Identifying Student Leaders from MOOC Discussion Forums through Language Influence

Seungwhan Moon **Saloni Potdar** **Lara Martin**
Language Technology Institute
Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA, 15213
{seungwhm, spotdar, ljmartin}@cs.cmu.edu

Abstract

Identifying and understanding the motivations of *student leaders* from Massively Open Online Course (MOOC) discussion forums provides the key to making the online learning environment engaging, collaborative, and instructive. In this paper, we propose to identify student leaders solely based on textual features, or specifically by analyzing how they *influence* other students' language. We propose an improved method of measuring language accommodation based on people's choice of words given a semantic topic of interest, and show that student leaders indeed coordinate other students' language usage. We also show that our proposed method can successfully distinguish student leaders from the two MOOC discussion forum datasets.

1 Introduction

One of the challenges Massively Open Online Courses (MOOCs) face is that they lack a physical medium that enables active real-time interaction between students and instructors, especially when compared to the offline learning environment. While online discussion forums in MOOCs play an important role in bridging this gap, the "massiveness" of the student size makes it hard for instructors to provide sufficient feedback or answers to students' questions in a timely manner.

It is often the *student leaders* who accommodate this situation by voluntarily helping other students and answering their questions in discussion forums. The student leaders encourage other students to participate in the discussion and make the online learning experience much more collaborative and engaging. Therefore, it is important to identify student leaders and understand their motivations, thus promoting more students to act like leaders. Identifying leadership in MOOCs also brings new insights to the multi-dimensional evaluation of students in online courses. This significantly builds upon previous literature that evaluates students taking MOOCs solely based on their task-oriented performance (Foltz and Rosenstein, 2013; Basu et al., 2013).

Identifying student leaders in MOOC courses is a challenging task, as illustrated in Figure 1. While most of the student leaders actively interact with other students in a large cluster of people, some student leaders only lead a small clique of students. Activeness of student participation cannot be a sole measure to identify student leaders, because there are a number of active 'questioners' who exhibit very different motivations from student leaders. This challenge inspires us to look closely at the *language* of the leaders in order to identify them.

The task of identifying leaders has been well studied in various domains, but the challenge is often unique to the specific property of an online network or a community. For example, a frequency-based data mining approach has been proven particularly successful for a social network with a strong visibility control (e.g. a friend network) and a discrete set of user actions (e.g. sharing of a post, etc.) (Goyal et al., 2008; Bodendorf and Kaiser, 2009; Shafiq et al., 2013). In their work, they identify leaders by tracking how a certain action gets shared and propagated among a given network of users. However, it is challenging to apply this approach for identifying leaders from MOOC discussion forums, because a visibility network of users or community actions are not clearly defined in MOOCs.

For an online community forum where the query information and use pattern are accessible, several studies have proposed to use the link structure and the topic information about users to identify opinion leaders (Li et al., 2013; Pal and Kon-

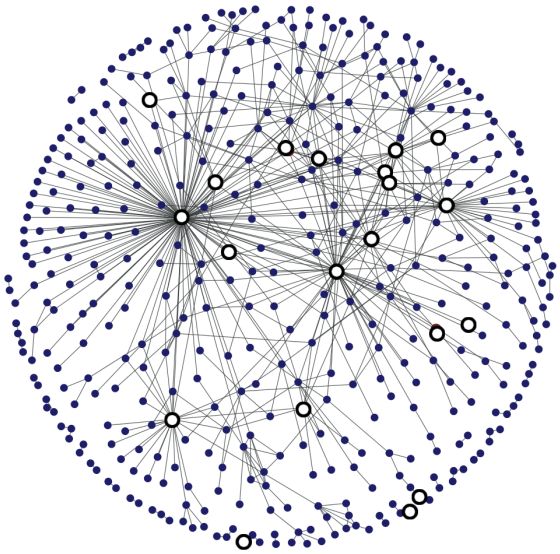


Figure 1: An interaction graph of the Python MOOC discussion forum where each node and edge represents a student and an interaction of two students within the forum (e.g. enough number of conversation exchanges above a threshold), respectively. Larger white nodes refer to the annotated student leaders. While most of the leaders are highly connected (actively interacting with other students), note that the white nodes may also appear in small cliques as well. Some of the highly connected nodes are not labeled as leaders, whom we refer to as active 'questioners'.

stan, 2010; Sharara et al., 2011). They employ features such as PageRank, HITS, and other non-linguistic features such as longevity (how long the person has stayed on the forum), etc., all of which serve as a cue in determining and identifying the extent of users' expertise and influence.

