Detection of Student Teacher's Intention using Multimodal Features in a Virtual Classroom

Masato Fukuda, Hung-Hsuan Huang and Toyoaki Nishida
Center for Advanced Intelligence Project, RIKEN, Kyoto University, Yoshida-honmachi, Sakyo-ku, Kyoto, Japan

Keywords: Intention Detection, Multimodal Interaction, Educational Application, User Assessment.

Abstract: The training program for high school teachers in Japan has less opportunity to practice teaching skills. As a new practice platform, we are running a project to develop a simulation platform of school environment with computer graphics animated virtual students for students' teachers. In order to interact with virtual students and teachers, it is necessary to estimate the intention of the teacher's behavior and utterance. However, it is difficult to detection the teacher's intention at the classroom only by verbal information, such as whether to ask for a response or seek a response. In this paper, we propose an automatic detection model of teacher's intention using multimodal features including linguistic, prosodic, and gestural features. For the linguistic features, we consider the models with and without lecture contents specific information. As a result, it became clear that estimating the intention of the teacher is better when using prosodic / non-verbal information together than using only verbal information. Also, the models with contents specific information perform better.

1 INTRODUCTION

According to the fundamental policy for development of educational human resources, published by Board of Education of Tokyo in 2007, the environment of school education problems is getting more complex and diverse with the evolvement of the whole society. Along with this situation, it is getting more and more difficult for teachers to adopt themselves to the problems which do not occur dozens of years ago. Trainees and novice teachers are demanded to accumulate their experience and to improve their knowledge before the actual deployment to schools.

In order to deal with the diversity of problems, student teachers need not only knowledge but also repeated practice to learn from experience. In Japan, however, the training of teachers mainly relies on classroom lectures in colleges and is compensated with the practice for a relatively short period, say only two to three weeks in actual schools. Even though there may be some chances for practicing teaching skills in the class of teacher-training course in the colleges, these practices are usually conducted by peer role-playings in small number of participants and are far from real situations where they have to face dozens of teenagers. The teacher-training programs in Japan obviously lacks the practice in teaching skill and the admission of classes. The result is, many young teachers left their jobs in the first year due to frustration and other mental issues.

In our ongoing project, we are developing virtual classroom systems based on multiple virtual agent as students for training or assessment on educational skills of student teachers(Fukuda et al., 2018)(Fukuda et al., 2017). Student teachers, or the trainees can interact with the virtual students in this immersive and realistic virtual classroom and practice their teaching and admission skills. The virtual students are operated by an operator the examination investigator from remote with a dedicated interface.

In order to improve the current Wizard-of-Oz (WOZ) prototype system to be an autonomous one, automatic and real-time assessment on the trainee is a mandatory function for the system. The virtual students can then automatically react to the trainee according to the assessment results. Automatic assessment can be considered as an association from the trainee's behaviors to the values of teaching skill metrics. Better understanding of the trainee's intention during the teaching practice will provide valuable information of the assessment process. The information of trainee's intention can be also used to drive the virtual students' reactions as a real-time interactive system. This paper presents an automatic detection model for the student teacher's intention using multimodal features (linguistics, prosody, and gesture).

Fukuda, M., Huang, H. and Nishida, T.
Detection of Student Teacher's Intention using Multimodal Features in a Virtual Classroom.
DOI: 10.5220/0007379901700177
In *Proceedings of the 11th International Conference on Agents and Artificial Intelligence (ICAART 2019)*, pages 170-177
ISBN: 978-989-758-350-6

This support-vector-machine (SVM) model is derived from the dataset gathered in actual teaching practice sessions with the virtual classroom prototype. The paper is organized as the follows: section 2 introduces the related works, section 3 describes the virtual classroom in more details and the procedure of data collecting experiment, section 4 describes the detection model, and finally section 5 concludes this paper.

