

協調作業におけるコミュニケーションに基づくグループ能力 計測ツールの開発

花 川 典 子

A Tool for Measuring Group Abilities Based on Communication in Collaborative Learning

Noriko HANAKAWA

和文抄録：

協調作業中のコミュニケーションに基づく協調学習を評価するためのツールを開発した。本ツールは2つのサブツールであるモニタリングツールとスコアリングツールから構成される。モニタリングツールは共同作業中に発生するコミュニケーションを記録し、コミュニケーションの話題に応じて作業の能力を評価するツールであり、スコアリングツールは作業の成果物、例えばレポートやプログラムを自動評価するツールである。両ツールともTF/IDF法やベクトル空間法の自然言語処理技術をつかったツールであり、日本語を解析する技術に基づく。実験の結果、本ツールは協調作業中の作業者を評価できることを確認した。さらに、本ツールでは、協調作業実施時間が連続しない場合や、Webサイトからの文章のコピーしたとき、さらにレポートの内容に一貫性がないときの正しい評価が困難であるという問題があることを確認し、Webサイト文書空間による類似文チェックやレポートの構造解析や意味解析が必要であることがわかった。

キーワード：

協調作業、コミュニケーション、モニタリング、スコアリング、自然言語処理技術

欧文抄録：

An evaluation tool for collaborative learning has been developed. The tool is based on communication which occurs during collaborative work. The tool consists of two sub-tools: a monitoring tool and a scoring tool. The monitoring tool can evaluate automatically students' behavior during collaborative work. The scoring tool can automatically score output products such as reports. The two tools use natural language techniques: TF/IDF method, vector space model method. As a result of experiments, we confirmed that the tool was able to provide rough evaluation of individual student's level of participation. In addition, we confirmed that the tool has a weakness in the evaluation's accuracy when students copy & paste sentences from web-sites to their reports.

1. はじめに

大学教育では、新しい知識獲得を行う上で協調学習が効果的であることは良く知られている[3]。協調学習とは、作業分担されている複数学生が協調作業を通してグループの目標を達成することにより知識を自ら獲得する学習方法である。協調学習は認知科学分野の研究成果に基づき提案された手法であり、その効果は多くの実験で確認されている。

しかし、大学での協調学習では学生個人の評価が難しいという問題点がある。大学での成績は学生個人に与えられる成績であり、協調学習でグループ作業したとしても、個人別の評価が必須であり重要となる。最も多く採用される評価方法は、協調作業の成果物を評価しその評価をグループのメンバー全員の成績とする方法である。この評価方法では、もし一人の学生はすべての作業を行い、他の学生がまったく何も作業しなかった場合でも、すべての学生は同じ評価という問題点があり、正確な学生個人の評価ができない。最も理想的な協調学習の学生個人評価方法は、学生個人の行動をすべて教員が観察する方法である。この観察によって個人別の作業貢献度を計測できる。しかし、この評価方法では教員が多くの学生の行動をすべて観察する必要がある、ゼミなどの少人数ならば可能であるが、受講生が40人から50人以上の通常授業で実施することは、教員の労力から考えると非常に困難である。

そこで、協調学習での学生個人を評価支援するツールを提案する。このツールはモニタリングツールとスコアリングツールの2つから構成される。モニタリングツールは協調作業中のコミュニケーションを自動収集し、その話題の発生継続時間からグループ能力を推測する。スコアリングツールは協調作業の成果物であるレポートを自動採点するツールである。両ツールを活用することで、教員は評価労力の増加なしに、協調学習での学生個人の正確な評価が可能となる。本稿の2章では、関連研究を紹介し、3章ではツールの詳細を説明する。ツールを使った評価実験の結果とそれに対する考察は4章で行い、5章にてまとめと今後の課題を述べる。

