Measuring the Power of Learning.®

Research Report
ETS RR–17-58

CPS-Rater: Automated Sequential Annotation for Conversations in Collaborative Problem-Solving Activities

Jiangang Hao

Lei Chen

Michael Flor

Lei Liu

Alina A. von Davier

December 2017

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

CPS-Rater: Automated Sequential Annotation for Conversations in Collaborative Problem-Solving Activities

Jiangang Hao,[1] Lei Chen,[1] Michael Flor,[1] Lei Liu,[1] & Alina A. von Davier[2]

1 Educational Testing Service, Princeton, NJ
2 ACTNext, ACT Inc., Iowa City, IA

Conversations in collaborative problem-solving activities can be used to probe the collaboration skills of the team members. Annotating the conversations into different collaboration skills by human raters is laborious and time consuming. In this report, we report our work on developing an automated annotation system, CPS-rater, for conversational data from collaborative activities. The linear chain conditional random field method is used to model the sequential dependencies between the turns of the conversations, and the resulting automated annotation system outperforms those systems that do not model the sequential dependency.

Keywords Automated annotation; sequential dependence; conversation; conditional random field

doi:10.1002/ets2.12184

Collaboration is an essential 21st-century skill for both academic and career success (Griffin, McGaw, & Care, 2012; Trilling & Fadel, 2009). In collaborative activities, team members' verbal communications can be used to probe their collaboration skills during the collaboration process. Annotating the communications with labels that reflect different collaboration skills is the first step to analyzing the collaboration process quantitatively. For computer-supported collaborations, current technology allows three possible types of communications: audio, video, and text chat. Annotating any types of these communication data is laborious, and an automated annotating system is highly desirable if one wants to scale up the study or provide real-time facilitation based on the communications.

The collaboration skills can be identified from the semantic meaning of the communication data. Audio and video data need to be transcribed into text before the annotation can be started.[1] Automated annotation of texts is a well-studied discipline in natural language processing (NLP). In the educational context, many automated annotations or scoring systems have been developed and applied to annotate essays, short constructed responses, dialog speech acts, and learning forum messages (Burstein, Leacock, & Swartz, 2001; Gianfortoni, Adamson, & Rosé, 2011; Leacock & Chodorow, 2003; Moldovan, Rus, & Graesser, 2011; Rosé et al., 2008). The basic working mechanism behind these automated annotation systems involves a quantitative representation of the text and a mapping of this representation to the labels/scores either via a simple linear regression or using more sophisticated machine learning methods (Chen, Fife, Bejar, & Rupp, 2016). Most of these approaches treat each response as independent of the others, which is sufficient for most of the aforementioned applications. However, if there are sequential dependencies among the responses, such as the communications in a collaborative activity, these methods will not be optimal, as they simply do not take advantage of the additional information from the sequential correlation. Proper modeling of the sequential dependency can help to improve the annotation accuracy for sequentially dependent responses. Multiple schemes have been suggested to leverage the sequential dependency for automated annotation problems in different applications; a review of these methods can be found in Dietterich (2002). On the basis of both theoretical and empirical comparison studies (Sutton & McCallum, 2012), it is suggested that the current state-of-the-art framework for modeling sequential dependency is the conditional random field (CRF; Lafferty, McCallum, & Pereira, 2001).[2]

Though CRF provides a general framework for modeling sequential dependencies, it does not spell out all the needed elements for a specific application. For example, the feature functions (see the next section for details) in the CRF could be optimized based on the particular properties of a specific application, and this optimization process is not directly transferable from one data set to another. Despite that CRF has a broad spectrum of applications in NLP, it has not been widely used to classify collaboration skills from conversations in collaborative activities. The closest applications of this kind are

