Computer-Supported Collaborative Learning
DOI 10.1007/s11412-012-9147-y

# The ACODEA framework: Developing segmentation and classification schemes for fully automatic analysis of online discussions

**Jin Mu · Karsten Stegmann · Elijah Mayfield · Carolyn Rosé · Frank Fischer**

Received: 14 December 2011 / Accepted: 25 April 2012

**Abstract** Research related to online discussions frequently faces the problem of analyzing huge corpora. Natural Language Processing (NLP) technologies may allow automating this analysis. However, the state-of-the-art in machine learning and text mining approaches yields models that do not transfer well between corpora related to different topics. Also, segmenting is a necessary step, but frequently, trained models are very sensitive to the particulars of the segmentation that was used when the model was trained. Therefore, in prior published research on text classification in a CSCL context, the data was segmented by hand. We discuss work towards overcoming these challenges. We present a framework for developing coding schemes optimized for automatic segmentation and context-independent coding that builds on this segmentation. The key idea is to extract the semantic and syntactic features of each single word by using the techniques of part-of-speech tagging and named-entity recognition before the raw data can be segmented and classified. Our results show that the coding on the micro-argumentation dimension can be fully automated. Finally, we discuss how fully automated analysis can enable context-sensitive support for collaborative learning.

**Keywords** Online discussion · Automatic content analysis · Text classification

J. Mu (✉) · K. Stegmann · F. Fischer
Ludwig-Maximilians-Universität München, Empirische Pädagogik und Pädagogische Psychologie,
Q1 Leopoldstraße 13, 80802 Munich, Germany
e-mail: jin.mu@psy.lmu.de

K. Stegmann
Universität Koblenz-Landau, Institut Erziehungswissenschaft/Philosophie, Bürgerstrasse 23, 76829
Landau, Germany

E. Mayfield · C. Rosé
Language Technologies Institute, Carnegie Mellon University, 5000 Forbes Avenue, 15213 Pittsburgh,
PA, USA

## Why should online discussions be coded automatically?

Online discussions have been widely used in the field of CSCL to foster collaborative knowledge construction. Learners work together to exchange ideas, negotiate meaning and formulate understanding (De Laat and Lally 2003). One important feature of online discussions is that this kind of communication produces a huge body of digital data as a byproduct of the interaction. Researchers are therefore confronted with the opportunity as well as the challenge of analyzing online discussions at multiple levels to understand the underlying mechanisms of group interaction (Strijbos et al. 2006), such as quality of argumentation, or social modes of interaction (Weinberger and Fischer 2006). A variety of multidimensional frameworks have been employed to apply appropriate analysis on dialogic argumentation (Clark et al. 2007). In this study, we focus specifically on analysis of what has previously been called micro-argumentation (Weinberger and Fischer 2006), with the idea of expanding to other dimensions of analysis in future work.

Evaluation of discussion quality consumes a huge amount of resources in research projects related to online discussions. In order to address this problem, Rosé and colleagues (2008) reported a series of experimental studies with about 250 online discussions (Stegmann et al. 2012; Stegmann et al. 2007) where about 25 % of all human resources in the research project were spent analyzing online discussions on multiple dimensions. Human coders had to be trained to annotate segments of these data using a multi-dimensional coding scheme that operationalized aspects of content as well as manner of argumentation and social modes of interaction. While uncovering findings related to how group knowledge construction works often make those efforts worth the time and energy they require, analyzing a huge body of online discussions by hand is an arduous task that slows down the progress of the research substantially. An automatic and thus faster classification of online discussions may affect the whole research process positively. One possible impact may be that an increasing number of researchers may be willing to analyze online discussions on multiple dimensions. Moreover, some of the resources made available through these automatic coding efforts may then be used to conduct follow-up studies or to try out additional pioneering approaches to data analysis.

Automatic classification may not only facilitate research on online discussions: It also allows for adaptive collaborative-learning support (Kumar and Rosé 2011; Kumar et al. 2007; Walker et al. 2009) to foster the quality of collaborative knowledge construction during online discussions (Gweon et al. 2006; Walker et al. 2009). Online discussions could be analyzed in real-time and instructional support measures like hints or scaffolds could be adapted to the quality of certain aspects of the collaboration. For example, learners who are unable to provide warrants and grounds for their claims may get offered scaffolding to construct better arguments. Learners who fail to relate their contributions to those of other learning partners may be explicitly asked to provide such connections.

Although various research approaches and corresponding computer-mediated settings have been developed to analyze discourse data automatically in the field of CSCL, it has proven to be challenging to realize the full potential of the newly introduced technologies. Actually, much current adaptive collaborative-learning support (ACLS) research is situated in the early stage of development, since the majority of the discourse analyses of collaborative conversations are currently still conducted “non-automatically” or “semi-automatically”. For instance, a non-automatic implementation of adaptive support for collaboration may be delivered only in the case when certain non-productive learning behaviors have been detected by an experimenter (Gweon et al. 2006). Kumar and colleagues (2007) as well as Wang and colleagues (2011) took further steps toward using machine learning to classify student utterances with an acceptable degree of reliability (Cohen's Kappa of 0.7 or higher). However, these efforts have mostly focused on detection of simple patterns that indicate task orientation. The purpose of the detection of these patterns was to make sure that students stay on the topics that are related to the learning task and providing resources to increase the conceptual depth of their discussion.

While this prior work was based on simple analyses of discussion data, recent advances in automating detailed content analyses have been accomplished by applying multi-dimensional categorical coding schemes, each dimension of which indicates certain sophisticated learning processes during collaborative learning (Dönmez et al. 2005; Rosé et al. 2008). These findings demonstrate that CSCL researchers can access various aspects of learning processes through automatically extracted diagnostic features from corpus data. Other recent work demonstrates that a linguistically motivated automatic analysis of social positioning in collaborative discussions can detect authoritativeness of stance of speakers relative to their partners with high correlations with human assessment ($r$=0.97) (Mayfield and Rosé 2011).

Nevertheless, further study is needed in order to address some important technical obstacles that still hinder the content analysis from being conducted in a fully automatic way. First of all, with the exception of approaches that have been applied to chat data (Howley et al. 2011; Kumar and Rosé 2011; Kumar et al. 2007), none of the analyses mentioned above were capable of dealing with the original 'raw' text contributed by the participants without segmentation by a human. Specifically, it is compulsory to divide raw data into units of analysis (segments). Errors at this stage can affect accuracy of coding in the later phases (Strijbos et al. 2006). Therefore the existing approaches must be considered to be semi-automatic, due to the requirement of manual segmentation. Secondly, developing a model that is capable of assigning a set of codes to unit fragments is a lengthy process itself. Finally, and possibly most importantly, the models trained in our prior work were highly context specific, and therefore demonstrated large performance drops when applied to data from other contexts. Therefore, despite the promise of earlier reported results, some crucial questions have emerged including the urgent need for the re-use of a coding scheme across diverse contexts, or in other words, developing context independent automated coding schemas to model similar behavioral patterns during online discussions.

Against this background, we developed a multi-layer framework, which has been optimized for fully automatic segmenting and context-independent coding using the previously introduced Natural Language Processing tool called SIDE (Mayfield and Rosé 2010a). What we offer is not simply a report on a use case of how to use SIDE in CSCL research. Rather we offer insights into what is required to adapt such a tool to make it appropriate for applying text classification technology in specific contexts within CSCL. In the remainder of the paper, we begin by providing an overview of the state-of-the-art in the application of NLP technologies in CSCL research. We offer an explanation of an important caveat in the use of automatic classification models in CSCL research, namely issues with the generality of trained models. We then present our methodological approach, which attempts to address this issue in a novel way. The key idea is to extract the semantic and syntactic features of each single word by using the techniques of part-of-speech tagging and named-entity recognition before the raw data can be segmented and classified on the desired dimensions (e. g., micro-argumentation). An evaluation demonstrating the extent to which we have been successful in this endeavor is also delivered with empirical evidence. Finally, we conclude with discussion of the limitations of our current work and plans for future research.

## Applying NLP technologies in CSCL

Natural Language Processing has long been used to automatically analyze textual data. The need for involving technology from NLP in the process of content analysis is growing in the presence of the Web and distance learning (Duwairi 2006). For instance, the NLP methods of content analysis have been developed for the automatic grading of essays (Duwairi 2006; Landauer 2003); and for intelligent and cognitive tutoring (Diziol et al. 2010; Rosé and Vanlehn 2005).