2. 関連研究

協調学習に関して、コンピュータでの支援方法、遠隔学習、学習環境などの多くの研究が提案されている。しかし、協調学習の学生個人の評価方法についての提案は協調学習方法の提案に比べて少ない。Elisらは協調学習での評価設備を提案している[1]。この評価設備と評価方法の利用範囲は特定の科目での特別なケースである。つまり、プログラミング科目のプログラミングのミスから評価する方法である。我々の評価ツールは科目や評価対象の限定がなく、広く一般に利用できるところがElisらの提案する評価方法とは異なる。また、Hazeyamaらはグループ学習のためのグループ構成システムを提案した[2]。このシステムは協調作業中の技術的知識の獲得をグループの知識として共有できるシステムである。さらにグループのメンバーにアンケ

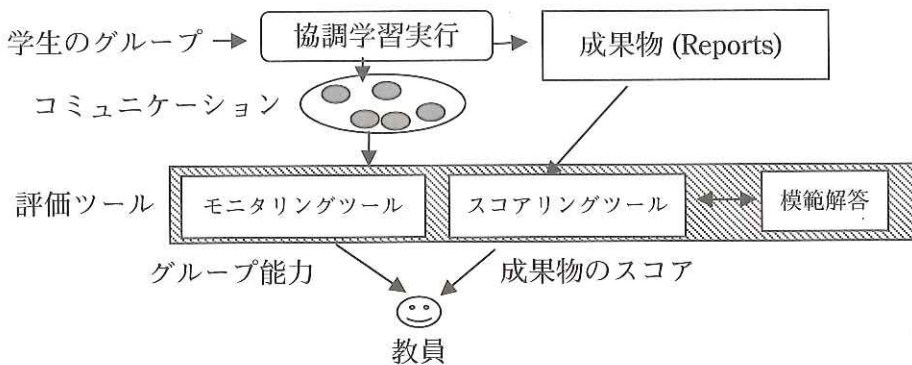


図1：ツールの概要

ートを実施する機能があり、協調作業での個人の貢献度を分析する機能もあり、協調作業中の個人の評価を行うことも可能である。しかし、アンケートに頼ることは客観性を失うことであり、作業者の主観中心の評価となる。我々のツールはアンケート等の主観に頼らずにコミュニケーションログを客観的に分析することで実現している。

3. 協調学習のための評価ツール

3. 1 評価の基本概念

協調学習での学生個人の評価はレポートやプログラム等の成果物の評価だけでなく、協調作業中の行動も評価対象とすべきであると考え。もし成果物だけで評価するならば、グループのメンバが同等に作業しなかったにもかかわらず、すべての学生の評価が同一になる。教員は成果物の評価と共に協調作業中の行動を協調学習の評価・成績に含める必要がある。これらの異なった観点の評価結果を用いて最終的に評価することで、お互いの評価の欠点を補うことができると考える。さらに、大人数講義の授業にも対応できるように、2つの方法での評価は効率的、かつ自動で実施されなくてはならない。少なくとも、評価ツールで自動収集し自動分析されたデータとその評価値は教員が最終成績を決定する際に役に立つものであることを目指す。

3. 2 協調学習の評価ツール

前述の基本概念に従って、評価ツールは2つの機能を持つ。一つは協調作業でのモニタリング機能、他方は成果物のスコアリング機能である。図1に本評価ツールの概要を示す。グループが協調作業を行う時、モニタリングツールにて学生のコミュニケーションログを自動的に収集する。コミュニケーションログを分析してグループの状態、例えば通常状態または混乱状態などを計測する。各コミュニケーションを発した学生を特定することで、学生が重要な問題解

決に対する貢献度合いを計測して個人の評価に利用する。また、コミュニケーションログの収集とグループ状態の計算はリアルタイムで実施されるので、協調作業の状況の時間的推移を知ることができる。一方、スコアリングツールは成果物であるレポートやプログラムをTD/IDF法やベクトル空間法の自然言語処理技術を利用して自動で評価するツールである。教員はこれらのツールから出力される評価結果に基づき、学生個人の最終成績を決定することができる。