Corresponding author: J. Hao, E-mail: jhao@ets.org

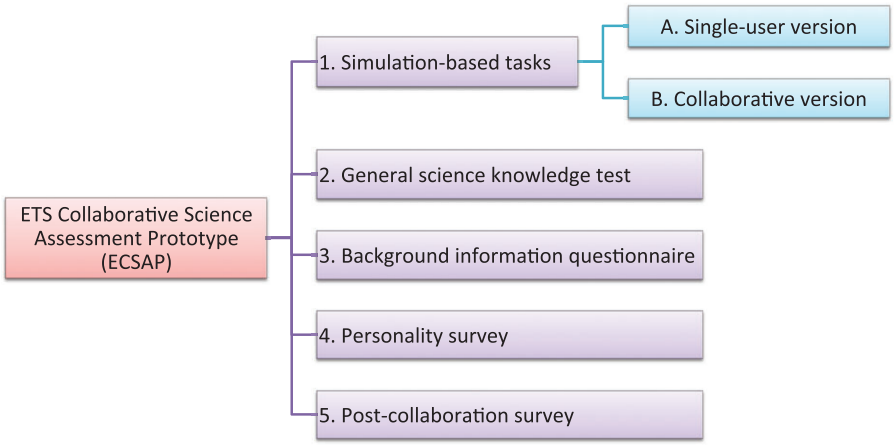

Figure 1 Assessment instruments used in the ETS collaborative science assessment prototype.

the classification of the dialog acts from live chats and tutorial dialogues (S. N. Kim, Cavedon, & Baldwin, 2010; Rus, Niraula, Maharjan, & Banjade, 2015) and the identification of the affects from human–human interactions (Siddiquie, Khan, Divakaran, & Sawhney, 2013). The major reason for CRF not being widely used to annotate collaborative skills is the lack of large-scale and annotated chat data from carefully controlled collaborative activities.

In this study, we applied CRF (more specifically, linear chain CRF) to model the sequential dependencies among chat communications and developed an automated annotation system, CPS-rater. The data set used in this study is from a large-scale online collaborative assessment prototype, the ETS collaborative science assessment prototype (ECSAP; Hao, Liu, von Davier, & Kyllonen, 2015, 2017; Liu, Hao, von Davier, Kyllonen, & Zapata-Rivera, 2015). The collaborative task in ECSCAP requires two participants to collaborate on a simulation-based task about volcanoes. Each team generated about 80 turns of chat communication throughout the task. In our completed data collection, we collected data from more than 500 dyadic teams, leading to a total of more than 40,000 lines of chat messages. Each turn of the conversations was annotated by human raters based on a coding rubric for collaborative problem-solving (CPS) skills (Liu et al., 2015). On the basis of this large annotated data set, Flor, Yoon, Hao, Liu, and von Davier (2016) have developed an automated annotation system by treating each turn of the conversations as an independent response. The current study is a further extension of the previous study in that it models the sequential dependency between the turns of conversations.

This report is organized as follows. We first introduce the chat data corpora and annotation. Then, we introduce CPS-rater by outlining the CRF framework, and we detail our tweaks of the feature functions. Finally, we compare the performance of CPS-rater with other nonsequential classification methods.

Data

ETS Collaborative Science Assessment Prototype

The ECSAP was developed to study the CPS skills in the domain of science. Figure 1 shows the five assessment instruments used in the ECSAP. A detailed description of each instrument is beyond the scope of the current report; we invite the interested reader to refer to Hao et al. (2017) for details. The collaborative conversations are the main data we are dealing with in the current report, and they were produced when dyadic teams took the collaborative version of the simulation-based task in the ECSAP. This simulation-based task was modified from an existing science assessment, Volcano Trialogue (Zapata-Rivera et al., 2014), which was designed to assess individuals' scientific inquiry skills. In the collaborative version, we added a chat window to the simulation, through which two participants collaborate to solve a set of problems in volcano science. A screenshot of the task is shown in Figure 2.

Data Collection and Annotation

We collected data through a crowdsourcing data collection platform, Amazon Mechanical Turk (Kittur, Chi, & Suh, 2008). We recruited 1,000 participants with at least 1 year of college education and teamed them randomly into dyadic teams to

Figure 2 Collaborative version of the simulation-based task used in the ETS collaborative science assessment prototype.

