Multimodal teaching analytics: Spanish language influence on machine learning models for teaching analytics tools

William N. Aaron Bork

August 5th, 2018

Abstract

This study in builds on previous work in Multimodal Learning Analytics (MMLA) that explored the feasibility of wearable sensors to train and evaluate machine learning models that automatically extract teacher instructional events and social planes of interaction in a classroom. This study introduces language differences as a new variable to determine if there are performance differences in teaching analytics models trained on Spanish language audio input versus models trained on English language audio input from the same teacher. 6 native-Spanish speaking teachers in Mexico wear sensor during class sessions to collected mobile eye-tracking, audiovisual recording, and accelerometry data. Class sessions are conducted in first Spanish then English. Data is used to build machine learning algorithms to train and evaluate machine learning models that automatically extract teacher instructional events and the social planes of interaction in a classroom. A quasi-experimental design is used to determine if different language input has an impact on model performance outcomes. Implications for this study will show if language makes a significant difference on performance of models which will call for further exploration into how researchers' models can be adapted to localized for linguistic differences to improve model performance. This work is towards the automatization of data processing and representation which can lead to useful tools providing teaching analytics for in-service and pre-service teachers to facilitate self-study.

Multimodal teaching analytics:

Spanish language influence

on machine learning models

for teaching analytics tools

Introduction

Teaching analytics is the field of helping teachers collect and act on data produced from their own teaching practices. Teaching analytics is a subfield of learning analytics, a field more centered on student behaviors than on teachers (Prieto, Sharma, Kidzinski, Rodríguez-Triana, & Dillenbourg, 2018). While much work has been done on student-focused learning analytics, less research exists on the teachers' action in the classroom. Within teacher analytics, a more focused subfield is forming know as multimodal teaching analytics (MMTA) and multimodal learning analytics (MMLA). "Multimodal" refers to a diverse range of data including video, audio, motion tracking, text, graphics, gestures and more (Prieto, et al.. 2018; Blikstein & Worsley, 2016; Ochoa & Worsley, 2016). These data are increasingly aggregated and analyzed with machine learning tools to find patterns in a teacher's teaching practice (Baker & Siemens, 2014; Berland, Baker, & Blikstein, 2014). An opportunity for new research exists about whether language differences in audio input will produce statistically different performance from current models for automatically identifying teacher actions and social planes in the classroom (Donnelly et al., 2016a; Donnelly et al., 2016b). This study explores this idea.

One application of this research is the exploration of creating tools to aid teacher professional development. While the data collected in multimodal teaching analytics is useful, the time and effort to transform that data into actionable information for teachers is burdensome (Prieto at al., 2018). Automatization of this data processing and representation can lead to useful tools for in-service and pre-service teachers to facilitate self-study. This study is situated within the line of research that is helping to develop the tools teachers and teacher-trainers of the future might use. More specifically, this study explores an open question of how audio input in the form of different languages influences the development of the machine learning models trained on that data (Donnelly et al., 2016a; Donnelly et al., 2016b).

Teachers' professional practices can be enhanced through appropriate professional development (Glickman, Gordon, & Ross-Gordon, 2014). In recent years, the usefulness of data-driven decision making in education has gained a foothold (Staman, Visscher, & Luyten 2013; Dunn, Airola, Lo, & Garrison, 2012). Useful formative feedback, in a human-friendly format, is important for improving teacher practices (Prieto, et al, 2018; Donnelly et al., 2016a; Donnelly et al., 2016b). Often, data about a teacher's teaching practice is collected through in-person and video observation. Additional educational professionals, such as instructional coaches, principals, or other teachers are necessary to collect these useful data. Provided that the requisite personnel are available, even the best collected data might go underutilized or even unused because of time limitations imposed upon the educational professionals involved (Staman et al., 2013. Current efforts to analyze data on a teacher's practice rely on labor and cost intensive processes which limit their practical application (Donnelly et al., 2016a; Donnelly et al., 2016b) which in turn limit teacher adoption (Staman et al., 2013). This is where automation can play a role.

