Linguistic Changes across Different User Roles in Online Learning Environment. What do they tell us?

Lavendini Sivaneasharajah, Katrina Falkner, Thushari Atapattu
The University of Adelaide
{lavendini.sivaneasharajah, katrina.falkner, thushari.atapattu}@adelaide.edu.au

ABSTRACT

In recent years, we have witnessed an increasing interest in online learning environments, particularly in Massive Open Online Courses (MOOCs). However, prevailing studies show that lower percentage of students complete their courses successfully in online learning environment. The vast amount of student data available in MOOC platforms enables us to gain insight into student learning behaviours. In this paper, we explore the idea of ‘student roles’, identifying linguistic change associated with roles that will later help us to understand students’ learning process in MOOCs. As an initial stage of this research, the study aims to categorise student roles (e.g. information seeker, information giver) using discourse analysis, and to further analyse the linguistic change for each student role with time. A multi-class classifier has been built to identify user roles with 82.20% F-measure. Further, our study on linguistic changes demonstrates that distinctive behaviors can be observed across different user roles. Prominent observations include discourse complexity, lexical diversity, level of information embeddedness and lexical frequency profile being high in information giver in comparison to information seeker and other user roles.

Keywords

MOOCs, Discussion forums, Student Role, Natural Language Processing, Machine Learning

1. INTRODUCTION

With the advent of Massive Open Online Courses (MOOCs) there has been an eruption in learning environment [10]. Students are increasingly seeking alternative learning mediums, with MOOCs increasingly looked upon as a valuable source of learning. As many of the MOOCs are freely available for students, it draws interest of thousands of learners.

According to the statistics, over 101 million learners are globally registered to study using MOOCs by the year of 2018 [16]. However, studies show that only one in every twenty students who enrol in MOOCs complete their studies successfully [9]. The participation in MOOCs seems complex with students’ enrolment for varying purposes and varying intentions [17]. Completion is not necessarily the only indicator of learning success. Knowing that students may enroll to courses for other purposes, we need to explore other perspectives of learning success beyond completion.

The primary problem aims to solve by this research is whether analysing the student role and its associated linguistic change can be used to understand student learning. We believe learner role can give us an indication on whether learning gain is important to measure learning success. We try to answer “Moving between roles is potentially an indication of learning gain”. This hypothesis has not been explored yet in prevailing literature.

As the studies discuss in this paper demonstrate a proof of concept, our initial stage is to identify user roles and a sample of linguistic indicators that are associated with these roles. Our overarching goal is to track these roles and their associated linguistic changes with time. And eventually, predicting the grades for student using these discourse features. We assume observing these roles and associated linguistic changes will eventually result in a deeper understanding of the student’s learning lifecycle.

Hecking et al. [5] identifies these roles with both linguistic and community-related features (e.g. votes, views). However, in a real time system, it is not realistic to wait for the community-related features to classify students into different roles as structural features can be generated throughout the course and they may change with time. Therefore, we intend to identify student roles in MOOC discussion forums solely based on a discourse analysis.

Few research studies [3; 4] have focused on linguistic changes that occur in online communities. Yet, linguistic changes have not been studied along with students’ role.

With reference to the aforementioned aspects, we aim to answer the following research questions: RQ1: Can we build a model that could predict user roles (information seeking, information giving, other) using linguistic only features? RQ2: Do linguistic indicators change significantly across user roles?

Answering these questions will result in identifying user roles and its associated linguistic changes in discussion forums. These studies may assist to understand students’ learning in online learning environment.

The contribution of this work includes a multi-class classification model that uses linguistic-only features to predict user roles in MOOC discussion forums. Since it uses linguistic-only features, our model can be applied to any online forums (e.g. technical forums) for role prediction. Further, we examine the linguistic indicators and its changing patterns associated with user roles at this stage with the intention of proposing a framework in future to understand students’ learning.