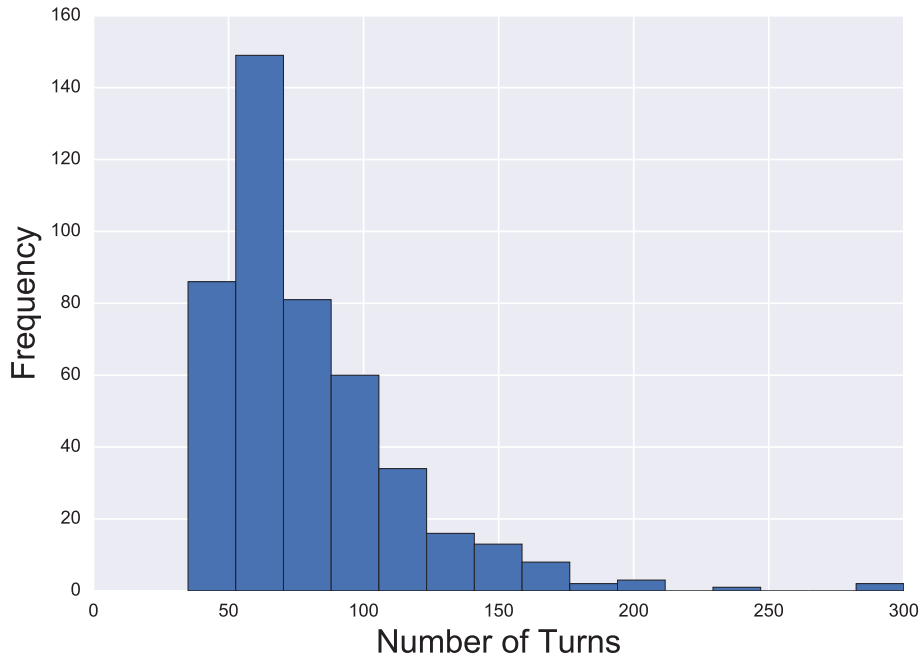

Figure 3 Distribution of number of turns of the conversations for each team.

take the collaborative version of the simulation task. After removing incomplete responses, we had complete responses from 482 dyadic teams. The responses include both conversations around and responses to the questions. When considering only the conversations, the average number of turns for each team is 80, and the average time for each session is 71 minutes. A distribution of the number of turns is shown in Figure 3.

Each turn of the chat conversations was annotated based on a CPS framework developed for the domain of science (Liu et al., 2015). The framework outlines four main categories of the CPS skills on which we would like to focus: sharing ideas, negotiating ideas, regulating problem-solving activities, and maintaining communication. Each of these categories has some subcategories, and the total number of subcategories amounts to 33.

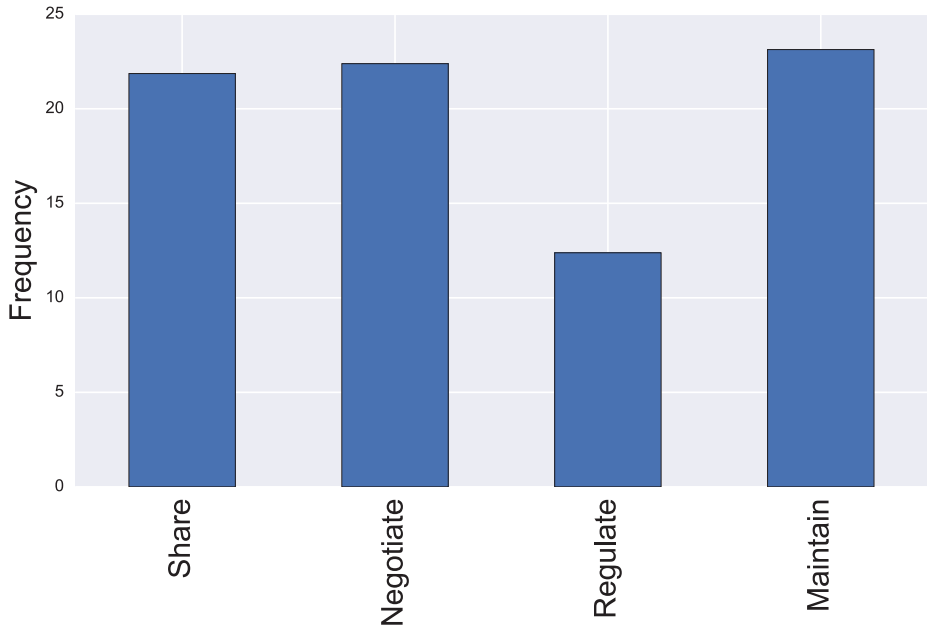

Figure 4 Average frequency of each category of collaborative problem-solving skills per team.