### History of NLP in support of learning technologies

In the last few years, researchers have begun to investigate various text classification methods to help instructors and administrators to improve computer-supported learning environments (Kumar and Rosé 2011; Romero and Ventura 2006). Text classification is an application of machine-learning technology to a structured representation of text, which has been a major focus of research in the field of NLP during the past decade. Typically, text classification is the automatic extraction of interesting, frequently implicit, patterns within large data collections (Klosgen and Zytkow 2002). Nowadays, text-classification tools are normally designed mainly for power and flexibility instead of simplicity (Romero and Ventura 2006), which can assess student's learning performance, examine learning behavior, and provide feedback based on the assessment (Castro et al. 2005). Consequently, most of the current text-classification tools are too complex for educators to use, and thus their features go well beyond the scope of what an educator might require (Romero and Ventura 2006).

Therefore, TagHelper (Dönmez et al. 2005) and its successor SIDE (Mayfield and Rosé 2010a) were developed to automate the content analysis of collaborative online discussions. As a publically available tool, TagHelper has been downloaded thousands of times in over 70 countries. Recently, application of TagHelper for automated tutoring and adaptive collaboration scripts have been extensively researched (Kumar and Rosé 2011). In order to make TagHelper tools accessible to the widest possible user base, default behavior has been set up in such a way that users are only required to provide examples of annotated data along with un-annotated data. TagHelper first extracts features like line length, unigrams (i.e., single words), bigrams (i.e., pairs of words that occur next to each other in the text), and part-of-speech bigrams (i.e., pairs of grammatical categories that appear next to one another in the text) from the annotated data. An interface for constructing rule-based features is also provided. In SIDE, more sophisticated support for extracting meaningful features is included, such as regular expressions, which are important in the area of information extraction and named entity recognition, which we make use of in the study reported in this paper. Recent work has also yielded approaches for automatic feature construction and support for error analysis (Mayfield and Rosé 2010b), which further enhances the ability to construct richer and more effective representations of text in preparation for machine learning. Tools such as TagHelper and SIDE then build models based on the annotated examples that it can then apply to the un-annotated examples. To get the best results, both tools allow users to switch easily between different machine learning algorithms provided by Weka (Witten and Frank 2005), such as Naïve Bayes, SMO, and J48.

Despite the effectiveness of applying TagHelper to analyze text-based online discussions, at least two challenges associated with the current NLP approach still need to be addressed. First, the automatic approach has so far only been demonstrated on annotated examples from corpora that come from a single scenario, and the generated model is quite context sensitive and case dependent, and has not been demonstrated to transfer well to online discussions with different topics. Second, the units of analysis (Weinberger and Fischer 2006) to be coded on multiple dimensions were identified by human analysts in our prior published work. Otherwise, the noise added by errors in the automatic segmentation leads to unsatisfactory coding results (Rosé et al. 2008). However, such an automatic segmentation is imperative as a precursor to investigating the use of text classification models for triggering the timing of real-time adaptive fading in our threaded discussion context. These two crucial issues motivated the investigation to explore whether the use of more advanced natural language processing technology can offer fully automatic and context-independent automation techniques for content analysis.

## Explanation of why generality of trained models is a problem

While automatic analysis of collaborative learning discussions is a relatively new area, analysis of social media such as blogs, discussion fora, and chat data has grown in popularity over the past decade and provides important insights to help us understand where issues regarding generality of trained models come from. In particular, results on problems such as gender classification (Argamon et al. 2003), age classification (Argamon et al. 2007), political affiliation classification (Jiang and Argamon 2008), and sentiment analysis (Wiebe et al. 2004) demonstrate how difficult stylistic classification tasks can be, and even more so when the generality is evaluated by testing models trained in one context on examples from another context. Prior work on feature engineering and domain adaptation has attempted to address this generalization difficulty. Here we review this extensive work, which demonstrates that while small advances towards generalization of trained models have been made, it remains an open problem in the field of language technologies, and thus in order to make practical progress in the field of CSCL, we must approach the problem in a more applied way by utilizing insights from our specific problem context, which we begin to describe in the following section.

One major challenge for training generalizable models is that there is typically a confound between topic distribution and whatever stylistic or structural variable is of interest. For example, the large body of work on analysis of gender based stylistic variation offers compelling examples that illuminate the reasons why generality of trained models is difficult to achieve (Argamon et al. 2003; Corney et al. 2002; Mukherjee and Liu 2010; Schler 2006; Schler et al. 2006; Yan and Yan 2006; Zhang et al. 2009). Gender based language variation arises from multiple sources. For example, within a single corpus comprised of samples of male and female language that the two genders do not speak or write about the same topics. Word based features such as unigrams and bigrams are highly likely to pick up on differences in topic rather than style (Schler 2006).

Recent work in the area of domain adaptation (Arnold 2009; Daumé III 2007; Finkel and Manning 2009) raises further awareness of the difficulties with the generality of trained models and offers insight into the reasons for the difficulty with generalization. One important issue is that variation in text feature distributions may be caused by multiple factors that are not independent and evenly distributed in the data. These confounding factors confuse learning algorithms because the multiple factors that lead to variation in the same textual features are difficult to tease apart. What exacerbates these problems in text processing approaches is that texts are typically represented with features that are at the wrong level of granularity for what is being modeled. Specifically, for practical reasons, the most common types of features used in text classification tasks are still unigrams (i.e., single words), bigrams (i.e., pairs of words that occur next to each other in the text), and part-of-speech bigrams (i.e., pairs of grammatical categories that appear next to one another in the text). Relying on relatively simple features keeps the number of extracted features manageable, which allows for efficient model learning. However, these approaches are prone to overfitting due to their simplicity and to the complicating factors mentioned above. This leads to large performance drops when a model trained on one domain is applied to another.

Specifically, when text is represented with features that operate at levels which are too fine-grained, features that truly model the target style or structural characteristics of interest are not present within the model. Thus, the trained models are not able to capture the style itself and instead make their predictions based on features that merely correlate with that style within that particular data set. This may lead to models that perform well within datasets that contain very similar samples of data, but will not generalize to different subpopulations, or even datasets composed of different proportions of the same subpopulations. Models employing primarily unigrams and bigrams as features are particularly problematic in this respect.

In recent years, a variety of manual and automatic feature engineering techniques have been developed in order to construct feature spaces that are adept at capturing interesting language variation without overfitting to content based variation, with the hope of leading to more generalizable models. PoS ngrams (i.e., sequences of grammatical categories that appear together in a text), which have frequently been utilized in genre analysis models (Argamon et al. 2003), are a strategic balance between informativity and simplicity. They are able to estimate syntactic structure and style without modeling it directly. In an attempt to capture syntactic structure more faithfully, there has been experimentation within the area of sentiment analysis on using structural features referred to as syntactic dependency features (Arora et al. 2009; Joshi and Rosé 2009). However, results have been mixed. In practice, the added richness of the features comes at a tremendous cost in terms of dramatic increases in feature space size. What has been more successful in practice is templatizing the dependency features (i.e., replacing specific words with categories in order to achieve better generalization). Templatizing allows capturing the same amount of structure without creating features that are so specific.

Syntactic dependency based features are able to capture more structure than PoS bigrams, however, they are still limited to representing relationships between pairs of words within a text. Thus, they still leave much to be desired in terms of representation power. Experimentation with graph mining from dependency parses has also been used for generating rich feature spaces (Arora et al. 2010). However, results with these features have also been limited. In practice, the rich features with real predictive power end up being difficult to find amidst a large number of useless features that simply add noise to the model. One approach in this direction has been a genetic programming technique which builds a strategic set of rich features. This approach has proven successful at improving the representational power of features above PoS bigrams, with only a modest increase in feature space size. Successful experiments with this technique have been conducted in the area of sentiment analysis, with terminal symbols including unigrams in one case (Mayfield and Rosé 2010b) and graph features extracted from dependency parses in another (Arora et al. 2010). Nevertheless, improvements using these strategic sets of evolved features have been very small even where statistically significant, and thus it is difficult to justify adding so much machinery for such a small improvement.