While some of the MOOC datasets provide this information, in this paper we only focus on the textual features of the MOOC discussion forums so that we can target general MOOC datasets. We show that we can identify leaders as role models who influence through language, and show how a community norm may form within a short life span of an online course via student leaders. We also propose a new approach to measure language accommodation which in our experiment furthers the previous literature on the subject.

The rest of the paper is organized as follows: Section 2 explains in detail the approach that we propose to identify leaders. Section 3 gives a brief overview of the two MOOC datasets from different courses, and we present our empirical evaluation in Section 4 on these datasets. Finally, we give our concluding remarks and proposed future work in Section 5.

2 Methods

It is well studied by the linguistics community that people tend to mimic the style of speech or choices of words made by the people that they are communicating with (Niederhoffer and Pennebaker, 2002). This phenomena is called *language coordination*, which is frequently observed especially when there are power differences within the conversation participants (Danescu-Niculescu-Mizil and Lee, 2011; Danescu-Niculescu-Mizil et al., 2012). We hypothesize that the power difference may arise within the students as well, especially through *dependence*: if a student A needs knowledge from a student B, and is thus dependent on B, this gives B a temporary power over A. As such, we identify a set of student leaders by how much other students accommodate their language when they converse with student leaders.

In order to measure students' language coordination towards student leaders, we take the similar approach proposed by (Danescu-Niculescu-Mizil et al., 2012). In their work, they provide a concise probabilistic coordination measure which defines language coordination from a speaker to a target on a set of function words. Specifically, they use 8 pre-defined categories and a total of 451 lexemes as a set of function words to track the language influence. Their proposed accommodation measure is shown to be successful in distinguishing the individuals of different power status. While this work bases its motivation from a specific line of work in the linguistics that defines particular function words as markers for influence, it does not fully capture the broad range of linguistic behaviors that are reported as language accommodation (Baxter and Braithwaite, 2008; Hall, 2008).

In this paper, we propose to measure language coordination based on people's choice of words, given a specific *theme*. Consider word clusters learned from a large corpus, where words are grouped by their semantic similarity. During a conversation between a speaker A and a target B, they can draw words from any cluster, which is analogous to choosing a topic or theme to discuss. Given a theme, people may choose any words from the chosen cluster, all of which have

a semantically similar meaning. However, if A follows B's specific choice of words given a cluster, we consider this action as evidence for language accommodation of A towards B. Based on the probabilistic analysis, we measure the overall language coordination for each conversation participant. Note that this definition of language accommodation can capture language coordination beyond the use of particular function words, and provide a way to analyze broader language influence that is unique to the community. Figure 2 shows the illustration of this approach.

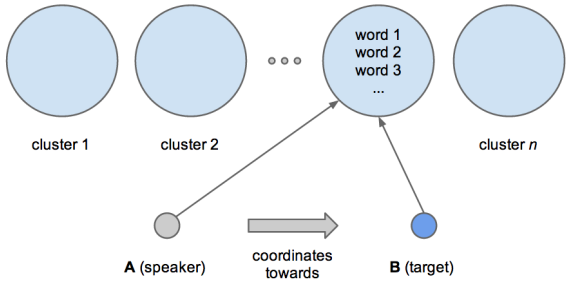


Figure 2: Language accommodation based on people's choice of words given a theme (cluster). Words are clustered based on their semantic similarity. If A (speaker) follows B (target)'s specific choice of word from a cluster, given all the other options of similar words within the same cluster, we define this action as language accommodation of A towards B.