Two human raters were trained on the CPS framework and then double-coded a subset of the discourse data (16% of the data). The unit of annotation was each turn of a conversation or each conversational utterance. All the coding was done at the subcategory level. On the basis of these subcategory labels and their mapping to the four main categories, a set of four-category labels were assigned later on. Given the fact that there are 33 categories in the initial annotation, it took a while for the two raters to achieve "stable" annotations. We noticed that the first 17 sessions from the 482 dyadic teams are less reliable and removed them from our final analysis. This left us with a total of 3,669 turns of conversation being labeled by two raters. The agreement of the human annotation as measured by the unweighted kappa is .617 for all 33 subcategories and .675 for the four main categories. Given that many subcategories of the CPS skills rarely appear in the conversations, we will stick to the four main categories of labels in this report. Figure 4 shows the average frequency of each of the four categories of CPS skills per team. In Table 1, we show the snippets of the annotated data from two different teams.

CPS-Rater

CRF provides a nice framework for modeling the dependency of sequential data. However, it is not necessarily an automated annotation system for conversations by itself. An additional set of wisdom on text preprocessing, feature selection, and sequence optimization is needed to create an automated annotation system for conversations in a collaborative task.

Conditional Random Field Framework

The annotation problem we consider here is one particular type of the more general classification problem in machine learning. Classifiers can be developed from both generative and discriminative perspectives. Generative models maximize the joint probability of the labels and data, whereas discriminative models maximize the conditional probability of the labels given the data. An example of the former is the naive Bayes classifier, and an example of the latter is the maximum entropy classifier.[3] As it is often difficult to model the probability of the data, discriminative models are generally more preferred over generative models (Sutton & McCallum, 2012).

A nonsequential classifier for text classification learns the mapping between each turn of the texts and its corresponding label and then applies the learned mapping to each turn of the new texts to predict the corresponding label. A sequential classifier, on the other hand, treats all the texts and their labels in a sequence as a whole and learns the mapping between all the turns of the texts and their labels together. It will apply the learned mapping to a sequence of new texts and predict

Table 1 Examples of Annotated Data From Two Different Teams

Topic	Chats	Label	Skill
First team			
IntroduceYourselves	Hello	3	Maintaining
IntroduceYourselves	Hey	3	Maintaining
Question1A	Chose b, cause its rocks cracking that cause the high frequency events	0	Sharing
Question1A	Yes, same here	1	Negotiating
Question1B	D sound right to you?	2	Regulating
Question1B	I couldn't remember, I thought it was C	2	Regulating
Question1B	You are right	1	Negotiating
QuestionsP2	A and B?	2	Regulating
QuestionsP2	Yes, that's what i got	1	Negotiating
QuestionsP3	52,431?	2	Regulating
QuestionsP3	I was only sure about 5 and 1 being first and last	0	Sharing
QuestionsP3	4 is probably second to last	0	Sharing
ExampleSeisQuestion1	A?	2	Regulating
ExampleSeisQuestion1	Picked a	0	Sharing
ExampleSeisQuestion2	Thoughts?	2	Regulating
ExampleSeisQuestion2	B?	2	Regulating
ExampleSeisQuestion2	Same	1	Negotiating
ExampleSeisQuestion3	Obviously c	0	Sharing
ExampleSeisQuestion3	C	0	Sharing
Second team			
IntroduceYourselves	Hi how are you?	3	Maintaining
IntroduceYourselves	I am fine. How about yourself?	3	Maintaining
IntroduceYourselves	Good glad to be able to work on this with you	3	Maintaining
IntroduceYourselves	I feel the same way.	3	Maintaining
IntroduceYourselves	I wonder what we will be doing.	2	Regulating
IntroduceYourselves	I was just about to ask what should we do now click next?	2	Regulating
IntroduceYourselves	I would wait a bit.	2	Regulating
IntroduceYourselves	Some thing dealing with volcanos, pretty sure.	0	Sharing
IntroduceYourselves	Tracking the I think	2	Regulating
IntroduceYourselves	Yes.	1	Negotiating
IntroduceYourselves	Have you ever seen a real volcano?	3	Maintaining
IntroduceYourselves	Nope, but that would be fun. You?	3	Maintaining
IntroduceYourselves	I saw Mt. St. Helens when I was younger.	3	Maintaining
IntroduceYourselves	Nice!	3	Maintaining
IntroduceYourselves	It was real pretty.	3	Maintaining
IntroduceYourselves	I bet there a little scary to I think too. A good mix of fun too!	3	Maintaining
IntroduceYourselves	Yes if you happen to live at the base of an active one.	1	Negotiating
IntroduceYourselves	I seen this show where these people would chase active volcanos talk about living life	3	Maintaining

Note. The topic column indicates the specific items around which the conversations focus within the simulation-based task.

their labels all at once. When there are dependencies among the turns, such as in a conversation, a distinct advantage for the sequential classifier is that it can make use of the dependency information to improve the accuracy of the annotation.