3. 3 モニタリングツール

モニタリングツールは協調作業中の学生の行動をコミュニケーションを用いて観測するツールである。本来のコミュニケーションとは、会話、会議内容、議事録、設計文書、メモ、メールログなどすべての人間相互理解のための意思疎通手段を示すが、ここでは作業中の学生同士の会話と限定する。コミュニケーションログはモニタリングツールにて自動収集され、グループ能力値が同一話題のコミュニケーションログの発生時間のばらつきにより計算される。つまり、特定の話題（特定の問題を解決しようとする話題）が長期間にわたり発生するならば、その話題の問題は的確に早期解決されないことを示す。反対に特定の話題が短期間に集中して発生しその後はその話題が発生しない場合は、その話題の問題が的確に集中的に議論されて速やかに解決されたことを示す。当然前者のグループ能力は低く、後者のグループ能力は高い。この理論の詳細は文献[4]に詳しく記載されており、基本的に良いコミュニケーションが良いプロダクトを生成し、悪いコミュニケーションがグループの混乱を導くという考え方である。さらに、モニタリングツールはコミュニケーションが誰の発言であったかを特定できるので、良いコミュニケーションに関わった学生の評価を上げ、悪いコミュニケーションに関わった学生または全く発言しない学生の評価を下げるができる。つまり、ひとりの学生がすべての作業を実施した場合、コミュニケーションの発生は極端に少なくなる。反対に他の学生は雑談をした場合、雑談コミュニケーションの発生時間のばらつきが長期間にわたり、能力値を下げることになる。このようにグループの能力、さらに学生個人の行動を把握するためにツールである。画面イメージを図2に示す。ひとつの楕円が同一話題のコミュニケーションのクラスタを示し、大きな円がコミュニケーションの発生が長期間にわたりばらついていることを示し、小さな円が集中してコミュニケーションが発生していることを示す。さらに、コミュニケーションを用いたグループ能力値の計測方法はオープンソースプロジェクトのコミュニケーションログを利用して評価実験された[4]。図3で示すように、6つのオープンソースプロジェクトで評価した結果、品質（単位ステップあたりのバグ密度）と -0.66 の相関関係があることが確認できた。オープンソースプロジェクトでは有志ボランティアが集まって開発していること、製品ではないので集中的な開発やバグ修正作業、議論ができないことを考慮すると、 -0.66 の相関係数は良い結果であると判断できる。したがって、本方法で計測したグループ能力値は通常状態や混乱状態などのプロジェクトの状態の傾向をある程度示すことができる。