Another direction is to construct template based features that combine some aspects of PoS ngrams in that they are a flat representation, and the backoff version of dependency features, in that the symbols represent sets of words, which may be PoS tags, learned word classes, distribution based word classes (such as high frequency words or low frequency words), or words. Such types of features have been used alone or in combination with sophisticated feature selection techniques or bootstrapping techniques, and have been applied to problems such as detection of sarcasm (Tsur et al. 2010), detection of causal connections between events (Girju 2010), or gender (Gianfortoni et al. 2011).

## The Automatic Classification of Online Discussions with Extracted Attributes (ACODEA) framework

Typical text classification for online discussions in CSCL is made to be applied by humans. These approaches rely strongly on implicit knowledge held by human coders (e.g., understanding sentences with misspelled words or wrong grammar) to reach an acceptable level of reliability. Text classification that should be applied automatically has to account for the more limited features that are usually used to train automatic classifiers. Our following framework supports the development of such classification schemes.

### Background on classification

Before delving into the specific processes of how the machine-learning tool operates, we further clarify the concepts that are to be classified. Witten and Frank (2005) detail how data can be associated with classes or concepts, which should be reproducible by tools for Natural Language Processing, intelligible to human analysts, and operational to be applied to actual examples. The starting point for understanding online-discussion analysis is to define the coding schemas. In choosing the coding schemas, the researcher needs to determine what sized segments (which range from single word, sentence, paragraph, to the entire message) match with the desired and target activities to be coded (Strijbos et al. 2006). Thus the first target concept to learn is to classify, at each word, whether a segment boundary occurs. Similar to an earlier segmentation approach (Rosé et al. 2008), the concept of segmentation is implemented as a “sliding window” consisting of a specific number of words. In this way, any segmentation is possible since the boundary between any neighboring pair of words is a possible site for a segment boundary. The second concept considered here is to sort each unit of analysis (segment) to one or more categories (dimensions of analysis). For instance, a specific sentence, utterance or message is classified according to quality of argumentation or social mode of interaction (Weinberger and Fischer 2006).

Each individual instance (word in the text to be segmented and then coded) provides an input to machine learning, which is characterized by a fixed and predefined set of features or attributes. Text classification often requires data transforming into appropriate forms (Han and Kamber 2006). Attribute construction (or feature construction), where new attributes are constructed and added from the given set of attributes, can help provide richer, more effective features for representing the text prior to text classification, consequently, ease the training of automatic classifiers as well.

### Overview of proposed approach

In this article we explore several enhancements to this machine-learning technology in order to overcome these challenges. For example, one promising direction to consider is the integration of information-extraction techniques for improving content analysis. Previous work on applying NLP in the field of CSCL, generally accepted raw text as input for segmenting and coding, and the features used for classification were very low-level and simplistic. The new approach used in the current study draws from techniques in information extraction, which allow the construction of a more sophisticated representation of the text to build the classification models on. Such technology includes named-entity recognition, which is an active area of research in the field of language technologies (*MUC6* 1995).

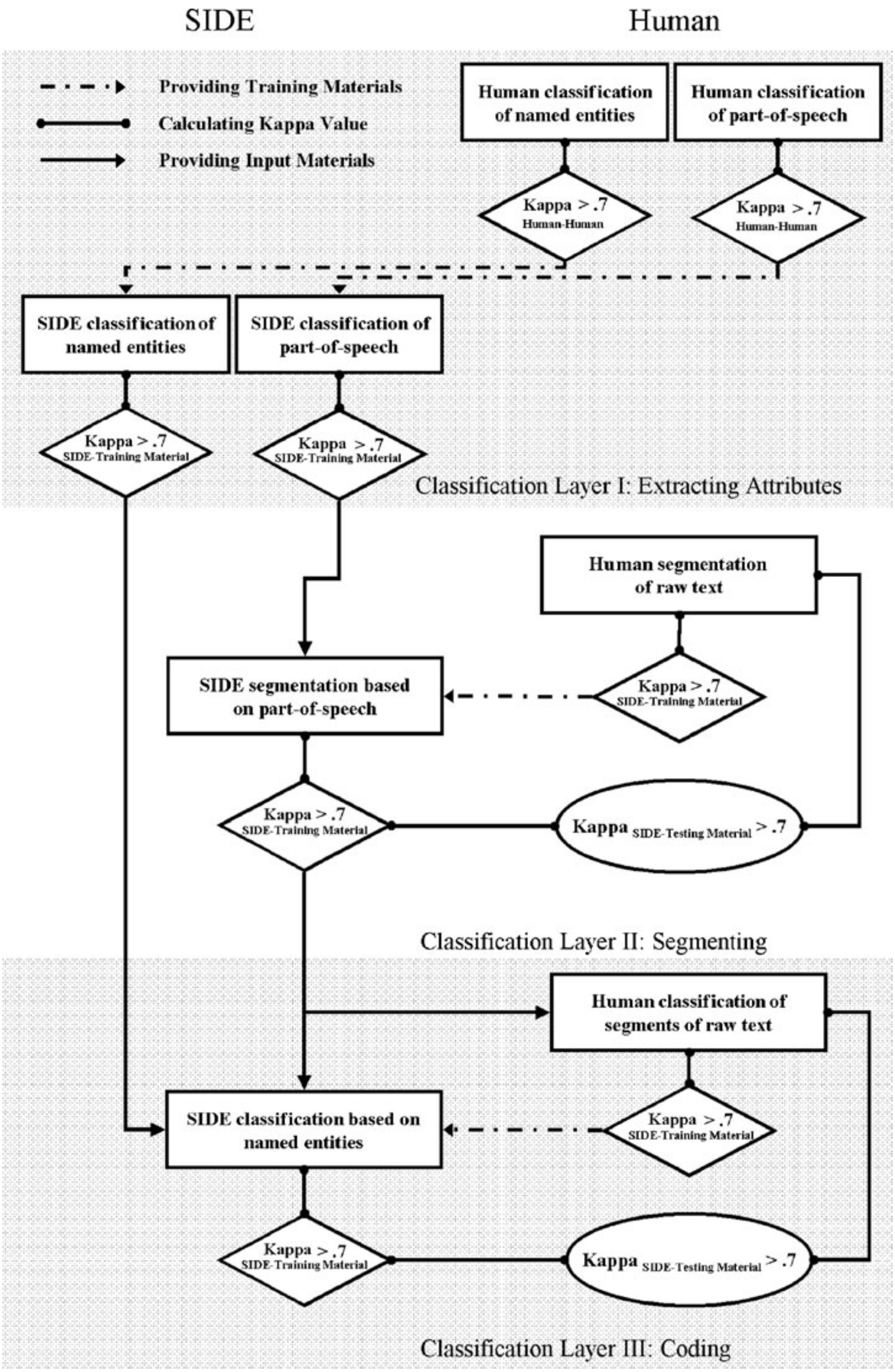
*Part-of-Speech tagging* (PoS) is the process of assigning a syntactic class marker to each word in a text (Brill 1992; Mora and Peiró 2007), therefore, a PoS tagger can be considered as a translator between two languages: the original language that has to be tagged and a "machine friendly" language formed by the corresponding syntactic tags, such as noun or verb. As Poel et al. (2007) proposed, PoS tagging is often only one step in a text-processing application. The tagged text could be used for deeper analysis. Instead of using PoS as the default generalized features, it makes sense to apply modified and specialized PoS categories and thereby to facilitate automatic segmentation if the unit of analysis is syntactically meaningful.

The goal of *Named-Entity Recognition* (NER) is to classify all elements of certain categories of "proper names" appearing in the raw text, into one of seven categories: person, organization, location, date, time, percentage, and monetary amount (*MUC6* 1995). Core aspects of NER are entity and mentions. Mentions are specific instances of entities. For example, mentions of the entity class "location" are New Brunswick, Rhodes, and Hong Kong. Therefore NER provides not only additional features based on extracted entities for each word, but also a more context-independent way to train automatic classifiers. The mentions of New Brunswick, Rhodes, Hong Kong are cities in, for example, the discussions about three past CSCL conferences, while Bloomington, Utrecht, and Chicago would have the same semantic function within discussions about three past ICLS conferences. As an initial step of pre-processing in information-extraction applications, an automatic classifier that had been trained with predefined entities (e.g. "location") instead of specific mentions (e.g. Hong Kong) might have more flexibility for modeling contextual information, potentially improving classification performance. More recently, there have been tasks developed to deal with different practical problems (IREX and CoNLL-2002), in which every word in a document must be classified into one of an extensive set of predefined categories, rather than only identifying names of people and locations.

