Machine Assessment of Student Discussion-Board Formal-Style Debates

Alexis Cooper[1], Jung Hee Kim[1] and Michael Glass[2]
[1]North Carolina A&T State University, Greensboro, USA
[2]Valparaiso University, Valparaiso, USA
adcoope2@aggies.ncat.edu
jungkim@ncat.edu
michael.glass@valpo.edu

Abstract: This paper explores machine identification of argumentative moves in asynchronous online student debates. Student debate is a technique for engaging with topics which have no clear answer. Online debates differ from less-structured discussion-board class discussions in that students are restricted to formal debating moves such as advancing a new argument, providing evidence, or rebutting an argument. They are also assigned to debate a particular side. This research explores whether the different argumentative skills in student debate can be distinguished from each other by reading the texts of the students' discussion. The data for this research are the postings from 20 week-long student debates, an assignment in an educational technology class. Utilizing exclusively the text the students wrote, machine classifiers were trained to recognize the argumentative role of a message. If machines can detect an average difference between these debate moves from the text of the debate, and if they can detect the side of the debate, then the students are likely exercising different argumentative skills. The experiment shows it is possible to identify the skills which make debating different than free-form discussion with accuracy significantly better than chance.

Keywords: argumentation-based learning, computer-supported collaborative learning, computer-supported discussion board, class debate

1. Introduction

This paper applies machine classifier techniques to asynchronous online debate assignments, with the aim of discovering whether debate-specific behaviours could be detected. Discussion-based assignments are suitable for ill-structured or complex topics, where students should employ critical thinking and there may not be clearly correct answers. Imposing debate structure on the student discussion, compared with less structured discussion, promotes critical thinking and argumentation skills, and increases the depth and engagement with the topic.

The debates studied here were administered as week-long online exercises (Jeong, 2006). Students posted their contributions to an online message board. Students were required to post arguments, rebuttals, and other argumentative moves. The broad research question underlying this research is: to evaluate the result. Is it possible to monitor whether students are indeed exercising different argumentative skills, as the assignment requires? If it isn't possible to discern any difference between the students' explanations and rebuttals, for example, that would call the purpose of the exercise into question. If a machine can discern the differences with accuracy better than chance, it could be useful as an aid for monitoring or assessing the online debates.

The research questions of this paper, therefore, are keyed to two ways in which debate assignments differ from free-form discussion:

- Can machine classifiers discern the different debate argumentative moves in the students' texts, with accuracy significantly better than chance.
- Can machine classifiers discern the side of a debate from the students' texts, with accuracy significantly better than chance.

A broader research question guiding these experiments is:

- Do machine classifiers help identify the textual characteristics of a good debate post.

The broader research question relates to understanding and assessing the quality of the text. The machine classifier models could help understand, for example, what constitutes a better argument post versus a better rebuttal post. Potentially they could inform instruction in writing better posts or could provide formative assessment to students. In service of this goal, these experiments are restricted to utilizing the texts of the posts themselves, eschewing other features which might improve classification accuracy. Discovering textual clues to the quality of a post guided the classifier design but is not answered in this research.

In this paper we describe the rationale for argumentation-based collaborative classroom assignments. We illustrate the debate data used in the experiment. We describe training and testing two different machine classifier models and evaluate the results. We discuss how these statistically significant results validate the salient characteristics of online asynchronous argumentation exercise. Building on this work for the future, we show how statistically significant classifiers could be put to future use assisting an instructor. We discuss how to improve the classifier accuracy. We also discuss how a similar experimental setup could be used to measure the engagement potential of a student's argument, judging whether a student's debate argument is likely to attract responses from other students.

2. Online asynchronous debates

2.1 Collaborative debating assignments

The educational intervention that is the focus of this study falls under the rubric of Collaborative Argumentation-Based Learning (CABLE) generally, and Argumentation-Based Computer-Supported Collaborative Learning (ABCSCL) for computer-supported versions (Noroozi et al., 2012). Collaborative argumentation involves students working together, as contrasted with, e.g., writing argumentative essays by themselves. Using collaborative argumentation as a pedagogical tool is an important trend "whereby learners gain a comprehensive understanding of a problem through the process of trying to convince each other" (Noroozi, 2020). A main pedagogical purpose is to help students engage with ill-structured or complex domains, where problems may not have clear-cut answers.

