

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

[10] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv, pp.1409.0473, 2014.

[11] O. Vinyals and Q. V. Le, " A Neural Conversational Mode," arXiv preprint arXiv:1506.05869, (ICML Deep Learning Workshop 2015), 2015.

[12] A. Weinberger and F. Fischer, "A frame work to analyze arugmetative knowledge construciton in computer-supported learning," Computer & Education, 46(1), pp. 71-95, 2006.

[13] S. Sukhbaatar, A. Szlam, J. Weston and R. Fergus, "End-to-end Memory Networks," Proceedings of the 28th International Conference on Neural Information Processing Systems, pp. 2440-2448, 2015.

[14] F. Scafino, G. Pio, M. Ceci, and D. Moro, “Hierarichical multi-dimensional classification of web documents with MultiWebClass,” International Conference on Discovery Science, pp.236-250, 2015.

 ISBN: 978-1-61208-619-4

法学の授業における反転学習とコンピュータ支援協調学習の事例研究

松村 佳記† 村上 康二郎†† 安藤 公彦††† 稲葉 竹俊†† 松永 信介†
東京工科大学 メディア学部† 教養学環†† クラウドサービスセンター†††

1. 研究概要

1.1. 研究背景

近年、反転学習と呼ばれる授業形態が注目されている。反転学習とは、旧来の講義のようにその場で先生から話を聞いて、後で復習するというスタイルではなく、事前に講義ビデオを視聴したり、クイズ問題に取り組むという能動学習を踏まえた上で、事後的に授業当日に臨むというものである。この学習スタイルは、昨今アクティブラーニングと呼ばれており、より質の高い議論や高度な学習内容の講義を展開する教育手法として着目されている[1]。

反転学習の利点は、従来学習者に任かされていた授業外学習の部分に対面授業と同様、教員の意図を大きく反映させられることにある。また授業外での学習（講義ビデオ視聴など）からの復習・応用などの新たな学習サイクルを生み出せるため、学習効果の向上が期待されている[2]。

この反転学習は、2010 年頃からブームとなった大学講義の録画ビデオをオンラインで無償提供する MOOC（Massive Open Online Course）等の動きに呼応している。MOOC の特徴は、然るべく学習した際には、大学からその科目に関する正式の修了証が得られるという点である。

1.2. 研究目的

本研究は、本学で開講されている法学の授業での反転学習の実践である。本学の一般教養科目である法学は履修者も多く、事前学習（予習）は大きな課題であった。

事前学習として講義ビデオの視聴と問題の解答を済ませた上で、当日の授業では、有意義なグループワークが展開されているかを検証する。具体的には、事前学習のクイズ問題の点数の結果によって、反転授業で行われるグループワークでのコメントの質を評価する。

A case study on the flipped learning and CSCL practice in a law classroom

† Yoshiki Matsumura, Shinsuke Matsunaga, School of Media Science, Tokyo University of Technology
†† Yasujiro Murakami, Taketoshi Inaba, Department of Liberal Arts, Tokyo University of Technology
††† Kimihiko Ando, Cloud Service Center, Tokyo University of Technology

2. 教材・システム概要

講義ビデオは Camtasia で編集を行い、70%を講義資料、30%を教員映像とする構成とした（図 1）。

図 1 講義ビデオ

クイズ問題は、本学の学習管理システムである Moodle を通じて制作した。問題内容は刑法に関するものであり、2 択式の 15 問を用意した（図 2）。

図 2 クイズ問題

当日の授業のグループワークは、システムが自動生成する 2～3 人組の中で行う。ハンドルネームを利用する仕様となっており、教室内の誰と議論しているかはわからない。発言をするときには、body に記載するメッセージと併せ、どういう趣旨の発言であるかを、挨拶・質問・提案・確認等からなる label 項目から選択することになっている（表 1）。

表 1 グループワークのコメントツール

id	body	Label
14893	(1)私は行為無価値論派です。理由は社会規範に反する行為を行ったのにもかかわらず、結果的に法律に触れなければ問題がないという判決は法に触れなければ何をしても良いという解釈が生まれるからです。 結果無価値論では、違法行為を働こうとした時点で罰することができるので)	提案

3. 評価実験

3.1. 実験概要

講義ビデオは 3 本あり、それぞれは約 10 分の構成である。授業の一週間前に、これらのビデオと関連するクイズ問題を学生に配信した。授業当日は、その場で提示される課題に対して、システムが設定したグループでの遠隔議論を展開した(図 3)。