Lavendini Sivaneasharajah, Katrina Falkner and Thushari Atapattu "Linguistic Changes across Different User Roles in MOOCs: What do they tell us?" In: *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, Anna N. Rafferty, Jacob Whitehill, Violetta Cavalli-Sforza, and Cristobal Romero (eds.) 2020, pp. 667 - 671

2. RELATED WORK

2.1 User role/ Post classification

Searle's taxonomy [15] has been widely used in literature and proven to be a most successful method in speech act classification. From this point, several classification mechanisms have been evolved based on Searle's taxonomy [1; 7].

Hecking et al. [5] have carried out post classification by integrating the categories that prevail in the existing research studies. The study presents three classes (information seeking, information giving and other). It used content-related features (e.g. phrases – "need help or helps you") and contextual features (e.g. position in the thread, number of votes) for classification purposes and obtained an average of 70% accuracy.

In a real time system, it is not realistic to wait for the contextual features to predict the given classes as they occur throughout the course and changes with time. Therefore, our study focuses solely on the linguistic aspects over contextual and structural features.

2.2 Linguistic change in online communities

An important facet of linguistic research is to identify the correlation between user lifespan and their language use [3; 13]. Given the rich recent work on linguistic analysis in different online communities [3; 6; 13; 14], research scholars also have attempted linguistic analysis in MOOC. Dowell et al. [4] have conducted a study on MOOC data to identify the conversion in learner's language and discourse characteristic with time. However, the research did not investigate the linguistic changes associated with each user role. To address this gap, we conducted several experiments using different linguistic features to discover discourse complexity, lexical diversity, number of embedded information and lexical frequency profile. Even though preliminary work on linguistic change has been conducted in other online communities, there is a lack of work conducted in MOOCs.

3. METHODOLOGY

3.1 Data set

We extracted a dataset from the AdelaideX[1] 'Introduction to Project Management' and 'Risk Management for Projects' courses offered in 2016 and 2017 respectively. A total of 9,497 user posts was extracted from 923 different users. We sampled 6000 posts from 'Project Management' for this study. We extracted user posts of students who have posted a minimum of six posts during the entire semester. Posts were manually annotated as information seeker (IS), information giver (IG) and other (O) user roles by two independent human evaluators. According to Cohens kappa, the high inter-rater agreement (k= 0.924) between the two annotators ensures the validity of the human annotation.

User role identification

We adopted machine learning techniques to build a multi-class classifier to predict user roles (IG, IS and O) for a given forum post using discourse features and linguistic features that were extracted using Linguistic Inquiry and Word Count (LIWC) tool[2]. We extracted multiple features to reflect several facets of the text.

We implemented multi-class classifiers using weka for role identification. All classifiers were tested using 10 Fold Cross-Validation to assess effectiveness.

The imbalanced data were handled using Synthetic Minority Oversampling TEchnique (SMOTE). Here we split the data into 70% for training and validation and 30% for testing. Then, we oversample the minority class on each training fold during cross validation. Then, validated the classifier on the remaining fold.

On the other hand, we also performed further analysis on role classification to explore the potential techniques that can be used to address this problem. We implemented multi-class text classification using Keras, a high-level neural network API.

We used existing pre-trained GloVe word embedding to convert the user posts to 100 dimension vectors. Then, built the model with one input layer, one embedding layer and one Long Short-Term Memory (LSTM) layer with 128 neuros and one output layer with three neurons.

3.2 Linguistic study

In our second study, we conducted several linguistic experiments (e.g. discourse complexity, lexical diversity) to understand the linguistic changes of each user role with time.

3.2.1 Discourse Complexity

According to an existing study by Crossley et al. [2], discourse complexity can be measured by several linguistic indicators. One possible way is using any given reading level measures. Therefore, we used Flesch Kincaid [8] , a reading level measure and used this measure to explore discourse complexity with time for each user. We also analysed the association between discourse complexity and student roles.