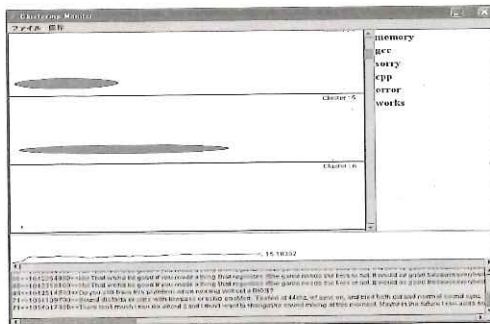


図2：モニタリングツールの画面イメージ

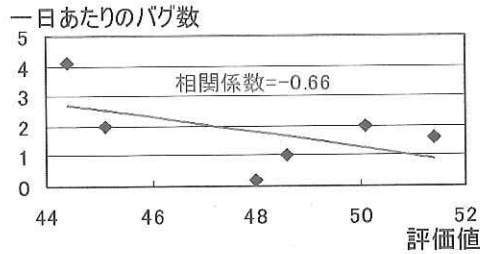


図3：モニタリングツールの評価

3. 4 スコアリングツール

スコアリングツールの目的はテキストファイルで書かれたレポートやプログラムを自動評価することである。大学教育では協調作業にてレポートを課することが多い。これらの場合、大量のレポート早急に評価する必要があるので、本ツールを導入する。図4はスコアリングツールの画面イメージである。スコアリングツールの基本的な考え方は、教員が模範解答のレポートを用意し、学生が作成したレポートとの間で文書類似度計算を行い、模範解答に近いレポートの評価を高くする方法である。まず、文書類似度計算では、自然言語処理技術を用いる。ここではTF/IDF法[6]によって模範解答から特徴語を抽出する。特徴語の数がベクトル空間法[5]の次元数となる。例えば、模範解答から200語の特徴語を抽出すると、200次元のベクトル空間を構築する。200次元のベクトル空間上に提出された個々のレポートのベクトルを決定し、模範解答ベクトルと各レポートベクトルの内積が文書の類似度となる。内積が小さければ、すなわちベクトルが近い方向を向いているならば模範解答に近い内容であり、内積が大きければ、すなわちベクトルが反対の方向を向いているならば模範解答には似ていないレポート内容となる。これがTF/IDF法とベクトル空間法による文書の類似度計算方法による評価方法のコンセプトである。

さらに、スコアリングツールは特徴語を自動抽出だけでなく手動で追加する機能も備える。これは模範解答から自動抽出された特徴語だけでは不十分な場合、教員が特別な語を手動入力できる機能であり、柔軟な採点が可能となる、また学生間の不正コピーを発見する機能も備える。提出されたレポート間での文書類似度を計算する機能であり、もし他のグループのレポートを不正コピーし、多少の変更を加えたとしても文書の類似度としては非常に高くなり不正を検出することが可能となる。スコアリングツールの模範解答との類似度で大まかなレポートの評価を下し、さらに不正コピーの可能性のあるレポートは慎重に吟味して教員はレポートの最終評価を決定することになる。

図5はスコアの正確さを検証するため、情報システム設計論での30レポートを評価した実験



図4：スコアリングツールの画面イメージ

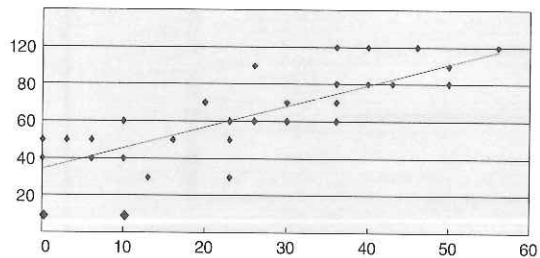


図5：スコアリングツールの実験結果

結果である。x 軸はスコアリングツールで計算したスコアを示し、y 軸は教員が手動でレポートを評価したスコアである。ツールの自動スコアと教員の手動スコアの相関係数は0.79となった。従って、スコアリングツールで極端にスコアが低いレポートは不可として手動評価対象からはずし、かつ、スコアリングツールで極端にスコアが高いレポートはすでにAとして手動評価対象からはずし、残りのレポートの評価を教員の手動で行うことが可能となる。これによって、教員の評価労力がかなり軽減されると予想できる。

4. 実験

4. 1 実験条件とその結果

スコアリングツールとモニタリングツールを含む協調学習評価ツールの有効性を検証するために、ツールの評価実験を行う。実験の条件を以下に示す。

- (1) 被験者は4グループであり、それぞれ2名の学生から構成される。
- (2) 2名の学生はレポート作成の協調作業を要求される。
- (3) レポートのテーマは「情報家電」であり、これについて調べてまとめること。
- (4) 調査を容易にするためにいくつかのキーワードを提供する。提供したキーワードは「アプライアンス」、「AMIDENアーキテクチャ」、「ヒューマンインタフェース」、「家電ミドルウェア」、「ホームネットワーク」、「ユビキタスコンピューティング」。
- (5) グループ内の2名学生は直接会話せずにMSN messenger[7]だけを利用して会話をする。
- (6) レポート作成の最大時間は8時間。
- (7) 調査の方法は、Web検索、文献、新聞記事、論文などを許可する。
- (8) IPSJ (Information Processing Society of Japan) の「情報家電」と題する特集記事を模範解答とする。