CRF provides a general framework for modeling the dependencies among the labels in a sequence from the discriminative perspective. The formal definition of CRF is as follows (Lafferty et al., 2001).

Definition

Let $G = (V, E)$ be a graph such that $\mathbf{Y} = (\mathbf{Y}_v)_{v \in V}$, so that $\mathbf{Y}$ is indexed by the vertices of G. Then $(\mathbf{X},\mathbf{Y})$ is a CRF in the case, when conditioned on $\mathbf{X}$, the random variables $\mathbf{Y}_v$ obey the Markov property with respect to the graph $p(\mathbf{Y}_v|\mathbf{X}, \mathbf{Y}_w, w \neq v) = p(\mathbf{Y}_v|\mathbf{X}, \mathbf{Y}_w, w \sim v)$, where $w \sim v$ means that w and v are neighbors in G.

Let us denote each turn of the chat messages as x_t and the corresponding CPS label as y_t, where t runs from 1 to T, with T as the length of the sequence. The x_t here is not necessarily a single number but a representation of the text in that

turn. If we further denote the sequence of $\{x_t\}$ and $\{y_t\}$ as $\mathbf{x}$ and $\mathbf{y}$, a linear chain CRF is defined as

$$p\left(\mathbf{y}|\mathbf{x}\right) = \frac{1}{Z\left(\mathbf{x}\right)} \prod_{t=1}^{T} \exp\left(\sum_{i=1}^{F} w_i f_i \left(y_{t-1}, y_t, \mathbf{x}, t\right) \right), \tag{1}$$

where $Z(\mathbf{x})$ is the normalization constant defined as

$$Z\left(\mathbf{x}\right) = \sum_{\mathbf{y}} \prod_{t=1}^{T} \exp\left(\sum_{i=1}^{F} w_i f_i \left(y_{t-1}, y_t, \mathbf{x}, t\right) \right). \tag{2}$$

The core part of Equation 1 is made of F feature functions, denoted as f_i, and the weight of each feature function is denoted as w_i, where the integer i runs from 1 to F. The feature functions consist of the transition feature functions of the entire observation sequence, denoted as $h_i(y_{t-1}, y_t, \mathbf{x}, t)$ for i running from 1 to M, and the state feature functions of the label at position t and the observation sequence, denoted as $s_i(y_t, \mathbf{x}, t)$ for i running from M to F. Here M is an integer dependent on the choice of the features. A major assumption of the linear chain CRF is that only adjacent labels will interact in the transition feature functions. Once the feature functions are set, the optimal labels corresponding to the sequence can be obtained by maximizing the conditional probability function, Equation 1.

Sequential Dependency

As the real power of the sequential modeling lies in the additional information from the sequential dependency of the data, we need to demonstrate that our data do show sequential dependency before we can be assured that the sequential modeling will help. Because only adjacent labels of the chats will be modeled in a linear chain CRF, we just need to examine the dependency of the adjacent pairs of the CPS labels in our data. A straightforward way to do this is by comparing the frequency (probability) of the adjacent label pairs against the label pairs from a randomly shuffled label list. We created 300 random realizations by shuffling the label list. Then we counted the consecutive pairs of the labels and compared them to those calculated based on the real label sequence. The results are shown in Figure 5. Several skill pairs' frequencies significantly deviate from the random realizations, which is an indication of sequential dependency among certain combinations of labels. By properly modeling the dependency into an automated annotation system, we can, in principle, improve the annotation accuracy.