With the support of current approaches in information extraction, the input to SIDE is assumed to be enhanced in a fully automatic way to be less context-dependent. In the following section, we will present the multi-layer framework for the development of classification schemes for automatic segmentation and coding.

Figure 1 is a flow-process diagram that illustrates how to apply our framework, the Automatic Classification of Online Discussions with Extracted Attributes (ACODEA), to achieve a fully automated analysis. Generally, there are three main layers in the proposed framework. The labeled rectangles represent the text classifications on the hierarchical layers, which are stacked with the pre-processing layer at the top, the segmenting layer at the middle and the coding layer at the bottom. The solid lines with arrows show how the output at the upper layer is the input for the lower layer. The dotted lines represent the information flow in one direction to offer the manually coded materials for training and testing the marching learning models. The Kappa values in diamond shapes are used to indicate a deciding point in the flow process where a test must be made to check the initial

Q2 **Fig. 1** Flowchart to outline the ACODEA framework with (**a**) hierarchical layers focusing on the converting procedural from input raw data to the final output coding results (**b**) parallel processing between human provider of training material and SIDE; and (**c**) branching points to decide if the reliability of the training materials or models are achieving the acceptable level (Kappa value 0.70 or higher) ▶


SIDE
Human
Providing Training Materials
Calculating Kappa Value
Providing Input Materials
Human classification of named entities
Human classification of part-of-speech
Kappa > .7 Human-Human
Kappa > .7 Human-Human
SIDE classification of named entities
SIDE classification of part-of-speech
Kappa > .7 SIDE-Training Material
Kappa > .7 SIDE-Training Material
Classification Layer I: Extracting Attributes
Human segmentation of raw text
SIDE segmentation based on part-of-speech
Kappa > .7 SIDE-Training Material
Kappa > .7 SIDE-Training Material
Kappa SIDE-Testing Material > .7
Classification Layer II: Segmenting
Human classification of segments of raw text
SIDE classification based on named entities
Kappa > .7 SIDE-Training Material
Kappa > .7 SIDE-Training Material
Kappa SIDE-Testing Material > .7
Classification Layer III: Coding

agreement of the training materials between human coders, as well as the reliability between machine SIDE and training material coded by hand. Oval shaped boxes signify the "ending" points of the process, if there is an expected agreement between SIDE and the additionally human-coded materials for testing the training models on the lays of segmenting and coding.

On the first layer of extracting attributes, the part-of-speech tagger and named entity recognition system are applied independently. We extract extra features from the text with the aim to construct a representation suitable for applying machine learning to, either the segmentation layer or the coding layer. The basic rules are to apply part-of-speech tagging and named entity recognition to extract features that are abstract enough to make interesting patterns apparent to machine learning algorithms and yield models that generalize well. On both the syntactic and semantic levels, rather than use predefined categories, we design customized sets of labels that extract information about the specific tasks or target activities we wish to classify. These labels align with behaviors that participants are expected to use during the discourse. In this case, each single word in the raw data for training must be pre-processed by human coders to extract the syntactic and semantic features. These annotated examples, which reach acceptable reliability, can then be used to train classifiers for all defined categories.

In addition, the entire architecture is structured to cascade from one layer to the next, incorporating information from the previous layers to improve the current classifier's performance. Extracting attributes on the syntactic level benefits from the use of off-the-shelf grammatical part-of-speech taggers, while the layer related to semantic representation benefits from the inclusion of named entities and techniques from information extraction. The output from these layers is used as the attributes for the final classification layers of segmentation and coding.

In this paper we propose that the problems introduced above, more specifically, the automatic segmenting and context-independent coding can be addressed by extracting abstract syntactic and semantic features beyond baseline feature spaces consisting of word-level representations such as unigrams and bigrams.

On the second layer, human coders have to classify the borders between the segments with raw data. These human coded examples are used to train the automatic segmentation by machine. However, the input to SIDE for generating the segmentation classifier is the set of preprocessed concepts from the syntactic attributes, instead of the raw text. By using the new technique of a sliding window, the segmentation model can be trained with high reliability (i.e., regarding the identification of borders between segments). This segmentation model can then be successfully applied to divide all the preprocessed data into the desired unit of analysis automatically.

Once the data is segmented, human coders have to classify all segments in the training data consistent with the dimensions defined for the coding layer. This is required to make sure both the human coder and SIDE classify the same segments whose boundary has already been identified by SIDE automatically.

This layered model is motivated by the idea that the initial layers allow the machine learning model trained at a higher level to learn more general patterns. However, there is a risk inherent in such an approach: Several classifications are made in a row, and thus errors on the different layers may be cascaded. Therefore, the final automatic classification must ultimately be checked against pure human coding to ensure reliability. We present such an evaluation in the following sections.

## Research questions

In the following, we will present a use case for this multi-layer framework. The main question addressed in this study is: how does the multilayer ACODEA framework perform in automatically analyzing discourse data? We divide this question into three sub-questions of interest:

RQ1: Can the first classification layer be automated with satisfactory reliability to extract syntactic and semantic attributes?

We expect that it is possible to achieve an acceptable level of agreement between automatically generated codes and human codes when we automate the classification on the layer of extracting the desired attributes (H1).

RQ2: Can the framework be applied towards the second segmentation layer successfully?

Regarding the layer of segmentation, we make the prediction that (H2a) the reliability between SIDE and Human coders is also at an acceptable level. Moreover, (H2b) segmenting based on pre-processed data by extracting the syntactic features is expected to outperform the approach of directly dividing the raw data into units of analysis. In addition one more hypothesis (H2c) about the effectiveness of the present approach is that the ACODEA framework can be applied to train context independent segmentation with sufficient reliability

RQ3: Can the automatic coding be implemented as the third layer of the multiple-layer classification with success?

With respect to the final goal of the present framework to fully automate the content analysis, we expect that (H3a) the performance of the NLP tool is satisfactory enough to achieve acceptable agreement (Kappa value 0.7 or even higher) with the human judgment. Compared with the coding process without extracting semantic features, ACODEA framework is hypothesized (H3b) to be capable of enhancing the reliability of automatic classification at the third layer. Furthermore, we expect that the framework can be also used to train context-independent classification with sufficient reliability.

## Method

### Participants and learning task

The composition of the training models for SIDE and the consequent evaluation of the innovative application of the NLP tool have lead to the implementation of a computer supported collaborative learning study for the discourse data collection. In the year of 2010, eighty-four (84) students of Educational Science at the University of Munich participated in this study. Students were randomly assigned to groups of three. Each group was randomly assigned to one of three experimental conditions. Even though the experimental treatments differ in the degree of receiving instructional scaffolding, learning tasks were the same across all groups. Learners were required to join a collaborative, argumentative online discussion and solve five case-based problems by applying an educational theory. The computer-based learning environment used in this experiment is a modified version of the one employed by Stegmann et al. (2007). The instructional scaffolding was implemented using S-COL (Wecker et al. 2010).

The chosen theory the students were applying in their discussions within the environment was Weiner's attribution theory (1985) and its application in education. The students individually read a lesson on attribution theory and a text introducing argumentation. In the collaborative learning phase, three problem cases from practical contexts were used as a basis for the online discussions. The case “Math” describes the attributions of a student with respect to his poor performance in mathematics. In the case “Class reunion” a math tutor talks about how he tries to help female students deal with success and failure in assignments. The case “Between-culture variance” describes differences in school performance between Asian and American/European students that were explained by attribution theory. Another two cases were used in the pre and post test, which mainly concern the factors that affect a student’s choice of a major at the university and student’s explanation for failure in the exam of “Text analysis”. In this empirical study the problem-based cases students are facing are designed to be varying crossing real-life studying contexts, therefore learners can apply the knowledge of attribution theory and argumentation skills to the various contexts (operationalized with the five cases mentioned above).

The multiple and complex conversations are arranged into threads within the discussion environment. Learners have the option either to start a new thread by posting a new message or reply to messages that had been posted previously. Replies can be oriented to the main topic, or to the reply posted by another member of the learning group, or even to someone's reply to a reply. Metadata features such as author, date, and post time are recorded by the environment. The learners enter the subject line and the body of the message themselves. When students reply to a previous message, the text of that message is also included within the body of the message, although that portion of text is marked in a different color to set it apart from the new message content.