3.2.2 Lexical Diversity

We calculated lexical diversity to measure the vocabulary usage in the given user posts. Measuring the level of lexical diversity requires to quantity how often different kind of words are used in text. According to the prevailing literature [12], lexical diversity can be measured using different formulas such as type-token ration (TTR), measure of textual lexical diversity (MTLD), vocd-D and many. Due to flaws in the traditional methods, we chose MTLD over other lexical diversity measures as MTLD avoids the adverse effects on text length in measuring the lexical diversity.

3.2.3 Lexical Frequency Profile

We have examined the Lexical Frequency Profile (LFP) associated with each user role to understand how well user has written his discourse. Initially, we extracted n-grams from lecture transcripts. We used CountVectorizer to tokenise the text and built a vocabulary list for lecture transcripts.

We created Lexical Frequency Profile for each user post with respect to the given vocabulary list using spaCy[3] , an advanced Natural Language Processing API. We created a Phrase Matcher Object and applied the matcher object on each user post to

[1] https://www.edx.org/school/adelaidex

[2] https://liwc.wpengine.com/

[3] https://spacy.io/

extract the keywords. Finally, we examined LFP for user roles and its pattern during a role change.

Since these linguistic indicators are normally distributed, we performed One-Way ANOVA to compare the mean value for each variable's distribution. These linguistic indicators are examined during role changes to understand whether there is a significant difference between user roles.

3.2.4 *Information Embeddedness*

Information embeddedness is one of the key elements that contributes towards student learning. In our study, information embeddedness can be defined as the number of information that can be extracted from any given discourse. This study attempts to find the level of information embeddedness using clause extraction.

Clause extraction has been used in a previous study [11] to determine the relationship between the clauses per sentence and language development. We develop a novel approach in which clauses are been extracted from parse tree using a rule-based approach.

Initially, a pipeline is being built with Part-Of-Speech (POS) tagging, lemmatisation using Stanford CoreNLP[4] to get the basic interpretation of a student post. Tree Annotation is used to extract a parse tree for a given sentence. Here, we divided a student's post into multiple sentences and identified the number of clauses embedded in each sentence. Initially, clause-level tags (e.g. SBAR) and word-level coordinating conjunction (e.g. CC) have been extracted from the parse tree. Then, we implemented a rule-based approach to extract the number of clauses.

4. Results

The experiment on role identification addresses information seeking, information giving and other role classification solely based on discourse analysis. We analysed the features extracted from LIWC. Further, we performed feature selection technique known as Recursive Feature Elimination with Cross Validation (RFECV) for feature ranking. We performed the feature ranking on 1200 user posts obtained from Risk Management course. According to the RFECV sixteen optimal number of features have been selected. We retrieved the features with highest ranking and fed these features to the classifier.

We conducted Multivariate Analysis of Variance (MANOVA) to measure the significance between linguistic features and user roles. Table 1 presents the top five variables that exhibit the largest effects size along with multivariate F value (Wilks' λ).

Table 1: MANOVA analysis of language variables

Feature	F	η^2
Words per Sentence	754.853*	0.201
Question Mark	505.057*	0.144
Article	493.305*	0.141
Interrogatives	385.516*	0.114
Personal pronouns	294.884*	0.090

*p<0.001

We implemented multiclass classifiers with different sets of algorithms using Weka. All these classifiers were tested using 10 Fold Cross-Validation to assess the accuracy. Among these, the Random Forest classification model performed best with 82.20 of F measure (see Table 2). Table 2 reports the accuracy, precision, recall and F-measures for different set of classifiers.