Feature Functions

The core parts of a linear chain CRF classifier are the feature functions. The choice of the feature functions will directly affect the performance of the classifier. As shown in the previous section, the feature functions consist of the transition features $h_i(y_{t-1}, y_t, \mathbf{x}, t)$ and the state features $s_i(y_t, \mathbf{x}, t)$. The former captures the sequential dependency of the labels, and the latter captures the relationship between the labels and the data. In our experiment, we chose a set of simple transition feature functions for the labels as the transition probabilities of the pairwise transitions from one label to another, that is,

$$h_i\left(y_{t-1}, y_t, \mathbf{x}, t\right) = P_{L(y_{t-1}),L(y_t)}, \tag{3}$$

where $L(y_t)$ and $L(y_{t-1})$ refer to the label classes corresponding to y_t and y_{t-1}, respectively. While for the state feature function, we chose the indicator functions defined as

$$s_i\left(y_t, \mathbf{x}, t\right) = \begin{cases} 1 & \text{if at position } (t),\ \text{token}_i \in x_t \\ 0 & \text{otherwise,} \end{cases}$$

where token_i is from a list of tokens we developed based on the conversation texts. As the chat conversations are full of slang words and irregular expressions, we first "regularize" all the texts using a contextually aware spell checker (Flor, 2012). For example, slang words and expressions such as "ya," "yea," "yeah," "yiss," "yisss," "yep," "yay," "yaaaay," and "yupp" are normalized to "yes" by using a dictionary of slang terms (Flor et al., 2016). Though character-level n-gram features can be used to address these irregular expressions, they will significantly increase the sparsity of the feature space; therefore we chose to do the correction based on our established text checker. On the basis of the cleaned text, we further selected a set of tokens, mostly words (unigram), word pairs (bigram), and some emotional symbols used in the chat, such as ":)" and ":(," all of which are considered to be informative, to reveal the CPS skills we defined.

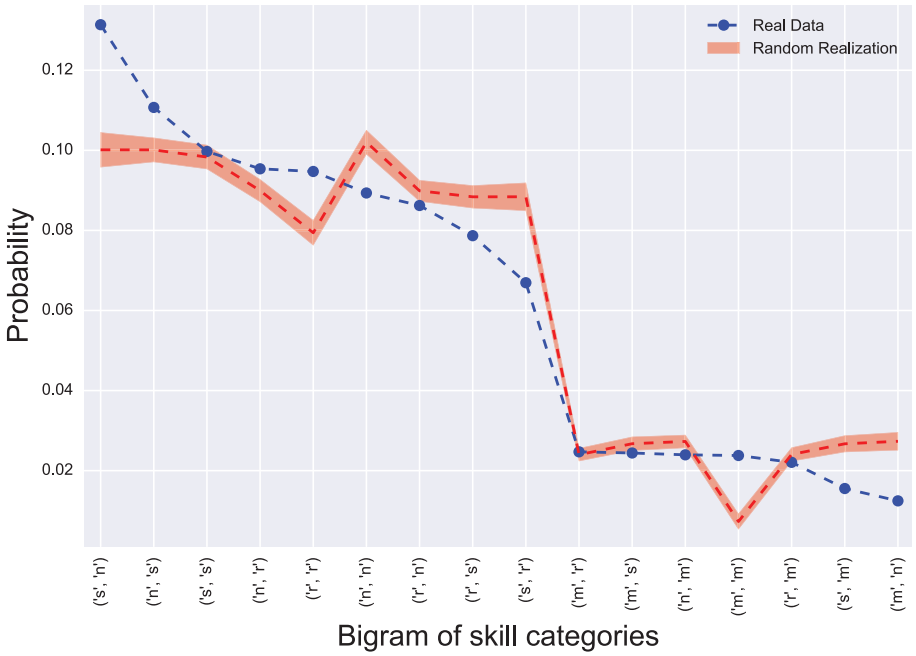

Figure 5 Comparison of the frequency of the consecutive pairs of labels between real data and the 300 realizations of the randomly shuffled label list. The red dashed line is the mean of 300 realizations, and the red band encloses the 95% confidence interval. The letters *s, n, r*, and *m* denote the four CPS skill *categories sharing ideas, negotiating ideas, regulating problem solving*, and *maintaining communications*, respectively.

Table 2 Average Performance of Different Classifiers Based on Eight Random Split-Half Cross-Validations

Method	Accuracy (%)	Cohen's kappa (unweighted)
Human – human	75.8	0.675
Baseline[a]	29.0	NA
Maximum entropy	66.9	0.551
Random forest	69.8	0.589
Naive Bayes	70.4	0.596
Linear SVM	71.9	0.619
Linear chain CRF	73.2	0.636

Notes. Each classifier is working under its optimal hyperparameters. For the classifiers other than linear chain conditional random field, we use the unigram and bigram features from the text. CRF = conditional random field. SVM = support vector machine. [a]The most frequent category.