To cluster words based on their syntactic and semantic similarity, we take the approach by (Mikolov et al., 2013a; Mikolov et al., 2013b) which maps words into high-dimensional vectors based on their statistical occurrence in relation to other words in a sentence. We then use the K-means clustering algorithm (MacQueen, 1967) to group the words by their Euclidean distance within the semantic space. To reduce the computational complexity, we pick the 20 most frequent clusters from the dataset that we analyze, and we use the words in those clusters as markers to track language coordination.

We then borrow the definition of language accommodation measure by (Danescu-Niculescu-Mizil et al., 2012), and define the language coordination of a speaker a towards a target b on a marker w_k (that belongs to a word cluster k) as follows:

$$C^{w_k}(a \rightarrow b) = P(E^{w_k}_{u_a \rightarrow u_b} | E^{w_k}_{u_b}) - P(E^{w_k}_{u_a \rightarrow u_b})$$

where a is the speaker that coordinates towards the target b, $E^{w_k}_{u_a \rightarrow u_b}$ is the event that the utterance of a exhibits a linguistic marker w_k in its reply to the utterance of b, and $E^{w_k}_{u_b}$ is the event that the utterance of b exhibits a marker w_k. The conversation set is defined over the exchanges that contain the words from a given cluster k.

In a thread-based discussion forum like the MOOC datasets, however, it is ambiguous to tell who is talking with whom. Therefore, we define the conversational exchange between b and a if b's post appears after a's post in the same thread.

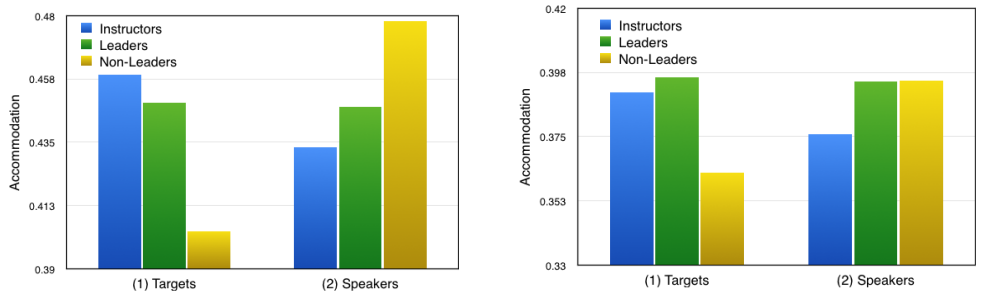
3 MOOC Dataset

In this section, we describe the two MOOC online discussion forum datasets we used in our studies. The datasets consist of the conversations from two courses from Coursera[1]: *Learn to Program: The Fundamentals* (Python) and *Introduction to Psychology as a Science* (Psychology). The Python course consists of 3,509 students, 7 instructors and 24,963 posts across 10 weeks. Each thread consists of replies and comments along with a username associated with it. The Psychology course spans over 12 weeks and has 1,982 students and 3 instructors. In our studies, we focus on the interaction between three groups of people: instructors (including professors and teaching assistants), student leaders, and non-leaders. In order to evaluate the performance of the proposed method on the MOOC discussion forums, we have hand-annotated leaders and non-leaders from a subset of the student pool.

4 Results and Discussion

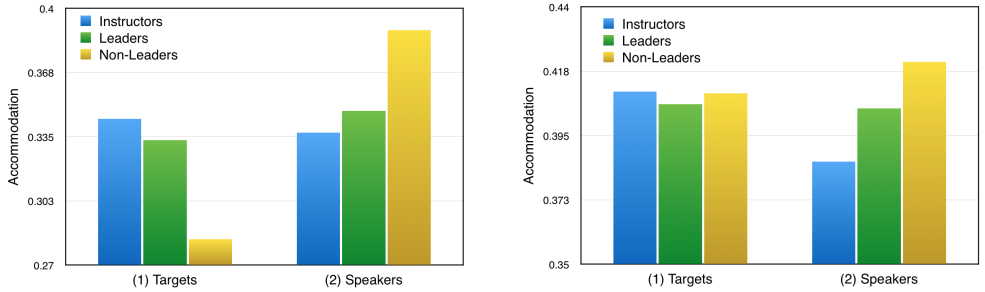
We test the following two hypotheses on language accommodation: (1) students coordinate more towards student leaders than towards non-student leaders (H_{target}), and (2) student leaders coordinate towards other students less than non-student leaders coordinate towards other students ($H_{speaker}$). Figure 3 shows the language accommodation of three different groups (instructors, leaders, and non-leaders) with other students that are not labeled as any group. We provide the results for the case when we apply our *cluster-based* accommodation measure to test H_{target} and $H_{speaker}$, and for when we use the function words as markers to track accommodation (Danescu-Niculescu-Mizil et al., 2012). For the *cluster-*