## Data source and procedure

We collected 140 conversation transcripts, each of which contained the full interaction from one group, and was targeted to a single scenario. Altogether, there are 74,764 words in the corpus. Two human coders analyzed almost one fifth of the raw data (equally distributed over five cases). About half of the human-coded data were used as the training materials on which a few automatic models can be built by SIDE, including the feature extraction, segmentation, and coding layers described earlier (Mayfield and Rosé 2010a). The left manually coded dataset were further used for material to test the training models. SIDE includes an annotation interface allowing for automatic and semi-automatic coding. To train such classifiers with SIDE we had to provide examples of annotated data. SIDE extracted multiple features from the raw data, like line length, unigrams, bigrams, part-of-speech bigrams, etc. Machine learning algorithms use these features to learn how to classify new data. As output, SIDE builds a model based on the human annotated data. This model can then be easily applied to classify un-annotated data, and then the assigned codes can be further reviewed on the annotation interface, which facilitates the process of humans correcting errors made by the automatic coding. Furthermore, SIDE employs a consistent evaluation methodology referred to as 10-fold cross-validation, where the data for training the models can be randomly distributed into 10 piles. Nine piles are combined to train a model. One pile is used to test the model. This is done 10 times so that each segment is used as a test set once. And then the performance values are averaged to obtain to final performance value.

## Statistical tests

The reliability of the coding was measured using Cohen’s Kappa value and percent agreement. Both of the indexes have been regarded as accepted standards for measuring coding reliability. Percent agreement is the most simple and most popular reliability coefficient (De Wever et al. 2006). Statistically, the inter-rater agreement is determined by dividing the number of codes, which is agreed upon by the total number (agree and disagree all inclusive) of codes. Supplemental criterion for success is reaching a level of inter-rater reliability with a gold standard as measured by Cohen's Kappa that is 0.7 or higher (Strijbos et al. 2006). Here it is worthwhile to further clarify that the present study was undertaken to evaluate different types of Kappa in the distinguishable phases, including (1) inter-rater agreement between human coders Kappa$_{\text{(Human-Human)}}$ to evidence the initial reliability of training examples; (2) inter-rater agreement generated by the 10-fold cross-validation (that is, 10 iterations of training and testing are performed and within each iteration a different fold of the selected data is coded by SIDE for validation while the remaining 9 equally sized folds are used for training.) to certify the internal reliability of the SIDE training models, The 10 results from comparing the coding between SIDE and manually coded training materials then can be averaged to produce a single estimation Kappa$_{\text{(SIDE-Training Material)}}$; and finally (3) the conclusive Kappa$_{\text{(SIDE-Testing Material)}}$ between SIDE and human coders calculated with the additional testing materials.

## Application of the framework

The layer described below is the core part of the architecture for extracting features from the text in order to construct a representation for it that is suitable for applying machine learning to, either for the coding layer or the segmentation layer.

(ia) Regarding the syntactic attributes: Each word in the computerized data can be pre-processed into multiple and syntactic categories. An example of such a tag is: Term, Verb, Property, Conjunction, Comma/Stop Symbol, and so on. These tags are a reduced version of the full tag set, making it more suitable for machine learning. Some stop words like Pronoun are clustered into the class of Other.

(ib) Regarding the semantic attributes, each single word in the text can fall into one of the multiple categories, either (a) Case, key words from problem space, (b) Theory, key words from the concerned conceptual space (actually, attribution theory in the present study), or (c) Extraneous theory, from the related educational theory. In addition, there are words that are important in reflecting the (d) evaluation either positive or negative among partners (which refers to key indicator of Counterargument), (e) Empty Message, and even (f) Other activities, can be extracted in this phase. All of the categories are chosen because they might support the coding on the classification layer. For instance, according to our learning task a claim would typically contain both case and theory information, while a ground mainly includes case information and a warrant only includes elaborations on attribution theory.

(ii) The unit of analysis was defined as a sentence or part of a compound sentence that can be regarded as 'syntactically meaningful in structure' (cf. Strijbos et al. 2006). For instance, according to these rules of segmentation, punctuation and the special words like 'and' are boundaries that can be used to segment compound sentences if the parts before and after the boundary are 'syntactically meaningful' segments. This size of segment has been proved to be reliable (Strijbos et al. 2006), and suitable for the coding dimension conducted in the current study. An entire process of extracting attributes, segmenting and coding of the selected example of argumentative discussion is illustrated in the Fig. 2.

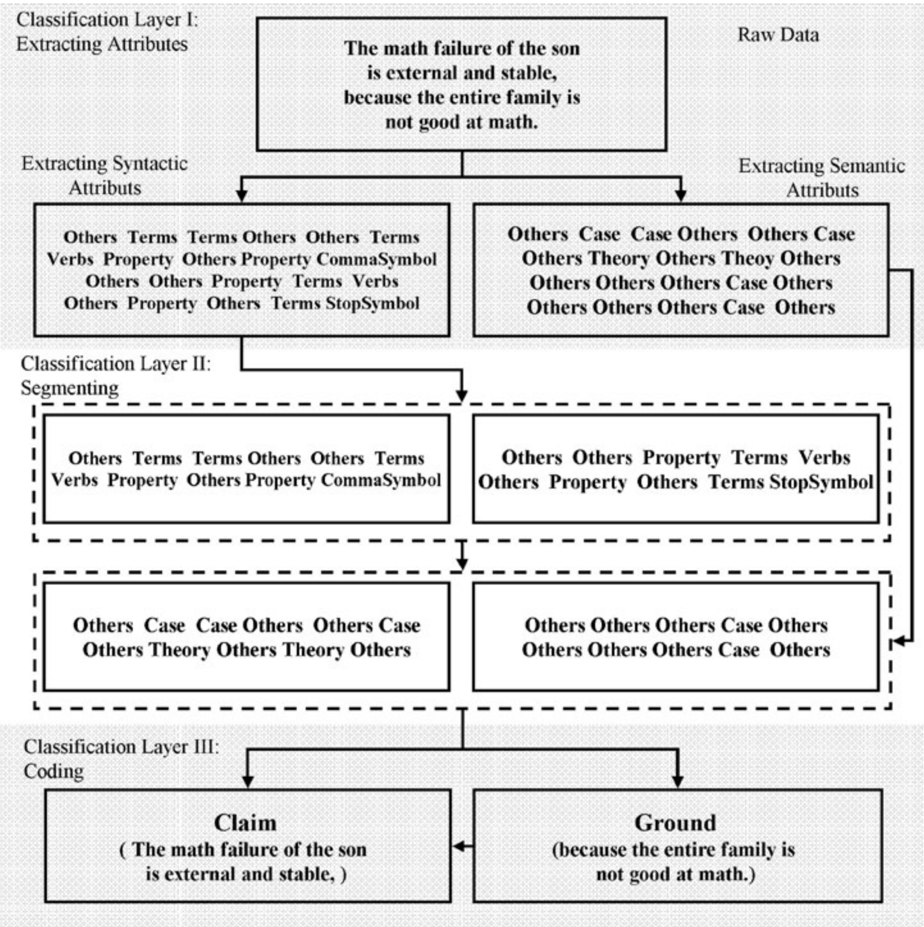

Classification Layer I:
Extracting Attributes
The math failure of the son is external and stable, because the entire family is not good at math.
Raw Data
Extracting Syntactic Attributs
Extracting Semantic Attributs
Others Terms Terms Others Others Terms Verbs Property Others Property CommaSymbol Others Others Property Terms Verbs Others Property Others Terms StopSymbol
Others Case Case Others Others Case Others Theory Others Theoy Others Others Others Others Case Others Others Others Others Case Others
Classification Layer II:
Segmenting
Others Terms Terms Others Others Terms Verbs Property Others Property CommaSymbol
Others Others Property Terms Verbs Others Property Others Terms StopSymbol
Others Case Case Others Others Case Others Theory Others Theory Others
Others Others Others Case Others Others Others Others Case Others
Classification Layer III:
Coding
Claim
( The math failure of the son is external and stable, )
Ground
(because the entire family is not good at math.)