2 RELATED WORKS

Studies of CG agents enabled emotional interactions between humans and computers using modalities that we use in daily life. There are other research groups who are developing virtual classroom systems for teacher training. All available system is wizard-of-oz (WOZ) style systems, that is, the virtual students are controlled by a human instructor rather than an autonomous behavior model. This comes from the difficulty in creating the model to generate realistic and believable behaviors of the virtual students in the diversity of the situations occur in schools. TeachLivE (Barmaki and Hughes, 2015) is a WOZ system where the trainee interacts with five virtual students whom are controlled by an operator and are shown on a large size display. Breaking Bad Behaviours (Lugrin et al., 2018)(Lugrin et al., 2016) is another teacher training system featuring a virtual reality (VR) environment. This is another WOZ style system where the instructor controls the level of overall disruption of a 24-student class with a slide bar and can assign one of six bad behaviors to individual virtual students. Neither of these systems have provided automatic assessment feature on the performance of the student teachers yet. Although the purpose and the requirements are different, public speaking is an activity related to teaching in a classroom. Automatic assessment method using multimodal features, prosody, gesture use, and eye contact is proposed in a environment with virtual audiences (Chollet et al., 2015)

In the tasks of classifying dialogue acts, multimodal features are also proved to improve the performance of the classifiers. Petukhova and Bunt (Petukhova and Bunt, 2011) investigated the possibility of automatic classification on dialogue acts with lexical and prosodic information. The RoboHelper project uses lexical, syntactic, utterance, gesture, and location features to classify dialogue acts based on an elderly-at-home data corpus(Chen and Eugenio, 2013)(Di Eugenio and Žefran, 2015). Ribeiro et al. (Ribeiro et al., 2015) investigated the influence of the context (N-gram with previous tokens) in predicting dialogue acts. Liu et al. (Liu et al., 2017) used a combination of CNN (convolution neural network) and RNN (recurrent neural network) to automatically extract context information in classifying dialogue acts.

3 MULTIMODAL DATA CORPUS OF TEACHING REHEARSAL

3.1 Virtual Classroom Training Platform

The setup of the virtual classroom prototype is shown in Figure 1. This system is comprised of two front ends, one is a simulated classroom for the trainee, the other one is the interface for the system operator / investigator. The virtual classroom is projected on a life-size large screen (roughly 100 inches) while one trainee stands in front of it and practice his/her teaching skill with the virtual students. The trainee's rehearsal is captured by a WebCam and is displayed at the operator's interface in real-time. The operator can control atmosphere of the virtual classroom using this interface in reacting to the performance of the trainee. Every virtual student's motion is always synchronized between the trainee interface and operator interface. That virtual classroom was created with Unity 3D game engine. For the CG models of students, we adopted Taichi Character Pack freely available at Game Asset Studio.

3.2 Data Collection Experiment

In order to collect the dataset as the ground truth for the detection model development, an experiment of actual rehearsal sessions was conducted. Nine students of our college (computer science and engineering) were recruited as the participants of the evaluation experiment. All of them are in the teacher-training course in our university or work as a cram school teacher with some degree of teaching experience. We issued the materials for high-school level mathematics to the participants and asked them to prepare a lecture up to 10 minutes prior to the experiment. The themes are about the computation of the area of circle or triangle. The subject, mathematics was chosen because all of the subject major in computer science, and we expect that more whiteboard-writing movements can be observed. During the experiment, the participants were organized to three groups. For each group, every participant made a lecture while the other two members in that group played the operator role at the same time.

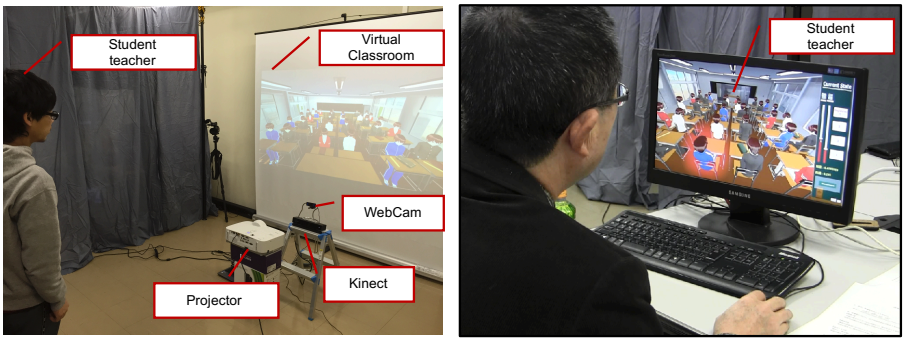

Figure 1: Setup of the virtual classroom prototype. Left: trainee interact with the virtual students projected on a large-size screen. Right: operator controls the virtual students with the view point from the rear of the virtual classroom.