Table 2: Results of classifier performance

Classifiers	Accuracy	Precision	Recall	F1	Cohen's Kappa
Naïve Bayes	71.28	74.40	71.30	71.00	0.5117
Random Forest	82.17	82.30	82.20	82.20	0.6955
Simple Logistic	79.35	79.60	79.40	79.40	0.6473
Logistic	79.43	79.70	79.40	79.50	0.6498
SMO	74.80	76.50	74.80	75.30	0.5770

We also performed, the text classification with Keras. As stated above we used GloVe 100 dimension vector to create the vector space for each user posts. We obtained 88.06 as test accuracy. We halt the model from further training to avoid over fitting.

The experiment on linguistic analysis uses different indicators to address the linguistic change associated with each user role.

According to the reading level measures (discourse complexity), the results indicates that if a particular user role can be seen in consecutive posts the level of complexity increases/decreases with minimum change and when there is a role change (e.g. IS → IG or IG→IS or O →IG) there is a dramatic change in discourse complexity. This trend is observed across our data set.

The mean value of Flesch Kincaid Grade Level measure for user roles are as follows: μ = 16.15±12.86(IG), μ= 8.77± 7.43(IS) and μ =5.30± 6.99 (O). High Flesch Kincaid score indicates the discourse is difficult to understand. This implies that discourse complexity decreases along with these user role changes whereas the readability of the text becomes easier with these role changes. The results from the One-Way ANOVA test show that there is a significant difference in mean values (discourse complexity) with p-value<0.001 among these user roles.

As stated above, lexical diversity of user posts were obtained via calculating 'Measure of Textual Lexical Diversity'. The mean value of MTLD for user roles are: μ = 60.845± 32.380 (IG), μ= 52.18± 39.59 (IS) and μ = 34.46± 46.55 (O). This indicates that lexical diversity of the user roles are decreasing along these role changes.

The results of Lexical Frequency Profile show that the number of lecture related keywords used in user post changes during a role change. For a given user, the number of keywords used in an information giving post increases – reach an optimal number and decreases with time whereas for an information seeking post it increases/decrease with time. Moreover, information giver uses more keywords from the lecture transcript than information seeker and other.

In information embeddedness factor, we extracted the clauses using a rule-based approach. Once the number of clauses been extracted using clause-level tags and rule-based approach, we compared them with user roles (IG, IS, O). Figure 1 shows the level of information embeddedness in a user posts (number of clause) changes with time for sample of three users.

[4] https://stanfordnlp.github.io/CoreNLP/

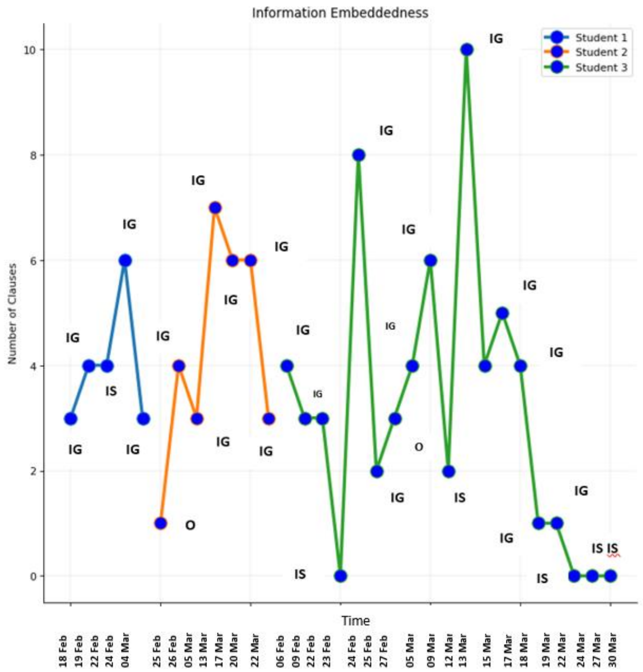

Figure 1: Information embeddedness across user roles with time

5. DISCUSSION

5.1 Can we build a model that could predict user roles using linguistic only features?

In the existing literature [5], user roles have been predicted based on the other contextual features (e.g. votes, views) which delays the predictions in a real time MOOC environment. Therefore, this study extends the line of research to construct a machine learning model to identify user roles (IS/IG/O) using linguistic-only features.