**Fig. 2** Application of the ACODEA framework (example)

(iii) Our coding layer was defined with respect to the approach of argumentative knowledge construction. Learners construct arguments in interaction with their learning partners in order to acquire knowledge about argumentation as well as knowledge of the content under consideration (Andriessen et al. 2003). Therefore on this layer we are mainly concerned with the following categories, based on the micro-argumentation dimension of the multidimensional framework developed by Weinberger and Fischer (2006):

(a) *Claim* is a statement that advances the position learners take to analyze a case with attribution theory.
(b) *Ground* is the evidence from a case to support a claim.
(c) *Warrant* is the logical connection between the grounds and claims that present the theoretical reason why a claim is valid. Consequently,
(d) *Inadequate Claim* should be differentiated in the coding, which concerns other related educational theory to explain a case.
(e) *Evaluation* is an expression of agreement or disagreement with a learning partner.

There are more technical dimensions to indicate the

(f) *Prompts*, which are the computer-generated prompts to structure the argumentative discourse, and

(g) *Other*, which cannot be sorted by any other dimensions, and finally

(h) *Empty Message* is the computer-generated message to report that the segment has no content.

## Results

Two coders created the training material for SIDE. The inter-rater agreement between two human coders was Cohen's Kappa$_{\text{(Human-Human)}}$=0.93 on the syntactic-attributes layer and Cohen's Kappa$_{\text{(Human-Human)}}$=0.97 on the semantic-attributes layer. In addition, the human coders achieved a high value of Cohen's Kappa$_{\text{(Human-Human)}}$=0.96 for the segmentation layer and Cohen's Kappa$_{\text{(Human-Human)}}$=0.71 for the coding layer. These results indicate acceptable human baseline performances for SIDE to be trained to analyze the un-annotated data regarding the extracted attributes, segmentation and coding layers.

RQ1: Can the first classification layer be automated with satisfactory reliability to extract syntactic and semantic attributes?

SIDE achieved an overall Cohen's Kappa$_{\text{(SIDE-Training Material)}}$=0.94 (Percent Agreement= 91.7 %) with the training material on the syntactic layer, and an overall of Cohen's Kappa$_{\text{(SIDE-Training Material)}}$=0.93 (Percent Agreement=91.0 %) on the semantic layer. An independent human coder (who created the testing material) and SIDE achieved an agreement of Cohen's Kappa$_{\text{(SIDE-Testing Material)}}$=0.92 (Percent Agreement=93.4 %) on the syntactic layer, and Cohen's Kappa$_{\text{(SIDE-Testing Material)}}$=0.84 (Percent Agreement=93.5 %) on the semantic layer. As shown in Table 1, the reliability of SIDE to analyze text on the syntactic and semantic layers is satisfactory across all five cases. Because the precision on the layer of extracting attributes greatly influences the performance of the steps further in the chain of linguistic treatments, the inter-rater reliability and agreement of the PoS tagger and named entity recognition is especially important.

RQ2: Can the framework be applied towards the second segmentation layer successfully?

Internal Cohen's Kappa$_{\text{(SIDE-Training Material)}}$=0.98 (Percent Agreement=99.6 %) was achieved by SIDE when it attempted to automatically segment the text . An overall model of segmenting was produced by the training material which is distributed evenly among the 5 cases, and has been pre-processed to extract syntactic attributes. A human coder and SIDE

**Table 1** Reliability of automatic attribute extraction within the 5 Cases (SIDE vs. testing material)

| Case | Syntactic-attributes layer I | | Semantic-attributes layer I | |
|---|---|---|---|---|
| | Cohen's kappa | Percent agreement | Cohen's kappa | Percent agreement |
| Overall cases | 0.94 | 91.7 % | 0.93 | 91.0 % |
| Major choice | 0.90 | 92.2 % | 0.95 | 98.1 % |
| Math | 0.85 | 88.1 % | 0.92 | 96.9 % |
| Class reunion | 0.91 | 92.9 % | 0.90 | 96.0 % |
| Between-culture variance | 0.93 | 94.4 % | 0.96 | 98.3 % |
| Text analysis | 0.88 | 90.3 % | 0.94 | 97.7 % |

achieved an agreement of Cohen's Kappa$_{\text{(SIDE-Testing Material)}}$=0.97 (Percent Agreement= 99.3 %). The inter-rater reliability of operating the overall model to the five different cases is displayed in the Table 2. The algorithm for segmenting generated on the base of the layer of extracting attributes achieved sufficiently higher Cohen's Kappa across the tested cases compared with the approach without extracting attributes.

In addition, to further prove the segmentation is context-independent, five distinct segmentation models based each on a single case have been verified by using two kinds of testing material, which is either consistent with trained models (e.g. testing the Math model with the text discussing on the case Math) or not (e.g. testing the Math model with the text of the other four cases). Only slight differences were found when the testing material was inconsistent with the training material (as shown in Table 3).

RQ3: Can the automatic coding be implemented as the third layer of the multiple classification with success?

SIDE achieved an internal Cohen's Kappa$_{\text{(SIDE-Training Material)}}$=0.77 (Percent Agreement= 81.3 %) using the extracted semantic attributes across all cases during training. The reliability across all cases comparing SIDE with a human coder (based on raw text) was sufficiently high (Cohen's Kappa$_{\text{(SIDE-Testing Material)}}$=0.81; Percent Agreement=84.5 %). As shown in Table 2, sufficient inter-rater agreement values were achieved for applying the overall model to all of the cases. It is also obvious that extracting semantic attributes substantially increases the agreement between human coders and SIDE. For example, the classification without extracting semantic attributes resulted in less acceptable kappa values of 0.47 for the case of Class reunion data and a kappa value of 0.53 for the case of Between-culture variance.

In order to provide additional statistical evidence for the assumed context-independent coding, five training models have been generated. The results of comparing the reliability of

**Table 2** Comparison without and with the layer of extracting attributes to automate the content analysis (SIDE vs. testing material)

| | Without extracting attributes | | With extracting attributes | |
|---|---|---|---|---|
| Segmentation layer II | Cohen's Kappa | Percent Agreement | Cohen's Kappa | Percent Agreement |
| Kappa $_{\text{SIDE-Training Material}}$ | 0.84 | 96.7 % | 0.98 | 99.6 % |
| Kappa $_{\text{SIDE-Testing Material}}$ | 0.86 | 97.0 % | 0.97 | 99.3 % |
| Major choice | 0.80 | 96.7 % | 0.95 | 99.1 % |
| Math | 0.86 | 96.6 % | 0.96 | 98.9 % |
| Class reunion | 0.87 | 97.0 % | 0.97 | 99.3 % |
| Between-culture variance | 0.90 | 97.7 % | 0.99 | 99.7 % |
| Text-analysis | 0.83 | 96.9 % | 0.98 | 99.6 % |
| Coding layer III | Cohen's Kappa | Percent Agreement | Cohen's Kappa | Percent Agreement |
| Kappa $_{\text{SIDE-Training Material}}$ | 0.70 | 75.6 % | 0.77 | 81.3 % |
| Kappa $_{\text{SIDE- Testing Material}}$ | 0.61 | 67.8 % | 0.81 | 84.5 % |
| Major choice | 0.63 | 71.2 % | 0.77 | 82.9 % |
| Math | 0.67 | 72.3 % | 0.78 | 82.6 % |
| Class reunion | 0.47 | 58.5 % | 0.76 | 81.0 % |
| Between-culture variance | 0.53 | 63.1 % | 0.85 | 87.5 % |
| Text-analysis | 0.68 | 75.0 % | 0.87 | 89.2 % |

**Table 3** Consistent vs. inconsistent models (training/testing) to automate the content analysis

| Training/testing models | Consistent | | Inconsistent | |
|---|---|---|---|---|
| Segmentation layer II | Cohen's Kappa | Percent Agreement | Cohen's Kappa | Percent Agreement |
| Major choice | 0.97 | 99.2 % | 0.95 | 98.8 % |
| Math | 0.96 | 98.9 % | 0.96 | 99.2 % |
| Class reunion | 0.94 | 99.0 % | 0.97 | 99.3 % |
| Between-culture variance | 0.98 | 99.5 % | 0.97 | 99.2 % |
| Text-analysis | 0.95 | 98.8 % | 0.95 | 98.9 % |
| Coding layer III | Cohen's Kappa | Percent Agreement | Cohen's Kappa | Percent Agreement |
| Major choice | 0.72 | 80.3 % | 0.66 | 72.0 % |
| Math | 0.73 | 78.3 % | 0.65 | 72.0 % |
| Class reunion | 0.59 | 81.9 % | 0.74 | 76.9 % |
| Between-culture variance | 0.74 | 78.4 % | 0.59 | 67.3 % |
| Text-analysis | 0.62 | 69.9 % | 0.74 | 78.5 % |

the models to code the selected texts which are either consistent with the training material or not, are further displayed in Table 3. It is important to note that we only observe a slight fluctuation of the reliability results when we test on different cases than we trained on. Nevertheless, the level of performance we achieved with the multi-layer approach is still acceptable (above or close to the cut-off value of 0.70). In certain contexts (namely the class reunion and text analysis contexts) the coding model indeed performed better with the inconsistent testing cases. The relative stability in the Cohen`s Kappa as well as Percent Agreement indicates the improved approach of automatic coding is adequate with respect to context-independence.