・実施日：平成 29 年 7 月 3 日（木）
・対象：東京工科大学「法学」授業履修者 154 名
・使用機器：ノート PC，スマートフォン
・評価視点：
　－ グループ間交流の質・量
　－ 事前学習の成績と提言の相関

図 3　反転授業の風景

3.2. 分析結果

まず、事前学習やクイズ解答と、授業中でのグループワークコメントの関連を確認した。図 4 は、事前学習のクイズとして課されている問題の正答数と授業当日のコメント数の相関のグラフである。想定されていたことではあるが、正答率が高い学生の発言が顕著であった。

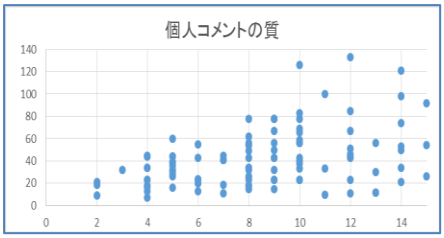


図 4　事前のクイズ解答数とコメント回数の相関

相関係数自体は、正のやや弱の約 0.4 であった。しかし、授業を観察している限り、学生たちの発言は活発で、今後の法学授業における反転学習の活用の展望が見える形となった。

大きな課題としては遅刻者対応である。システムとしてはグループを自動生成できるものの、遅刻者への対応は手動処理となっている。

4. まとめ

本稿では、東京工科大学の法学科目を対象とした反転学習とコンピュータ支援協調学習の効果検証について論じた。大・中規模の授業での事前学習は、当日の授業での効率化を図る上で欠かせない。

本実践において、学生は反転学習という新たな経験に触れ、積極的に学習に取り組んでいた。ただ、慣れも無いせいか、戸惑いがあったのは事実である。また、2 択というクイズ制約も再考の余地がある。記述式となると学生の抵抗感が高まる危惧があったために極力シンプルにしたが、事前学習との相関を測る意味においては、4 択程度の自由度はあっても然りかと感じた。しかし、今回の取り組みが授業効率に寄与したことは、分析結果からも顕著に表れている。とりわけ、予習の意識を学生に植え付けられたのは意義深い。これは、実際のデータ分析においても裏付けられた。通常の授業形態と比較すると、事前学習を踏まえて、学生は集中してグループ課題に取り組み、有意義に学習を進めることができたと言える。

数学や物理などの自然科学分野では、事前学習に困難がきたすこともあるが、人文・社会系の科目の事前学習から繋がる反転学習の効果は高いと言える。今回は法学の授業での取り組みであったが、他の人文・社会系の科目への展望を開くことが今後の課題と考える。

参考文献

[1] 森 朋子, 溝上 慎一, アクティブラーニング型授業としての反転授業(理論編), ナカニシヤ出版, 2017

[2] 宗村 広昭, 鹿住 大助, 小俣 公司, 反転授業における講義ビデオの視聴行動と成績の関係性, 日本教育工学会論文誌, 2017

Moodle 環境を活用した反転学習用 CSCL システムの開発

Development of the Moodle-based CSCL system for flipped learning

松永 信介* 安藤 公彦† 稲葉 竹俊‡
Shinsuke MATSUNAGA Kimihiko ANDO Taketoshi INABA

1. はじめに

近年，反転学習とよばれる学びのスタイルが注目されている．授業を受けてから復習へ移行するという旧来の学習形態ではなく，事前に課せられた予習に取り組んだ上で当日の授業に臨むという新たな学習形態である．効果的・効率的なアクティブラーニングの展開に繋がるという意味で期待が大きく，21 世紀の学習観のトレンドになりつつある．アクティブラーニングは，昨今の教育改革の一環として登場してきた概念であり，受動的な学びの姿勢を改め，主体的・能動的に取り組む学びのスタイルのことである．個における効果的な知識の定着や想像力の醸成はもちろんのことであるが，グループワークや討論などの他者との共同活動を通じて協調性・社会性を育むこともそのねらいとしている[1][2].