MMLA & Automation

Multimodal Learning Analytics (MMLA) "captures, integrates and analyzes learning traces from different sources to obtain a more holistic understanding of the learning process, wherever it happens (Spikol et al., 2017, p. 518). MMLA are collected through a diverse set of technologies and sensors, then analyzed with machine learning techniques (Spikol et al., 2017). Work with rich multimodal data has been done with a wide range of data types including facial expression recognition (Kim, Lee, Roh, & Lee, 2015), speech recognition (Dhall, Ramana Murthy, Geocke, Joshi, & Gedeon, 2015), smartphone sensors (Han, Vinh, Lee, & Lee, 2012), classroom motion tracking (Irvin, Crutchfield, Greenwood, Kearns, & Buzhardt, 2017), wearable sensors (Prieto, Sharma, Dillenbourg, Rodriguez-Triana, 2016), and more.

The large databases built by MMLA must be processed and transformed into useful data for educators. Data processing of MMLA has evolved in the field from earlier simple machine learning algorithms to now deep and recurrent neural networks (Prieto et al., 2018). MMLA has been turning an eye towards teacher practices thus creating a space for Multimodal Teaching Analytics (MMTA). Machine learning is used to build models which can automatically analyze classroom practice data and provide actionable information to education professionals (Donnelly et al., 2016b).

Recent Studies

Early work in MMLA/MMTA has built upon traditional methods of data collection like observations, audio and video recordings, teacher self-reflections, and even student feedback (Prieto et al., 2018). New technology has expanded the types of data that can be collected and analyzed from a teacher's teaching practice. Efforts have also focused on how the technology can be used to study phenomena and problems in authentic classroom settings with minimal

disruption to teaching and learning environments (Prieto et al., 2018, Berland et al., 2014; Fischer, Wild, Sutherland, & Zim, 2014; Roschelle, Dimitriadis, & Hoppe, 2013). This is especially difficult in the noisy, complex environments of the modern classroom (Donnelly et al. 2016a; 2016b; D'Mello et al, 2015).

More recent research has been exploring the bounds of which data types to collect and improving the accuracy of models. Irvin et al. (2018) demonstrated the feasibility of motion tracking in a classroom space to collect data on how a student spends their time a various learning centers. Data was then displayed with heat maps to aid visualization. Irvin et al. (2018) suggested automated collection, processing, and reporting of these data allows for more precise data-based decisions by the teacher and school psychologist to quickly identify where a learner is spending their time in the preschool classroom to determine if the child is meeting developmental and learning goals.

In a series of studies, Prieto et al. (2018; 2016) used mobile eye-tracking, audiovisual recording, and accelerometry data from sensors worn by teachers to build models through machine learning algorithms that could extract a set of teaching activities (explanation; monitoring; questioning; repairs; other) as well as social planes (class-wide; small group; individual; other) from live-teaching sessions. Prieto et al. (2018) concluded that the machine learning models can be successfully trained with MMLA data to identify certain aspects of teaching practice with reasonable accuracy and much faster than a human coder could. This present study is based heavily on this work.

Donnelly et al. (2016a; 2016b) focused on different audio-only input sources (teacher mic and classroom mic) in a series of studies to automatically identify certain instructional events (Question & Answer; Procedures & Directions; Supervised Seatwork; Small Group Work; and

Lecture). Researchers' results demonstrate the utility of audio-only streams for detection of certain instructional events to provide teachers with personalized formative feedback about their use of class time.

Spoken Languages

MMLA/MMTA typically includes video and audio among other data types. Natural language is one of the features extracted from an audio or video stream. Since language is an important operational tool in the classroom, for both students and teachers, there is value in incorporating natural language into teacher analytic models. The bulk of research in the field is conducted with English language data which results in future tools built for English language users. Language localization might increase adoption of these tools with non-English speaking users. Additionally, non-language localized tools might present different and negative performance results. Some researchers believe that localization of machine learning models can be done by tagging high-level linguistic features that share similar linguistic structures (Donnelly et al, 2017). However, it is an open question if choice of spoken language influences the outcome of machine learning model performance in automatically extracting teacher actions from a given session. This linguistic relationship is the focus of this study.