3.3 Teaching Intentions

In the experiment, each participant made a lecture between five to 10 minutes, and the average duration was 457.2 seconds. In order to analyze student teacher's intention during the rehearsal sessions, an intuitive analysis unit is an *utterance*. The utterances are then segmented from the original audio track with a silent period, 200 ms as boundaries. Table 1 shows the summary of the utterances of all of the participants. This indicates that the subjects spoke less than half of the time (41.5%) during the experiment. While during other time, they may write something on the whiteboard or waited the virtual students to react to his / her instructions.

Table 1: Summary of the utterances collected in the data corpus.

Participant	Number	Duration	S.D.
P1	143	1.34	0.88
P2	246	1.46	0.93
P3	236	1.24	1.18
P4	250	1.55	1.21
P5	152	0.90	1.49
P6	138	1.49	1.19
P7	122	1.49	0.83
P8	86	1.90	1.24
P9	120	1.47	1.27
Average	166	1.43	1.13

According to our knowledge, there is no available annotation scheme dedicated to teaching intentions. Therefore, a dedicated scheme has to be defined for further analysis on the corpus. The lectures conducted in a classroom is practically an interactive communication between the teacher and the students. Such an annotation scheme can be considered as a set of domain-specific dialogue acts. There have been a number of dialogue act coding schemes proposed in the field of computational linguistics and dialogue systems. For example, SWBD-DAMSL (Dialog Act Markup in Several Layers) (Core and Allen, 1997)(Jurafsky, 1997), MRDA (Meeting Recorder Dialog Act) (Shriberg et al., 2004), and AMI (Augmented Multi-party Interaction) (Carletta et al., 2006). However, the schemes are usually bound to a specific data corpus (e.g. MRDA for the ICSI corpus (Janin et al., 2003)) and task. Although the schemes like DAMSL is considered for general purpose and is widely used, there was no standardized scheme for general dialogue act annotation.

Standardization facilitates the sharing of data and data process methods. Recently, there is a researcher oriented activity in promoting a standardized dialogue act scheme for general purpose as an international ISO standard, ISO 24617-2 (Bunt et al., 2017). This standard inherited the existing DIT++ scheme (Bunt, 2009)and simplified it. This ISO standard defines a multi-dimension scheme with eight dimensions: Auto-Feedback, Allo-Feedback, Time Management, Turn Management, Own Communication Management, Partner Communication Management, and Social Obligations Management in addition to the task dimension. Upon these dimensions, 30 dimension-specific communicative functions and 26 general-purpose ones are defined.

In this work, we basically followed the definitions of DIT++ / ISO 24617-2 with the modifications in reflecting the nature of classroom lectures. The main modifications were made on the task dimension. Since the interaction occur during a lecture is basically the *information providing* from the teacher to the students. This communicative function is supposed to be the majority of the teacher's utterances by the nature of a lecture, it should be divided to more detailed classes to deeper insight of the interaction. How the student teacher explains the lecture materials to the students is an important factor in evaluating the student teacher's skill. Considering the subjects taught in high schools, we assume that the explanation can

be classified into two categories. The materials which can be sufficiently explained orally, and the materials which will be more comprehensive if supplemented with pictorial information. The resulted definitions of teaching intention are listed as the follows:

Information-Providing-Oral: the teacher is explaining some concept or is describing some truth which is sufficiently comprehensive only by oral explanation. Most parts of the subjects like Japanese language or history are supposed to be taught in this way.