本稿では，一般教養の人文社会系の「法学」と「心理学」の 2 科目を対象に，本学の学習管理システムである Moodle を活用して実施した反転学習の成果ならびにそれを受けての課題・考察について報告する．

2. システム・実施環境

事前学習と授業当日の学習はともに，本学が 2014 年度より運用している基盤学習管理システム Moodle のもとで展開することとした．

以下に，事前学習用として Moodle に載せる教材（講義ビデオやクイズ問題），授業当日の学習用として Moodle と連動して機能する CSCL について述べる．

2.1 ミニ講義ビデオとクイズ問題

法学・心理学ともに，事前学習として課されるのは，Moodle 上にアップされた 10 分程度のミニ講義ビデオ 3 本の視聴と，その講義内容に関連する数問の確認クイズへの解答である．

ミニ講義ビデオは，録画映像をそのまま流すのではなく，講師の解説と同期するように編集された講義資料を併用する仕様とした．具体的は，画面を 7：3 で縦に 2 分割し，左側に講義資料が掲載され，残りの右側にワイプのように講師映像が映される設計とした（図 1）．なお，ビデオの編集には Camtasia Studio を用いた．

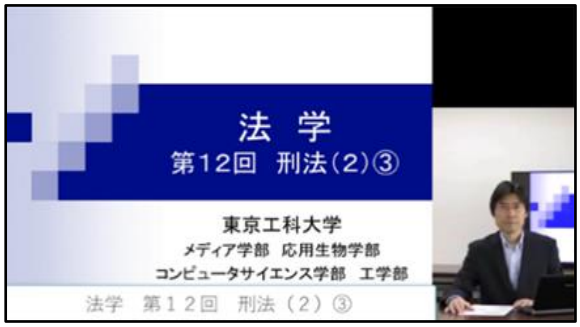


図 1　法学のミニ講義ビデオ

クイズ問題は，Moodle のクイズモジュールを活用して制作した．法学は刑法に関する 2 択式の 15 問，心理学は内発的動きづけに関する多肢選択式の 17 問をそれぞれ用意した（図 2）

図 2　心理学の 4 択クイズ問題

2.2 CSCL

授業当日は，Moodle と連動して機能する CSCL を活用した．この CSCL 環境で各学生が事前に準備することはシンプルであり，ハンドルネームの設定と簡単なアンケートへの回答である．システムは，その登録情報をもとに，2～3 人のグループを自動生成する．そして，ここから所定の課題テーマに関するクラウド上のグループワークが始まる．

各学生は教室内の誰と議論しているかはわからないが，“挨拶・確認・質問・提案・報告”という 5 種の発言趣旨カテゴリーから適切と思われるものを一つ選択した上でグループメンバーにメッセージを送る．システムはそれを受けて，図 3 のような発言ログを残す．

＊東京工科大学メディア学部
School of Media Science, Tokyo University of Technology
† 東京工科大学クラウドサービスセンター
Cloud Service Center, Tokyo University of Technology
‡ 東京工科大学教養学環
Department of Liberal Arts, Tokyo University of Technology

id	body	label
14893	(1)私は行為無価値論派です。理由は社会規範に反する行為を行ったのにもかかわらず、結果的に法律に触れなければ問題がないという判決は法に触れなければ何をしても良いという解釈が生まれるからです。 結果無価値論では、違法行為を働こうとした時点で罰することができるので)	提案