MSN messengerで収集したコミュニケーションログの例を図6に示す。コミュニケーションログは、シーケンス番号、コミュニケーション発生時間、コミュニケーション内容、発言者ID



図 6：実験でのコミュニケーションログ

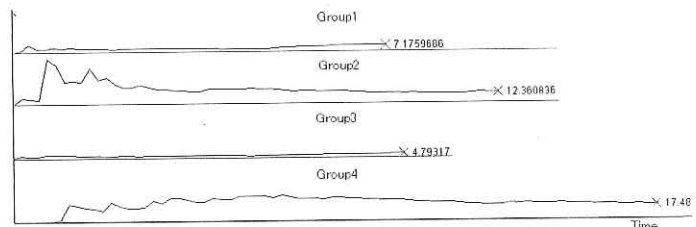


図 7：モニタリングツールでのグループ能力値の変化

の情報を含む。コミュニケーションログを利用して、モニタリングツールで解析し同一話題のクラスタを生成する(図2の参照)。クラスタサイズ(図2の楕円の大きさ)がコミュニケーションの発生時間に基づき計算される。このクラスタサイズがグループの能力値となる。能力値の高いクラスタ、すなわち小さいサイズのクラスタでコミュニケーションした学生の評価が高く、能力値の低いクラスタ、すなわち大きいサイズのクラスタでのコミュニケーションした学生の評価が低くなる。時系列に計算した能力値の変化をプロットしたグラフを図7に示す。図7のx軸が時間を示し、y軸が能力値を示す。y軸の大きい値が能力値の低さ、すなわち混乱している状態を示し、y軸の小さい値が能力値の高さ、すなわち落ち着いて正常に進行している状態を示す。図7では、グループ4がy軸の値が高い時期が長く続いている、すなわち混乱している状態が長く続いている状況を示し、グループ1とグループ3は比較的スムーズに作業が進行している状況が把握できる。また、グループ2は当初は非常に混乱していたが、次第に落ち着いた状況へ変化していったことが分かる。

実験結果を表1に示す。表1にはSocre列、GroupAbility列、TotalResult列でそれぞれ、スコアリングツールでの評価値、モニタリングツールでの評価値、総合判定値をそれぞれ示す。ただし、それぞれの項目にはManual値をもち、比較するために教員が手動で評価した結果を提示した。ツールでの評価と教員の手動での評価の差異については次のセッションにて考察する。

ツールでの総合判定値はグループ2が最も高く(4.7)、グループ4が最も低い(2.8)。さらにモニタリングツールで計測したグループ能力値は実験期間全体での能力値を示し、高い値が混乱などの悪いグループ状態を示す。スコアリングツールで計測したスコアは「情報家電」をテーマとしたレポートの評価スコアであり、高い値がよいレポートを示す。グループ能力値とスコア値での高い値の意味が異なるので注意して表1を参照してほしい。2つの値は協調学習の評価に関する異なった視点から求められたものである。例えば、グループ4のレポートスコアは高い(42)が、グループ能力値(15.0)は他のグループの能力値(10.9, 9.7, 9.0)よりも悪い。グループ4の協調作業は混乱した状況下で進行したが、結果として良いレポートを作成したことになる。混乱した状況が大きく影響して総合判定では最も悪い結果(2.8)となった。

4. 2 考察

考察では、実験で得たスコア値とグループ能力値と総合判定の正しさを検証する。実験結果の正確性を評価するために、まず実験中に収集したコミュニケーションログと成果物であるレポートを教員が手動で分析、評価した。表1のManualのカラムは教員が手動で決定した評価値を示す。