Information-Providing-Pictorial: the teacher is explaining some concept which is more comprehensive if pictorial information is also provided. For example, the shape / size of something or the spacial relationship of objects. Many concepts of the subjects like mathematics or physics are supposed to be taught in this way. Non-verbal behaviors like drawing a diagram on the whiteboard or performing iconic gestures are often accompany with this intention.

Instruct: the teacher is instructing the students to do some actions. For example, "wake up" .

Auto-Feedback: the teacher is giving feedbacks to the requests from the students. For example, "yes", "what", and "OK, I see".

Allo-Feedback: the teacher is requesting feedbacks from the students. For example, "is there any question" .

Communication Management: the teacher is saying something not directly relevant to the subject but is trying to manage the flow of lecture or maintain the interests of the students. Fillers and jokes are also included in this intention.

By following the definitions above, one of the authors annotated the whole corpus. Table 2 shows the annotation results.

Table 2: Distribution of the intention labels. Each intention class is represented by an abbreviation. The columns "**Num.**", "**Dur.**", and "**S.D.**" denote the number of instances, duration in seconds, standard deviation of duration and percentage, respectively.

Intention	Num.	Dur.	S.D.	%
Info-Pictorial	570	1.39	1.14	38.2
Info-Oral	342	1.71	1.39	22.9
Auto-Feedback	195	0.72	0.57	13.1
Instruct	144	1.52	1.33	9.6
Allo-Feedback	138	1.00	0.90	9.2
Commun-Man	104	0.86	0.94	7.0
Total	**1,493**			

4 AUTOMATIC DETECTION WITH MULTIMODAL FEATURES

This section describes the multimodal model in automatically classifying teaching intentions described in last section. Multimodal features and context information have been shown effective in classifying dialogue acts in previous studies(Chen and Eugenio, 2013)(Di Eugenio and Žefran, 2015)(Liu et al., 2017)(Petukhova and Bunt, 2011)(Ribeiro et al., 2015). In our specific context, lecture rehearsal, the student teachers are supposed to make explanations ton the lecture materials, write text or draw diagrams on the whiteboard. In order to capture the student teacher's intentions in these activities, the contents of the materials, the voice tone in the presentation, and the use of body postures, gestures and whiteboards are supposed to be relevant. Therefore, linguistic, prosodic, and gestural features are adopted in the multimodal learning process.

4.1 Linguistic Features

In analyzing the contents of the lecture rehearsal sessions, the technique, bag of words (BoW) is adopted. First, the words used in the data corpus are extracted with Japanese morphological analyzer, Mecab (Kudou, 2013) from the transcripts of the corpus. With this analysis, Japanese words are segmented and converted to their basic forms. The part-of-speech information can also be obtained as the metadata of extracted words. As the results, 604 words are extracted as the initial set of candidate words. In order to compose effective BoW vectors of words, the words representative the characteristics of a certain data class and distinguish it from other classes should be selected from the candidate words. Most common and not-meaningful words in a language, or so-called *stop words* are therefore needed to be filtered out from the set of candidate words. In the case of Japanese language, there is not good general purpose dictionary of stop words, therefore, we filter out the categories of parts of speech which do not convey significant information (particle, conjunction, auxiliary verb, the attributive form of a verb, and adverb). A customized Mecab-compatible dictionary which is frequently updated with new Japanese words, mecab-ipadic-NEologd [1] is adopted for the word reference. This is because the younger generation often use abbreviated forms and slang in their casual conversation. The size of candidate word is then reduced to

[1] https://github.com/neologd/mecab-ipadic-neologd

452 words after this filtering.

In order to determine the "important words" which distinguishes one data class from the others, the popular technique in text process, TF-IDF (term frequency - inverse document frequency) is adopted. TF-IDF algorithm weights the words which frequently appear in one document but less frequently appear in other documents as the *important words* in distinguishing that document from the others. By taking account into the context of this work, we then consider the collection of the sentences belonging to one teaching intention as a "document" and apply TF-IDF algorithm on the six documents to score the candidate words.