The Present Study

It is within the language space described in the previous section that this study is situated. This present study builds upon previous work on the feasibility of wearable sensors to train and evaluate machine learning models to automatically extract teacher instructional events and social planes of interaction in a classroom (Prieto et al., 2018; 2016). This study replicates their work but uses both Spanish and then English language audio input at different times from the same teacher. The aim is to investigate whether language differences in audio input will produce

statistically different performance from current teaching analytics models for automatically identifying teacher actions and social planes in the classroom. This study investigates a stated direction for new research from Donnelly et al. (2016a; 2016b) who write, "We have also only tested our system in English language classrooms. Given the proliferation of [automatic speech recognition] ASR for many languages, we anticipate our approach will largely extend to other languages, provided an adaption be made to the natural language processing features to suit other languages" (p. 51).

Research Question

Are there performance differences in teaching analytics models trained on Spanish language audio input versus models trained on English language audio input from the same teacher?

Design Synopsis

A quasi-experimental design is used to answer this research question. Spanish language audio input is the independent variable (the treatment) that might influence the dependent variable of median F1 performance scores of the machine learning models. F1 scores are a way to determine how well the machine learning model can identify and match the teaching actions that a human hand-coded via video observation. The control is the median F1 performance scores of the machine learning models trained on English language audio input. The results will come from the difference between the median F1 performance score for teaching analytics models trained on Spanish language audio input versus models trained on English.

Implications for this study might show if language makes a significant difference on the performance of models which would call for further exploration into how researchers' models can be localized for linguistic differences to improve model performance. This is especially

useful for teachers who teach in languages other than English or teach in more than one language. For example, in a bilingual school or international school. This work is towards the automatization of data processing and representation which can lead to useful tools providing teaching analytics for in-service and pre-service teachers to facilitate self-study.

Method

Participants

Participants will be 6 Mexican teachers in Monterrey, Nuevo Leon, Mexico. Two teachers will be in the elementary level, two in middle school, and two in the high school level. All participants are bilingual native Spanish speakers with native or near-native English.

Sampling will draw from a pool of both private and public schools in the city of Monterrey that teach in English at least 50% of the time as listed by the Mexican Secretariat of Public Education (Secretaría de Educación Pública). There are less than 10.

To build the sample, researchers will contact each individual school by email to explain the study and ask for a preliminary meeting with school officials. Schools with consenting leadership are asked to identify teachers who are teaching at least once class in Spanish and one in English. A pool of interested candidate teachers is created. From this pool, random sampling will denote the 6 participants. Consenting teachers will need to obtain an informed consent document from every student in a given class session scheduled to be used for data collection. Although students are not the subjects of this study they might appear in audio and video data incidentally.

Data collection will begin in January 2018 and last 1 academic semester ending in June 2018. Participants receive no compensation for their participation.

Measures

This study sets the independent variable as Spanish language. This variable is measured simply as "Spanish Language". The dependent variable is the median F1 performance scores of the machine learning models trained on Spanish language audio input. The control is the median F1 performance scores of the machine learning models trained on English language audio input. English language is an appropriate control measure as this will allow other researchers to replicate the experiment against a common lingua franca.

F1 scores are a measure of a test's accuracy that use both the precision and the recall of the test to compute a score (Sasaki, 2007). F1 scores range from 1 (perfect precision and recall) to 0 (worse precision and recall). They are often used in the field of machine learning (Tjong, Kim Sang & De Meulder, 2003). The F1 scores in this study determine how well the machine learning models can identify and match the teaching actions that a human hand-coded via video observation. Thus, the median F1 scores for the modeled trained on Spanish language input will either be higher, lower, or equal to the F1 scores for the models trained on English language input.

Validity of the independent variable measure (Spanish Language) is a naturally strong choice for measuring which language is being used. It is an accurate representation of the construct being measured. Reliability of the independent variable measure (Spanish Language) is bolstered by the fact that only Spanish will be used during class sessions taught in Spanish.

The dependent variable measure, median F1 scores, is an appropriate choice to measure performance scores of the machine learning models trained on either Spanish or English language audio input. Median F1 scores are accepted as a measure in related work on teaching activity extraction (Prieto et al., 2018; Donnelly et al., 2016a; Donnelly et al., 2016b). This

improves reliability of the measure and allows comparison between this study and previous studies in the field. Validity is strengthened as median F1 scores provide a harmonic average of precision and recall instead of a simple accuracy score which is misleading (Prieto et al., 2018).