A purpose of applying debate as a classroom intervention is that argumentation produces constructive interaction (Andriessen, 2006; Baker, 1999). Unlike formal debate competitions, collaborative argumentation is not about winning a debate. Neither is it about changing minds, though it is possible to chart students' evolving conceptions and map argumentative structure (Baker, 2009). Group collaborative debates are knowledge co-construction. The students argue to learn (Andriessen, 2006), each student adding a bit to the overall understanding of the proposition at hand.

Another advantage of debate over free-form discussions is that social processes interfere less with argumentation processes. In unstructured discussion people can be reluctant to counter each other directly, due to issues of politeness, losing face, or fear of starting a fight (Andriessen, 2006).

Knowledge construction dialogues can be described in multiple dimensions, e.g. epistemic, social, participation, and argumentation (Weinberger and Fischer, 2006). The argumentation structure of the discussion is the collection of claims, warrants, refutations, etc. which construct knowledge. The social dimension is the structure and modes of people responding to each other, for example asking questions, advancing an idea, agreeing or disagreeing, or building consensus. The participation dimension reflects how frequently and how much each participant is contributing, and whether participation is homogeneous across the group. The epistemic dimension relates to whether contributions are on or off task, how participants connect theories to the task at hand, and other issues.

2.2 Scripts for debating assignment

The debate assignment is guided by a script. Scripts are an instructional support which "provide some instructions for learners regarding how group members should collaborate and complete tasks" (Noroozi, 2020). The instructions for a collaborative debate regulate the structure of the argumentative aspects of the discussion and reduce the complexity of the discussion in the non-argumentation dimensions. All posts should be on task, learners are instructed to post a minimum number of times, and the permissible varieties of interaction are restricted.

If the participants are following the assignment, their posts will be on-task and restricted to certain argumentative moves and participants will consistently argue one side or the other.

2.3 Argumentative roles and sides of participant statements

In the debates studied here the different argumentative roles a statement can fill are: an original proposition, an argument (meaning a statement making a claim), a rebuttal to a previous statement, an explanation of a previous statement, or evidence for a previous statement (Jeong, 2006). Restricting student dialogue moves to a set of argumentation/debating moves is a form of scaffolding which produces better results than simply telling students to discuss the issue (Cho and Jonassen, 2002; Jonassen and Kim, 2010). Furthermore, by assigning students to different sides of the debate, students are encouraged to form strong arguments for ideas they would naturally dislike and thus avoid engaging (Nussbaum and Kardash, 2005).

A representative broad categorization of argumentative roles is: claim, data, warrant, backing, rebuttal, and qualifier (Toulmin, 1958). Common frameworks used by researchers for modelling and for computer processing of student discussion, such as (Weinberger and Fischer, 2006), contain more distinctions.

The categories used in these debates are simpler than the frameworks used by researchers, designed for use by students. Students were required to be self-aware, labelling each of their posts according to its role. The simplified categories in these debates were refined from earlier experiments, using similar discussion topics, until students showed a high level of labelling accuracy (Jeong, 2004; Jeong, 2006).

2.4 Online asynchronous collaborative debate

In the online debates studied here the instructor posted the original proposition to a discussion board. Students posted arguments in favour of or opposed to the proposition. Other students could post responses, where a chain of responses appears as a discussion board thread. Students could respond to any post already on the discussion board, so there was not strict turn-taking as happens in a formal debate or a verbal conversation. The debate took place over the course of a week.

The discussion board provides a useful representational tool for visualizing the structure of an argument. The visual display of the discussion shows the attachments of posts responding to posts, where each post is labelled with a title, the debate side, and argumentative role. Verbal discussions, by contrast, occur in linear time with alternating turns. The structure of argument discussions is not well-matched to linear presentation. Linear presentation hinders broadening and deepening the discussion (Rummel and Spada, 2005). Tools which visualize debate structure and content are believed to help participants stay on track, consider all aspects of the problem, address different branches of the evolving argument selectively, and in other ways argue more effectively (Noroozi, 2020).

+PROP: Reply to this thread to post arguments to SUPPORT the claim "The Country NED should develop a national distance education program".