## Discussion

This paper proposes a systematic framework called ACODEA (Automatic Classification of Online Discussions with Extracted Attributes), which has been applied successfully for the design, implementation and evaluation of a methodology for automatic classification of a large German text corpus. Due to the extracted syntactic and semantic features, ACODEA allows a bottom-up specification of the in-depth information contained within the discourse corpus and it is therefore more precise and reliable than the traditional approach without extracting features during a pre-processing phase before content analysis. More importantly, it provides insights to identify the unit fragments automatically, when the inputs for segmentation consist of computer-friendly syntactic symbols. Also, the acceptable reliability of automatic segmentation and coding across various contexts in the current study offers hope that the resulting classification models can be quickly adapted for a new knowledge domain by adding a simple specification of the semantic/syntactic attributes at the pre-processing layer.

As compared to previous work, the ACODEA framework introduced here has made substantial headway towards addressing the most challenging methodological problems with respect to full automation and context independence. Besides the above-mentioned contributions to automated discourse analysis, we also show significant progress in bridging the gap between the approved methodological improvements on the one hand and the inadequate practical application on the other hand. Automating detailed content analyses is not a novelty in CSCL research, which has been conducted for many intentions (Walker et al. 2009). At present, however, the number of studies demonstrating the explicit specifics of an analysis method is limited. In this respect it would be challenging — for the novices who lack extensive computational linguistics knowledge — to implement the existing analysis framework in novel situations. This paper can also serve as an example of the process that is required to use such technology in a CSCL context in order to serve as a model that other researchers can follow in their own work.

A focus on transfer of learning from one context to another, which has witnessed a great increase in attention in recent years in the field of CSCL, is defined as the expectation that the knowledge/skill previously acquired in carrying out a cognitively complex learning task can be applied in a context different from the learning context. As elaborated in the current study, our goal was to perform analysis of discourses that not only occurred during the collaborative-learning phase of our study, but also that took place before and after the intervention in the transfer cases. From this point of view, one of the most important issues raised in our work is the ability of the improved NLP tool to transfer what it learned from previous domains to new contexts. The contribution of this study is also more efficient than previous work in adaptation across widely disparate domains, e.g., from newswire to biomedical documents (Daumé III 2007). We demonstrate that models trained in one context can be effectively be applied to other target contexts which share some likeness to the original resource context.

Nevertheless, it must be acknowledged that the work presented in this paper is somewhat tailored to the specific analysis associated with the multi-dimensional coding scheme developed in our earlier work. As mentioned above, domain adaptation of text-classification models in the general case is still an open research problem. In other words, the developed approach is merely context-free for analyzing a specific discussion activity, which is assumed to be valuable for learning (e.g., micro-argumentation in this case). The preprocessing steps of PoS tagging and named-entity extraction make strong assumptions about what characteristics of the texts vary from context to context in terms of the five cases investigated in the current study. Hence, the specific coding schemas can only be applied in particular contexts, in which the underlying mechanism of the concerned learning activities is particularly similar and specialized (e.g., epistemic activities embedded in the argumentative knowledge construction). Depending on the domain as well as the type of target learning process, different sets of categories for the layer of extracting semantic attributes may be used, for instance, aiming at the maximum performance of problem solving or thought–provoking questioning.

Considerable efforts in terms of time and other research efforts have been spent on exploring the application of the state-of-the-art text-classification technology to enable content analysis in a fully automatic and reliable way. The present ACODEA framework is critical not only to help researchers to speed up their projects through removing time-consuming tasks, such as segmenting and coding; it may also essentially change the way we design learning environments and scaffold the desired collaborative learning. Specifically, automatic analysis of online discussion could provide instructors with the capability to monitor the learning progress occurring in real-time, to indicate what specific and personalized need should be addressed. In this way, a fully automatic system could enable adaptive intervention for collaborative learning, which is assumed to be more efficient in promoting higher order thinking or collaborative behavior, than the static, one-size-fits-all interventions (Gweon et al. 2006; Kumar et al. 2007).

In addition, one interesting issue should still be further investigated to enrich our coding schemas involving argumentative knowledge construction. Specifically, it would be useful to be able to assess how “strong” the argumentation is, rather than only how structurally complete the argumentation is, as we have done so far. From an epistemic perspective, an appropriate argument is more than a simple pile-up of information from problem and conceptual space, which includes a structurally appropriate connective between specific case and concerned theory. One possibility is that in the pre-processing step, the keywords from case and theory, which are correctly connected corresponding to an expert model, can be weighed automatically. This way, scaffolds provided by an adaptive collaboration script assisted by the automated and customized approach of qualitative content analysis can be much more powerful in its facilitation role, supporting valuable learning processes.

To sum up, it is obvious that the development of ACODEA is a process of breaching scientific boundaries of multiple research domains ranging from education, psychology and computer science to linguistics. Our results can be seen as evidence of progress through interdisciplinary research in the field of CSCL. Empirical evidence in the present study further suggests that the multi-layer content-analysis approach elaborated upon here, along with the outlined steps to be customized for different contexts and alterative coding dimensions of interest, will further stimulate additional and interesting research in the field of CSCL.

## References

Andriessen, J., Baker, M., & Suthers, D. (2003). Argumentation, computer support, and the educational context of confronting cognitions. In J. Andriessen, M. Baker, & D. Suthers (Eds.), *Arguing to learn: Confronting cognitions in computer-supported collaborative learning environments* (pp. 1–25). Dordrecht: Kluwer Academic Publishers.

Argamon, S., Koppel, M., Fine, J., & Shimoni, A. R. (2003). Gender, genre, and writing style in formal written texts. *Text - Interdisciplinary Journal for the Study of Discourse, 23*(3), 321–346. doi:10.1515/text.2003.014.

Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2007). Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday 12*(9).

Arnold, A. O. (2009). *Exploiting domain and task regularities for robust named entity recognition.* PhD thesis, Carnegie Mellon University.

Arora, S., Joshi, M., & Rosé, C. P. (2009). *Identifying types of claims in online customer reviews*. Paper presented at the Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers (pp. 37–40), Boulder, Colorado, USA.

Arora, S., Mayfield, E., Rosé, C. P., & Nyberg, E. (2010). *Sentiment classification using automatically extracted subgraph features*. Paper presented at the Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text (pp. 131–139), Los Angeles, California, USA.

Brill, E. (1992). *A simple rule-based part of speech tagger*. Paper presented at the Proceedings of the Third Conference on Applied Natural Language Processing (pp. 152–155), Trento, Italy.

Castro, F., Vellido, A., Nebot, A., & Minguillon, J. (2005). *Detecting atypical student behaviour on an e-learning system*. Paper presented at the Simposio Nacional de Tecnologas de la Informacin y las Comunicaciones en la Educacion (pp. 153–160), Granada, Spain.

Clark, D., Sampson, V., Weinberger, A., & Erkens, G. (2007). Analytic frameworks for assessing dialogic argumentation in online learning environments. *Educational Psychology Review, 19*(3), 343–374. doi:10.1007/s10648-007-9050-7.

Corney, M., de Vel, O., Anderson, A., & Mohay, G. (2002). *Gender-preferential text mining of e-mail discourse*. Paper presented at the the 18th Annual Computer Security Applications Conference (pp. 21–27), Las Vegas, NV, USA.

Daumé III, H. (2007). *Frustratingly easy domain adaptation*. Paper presented at the the 45th Annual Meeting of the Association of Computational Linguistics (pp. 256–263), Prague, Czech Republic.

De Laat, M., & Lally, V. (2003). Complexity, theory and praxis: Researching collaborative learning and tutoring processes in a networked learning community. *Instructional Science, 31*(1), 7–39. doi:10.1023/a:1022596100142.