Performance Comparison

In this study, multiple machine learning classifiers have been used. On the basis of eight runs of random split-half Monte Carlo cross-validation, the comparison of the performance is shown in Table 2. The linear chain CRF-based classifier outperforms all other major nonsequential classifiers used in this study.

Discussion

We report herein a sequential automated annotation system for collaborative communications based on the linear chain CRF CPS-rater. We applied it to the conversational chats generated from a collaborative task in the ECSAP. In our model, we consider only the dependency from adjacent turns of conversations. Though this may not capture the longer range dependency in the sequence, it already outperforms nonsequential methods, such as support vector machines. It is worth noting that such a modeling scheme can potentially be applied to annotating general short constructed responses from scenario-based tasks, where the responses to different items may be correlated.

Though it is plausible to expect that modeling more complex dependency and choosing more sophisticated feature functions may potentially further improve the performance, we also caution that overmodeling the dependency and tweaking the feature functions may reduce the generalizability of the trained model based on our current data. In an ongoing work, we explore other sequential modeling methods, including deep learning-based methods, and will report the findings in the near future.

Notes

1 Note that the video and audio communication data can yield additional information, such as paralinguistic features or affects.
2 It is worth noting that the recent progress in deep learning, especially the combination of convolution neural network (CNN) and recurrent neural network (RNN), has been shown to outperform most of the traditional approaches (which utilize human-engineered feature representation and machine learning mapping) for a number of applications, such as speech recognition, text annotation, and image tagging (Bertero & Fung, 2016; Y. Kim, 2014; Li & Wu, 2016; Shen & Lee, 2016; Wang et al., 2016). However, the price for the increased predictive accuracy is the decreased interpretability of the feature representation used in deep learning methods. Though this may be acceptable for many practical applications, it becomes very challenging for applications in educational assessments where the interpretability of the scoring elements is needed and valued.
3 The maximum entropy classifier is generally referred to as multinomial logistic regression in the statistics community.

References

Bertero, D., & Fung, P. (2016). A long short-term memory framework for predicting humor in dialogues. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 130 – 135). Retrieved from http://aclweb.org/anthology/N/N16/N16-1016.pdf

Burstein, J., Leacock, C., & Swartz, R. (2001). Automated evaluation of essays and short answers. In *Proceedings of the Fifth International Computer Assisted Assessment Conference, Loughborough University, UK* (pp. 41 — 54). Retrieved from https://dspace.lboro.ac.uk/2134/1790

Chen, J., Fife, J. H., Bejar, I. I., & Rupp, A. A. (2016). *Building e-rater ® scoring models using machine learning methods* (Research Report No. RR-16-04). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12094

Dietterich, T. G. (2002). Machine learning for sequential data: A review. In T. Caelli, A. Amin, R. P. W. Duin, D. de Ridder, & M. Kamel (Eds.), *Structural, syntactic, and statistical pattern recognition: Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)* (pp. 15 – 30). Berlin, Germany: Springer. https://doi.org/10.1007/3-540-70659-32

Flor, M. (2012). Four types of context for automatic spelling correction. *Traitement Automatique des Langues, 53*(3), 61 – 99.

Flor, M., Yoon, S.-Y., Hao, J., Liu, L., & von Davier, A. (2016). Automated classification of collaborative problem solving interactions in simulated science tasks. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 31 – 41). Retrieved from http://aclweb.org/anthology/W16-0504

Gianfortoni, P., Adamson, D., & Rosé, C. P. (2011). Modeling of stylistic variation in social media with stretchy patterns. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties* (pp. 49 – 59). Stroudsburg, PA: Association for Computational Linguistics.

Griffin, P., McGaw, B., & Care, E. (2012). *Assessment and teaching of 21st century skills*. New York, NY: Springer. https://doi.org/10.1007/978-94-007-2324-5

Hao, J., Liu, L., von Davier, A., & Kyllonen, P. (2015). Assessing collaborative problem solving with simulation based tasks. In O. Lindwall, P. Hakkinen, T. Koschmann, P. Tchounikine, & Ludvigsen, S. (Eds.), *Exploring the material conditions of learning: Computer supported collaborative learning (CSCL) conference 2015* (Vol. *1*, pp. 544 – 547). Gothenburg, Sweden: The International Society of the Learning Sciences.