まず、スコアリングツールでのレポートのスコアの正しさについて考察する。ツールでの自動評価と教員の手動評価でもっとも差が大きかったのはグループ1である。自動評価されたスコアが36で悪いほうから2番目のスコアであったが、教員の手動評価では75と最も良いレポートと評価されている。グループ1のレポートを詳細に調べると、他のレポートとの差異が明らかになった。他のレポートは実験開始時に提示された6つのキーワード（4.1の実験の条件を参照）ごとに箇条書きされたものであるのに対して、グループ1は調査項目を調べた上で全体をまとめて文章を作成している。特にグループ2とグループ3のレポートは、キーワードに関して調べた内容間には文章全体としての脈絡はなく、単にキーワードの説明を羅列しただけで、レポートとしての「はじめに」、「おわりに」等も存在しなかった。対照的に、グループ1は「はじめに」や「おわりに」等のレポートの構成を忠実に守り、さらに箇条書きを利用せずに全体を「情報家電」のテーマでまとめ、さらに各キーワードに関する説明も全体の文章の脈絡のなかで説明されている。このグループ1だけが低く評価された点が、ベクトル空間法とTF/IDF法を利用した文書類似度計算による自動評価の限界を示す。文章の脈絡や一貫性、論理性は本技術を用いた文書類似度計算では求めることができない。しかし、スコアリングツールはレポートの脈絡や構成を評価すべきと考えられる。少なくとも、レポートが「はじめに」や「おわりに」等が存在せず、単なる「箇条書き」から構成されている場合は、レポート形式不備の評価は行うべきである。さらに、Webサイトからコピー＆ペーストで関連する文を切り貼りしただけで文章全体の脈絡のないレポートも、本スコアリングツールでは良い評価となる。自然言語処理技術のTF/IDF法とベクトル空間法の利用だけでは重要な限界が明らかに確認された。

次にモニタリングツールで計測したグループ能力値について考察する。グループ1、グルー

表1：実験結果

	Score		Group Ability		Total result	
	by Scoring tool	Manual	by Monitoring tool	Manual	By tool	Manual
Group1	36	75	10.9	8	3.3	A
Group2	46	60	9.7	10	4.7	B+
Group3	34	60	9.0	11	3.7	B
Group4	42	65	15.0	12	2.8	B

ブ2とグループ3のモニタリングツールでのグループ能力値がほぼ同じであるが、グループ4のグループ能力値だけが高い（従って状況は混乱して、評価は悪い）。そこでグループ4のコミュニケーションログを詳しく検証すると、他のグループとの決定的な2つの差異が明らかになった。ひとつめはグループ4の2名の学生は連続して実験を実施したわけではなく、午前中に自宅で実験を実施し、その後、登校して授業を受講し、帰宅後の深夜に実験を再開している。コミュニケーションログの発生時間によって機械的にコミュニケーションクラスタの大きさを計算しているので、このように大きな間隔が開いてしまうと、正確なコミュニケーションクラスタの大きさが計算できなくなる。従って、同一話題コミュニケーションが午前から深夜にかけて非常に長い時間継続されたとみなされ、特定問題が早期に解決できないとみなされ、結果としてグループ能力値が低く計算されてしまった。二つ目の差異は、自宅で実験を行ったため、コミュニケーションの中に食べ物やレジャーに関する多くの雑談が含まれていたことである。食べ物などの雑談が継続的に続くと、グループ能力値を低く計算される。ただし、雑談であっても短時間で終了し、その後その話題が二度と出ない場合、能力値は低く計算されない。

表2に期待された評価結果と自動評価した結果を示し、ツールでの評価の正しさのまとめを行う。Expected Result列は期待する結果を示し、Result by Monitoring tool列とResult by Scoring tool列は、この実験で得たそれぞれのツールでの自動評価結果を示す。例えば、グループ内で雑談（Having chat）が多く、問題解決ができない（Not resolving problems）場合、評価ツールはモニタリングツールで悪い評価をされ、期待された評価（Low）と同等になることを示す。また、グループ内の作業分担が良い時は、モニタリングツールでよい評価され、これも期待される評価（High）と一致する。しかし、レポートがWebサイトの文のコピー＆ペーストで単純に文を切り貼りしただけの場合（Executing Copy&Paste）でも、スコアリングツールとモニタリングツールともに良い評価と計算され、期待された評価（Low）とは相反する結果となる。同様に箇条書きだけのレポートについてもスコアリングツールでは良い評価となり、期待される評価（Low）とは異なる。さらに、レポートの文章の脈略や一貫性がある場合、良い