Procedures

Procedures and technical specification closely replicate those outlined in Prieto et al. (2018) to improve comparison to prior research. During teaching sessions, each teacher will be outfitted with SMI eye-tracking glasses to collect gaze data at 60 Hz. These glasses also record a teacher's subjective video at 24 FPS in HD resolution, plus an audio stream. Teachers will also carry a smartphone with an application that records three-axis accelerometer signals for tracking movement in the classroom.

Lesson planning. Teachers write a lesson plan for all 4 sessions. They will hand-code these lesson plans prior to teaching the lesson using the set of codes detailed later in this section. Lesson plans will be later used to compare planned activity versus human-coded activity versus machine learning extracted activity. Each lesson will be similar in structure, though the content may differ each session. Lessons should be planned for 30 - 60 minutes of activity. Teachers then conduct their planned lesson as they normally would. Class sessions will take place as normally scheduled by individual school schedules.

Data coding. Each participant teacher will participate in 4 sessions. This will result in 24 recorded sessions with 8 sessions at each level (elementary, middle, and high school). After all sessions are run, the audio + video streams will be manually coded. Two teachers (one who taught the session and one who did not) and two researchers (one who observed the session and one who did not) will come to agreement on codes for each teaching session. This coding will determine the "grounded truth" – an agreement between the teacher and researchers of what

transpired in each teaching session – to later test the machine learning models against. Four-way inter-rater reliability will help control coding bias.

For machine learning training, each data source will be partitioned into 10 second episodes using slide windows with an overlap of 5 seconds. From each 10 second window, features will be extracted for the eye-tracking data, accelerometer data, audio data, video data.

Table 1

Data Coding Tags

Coding tags for teaching activities will be:	Coding tags for social planes will be:
• EXPlanation - teacher delivering content in lecture style; • MONitoring - checking around the classroom while students work on a task; • REPairs - teacher answers to a student question or doubt; • QUEstioning - teacher tries to assess the student(s) knowledge orally; and • Other activities not included above.	• INDividual • small-GRouP • CLaSs-wide social planes, or other plane of interaction (i.e., not socially relevant or no social interaction).

Machine learning. A "personalized model" and a "general model" will be trained. A personalized model is trained on data from just a single teacher. A general model is trained on data from all the teachers in the sample. To remain consistent with the procedures of the replicated study, researchers will train the models on a leave-one-out method. The R programming language will be used for alignment and pre-processing of data. The "randomForest" R package will be utilized. Video feature extraction will be performed by Lua

implantation of VGG-19. Audio feature extraction will be done with openSMILE toolkit. Neural network models are created using the Keras library in the Python programming language. Source code for these processing and modelling tools are freely available online at the study's dedicated Github repository: https://github.com/lprisan/paper-JCAL-multimodal-teaching-analytics.

To extract the teaching activity, Markov Chain-Enhanced Random Forest classification algorithms will be used on the data. Prieto et al. (2018) found in their study that this choice "proved to have the most robust performance, both on predicting the teaching activity and the social plane of interaction" (p. 198). Their F1 scores are shown below. These scores are the baseline scores the current study will use to assess validity of its own dependent variable F1 scores.

Table 2

Median F1 Performance Scores from Prieto et al. (2018)

Personalized Model Performance	General Model Performance
• Teaching activity median F1 = 0.741 • Social plan median F1 = 0.834	• Teaching activity mean F1 = 0.716 • Social plane F1 = 0.837

Finally, to aid visual presentation of the data, a graph will be produced that lays out the actions and social planes coded by (a) teachers in their lesson plans, (b) human-coded via video, and (c) automatically detected by the machine learning models. This 3-layer graph, modelled from Prieto et al., (2018) will help communicate the findings visually to those without intimate knowledge of the field.

Data Analyses

This study is designed to determine if there are performance differences in teaching analytics models trained on Spanish language audio input versus models trained on English language audio input from the same teacher. Assuming normally distributed results, a Chi-square test will demonstrate if a significant difference exists at the $p < .05$ level between the expected and observed frequencies of the two dependent variables to be compared. The dependent variables analyzed in this test are median F1 scores of models trained on Spanish language input and the median F1 scores of models trained on English language input. Other statistical tests will be used as well. A T-test will help assesses whether the means of the two groups are statistically different from each other. If the results are not normally distributed, a Mann–Whitney U test can be applied for a means to a similar end.