> **+ARG1:** The following facts appear to be barriers to an advanced DE program in NED: 1. no mention of computer ownership but only 65% have telephones, not a good sign 2. very few people are educated or are getting educated However, **there is a large immigrant population that is interested in learning the culture and history. Making such course available should smooth assimilation into the culture** for the 1% of new NEDites that arrive each year. (Student 1)
>
> > **-BUT: Language is already a barrier for these immigrants** as most of them speak only a little NEDESE. Assuming instruction was provided in a universal language common to these immigrants, would they be able to handle the complexities of undertaking distance education, on top of taking instruction that might not comfortably be in their native language? (Student 2)
> >
> > **-BUT:** Studies show that most persons (60%) who are interested in distance education have either attended some college or at least have an AA degree. Therefore, **although the immigrants may want to learn the culture and history, they may not be well suited for learning at a distance**. (Student 3)

Figure 1: Extract from debate, showing proposition, argument, and two rebuttals. Emphasis added

3. Debate data

Figure 1 illustrates an extract from the debate transcripts used in this study. The debates were administered in a graduate educational technology class at Florida State University (Jeong, 2006). Each debate starts with two proposition posts from the instructor: one in support and another in opposition of the main topic. There is one record for each student post thereafter. There are debate logs from five class cohorts. Each cohort participated in four debates, totalling 20 debates with 2,068 posts in the entire dataset. Students typically posted about 6 times in each debate.

Each debate posting contains the side of the debate that the student represented, a message type label, a subject title, the text of the post, an anonymized student ID, a timestamp, and information to identify which post it was responding to. The two sides are "+" and "-". The message types are PROP, ARG, BUT, EXPL, EVID, AMEND, or RETRACT. The PROP posts are only posted by the instructor of the course. The next four message types are argumentative roles: state an argument (ARG), challenge a previous post (BUT), provide an explanation to another student's post from the same side (EXPL), or provide evidence for a previous post in the debate (EVID). AMEND and RETRACT student postings are not included in this study.

The ARG posts are numbered sequentially. The first argument on each side. of the debate is ARG1. Figure 1 shows a proposition, an argument in favour of the proposition from one student, and two rebuttals to that argument from two others

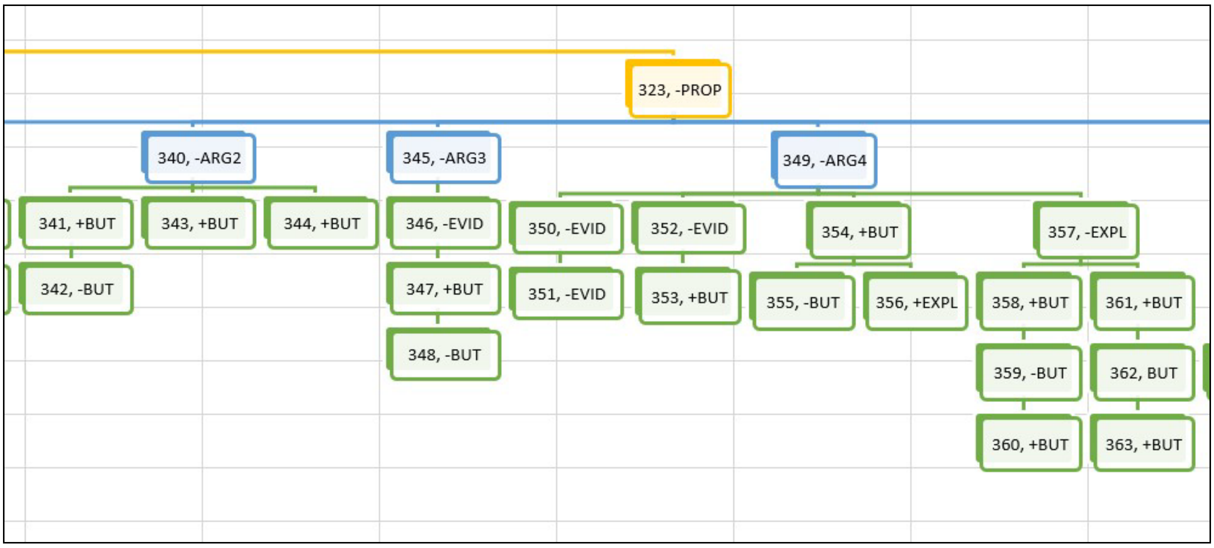

Figure 2: Tree structure of debate postings, showing three arguments from the opposing side with responses

Figure 2 shows the structure of postings from part of a debate. Notice that some arguments, e.g. ARG4, attract chains of back-and forth with rebuttals, evidence, and explanations. Some arguments attract less engaged discussion, and 17% of argument moves attracted no response.