To address RQ1, we conducted an experiment on IS/IG/O post classification. The success of our approach with 82.20% of F-measure and the text analysis performed using Keras with 88.06 accuracy demonstrates that simple linguistic features can be used in role predictions in real time.

According to the feature space used in this classification, it is evident that sixteen identified linguistic features (see Table 1- we presented the first five features that holds largest effects size due to the page limit) can distinguish IG, IS, O posts. For example, information-seeking posts contain high amount of negative emotions comparatively to information giving posts. Likewise, the number of question marks is high in information seeking posts compared to information giving user posts.

5.2 Do linguistic indicators change significantly across user roles?

To investigate RQ2, we conducted several linguistic experiments to explore the linguistic change across user roles.

The results of our linguistic experiments demonstrate that the readability level (µ of Flesch Kincaid Grade Level) of the information giver is high (i.e. discourse complexity is high) when compared to information seeker and other user roles. This implies that there is a high dramatic change in the linguistic complexity during a role change. One possible reason can be information givers tend to include words that are more complex and provide extensive information when comparing to other user roles.

We further anlaysed this results by manually analyzing random user posts retrieved from the data set. According to the sample user posts given below, information givers try to elaborate their information with more complex words than information seeker.

Information Giver- *"Great use of the likelihood/impact scale! You might also want to use the PESTLE framework to identify broader areas of potential concern..."*

Information Seeker- *"That was great can i please gain form you, the Challenges you faced during you first project"*

Similarly, the results obtained for lexical complexity shows that lexical complexity is higher for information giver. According to the above sample user posts, it is vital that the vocabulary usage is higher in information giver than information seeker.

The trend in the lexical frequency profile shows that information giver uses more keywords from lecture transcripts at the beginning of the course, reaches an optimal point and decreasing afterwards. One possible reason could be that they are enthusiastic to share the lecture related information during the start of the course and it increases with time. Further, the reason to decrease the amount of content-related keywords from the lecture transcript at the end of the course might be because they elaborate concepts in their own words or uses related keywords from other resources as they progress.

We can observe two kind of trends in information seeker. First trend is they use more keywords as they progress. The reason could be, they might not know the content at the beginning but with time, they know the course related keywords. Other trend is they use less keywords with time. The reason might be they try to change their role from the information seeker. Further, we hope to do a meticulous analysis to explore these patterns with the intention of discovering the exact reasons behind them.

In summary, we have achieved the aim of our study as the classifications is purely built upon the idea of utilising linguistic-only features. Further, to understand student learning, we explored RQ2 by examining the different linguistic indicators. These linguistic indicators will have a great potential to understand a user's behavior in any kind of discussion forum.

6. CONCLUSION

We have presented a multi-class user role classification in MOOC discussion forums using linguistic-only features with the intention of eliminating the drawbacks (e.g. contextual features) that exist in previous studies. Our model performed well comparing to base line model with 82.20% of F-measure

On the other hand, our linguistic study gives us a clear differentiation on linguistics aspects associated with each role. The level of information embeddedness, and discourse complexity and lexical diversity of information giver is high compared to information seeker and other. As a proof of concept, our technique demonstrated the potential of identifying the linguistic behaviors for each user role.

This novel approach holds a great promise for user role classification and the associated linguistic behavior in MOOC discussion forums. Additionally, we believe that tracking these role changes and associated linguistic changes will help to understand the student learning in MOOC discussion forums.

7. REFERENCES

[1] Bhatia, S., Biyani, P., and Mitra, P., 2012. Classifying User Messages For Managing Web Forum Data. In *Proceedings of the Proceedings of the 15th International Workshop on the Web and Databases* (2012), 13–18.