Discussion

Limitations and Delimitations

This study will inherit the limitations of the replicated study while adding its own limitations regarding language. A small sample of 6 participants will produce small scale and variety in the dataset, which limits generalizability. Human coding of the teaching sessions introduces bias. The study controls for this by getting 4-way inter-rater reliability between 4 coders. Language audio data will only be collected in a single version of Spanish (Mexican Spanish) and a single regional dialect ("Regio" – people from Monterrey). Another limitation is the English levels of Mexican teachers vary from native (grew up bilingual) to non-native (learned later in life). Another limitation is that findings will only be applicable to Spanish/English language pairings. Generalizability is further limited because this study will take place in bilingual or English-majority schools. A study taking place in an English-majority

school in Monterrey, Mexico is likely to be a place of a higher socio-economic status (SES). This might introduce unseen effects of high SES culture, which influences language use, which in turn effects the data used to train the teaching analytics models. Another limitation is that a bilingual Spanish/English teacher might naturally speak in both languages without realizing it. Spanish utterances in an English class session, and vice versa, could affect the data set.

While the human aspects of the study introduce limitations, the digital tools and methods for training the machine learning models aid to bolster generalizability of findings as the tools and methods are freely available for all to download from the previous study's dedicated GitHub repository. Standardization in the non-linguistic aspects of the study makes it easier to make conclusions about how language influences the models currently being used in the field.

Implications

The results of this study might demonstrate if there are performance differences in teaching analytics models trained on Spanish language audio input versus models trained on English language audio input from the same teacher. This is especially useful information for teachers who teach in languages other than English or teach in more than one language. For example, in a bilingual school or international school. If a language induced performance difference exists, the results suggest a need to localize machine learning models to account for linguistic differences. If a language induced performance difference does not exist, the results suggest there is no need to localize machine learning models to account for linguistic differences. Either finding could strengthen the line of research that aims to develop tools to automate a process in the field of teaching analytics. This work is towards the automatization of data processing and representation which can lead to useful tools providing teaching analytics for in-service and pre-service teachers to facilitate self-study.

Future Research

Future research might include replications of this study with other language pairings, or use of more cost-effective, discrete, and private data collection means. A more diverse set of teaching contexts could be studied. Exploring if certain content areas lend themselves better to teaching analytics is yet another direction. Another extension of this study is to increase the teacher actions being studied here. Yet another future direction is using more rigorous machine learning model evaluations beyond leave-one-out methods.

References

Baker, R., & Siemens, G. (2014). Educational data mining and learning analytics. In R. K. Sawyer (Ed.), The Cambridge handbook of the learning sciences (pp. 253-272). doi.org/10.1017/CBO9781139519526.016

Berland, M., Baker, R.S., Blikstein, P. (2014). Educational Data Mining and Learning Analytics: Applications to Constructionist Research. Tech Know Learn 19. p. 205–220. doi:10.1007/s10758-014-9223-7

Blikstein, P., & Worsley, M. (2016). Multimodal learning analytics and education data mining: Using computational technologies to measure complex learning tasks. Journal of Learning Analytics, 3(2), 220–238.

Dhall, A., Ramana Murthy, O. V., Goecke, R., Joshi, J., & Gedeon, T. (2015). Video and image based emotion recognition challenges in the wild: EmotiW 2015. *In Proceedings of the 17th ACM International Conference on Multimodal Interaction. ICMI'15.* ACM, p. 423–426.