Table 1 shows the counts and proportions of the different argumentative roles of the debate postings. The prompt instructed the students that each post would contain only one argumentative role. Posts for each side were nearly evenly split: 1,037 (50.3%) for the opposing side of a debate, and 1,023 posts for the supporting side.

Table 1: Proportions of different argumentative roles

Role	Number	% of Total
ARG argument	393	19
BUT rebuttal	961	47
EXPL explanation	296	14
EVID evidence	370	18
PROP proposition	40	2

4. Machine classifier experiments

4.1 Classifier models

We trained classifiers to label a) the side of the debate a post represented, and b) the argumentative role of the post. Each classifier is a support vector machine (SVM) with a radial basis kernel. Class weighting was applied during training to compensate for unbalanced frequency of categories. Weighting during training reduces the tendency of the trained classifier to err toward over-predicting the most frequent categories.

There are five categories for the argumentative role of a post. The Scikit-Learn SVM-SVC algorithm implements multi-category classifiers by training binary classifiers for every one-on-one pair of categories. Then it combines the results. The side of the debate is a simple two-category model.

Each model was presented with tokenized text of the student's post, prepared by the NLTK text tokenizer. The vocabulary was restricted to the 1,500 most frequent tokens, out of over 9,000 unique tokens. We applied a random 60%/40% training/evaluating split and an 80%/20% split. Models were programmed in Python using Scikit-Learn, Keras, and NLTK libraries.

Table 2: Evaluation of classifiers

Target	Train/Test Split	Kappa	Accuracy
Role	60/40	0.25	0.48
Role	80/20	0.29	0.50
Side	60/40	0.26	0.63
Side	80/20	0.26	0.63

4.2 Evaluation

Table 2 shows the results of applying the trained classifiers to the evaluation data. The trained classifiers were evaluated using the Kappa statistic (Di Eugenio and Glass, 2004). Kappa is commonly used for comparing two sets of categorial ratings. We compared the argumentative role and the side of the debate, as predicted by the classifier, against ground truth labels from the evaluation data. Kappa = 0 means the classifier agreed with the ground truth with the same accuracy as expected by independent random guessing, where guesses are distributed according to frequency of the categories. Kappa = 1 means the classifier was in perfect agreement with the labels in the data. Kappa = 0.25 means the classifier correctly captured 25% of the erroneous cases a random classifier would have missed.

A random classifier that predicted argument roles according to their frequencies in Table 1 would have accuracy about 0.31, this accuracy was improved to 0.50. Random prediction of side would have an accuracy of about 0.50, this accuracy was improved to 0.63. All the experiments performed much better than chance. A chi-square test on the four evaluation cases in Table 2 show all results are significant with $p < 0.01$.

5. Results

The two research questions are answered in the affirmative. The hypothesis that debate-specific behaviours can be recognized in the text of the student debates is borne out.

The result shows unambiguously that within our dataset of 20 different online asynchronous student debates there are differences among the different types of debating argumentation roles. There are also differences in arguing for and against a proposition. The differences can be discerned by a relatively simple machine classifier, using only the text of what the students wrote.

Individual posts are not reliably classified in our experiment. However statistical tests show the improvement over a random classifier is not an accident. Therefore, a classifier applied to a collection of posts should be able to provide information for estimating the frequencies of each category within the aggregate.

6. Discussion

6.1 Validating the computer-based collaborative debate assignment

The main value to this work is to show characteristics of the assignment that are particular to collaborative debating can be detected in the students' work.

The script encourages the student to utilize certain types of argumentative moves. Students are assigned to take sides. Both these activities have been shown in general to be pedagogically beneficial, both differ from free-form discussion. The computer classifiers can find distinctions within the student data based on these activities.

It is reasonable to conclude that student debate behaviour is indeed influenced by the scaffolding script. In the future computer classifiers could be used to help assess and monitor the activity.

6.2 Accuracy of classifiers

Classifier accuracies of about 0.50 (role) and 0.63 (side) mean that categorizing individual posts would not be accurate enough for assessing individual posts. In section 6.3 of this paper, we discuss how accuracy could be improved.