図 3　発言ログ

右端の label は上記 5 種のどのカテゴリー趣旨で発言しているかの情報であり，発言内容そのものは中央の body フィールドに記録される．

3. 評価実験

3.1 概要

各科目の実施日および履修者数は次の通りである．なお，事前学習用の講義ビデオとそれに付随するクイズ問題は各実施日の一週間前に開示した．

■ 科目：法学（前期）
・実施日：2017 年 7 月 3 日
・履修者数：154 名

■ 科目：心理学（後期）
・実施日：2017 年 12 月 8 日
・履修者数：54 名

■ 科目：法学（後期）
・実施日：2017 年 12 月 14 日
・履修者数： 319 名

使用機器はノート PC あるいはスマートフォンである．また，主な検証ポイントとして

- E1：事前学習の成績と授業当日の発言回数との相関
- E2：授業当日のグループワークの質と量

を設けた．

3.2 結果

3.2.1 法学（前期）

システムの試験運用としての調査実験であり，E2 の評価基準を設定する目的で実施した．この時点での仮評価基準は次の通りである．

・A 評価（※“挨拶”は除く）
　label 選択が適切で，具体例なども挙げて議論の流れに沿った body である．

・B 評価（※“挨拶”は除く）
　label 選択は不適切と思われるが，body は具体例を伴い，議論の流れに沿っている．

・C 評価
　具体例こそないが，課題に関係する body で，議論を促進する内容である．label 選択の妥当性は考慮しない，

・D 評価
　課題に関係する要素が body にはないが，議論を遮断する発言ではない．label 選択の妥当性は考慮しない，

・E 評価（※“挨拶”は除く）
　課題や議論の流れに沿わない，あるいはマイナス指針の body である．label 選択の妥当性は考慮しない．

上記の基準をもとに，システムに反映された 566 発言のカテゴリーと評価を整理したものが下表 1 である．

表 1　発言のカテゴリーと評価（法学（前期））

	A	B	C	D	E	小計
挨拶	−	−	36	25	−	61
確認	5	20	151	17	2	195
質問	−	6	57	10	1	74
提案	−	17	77	18	1	113
報告	5	13	83	21	1	123
小計	10	56	404	91	5	556

label は適度なバラツキがあり，最も多いカテゴリーは“確認”であった．議論のまとめに至る“提案・報告”もそれぞれ全体の 4〜5 分の 1 を占め，適度な分布になったものと考える．一方，評価基準に関しては，A と E は該当する学生数が少なく，要調整という課題を残した．

3.2.2 心理学（後期）

後述する 3.2.3（法学（後期））とともに，評価基準を次のようにシンプルに変更した．変更の要点としては，3.2.1 の A はそのまま保持し，旧 B と旧 C を併合，また旧 D と旧 E を併合する方針である．この方針のもとに新たに設定した評価基準 A・B・C は次の通りである．なお，“挨拶”に該当する発言は概ね冒頭でのやりとりに限られ，議論が活性化する前の会話に過ぎないので，一律に C 評価とした．

・A 評価（※“挨拶”は除く）
　label 選択が適切で，具体例なども挙げて議論の流れに沿った body である．

・B 評価（※“挨拶”は除く）
　具体例こそないが，課題に関係する body で，議論を促進する内容である．label 選択の妥当性は考慮しない．

・C 評価
　課題に関係する要素が body にはないが，議論を遮断する発言ではない．label 選択の妥当性は考慮しない．

表 2 は，事前調査をもとに策定した発言の種別とその分布割合である．事前調査で確認されなかった A 評価の“質問”と“提案”については，発言ログを精査した結果，それに該当すると思われるものがいくつか確認できたので，確率としては 0 とはせず，“質問”や“提案”の半分とした．なお，この差分調整は最も人数の多かった“確認”の B 評価（旧 C 評価）のところで行っている．下表 2 は，そのような微調整を行った上での期待確率分布である．

表 2　発言カテゴリーと評価の想定確率分布

	A	B	C
挨拶	−	−	0.110
確認	0.008	0.290	0.034
質問	0.004	0.113	0.020
提案	0.004	0.170	0.034
報告	0.008	0.173	0.173

（※ 各数値は小数点以下第 4 位を四捨五入）

また併せて，各発言のカテゴリーとその質に関する素点換算表を作成した（表 3）．能動的発言カテゴリーである“提案”の中で質の高いもの(A 評価)を基準値 1 とし，その他をほぼ 0.2 刻みの差分で定めた．なお，“挨拶”についてはC評価しか設定していないので，このルールとは関係なく，微たる加点として 0.1 に設定した．

表 3　発言の素点換算表

	A	B	C
挨拶	−	−	0.1
確認	0.6	0.4	0.2
質問	0.8	0.6	0.4
提案	1	0.8	0.6
報告	1.2	1	0.8

まず，検証視点 E1 に関してであるが，図 4 のような散布図を得た．横軸が事前学習のクイズ問題の点数（満点 17）で，縦軸が授業当日の発言回数である．

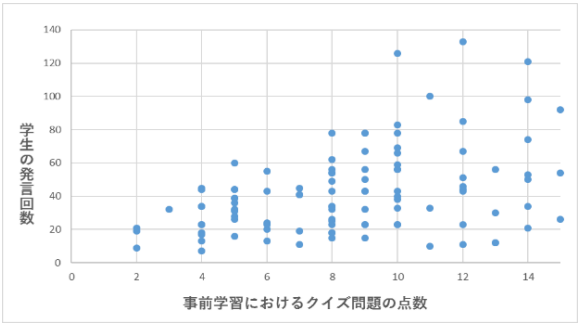


図 4　心理学（後期）の散布図

相関係数は約 0.6 であり，やや弱いが正の相関を確認した．ただ，発言ログを精査する限り，事前学習のクイズの正答数が大きい学生が授業当日にあまり発言しておらず，必ずしも実態を反映していないようにも思え，発言を喚起する施策や仕組みの必要性の課題を残した．