De Wever, B., Schellens, T., Valcke, M., & Van Keer, H. (2006). Content analysis schemes to analyze transcripts of online asynchronous discussion groups: A review. *Computers in Education, 46*(1), 6–28. doi:10.1016/j.compedu.2005.04.005.

Diziol, D., Walker, E., Rummel, N., & Koedinger, K. (2010). Using intelligent tutor technology to implement adaptive support for student collaboration. *Educational Psychology Review, 22*(1), 89–102.

Dönmez, P., Rosé, C., Stegmann, K., Weinberger, A., & Fischer, F. (2005). *Supporting CSCL with automatic corpus analysis technology*. Paper presented at the Proceedings of th 2005 Conference on Computer Support for Collaborative Learning: Learning 2005: The Next 10 Years! (pp. 125–134), Taipei, Taiwan.

Duwairi, R. M. (2006). A framework for the computerized assessment of university student essays. *Computers in Human Behavior, 22*(3), 381–388.

Finkel, J., & Manning, C. (2009). *Hierarchical bayesian domain adaptation*. Paper presented at the Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 602–610), Boulder, Colorado, USA.

Gianfortoni, P., Adamson, D., & Rosé, C. P. (2011). *Modeling stylistic variation in social media with stretchy patters*. Paper presented at the First Workshop on Algorithms and Resources for Modeling of Dialects and Language Varieties (pp. 49–59), Edinburgh, Scotland, UK.

Girju, R. (2010). *Towards social causality: An analysis of interpersonal relationships in online blogs and forums*. Paper presented at the the Fourth International AAAI Conference on Weblogs and Social Media (pp. 251–260), Montreal, Quebec, Canada.

Gweon, G., Rosé, C., Carey, R., & Zaiss, Z. (2006). *Providing support for adaptive scripting in an on-line collaborative learning environment*. Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 251–260), Montreal, Quebec, Canada.

Han, J., & Kamber, M. (2006). *Data mining: Concepts and techniques*. San Mateo: Morgan Kaufmann Publishers.

Howley, I., Mayfield, E., & Rose, C. P. (2011). *Missing something? Authority in collaborative learning*. Paper presented at the Connecting Computer-Supported Collaborative Learning to Policy and Practice: CSCL2011 Conference (pp. 366–373), Hong Kong.

Jiang, M., & Argamon, S. (2008). *Political leaning categorization by exploring subjectivities in political blogs*. Paper presented at the the 4th International Conference on Data Mining(pp. 647–653), Las Vegas, Nevada, USA.

Joshi, M., & Rosé, C. P. (2009). *Generalizing dependency features for opinion mining*. Paper presented at the Proceedings of the ACL-IJCNLP 2009 Conference Short Papers (pp. 313–316), Suntec, Singapore.

Klosgen, W., & Zytkow, J. (2002). *Handbook of data mining and knowledge discovery*. New York: Oxford University Press.

Kumar, R., & Rosé, C. (2011). Architecture for building conversational agents that support collaborative learning. *IEEE Transactions on Learning Technologies, 4*(1), 21–34. doi:10.1109/tlt.2010.41.

Kumar, R., Rosé, C., Wang, Y.-C., Joshi, M., & Robinson, A. (2007). *Tutorial dialogue as adaptive collaborative learning support*. Paper presented at the Proceeding of the 2007 Conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work (pp. 383–390).

Landauer, T. K. (2003). Automatic essay assessment. *Assessment in Education: Principles, Policy & Practice, 10*(3), 295–308. doi:10.1080/0969594032000148154.

Mayfield, E., & Rosé, C. (2010a). *An interactive tool for supporting error analysis for text mining*. Paper presented at the Proceedings of the NAACL HLT 2010 Demonstration Session (pp. 25–28), Los Angeles, California.

Mayfield, E., & Rosé, C. (2010b). *Using feature construction to avoid large feature spaces in text classification*. Paper presented at the Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation (pp. 1299–1306), Portland, Oregon, USA.

Mayfield, E., & Rosé, C. P. (2011). *Recognizing authority in dialogue with an integer linear programming constrained model*. Paper presented at the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1 (pp. 1018–1026), Portland, Oregon.

Mora, G., & Peiró, J. A. S. (2007). *Part-of-speech tagging based on machine translation techniques*. Paper presented at the Proceedings of the 3rd Iberian Conference on Pattern Recognition and Image Analysis, Part I (pp. 257–264), Girona, Spain.

MUC6. (1995). *Paper presented at the the sixth message understanding conference*. Maryland: Columbia.

Mukherjee, A., & Liu, B. (2010). *Improving gender classification of blog authors*. Paper presented at the Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (pp. 207–217), Cambridge, Massachusetts.

Poel, M., Stegeman, L., & op den Akker, R. (2007). A support vector machine approach to dutch part-of-speech tagging. In M. R. Berthold, J. Shawe-Taylor, & N. Lavrac (Eds.), *Advances in intelligent data analysis VII* (Vol. 4723, pp. 274–283). Berlin: Springer Verlag.

Romero, C., & Ventura, S. (2006). *Data mining in e-learning*. Southampton: Wit Press.

Rosé, C., & Vanlehn, K. (2005). An evaluation of a hybrid language understanding approach for robust selection of tutoring goals. *International Journal of AI in Education, 15*(4), 325–355.

Rosé, C., Wang, Y.-C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., & 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, 3*(3), 237–271. doi:10.1007/s11412-007-9034-0.

Schler, J. (2006). Effects of age and gender on blogging. *Artificial Intelligence, 86*, 82–84.

Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. (2006). *Effects of age and gender on blogging*. Paper presented at the Proc. of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs (pp. 199–205), Stanford, California, USA.

Stegmann, K., Weinberger, A., & Fischer, F. (2007). Facilitating argumentative knowledge construction with computer-supported collaboration scripts. *International Journal of Computer-Supported Collaborative Learning, 2*(4), 421–447. doi:10.1007/s11412-007-9028-y.

Stegmann, K., Wecker, C., Weinberger, A., & Fischer, F. (2012). Collaborative argumentation and cognitive elaboration in a computer-supported collaborative learning environment. *Instructional Science, 40*(2), 297–323. doi:10.1007/s11251-011-9174-5.

Strijbos, J.-W., Martens, R. L., Prins, F. J., & Jochems, W. M. G. (2006). Content analysis: What are they talking about? *Computers in Education, 46*(1), 29–48. doi:10.1016/j.compedu.2005.04.002.

Tsur, O., Davidov, D., & Rappoport, A. (2010). *ICWSM—a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews*. Paper presented at the the Fourth International AAAI Conference on Weblogs and Social Media (pp. 162–169), Washington, DC, USA. http://staff.science.uva.nl/~otsur/papers/sarcasmAmazonICWSM10.pdf

Walker, E., Rummel, N., & Koedinger, K. (2009). CTRL: A research framework for providing adaptive collaborative learning support. *User Modeling and User-Adapted Interaction, 19*(5), 387–431.

Wang, H.-C., Rosé, C., & Chang, C.-Y. (2011). Agent-based dynamic support for learning from collaborative brainstorming in scientific inquiry. *International Journal of Computer-Supported Collaborative Learning, 6*(3), 371–395. doi:10.1007/s11412-011-9124-x.

Wecker, C., Stegmann, K., Bernstein, F., Huber, M., Kalus, G., Kollar, I., & Fischer, F. (2010). S-COL: A copernican turn for the development of flexibly reusable collaboration scripts. *International Journal of Computer-Supported Collaborative Learning, 5*(3), 321–343. doi:10.1007/s11412-010-9093-5.

Weinberger, A., & Fischer, F. (2006). A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. [Journal]. *Computers in Education, 46*, 71–95.

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review, 92*(4), 548–573. doi:10.1037/0033-295x.92.4.548.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., & Martin, M. (2004). Learning subjective language. *Computational Linguistics, 30*(3), 277–308. doi:10.1162/0891201041850885.

Witten, L. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*. San Francisco: Elsevier.

Yan, X., & Yan, L. (2006). *Gender classification of weblog authors*. Paper presented at the the AAAI Spring Symposium Series Computational Approaches to Analyzing Weblogs(pp. 228–230), Stanford, California, USA.

Zhang, Y., Dang, Y., & Chen, H. (2009). *Gender difference analysis of political web forums: An experiment on an international Islamic women's forums*. Paper presented at the Proceedings of the 2009 IEEE International Conference on Intelligence and Security Informatics (pp. 61–64), Richardson, Texas, USA.