The classifiers based on word tokens constructed in this experiment do not appear to be accurate enough to use for discovering associations between text features and argumentative categories. The broader question behind our experiments, machine detection of text features characteristic of the quality of the posts, therefore, remains unanswered.

7. Future work

7.1 Aggregate machine classifiers as instructor tools

The result suggests that an automated assessment tool based on machine classifiers could assist instructors in monitoring the debates. Recall that debates happen asynchronously over time. Students in these debates logged in and contributed multiple times over the course of a week. The debate structure is not linear. New posts often respond to other posts in the middle of a tree structure. It is difficult for an instructor to keep abreast of the developing debate.

During the week that the debate accretes, automated classifiers applied to the developing collection of posts could provide guidance as to which aspects of the debate might deserve an instructor's attention.

The automated classifier labels would be compared to the participant's own labels, tabulating disagreements. A debate where the classifier applied to the aggregate of all posts differed more often than expected might benefit from attention. In a similar vein, disagreements could be measured according to each type of argumentative move, or according to the side of the debate, to identify possible problems. An instructor reading the debates might concentrate on the student posts that did not agree with the automatic classification, a set which is likely to be enriched with students who are unskilled in the argumentative moves or are not following the spirit of the assignment. Concentrating on the posts correctly classified is more likely to find the exemplary ones.

7.2 Machine identification of engagement

In asynchronous online debates students choose which posts to respond to. An argument which attracts no responses does less to advance the purpose of the assignment than an argument which attracts many responses. Thus, it will be useful to predict the amount of engagement an argument is likely to attract.

Preliminary work by Cooper (2022) has found machine classifiers can predict better than chance whether a post will be followed by any responses, based solely on the text of the post. There are other explanatory factors beyond the text of the post. For example, Cooper (2002) also found that posts later in the debate are less likely to attract responses.

There are several quantitative measures of engagement that could be modelled. Some possibilities are a) how many direct responses a post receives, b) how many different students choose to engage with an argument, and c) the length of the sequence of turns which follows an argument (Cooper, 2022). Argument posts are intended to start a thread of debate, it might be productive to study the engagement potential of argument posts separately from the other debate moves.

The ability to predict engagement could also help discover the textual aspects that produce more engaging posts. Jeong (2005) has shown, for example, that intensifier words (e.g "very," "never," "only") and qualifier words (e.g., "I think," "probably") change the likelihoods of different types of responses.

7.3 Improved classification accuracy

This work utilized only the text of the student posts. For better accuracy more information can be found from utilizing additional information, viz: a) non-text structural features and b) the discussion context of a post.

The non-text features of a post are useful for classification because side and argumentative move are not independent of the structural context. The post attached to an original proposition is most likely an argument and an argument is most often followed by a rebuttal. There are other statistical relationships among the various argumentative moves and sides of an argument (Jeong, 2005). Adding the structural context of a post to the classifier's input data could thus be used to improve classification accuracy.

Examining the text of adjacent turns has also been fruitful in improving classifier performance. For example, a study modelling argumentative dialogue acts in student collaborative discussions utilized a context of several consecutive dialogue turns (Rosé et al., 2008). In a tree-structured discussion, classifying a post would also include as context the text of the predecessor (parent node) of the targeted post and possibly the responses (child nodes) also.

It may also be productive to apply other natural language processing technologies. Experiments with some newer techniques did not improve accuracy compared to the word tokens and SVM utilized in this paper (Cooper, 2022). However BERT (Bidirectional Encoder Representations from Transformers) has quite fruitfully been applied toward machine modelling of student conversations, so there are likely avenues for improvement there (Dowell and Kovanovic, 2022).

8. Conclusions

Online collaborative debating assignments can checked by computer to see whether students are following the debate instructions. This work shows that machine classifiers are able verify that students are engaging in the valuable behaviours which make argumentation-based collaborative learning distinct from free-form discussion.

We trained computer text classifiers to validate a set of computer-based asynchronous collaborative debate classroom exercises. Students posted their debate contributions to an online message board over the course of a week. Debate exercises differ from free-form student discussion by restricting participant contributions to a set of argumentative debate moves, and by assigning participants to a particular side of the debate. We achieved statistically significant accuracy in machine identification of these two characteristics, using only the text of the student posts. This machine result validates, that participants behaved in ways that are salient to the format of a debate exercise.

This result points the way to potentially useful machine aids for instructors in administering and teaching online debate exercises.