For capturing the contents of the materials, an issue needs to be considered is the generality of the model. If more contents specific information is adopted, the accuracy is supposed to get higher, but the generality will become lower, due to the overfitting of the training data. In considering the trade-off of this issue, we analyzed the accuracy with varying weights of content-specific features. The next step is then to determine the word clusters in the sense of their meanings in the distributed representations generated by Word2vec (Mikolov et al., 2013) models. A CBOW (Continuous-Bag-of-Words) model is constructed from the word set in last step with window size 15. In order to distinguish general words from context-specific words, we then applied 2-Nearst-Neighbors clustering to the word vectors. Table 3 shows the top-12 results of the clustering in the order of TF-IDF values . Although not completely separated, contents related words (mathematics) are basically assigned to cluster B (66 words) while the other words are basically assigned to cluster A (386 words). It can be noticed that the top word in cluster B, filler "uh" is not a mathematic relevant word. This seems to come from the following situation that is frequently observed in the corpus: when the student teachers are starting to draw something on the whiteboard while explaining a mathematical concept, they often started with a "uh" before the drawing itself. This co-occurrence may cause this result.

The next issue is the determination of the size of the bag of words, the larger the size, the accuracy is supposed to be higher, but the model will be more fitting to the contents, that is, the model will be less general. For determine the element words in the BoW vector, all the candidate words are sorted in separate queues of each intention class in the order of TF-IDF values. The words are selected in the round-robin and top-first manner, i.e. the word which has highest TF-IDF value among the six words in the head of the six queues of the intention classes, one by one.

Table 3: 2-NN clustering results of the words.

Cluster A		Cluster B	
Word	TF-IDF	Word	TF-IDF
yes	0.89	uh	0.76
no problem	0.68	(number)	0.56
sleep	0.60	area	0.48
understand	0.44	multiply	0.29
wake up	0.44	cm	0.25
request	0.33	circle	0.23
good	0.30	rectangular	0.20
wrong	0.28	here	0.17
thank	0.21	half	0.17
(misbehavior)	0.21	become	0.16
do	0.16	high	0.14
sorry	0.16	matter	0.13

The selection from a certain queue stops after n words are selected from that queue. Consequently, the size of the BoW vector is n x 6 where n words are selected from the queue of each intention class. We then conducted an experiment to determine an appropriate n. Two types of BoW vectors are generated and are used to train SVM (Support Vector Machine) models to classify intention classes. One is so-called L_f model where the full set of candidate words (cluster A + B) are used, and the L_g model where only the words in cluster A are used. The full set L_f model is supposed to be more content dependent , and the L_g model is supposed to be more general. Figure 2 depicts the performance of these two models regarding to the value of n. The experiments were conducted with the same procedure as the ones described in later section. The results are cross validated in leave-one-subject-out manner, and F-measure values are the micro average of the results. From the observation on the results, L_f models always performs slightly better than the L_g models, the performance increases when n increases but get to level when n is large enough. Finally, the following features are selected as the linguistic feature set:

- Uni-grams of part of speech in five categories, noun, verb, adjective, interjection, and filler.
- BoW vector selected in the procedure described above with the size, 48 words ($n = 8$).

4.2 Prosodic Features

Voice prosody conveys nuance in additional to language. The same sentence can be interpreted into multiple ways when it is spoken in different tones. For example, when someone raise the end of one utterance, it probably should be interpreted as a question. Therefore, prosodic information is expected to be also effective in the classification of teaching intentions. The prosodic features used are extracted with

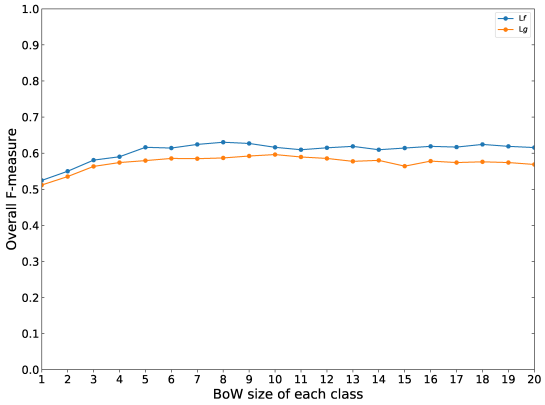

Figure 2: Classification performance of L_g and L_f models regarding to the value of the BoW size of each class (n).

the phonetic analyzer, Praat (Paul and David, 2018) at the sampling rate 100 Hz. As a result, the following features are adopted in the prosody feature set.