Hao, J., Liu, L., von Davier, A. A., & Kyllonen, P. (2017). Initial steps towards a standardized assessment for CPS: Practical challenges and strategies. In A. A. von Davier, M. Zhu, & P. C. Kyllonen (Eds.), *Innovative assessment of collaboration* (pp. 135 – 156). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-33261-19

Kim, S. N., Cavedon, L., & Baldwin, T. (2010). Classifying dialogue acts in one-on-one live chats. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 862 – 871). Stroudsburg, PA: Association for Computational Linguistics.

Kim, Y. (2014). *Convolutional neural networks for sentence classification*. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1746 – 1751). Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1181

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with mechanical Turk. In *Proceedings of the Sigchi Conference on Human Factors in Computing Systems* (pp. 453–456). New York, NY: ACM. https://doi.org/10.1145/1357054.1357127

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning ICML* (Vol. 1, pp. 282–289). San Francisco, CA: Morgan Kaufmann.

Leacock, C., & Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers and the Humanities, 37*, 389–405. https://doi.org/10.1023/A:1025779619903

Li, W., & Wu, Y. (2016, December). Multi-level gated recurrent neural network for dialog act classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical papers* (pp. 1970–1979). Retrieved from http://aclweb.org/anthology/C16-1185

Liu, L., Hao, J., von Davier, A. A., Kyllonen, P., & Zapata-Rivera, D. (2015). A tough nut to crack: Measuring collaborative problem solving. In *Handbook of research on technology tools for real-world skill development* (pp. 344–359). Hershey, PA: Information Science Reference.

Moldovan, C., Rus, V., & Graesser, A. C. (2011). Automated speech act classification for online chat. In S. Visa, A. Inoue, & A. L. Ralescu (Eds.), *Proceedings of the 22nd Midwest Artificial Intelligence and Cognitive Science Conference 2011* (pp. 23–29). Retrieved from http://ceur-ws.org/Vol-710/paper22.pdf

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*, 237–271. https://doi.org/10.1007/s11412-007-9034-0

Rus, V., Niraula, N. B., Maharjan, N., & Banjade, R. (2015). Automated labelling of dialogue modes in tutorial dialogues. In *The Twenty-Eighth International Florida Artificial Intelligence Research Society Conference* (pp. 205–210). Retrieved from https://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS15/paper/view/10450

Shen, S.-S., & Lee, H.-Y. (2016). Neural attention models for sequence classification: Analysis and application to key term extraction and dialogue act detection. In *Proceedings of Interspeech 2016* (pp. 2716–2720). Baixas, France: ISCA. https://doi.org/10.21437/Interspeech.2016-1359

Siddiquie, B., Khan, S., Divakaran, A., & Sawhney, H. (2013). Affect analysis in natural human interaction using joint hidden conditional random fields. In *IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1–6). New York, NY: IEEE.

Sutton, C., & McCallum, A. (2012). An introduction to conditional random fields. *Foundations and Trends in Machine Learning, 4*, 267–373. https://doi.org/10.1561/2200000013

Trilling, B., & Fadel, C. (2009). *21st century skills: Learning for life in our times*. New York, NY: John Wiley.

Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., & Xu, W. (2016, June). CNN-RNN: A unified framework for multi-label image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2285–2294). New York, NY: IEEE. https://doi.org/10.1109/CVPR.2016.251

Zapata-Rivera, D., Jackson, T., Liu, L., Bertling, M., Vezzu, M., & Katz, I. R. (2014). Assessing science inquiry skills using trialogues. In S. Trausan-Matu, K. E. Boyer, M. Crosby, & K. Panourgia (Eds.), *Intelligent tutoring systems. ITS 2014* (pp. 625–626). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-07221-084

Suggested citation:

Hao, J., Chen, L., Flor, M., Liu, L., & von Davier, A. A. (2017). *CPS-rater: Automated sequential annotation for conversations in collaborative problem-solving activities* (Research Report No. RR-17-58). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12184

Action Editor: Keelan Evanini

Reviewers: Isaac Bejar and Vikram Ramanarayanam