Acknowledgements

The classroom debates that are the subject of this study were developed and administered by Dr. Allan Jeong of Florida State University. We are grateful to him for sharing his data and for guidance during the research.

References

Andriessen, J. (2006) "Arguing to Learn", in Sawyer, R.K. (ed.) *Cambridge Handbook of the Learning Sciences*, Cambridge University Press, Chapter 26.

Baker, M. J. (1999) “Argumentation and Constructive Interaction”, in Rijlaarsdam, G. and Espéret, E. (series eds.), Coirier, P. and Andriessen, J. (vol. eds.) Studies in Writing: Vol. 5. *Foundations of Argumentative Text Processing*, pp 179–202.

Baker, M. J. (2009) “Intersubjective and Intrasubjective Rationalities in Pedagogical Debates: Realizing What One Thinks”, in Schwarz, B., Dreyfus, T. and Hershkowitz, R. (eds.) *Transformation of Knowledge Through Classroom Interaction*, Routledge, Chapter 9.

Cho, K. L. and Jonassen, D. H. (2002) “The Effects of Argumentation Scaffolds on Argumentation and Problem Solving”, *Educational Technology Research and Development*, Vol. 50, No. 3, pp 5–22.

Cooper, A. (2022). *Gauging Student Participation in Online Asynchronous Debates*, M.S. Thesis, Dept. of Computer Science, North Carolina Agricultural and Technical State University, pp 34–45.

Di Eugenio, B. and Glass, M. (2004) “The Kappa Statistic: A Second Look”, *Computational Linguistics*, Vol. 32, No. 1, pp 95–101.

Dowell, N. and Kovanovic, V. (2022) “Modeling Educational Discourse with Natural Language Processing”, in Lang, C., Siemens, G., Wise A. F., Merceron, A. and Gasevic, D. (eds.) *The Handbook of Learning Analytics*, 2nd ed., SOLAR, Chapter 11, pp 105–119.

Jeong, A.C. (2004) “The Combined Effects of Response Time and Message Content on Growth Patterns of Discussion Threads in Computer-Supported Collaborative Argumentation”, *Journal of Distance Education* Vol. 19, No. 1, pp 36–53.

Jeong, A.C. (2005) “The Effects of Linguistic Qualifiers and Intensifiers on Group Interaction and Performance in Computer-Supported Collaborative Argumentation”, *International Review of Research in Open and Distance Learning*, Vol. 6, No. 3.

Jeong, A.C. (2006) “The Effects of Conversational Language on Group Interaction and Group Performance in Computer-Supported Collaborative Argumentation”, *Instructional Science*, Vol. 34, No. 5, pp 367–397.

Jonassen, D. H. and Kim, B. (2010) “Arguing to Learn and Learning to Argue: Design Justifications and Guidelines”, *Educational Technology Research and Development*, Vol. 58, No. 4, pp 439–457.

Noroozi, O. (2020) “Argumentation-Based Computer Supported Collaborative Learning (ABCSCL): The Role of Instructional Supports”, *European Journal of Open Education and E-learning Studies*, Vol. 5, No. 2, pp 16–32.

Noroozi, O., Weinberger, A., Biemans, H. J., Mulder, M. and Chizari, M. (2012) “Argumentation-Based Computer Supported Collaborative Learning (ABCSCL): A Synthesis of 15 Years of Research”, *Educational Research Review*, Vol. 7, No. 2, pp 79–106.

Nussbaum, E. M. and Kardash, C. M. (2005) “The Effects of Goal Instructions and Text on the Generation of Counterarguments During Writing”, *Journal of Educational Psychology*, Vol. 97, No. 2, pp 157–169.

Rosé, C., Wang, Y. C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A. and Fischer, F. (2008) “Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning”, *International Journal of Computer-Supported Collaborative Learning*, Vol. 3, No. 3, pp 237–271.

Rummel, N. and Spada, H. (2005) “Learning to collaborate: An instructional approach to promoting collaborative problem solving in computer-mediated settings”, *The Journal of the Learning Sciences*, Vol. 14, pp. 201–241.

Toulmin, S. (1958) *The Uses of Argument*, Cambridge University Press, Cambridge.

Weinberger, A. and Fischer, F. (2006) “A framework to analyze argumentative knowledge construction in computer-supported collaborative learning”, *Computers and Education*, Vol. 46, No. 1, pp 71–95.