- Average of pitch (F_0)
- Average of intentsity
- Range of pitch (difference between maximum and minimum)
- Range of intensity
- Ratio between the pitch of second and first half of a sentence
- Ratio between the intensity of second and first half of a sentence

4.3 Gestural Features

Among the proposed teaching intentions, Providing-Information (pictorial) is the explanation about abstract concept with shape and size, it is supposed to be more comprehensive if complemented with iconic gestures or the drawings on the whiteboard. Therefore, a feature set capturing the gestural behaviors of the student teachers should provide the cues in identifying this intention class. 18 joint positions in two-dimensional coordinates (X, Y) were extracted with the tool, OpenPose [2] at the same frame rate of the video corpus (29.97 fps). In order to distinguish intentional gestures and whiteboard drawings from noises, the area scanned by the ends of two hands (the wrists) is supposed to provide effective cues. Finally, the following features are chosen in this part where all coordinate values are transformed to a unified coordinate system (the center of the waist as the origin).

- Maximum distance and the standard deviation of two wrists in horizontal direction
- Maximum distance and the standard deviation of two wrists in vertical direction
- Maximum distance and the standard deviation of two elbows in horizontal direction
- Maximum distance and the standard deviation of two elbows in vertical direction
- Maximum distance and the standard deviation of two shoulders in horizontal direction

4.4 Automatic Classification Model

As shown in Table 2, the distribution of the intention classes is imbalanced. We then oversampled smaller classes with SMOTE (Chawla et al., 2002) algorithm and under-sampled the larger classes while keeping the total weight (amount) of the dataset both in training and testing phases. A support vector machine (SVM) classification model was then built with the feature sets described above. All of the experiments were conducted with a customized program developed with Weka API (Garner, 1995)and the SMO implementation of SVM. The Pearson VII function (PUK) kernel (Zhang and Ge, 2013) was adopted because it achieved best performance among the available kernels. The complexity parameter C was explored from 1.0 to 10.0 at the step size 1.0 and was determined to be 4.0. The kernel parameter ω and σ were were tested between 1.0 and 0.1 and were determined to be 0.1 both. Normalization and standardization trials on the features were conducted and standardization achieved better performance. Figure 3 depicts the performance comparison of all combinations of available feature modalities with the L_g word set (without content dependent words). The vertical axis is micro average of the F-measure scores of all teaching intention classes. From the results, the following facts can be discovered:

- Generally, the models with multiple modalities perform better.
- When compare linguistic, prosodic, and gestural feature sets, linguistic feature set is always dominating the other two feature sets. From the nature in this classification task, identifying the intention of the teacher's utterances, verbal information is expected to be most effective.
- The effectiveness of gestural features varies a lot among intention classes (from less than 0.1 to 0.5), they are exceptionally ineffective in identifying Information-Providing-Oral and Auto-Feedback. This implies that the student teachers did not tend to use gestures or the whiteboard when they utter in these two intentions.

[2] https://github.com/CMU-Perceptual-Computing-Lab/openpose

- Prosodic feature set is not as effective as the linguistic feature set, but it contributes to improvement of the detection performance of each intention class.
- Gestural feature set is much more effective in pictorial explanation than oral explanation of Information-Providing. This coincides to the definition of these two intention classes.
- In the worst performing class, Instruct, adding both prosodic and gestural features can greatly improve the performance.