表2：期待された結果とツールでの評価結果の比較

Condition of group and report	Expected results	Results by Monitoring tool	Results by Scoring tool
Executing "Copy & Paste"	Low	High	High
Itemizing	Low		High
Having chat	Low	Low	
Not resolving problems	Low	Low	
Making good consistency as reports	High		Not available
Sharing work well	High	High	

評価が期待されるが、スコアリングツールはそれらを評価することができない。

表2に示すように、提案した本ツールでは協調学習の実行状況やレポートの内容に応じて正しく評価できる範囲が限定される。期待される評価と異なる評価を計算する場合には、本ツールに新しい機能を追加する必要がある。例えば、Webサイトのコピー＆ペーストで作成された文章であるチェックは、Webサイトからあらかじめ関連するサイトの文章を自動収集してWebサイトチェック用の文書空間を構築し、それらの中に含まれる文章との類似チェックを行うことによって、Webサイトからのコピー＆ペーストの有無を自動評価へ加味することができる。さらにレポートとして「はじめに」や「おわりに」などの構成チェック機能を追加することで、少なくとも箇条書きだけのレポートの評価を低くすることが可能となる。最も大きな問題は文章の脈絡や一貫性の評価の問題であるが、これは意味解析などの技術を新しく導入することで改善されると期待できるが、やはり教員が目を通す必要性はある。従って、本評価ツールは協調学習を完全に自動評価するという利用方法よりは、教員の評価の支援をして、より少ない労力にてより正確な評価を得ることを目的として本評価ツールを利用することが有用な利用方法と思われる。

5. まとめ

協調学習での学生個人を評価するための評価ツールを作成した。本ツールはモニタリングツールとスコアリングツールで構成され、モニタリングツールは協調作業中の学生の会話などのコミュニケーションから学生個人の行動を観察し、グループ能力値として計算する。スコアリングツールは協調作業で作成された成果物を自動評価するツールであり、模範解答との文書類似度を自然言語処理のTF/IDF法とベクトル空間法に基づいて計算し、模範解答との類似度をレポートの評価とする自動評価ツールである。2つの異なる観点から評価することによって、協調学習中の学生個別の評価が可能となった。しかし、Webサイトのコピー＆ペーストした文章や、箇条書きのみのレポートや一貫性のないレポートを正確に評価することができず、また無意味な長期間の作業中断がある場合も正確に評価できないことがわかった。これらの問題を解決するために、Webサイトの文書空間構築し類似文章を検索する機能や、レポート構造解析、意味解析などの新規機能を組み込むことを将来の課題とする。

参 考 文 献

- [1] Ellis, T.J., Hafner, W. "Peer Evaluations of Collaborative Learning Experiences Conveyed Through an Asynchronous Learning Network" , HICSS '05. Proceedings of the 38th Annual Hawaii International Conference on, 03-06, pp.5b- 5b, 2005.
- [2] HAZEYAMA,N., SAWABE,S., KOMIYA,S., "Group Organization System for Software Engineering Group Learning with Genetic Algorithm" IEICE Vol.E85-D No.4 pp.666-673 2002.
- [3] Slavin, E. Robert. Cooperative Learning: Theory, Research, and Practice. Needham Heights, MA: Ally and Bacon (1995).
- [4] Noriko Hanakawa, Kimiharu Okura, "A project management support tool using communication for agile software development", Proc. of the11th Asia-Pacific Software Engineering Conference (APSEC2004), pp.316-323, Dec. 2004.
- [5] Salton,G. and McGill, M. J.: "Introduction to Modern Information Retrieval", McGraw-Hill (1983)
- [6] <http://mikilab.doshisha.ac.jp/dia/research/report/2004/0812/001/report20040812001.html>
- [7] <http://messenger.msn.co.jp/>

(2006年 3 月 6 日受理)