次に，評価視点 E2 に関してであるが，素点換算表をもとに全 605 発言を分類・算出した結果，下表 4 のような累計結果となった．

表 4　心理学（後期）の素点合計

	A	B	C	小計
挨拶	−	−	12.9	12.9
確認	25.2	55.2	24.2	104.6
質問	4.8	21.0	2.8	28.6
提案	26.0	40.0	7.8	73.8
報告	9.6	10.0	16.0	35.6
小計	65.6	126.2	63.7	255.5

表 2，3 に基づく期待値は約 0.65 であるが，表 4 に基づく全発言の平均評価は約 0.42（=255.5／605）となり，数値上で期待値を下回った．

3.2.3　法学（後期）

3.2.2 で述べた心理学と同様の手法・基準で実践した．なお，前期と同様，クイズ問題は 2 択の 15 問であり，ここは心理学とは理解度確認の問題とは難易度は異なる．

まず，検証視点 E1 に関してであるが，図 5 の散布図を確認した．横軸と縦軸のラベルは図 4 と同様である。相関係数は約 0.4 と心理学の授業を下回ったが，履修者数対比での課題取組みは全体的に低かったので，その影響もあると推察される．

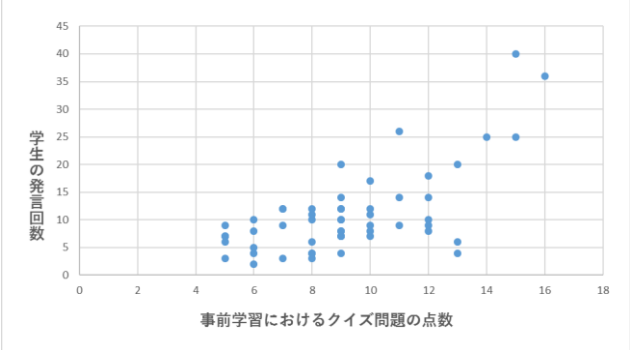


図 5　法学（後期）の散布図

次に，心理学と同様に，検査視点 E2 を確認した．素点換算表をもとに全 3780 発言を分類・算出した結果，下表 5 のような累計結果となった．

表 5　法学（後期）の素点合計

	A	B	C	小計
挨拶	−	−	83.16	83.16
確認	1.816	228.81	23.63	254.26
質問	24.8	24.069	0.9	49.796
提案	0.60	53.21	2.55	56.36
報告	0.38	49.997	119.2	169.58
小計	27.60	356.09	229.4	613.45

表 2，3 に基づく期待値は約 0.65 であるが，表 5 に基づく全発言の平均は約 1.01（=613.45／605）となり，数値上で期待値を上回った．

4. まとめ

4.1 成果

本稿では，本学で開講している人文社会系科目の法学と心理学を対象とした反転学習を取り上げ，その効果と支援環境である CSCL について論じた． 学習効果は数値上にも表れており，現時点で比較可能なデータのある法学（前期）では，期末試験の素点（60 点満点）が，昨年度の 24.1 点から 25.7 点へと増えた． また，教員からも従来の授業に比べるとより質の高い授業展開ができたというコメントを得た．

4.2　課題

課題テーマが一律であったため，本来議論するグループメンバーではなく隣の友人に相談する状況が散見された．その都度忠告することで，そのような相談は減少したが，潜在的な問題と考えられる．

4.3 展望

今回の試みは人文社会系の科目が対象あったが，今後は数学や物理などの数理・自然科学系の科目で実践してみて，その効果の差異を見るのは興味深い．また，AI を活用して隣の友人とのテーマの共有がないシステム環境にできればと考えている．

謝辞

本研究の一部は，日本学術振興会の科学研究費補助金（課題番号: 16K04848）の助成により行われた．

参考文献

[1] 森朋子，溝上慎一，アクティブラーニング型授業としての反転授業(理論編)，ナカニシヤ出版，2017

[2] 稲葉竹俊，奥正廣，工藤昌宏，鈴木万希枝，村上 康二郎，プロジェクト学習で始めるアクティブラーニング入門 ～テーマ決定からプレゼンテーションまで～，コロナ社，2017