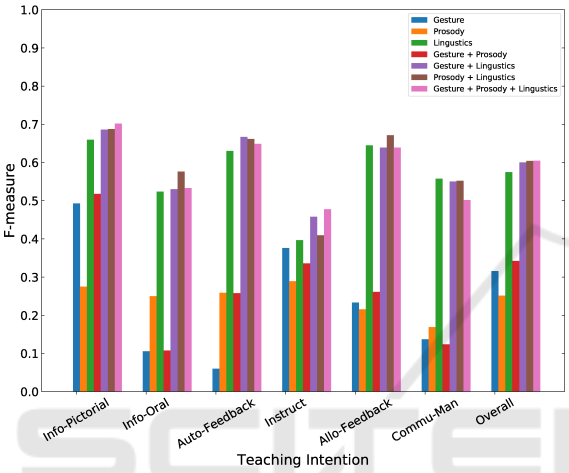

Figure 3: Comparison of the performance of all possible combinations of feature sets in all teaching intention classes with general BoW vector.

The confusion matrix of the classification results with full feature sets is shown as Table 4. It can be observed that there was difficult in distinguishing the two Information-Providing intention classes. From the definitions of these two classes, they are close to each other and are expected to be distinguished with gestural features. By further improving gestural features to include more detailed information, the performance may be improved. On the other hand, there were no strong tendency found in other classes.

For reference, the performance of all possible combinations of feature sets in all teaching intention classes with contents dependent BoW vector is depicted in Figure 4. With the features more fitting to the lecture subjects, the performance is slightly better in all models with linguistic feature set. On the other hand, the tendency in strength and weakness is very similar to the models with general BoW vector.

Table 4: Confusion matrix of the classification results with full feature sets. The rows are true classes and the columns are classified classes.

	A	B	C	D	E	F
A: Commu-Man	91	3	15	13	19	18
B: Allo-Feedback	7	93	7	17	11	11
C: Auto-Feedback	10	17	84	11	8	4
D: Instruct	3	6	14	77	11	16
E: Info-Oral	33	6	5	23	145	70
F: Info-Pictorial	60	20	0	21	131	413

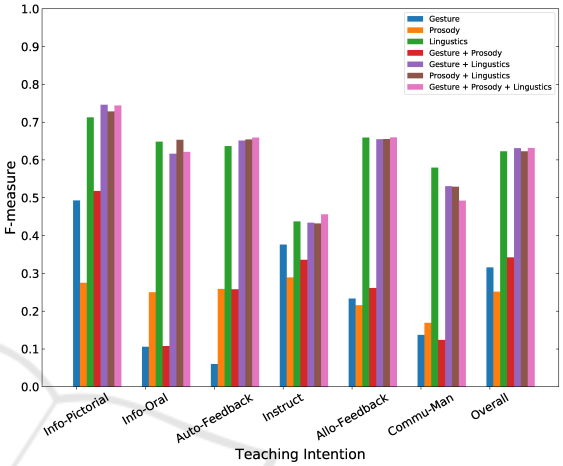

Figure 4: Comparison of the performance of all possible combinations of feature sets in all teaching intention classes with contents dependent BoW vector.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we present a SVM model using multimodal features in classifying the intentions of student teachers in the lecture rehearsals in a virtual classroom system. The linguistic, prosodic, and gestural feature sets are adopted. It is found that although verbal features are most powerful in this task, non-verbal features do improve overall performance, especially in the intention classes where the student teachers often use gestures or the drawings on the whiteboard to explain abstract concepts in the subject like mathematics.

In the future, we would like to introduce the intention dimension in the automatic assessment of the performance of student teachers. The temporal distribution of the intentions and the co-occurrences of the intentions and the events happen in the classroom are expected to provide valuable hints in the assessment. After that, we plan to realize a fully autonomous virtual classroom system and conduct evaluation experiments in practical use.

REFERENCES

Barmaki, R. and Hughes, C. E. (2015). Providing Real-time Feedback for Student Teachers in a Virtual Rehearsal Environment. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 531–537, New York, NY, USA. ACM.

Bunt, H. (2009). The DIT++ taxonomy for functional dialogue markup. In *AAMAS 2009 Workshop, Towards a Standard Markup Language for Embodied Dialogue Acts*, pages 13–24.