Donnelly, P.J., Blanchard, N., Olney, A.M., Kelly, S., Nystran, M., D'Mello, (2017). Words matter: automatic detection of teacher questions in live classroom discourse using linguistics, acoustics, and context. *LAK '17 Proceedings of the Seventh International Learning Analytics & Knowledge Conference* p. 218-227. doi: 10.1145/3027385.3027417

Donnelly, P.J., Blanchard, N., Samei, B., Olney, A.M. Sun, X., Ward, B., Kelly, S., Nystran, M., D'Mello, S.K. (2016). Automatic teacher modeling from live classroom audio. *UMAP '16 Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*. p. 45-53. 10.1145/2930238.2930250

Donnelly, P.J., Blanchard, N., Samei, B., Olney, A.M. Sun, X., Ward, B., Kelly, S., Nystran, M., D'Mello, S.K. (2016). Multi-sensor modeling of teacher instructional segments in live classrooms. *ICMI 2016 Proceedings of the 18th ACM International Conference on Multimodal Interaction*. p. 177-184. 10.1145/2993148.2993158

D'Mello, S.K., Olney, A.M., Blanchard, N., Samei, B., Sun, X., Ward, B. and Kelly, S. 2015. Multimodal Capture of Teacher-Student Interactions for Automated Dialogic Analysis in Live Classrooms. Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (2015), 557–566.

Dunn, K.E., Airola, D.T., Lo, W.J., Garrison, M. (2012). What teachers think about what they can do with data: Development and validation of the data driven decision-making efficacy and anxiety inventory. Contemporary Educational Psychology 38(1). p. 87-98. doi: https://doi.org/10.1016/j.cedpsych.2012.11.002

Fischer, F., Wild, F., Sutherland, R., & Zim, L. (2014). Grand challenge problems from the Alpine Rendez-Vous. In Grand challenges in technology enhanced learning. SpringerBriefs in Education (pp. 3–71). Cham, Switzerland: Springer International Publishing.

Glickman, C.D., Gordon, S., & Ross-Gordon, J.M. (2014). SuperVision and instructional leadership - A developmental approach. Pearson Education.

Han, M.; Vinh, L.T.; Lee, Y.-K.; Lee, S. (2012). Comprehensive Context Recognizer Based on Multimodal Sensors in a Smartphone. *Sensors* 2012, *12*, 12588-12605. doi: 10.3390/s120912588

Irvin, D.W., Crutchfield, S.A., Greenwood, C.R., Kearns, W.D., Buzhardt, J. (2018). An automated approach to measuring child movement and location in the early childhood classroom. *Behavior Research Methods, 50(*3), 890–901. doi: 10.3758/s13428-017-0912-8

Kim, Bo-Kyeong, Lee, Hwaran, Roh, Jihyeon, and Lee, Soo-Young. (2015). Hierarchical committee of deep CNNs with exponentially-weighted decision fusion for static facial expression recognition. *In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15)*. ACM, New York, NY, USA, p. 427-434. DOI: http://dx.doi.org/10.1145/2818346.2830590

Ochoa, X., & Worsley, M. (2016). Editorial: Augmenting learning analytics with multim. Journal of Learning Analytics, 3(2), 213–219.

Prieto, L.P., Sharma, K., Kidzinski, Ł., Rodríguez-Triana, M.J., Dillenbourg, P. (2018). Multimodal teaching analytics: automated extraction of orchestration graphs from wearable sensor data. Journal of Computer Assisted Learning. 0266-4909. https://doi-org.proxy1.cl.msu.edu/10.1111/jcal.12232

Prieto, L,P., Sharma, K., Dillenbourg, P., Rodríguez-Triana, M. J. (2016). Teaching analytics: towards automatic extraction of orchestration graphs using wearable sensors. LAK '16 Proceedings of the Sixth International Conference on Learning Analytics & Knowledge. 148-157 doi: 10.1145/2883851.2883927

Sasaki, Y. (2007). The truth of the F-measure. Teach Tutor mater, p. 1–5.

Spikol, D., Prieto, L., Rodríguez-Triana, M., Worsley, M., Ochoa, X., Cukurova, M.,Vogel, B., Ruffaldi, E., Ringtved, U.L. (2017). Current and future multimodal learning analytics data challenges. *LAK '17 Proceedings of the Seventh International Learning Analytics & Knowledge Conference* p. 518-519. doi:10.1145/3027385.3029437

Staman, L., Visscher, A.J., Luyten, H., (2014). The effects of professional development on the attitudes, knowledge and skills for data-driven decision making. Studies in Educational Evaluation (42). p. 79-90.

Tjong Kim Sang, E. F., De Meulder, Fien. 2003. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. CONLL -03 Proceedings of the seventh conference on natural language learning at HFT-NAACL 2003 - Volume 4 p. 142-147.