[2] Crossley, S.A., Greenfield, J., and McNamara, D.S., 2008. Assessing Text Readability Using Cognitively Based Indices. *TESOL Quarterly 42*, 3, 475-493. DOI= http://dx.doi.org/10.2307/40264479.

[3] Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., and Potts, C., 2013. No country for old members: user lifecycle and linguistic change in online communities. In *Proceedings of the Proceedings of the 22nd international conference on World Wide Web* (Rio de Janeiro, Brazil, 2013), ACM, 307-318. DOI= http://dx.doi.org/10.1145/2488388.2488416.

[4] Dowell, N.M.M., Brooks, C., Kovanović, V., Joksimović, S., and Gašević, D., 2017. The Changing Patterns of MOOC Discourse. In *Proceedings of the Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale* (Cambridge, Massachusetts, USA, 2017), ACM, 283-286. DOI= http://dx.doi.org/10.1145/3051457.3054005.

[5] Hecking, T., Chounta, I.-A., and Hoppe, H.U., 2016. Investigating social and semantic user roles in MOOC discussion forums. In *Proceedings of the Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (Edinburgh, United Kingdom, 2016), ACM, 198-207. DOI= http://dx.doi.org/10.1145/2883851.2883924.

[6] Huffaker, D., Jorgensen, J., Iacobelli, F., Tepper, P., and Cassell, J., 2006. Computational measures for language similarity across time in online communities. In *Proceedings of the Proceedings of the HLT-NAACL 2006 Workshop on Analyzing Conversations in Text and Speech* (New York City, New York, 2006), Association for Computational Linguistics, 15-22.

[7] Kim, S.N., Wang, L., and Baldwin, T., 2010. Tagging and linking web forum posts. In *Proceedings of the Proceedings of the Fourteenth Conference on Computational Natural Language Learning* (Uppsala, Sweden, 2010), Association for Computational Linguistics, 192-202.

[8] Klare, G.R., 1974. Assessing Readability. *Reading Research Quarterly 10*, 1, 62-102. DOI= http://dx.doi.org/10.2307/747086.

[9] Koller, D., Ng, A., and Chen, Z., 2013. Retention and Intention in Massive Open Online Courses: In Depth (2013). ' 'from https://er.educause.edu/articles/2013/6/retention-and-intention-in-massive-open-online-courses-in-depth

[10] Loya, A., Gopal, A., Shukla, I., Jermann, P., and Tormey, R., 2015. Conscientious Behaviour, Flexibility and Learning in Massive Open On-Line Courses. In *Proceedings of the Procedia - Social and Behavioral Sciences* (2015 2015), 519-525. DOI= http://dx.doi.org/10.1016/j.sbspro.2015.04.686.

[11] Lu, X., 2011. A Corpus-Based Evaluation of Syntactic Complexity Measures as Indices of College-Level ESL Writers' Language Development. *TESOL Quarterly 45*, 1, 36-62.

[12] McCarthy, P.M. and Jarvis, S., 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods 42*, 2, 381-392.

[13] Nguyen, D. and Rosé, C.P., 2011. Language use as a reflection of socialization in online communities. In *Proceedings of the Workshop on Languages in Social Media* Association for Computational Linguistics, 76-85.

[14] Postmes, T., Spears, R., and Lea, M., 2006. The Formation of Group Norms in Computer-Mediated Communication. *Human Communication Research 26*, 3, 341-371. DOI= http://dx.doi.org/10.1111/j.1468-2958.2000.tb00761.x.

[15] Searle, J.R., 1976. A Classification of Illocutionary Acts. *Language in Society 5*, 1, 1-23.

[16] Shah, D., 2018. By The Numbers: MOOCs in 2018 (2018). ' 'from https://www.classcentral.com/report/mooc-stats-2018/

[17] Zheng, S., Rosson, M.B., Shih, P.C., and Carroll, J.M., 2015. Understanding student motivation, behaviors and perceptions in MOOCs. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, 1882-1895.