Bunt, H., Petukhova, V., Traum, D., and Alexandersson, J. (2017). Dialogue Act Annotation with the ISO 24617-2 Standard. In Dahl, D. A., editor, *Multimodal Interaction with W3C Standards: Toward Natural User Interfaces to Everything*, pages 109–135. Springer International Publishing, Cham.

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., and Wellner, P. (2006). The AMI Meeting Corpus: A Pre-announcement. In *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction*, MLMI'05, pages 28–39, Berlin, Heidelberg. Springer-Verlag.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Oversampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Chen, L. and Eugenio, B. D. (2013). Multimodality and dialogue act classification in the RoboHelper project. In *SIGDIAL 2013 - 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Proceedings of the Conference*, pages 183–192. Association for Computational Linguistics (ACL).

Chollet, M., Wörtwein, T., Morency, L.-P., Shapiro, A., and Scherer, S. (2015). Exploring Feedback Strategies to Improve Public Speaking: An Interactive Virtual Audience Framework. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pages 1143–1154, New York, NY, USA. ACM.

Core, M. G. and Allen, J. (1997). Coding dialogs with the DAMSL annotation scheme. In *AAAI Fall Symposium on Communicative Action in Humans and Machines*, volume 56. Boston, MA.

Di Eugenio, B. and Žefran, M. (2015). The RoboHelper Project: From Multimodal Corpus to Embodiment on a Robot. In *2015 AAAI Fall Symposium Series*.

Fukuda, M., Huang, H.-H., Kuwabara, K., and Nishida, T. (2018). Proposal of a Multi-purpose and Modular Virtual Classroom Framework for Teacher Training. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 355–356. ACM.

Fukuda, M., Huang, H.-H., Ohta, N., and Kuwabara, K. (2017). Proposal of a Parameterized Atmosphere Generation Model in a Virtual Classroom. In *Proceedings of the 5th International Conference on Human Agent Interaction*, HAI '17, pages 11–16, New York, NY, USA. ACM.

Garner, S. R. (1995). Weka: The waikato environment for knowledge analysis. In *Proceedings of the New Zealand Computer Science Research Students Conference*, pages 57–64.

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., and Wooters, C. (2003). The ICSI Meeting Corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 1, pages I–I.

Jurafsky, D. (1997). Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual. *Institute of Cognitive Science Technical Report*.

Kudou, T. (2013). MeCab: Yet Another Part-of-Speech and Morphological Analyzer [Computer Software] Version 0.996.

Liu, Y., Han, K., Tan, Z., and Lei, Y. (2017). Using Context Information for Dialog Act Classification in DNN Framework. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2170–2178.

Lugrin, J.-L., Charles, F., Habel, M., Matthews, J., Dudaczy, H., Oberdörfer, S., Wittmann, A., Seufert, C., Porteous, J., Grafe, S., and Latoschik, M. E. (2018). Benchmark Framework for Virtual Students' Behaviours. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '18, pages 2236–2238, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Lugrin, J.-L., Latoschik, M. E., Habel, M., Roth, D., Seufert, C., and Grafe, S. (2016). Breaking Bad Behaviors: A New Tool for Learning Classroom Management Using Virtual Reality. *Frontiers in ICT*, 3.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*.

Paul, B. and David, W. (2018). Praat: Doing phonetics by computer [Computer Software] Version 6.0.40.

Petukhova, V. and Bunt, H. (2011). Incremental Dialogue Act Understanding. In *Proceedings of the Ninth International Conference on Computational Semantics*, IWCS '11, pages 235–244, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ribeiro, E., Ribeiro, R., and de Matos, D. M. (2015). The Influence of Context on Dialogue Act Recognition. *arXiv:1506.00839 [cs]*.

Shriberg, E., Dhillon, R., Bhagat, S., Ang, J., and Carvey, H. (2004). The ICSI meeting recorder dialog act (MRDA) corpus. Technical report, INTERNATIONAL COMPUTER SCIENCE INST BERKELEY CA.

Zhang, G. and Ge, H. (2013). Support vector machine with a Pearson VII function kernel for discriminating halophilic and non-halophilic proteins. *Computational Biology and Chemistry*, 46:16–22.