[1] https://www.coursera.org, one of the leading MOOC providers



(a) Python: Cluster-based

(b) Python: LIWC-derived Function Words



(c) Psychology: Cluster-based

(d) Psychology: LIWC-derived Function Words

Figure 3: The aggregated language accommodation measurement using (a), (c): cluster-based and (b), (d): LIWC-derived lexemes, (1) from students towards each target class (testing H_{target}) and (2) from each speaker class towards students (testing $H_{speaker}$), for the Python and the Psychology datasets.

based method, we use `word2vec`[2] which provides the word vectors trained on the *Google News* corpus (about 100 billion words) (Mikolov et al., 2013b). Table 1 directly shows the difference between the two methods.

Figure 3 shows that student leaders influence other students' language more than non-leaders do ($p < 0.05$), supporting our first hypothesis H_{target}. It can also be seen that the language of non-leaders coordinates towards that of other students more than the language of student leaders does ($p < 0.05$), supporting our second hypothesis $H_{speaker}$. Note that instructors and leaders exhibit almost the same behavior in terms of language accommodation. These results coincide with the observation that student leaders and instructors play a similar role in discussion forums. In addition, while both word cluster-based and LIWC-derived methods support our hypotheses, the distinction seen is more significant in the result from our cluster-based method (summarized in Table 1). These results indicate that the proposed method of measuring accommodation can capture the language influence more accurately than the previous method.

Based on our proposed measure of language accommodation, we were able to see how language influence is accumulated throughout the lifetime of the community. Figure 4 shows that the language coordination of students towards student leaders decreases as the course progresses, eventually converging to the level of language coordination from students to non-student leaders. The same convergence behavior can be observed from the language coordination of student leaders and non-leaders towards students as well. This result indicates that the distinction between students and non-student leaders becomes less significant in terms of their language influence. This result can also be interpreted as a community norm being formed throughout the course, which was initiated by student leaders at first. While MOOC courses have a relatively short lifespan, the results make intuitive sense because they often include technical jargon (e.g. the programming related words for Python MOOC course) which can be quickly learned by community members.

Table 2 shows the prediction accuracy on the task of differentiating between a student leader and a non-leader given a set of conversation exchanges between two people (a,b) with different status. We used the following features as input to an SVM classifier. *Cluster* uses the binary fea-

[2] `https://code.google.com/p/word2vec/`

		Δ Accommodation (%)	
		Cluster	LIWC
(a)	ΔC_{target}	**4.58**	3.35
	$\Delta C_{speaker}$	**-3.04**	-0.01
(b)	ΔC_{target}	**5.01**	-0.38
	$\Delta C_{speaker}$	**-4.09**	-1.62

Table 1: The difference in language accommodation measure between leaders and non-leaders for each method (cluster-based, LIWC-derived function words) on (a) Python and (b) Psychology MOOC datasets. ΔC_{target} refers to the students' language accommodation towards leaders subtracted by their language accommodation towards non-leaders. $\Delta C_{speaker}$ refers to the leaders' language accommodation towards students subtracted by non-leaders' language accommodation towards students. Higher absolute value of ΔC indicates that the method can distinguish leaders and non-leaders better.

tures that indicates whether a coordinates towards b more than b towards a on each marker from the word cluster-based method. *LIWC* uses the binary features as well, using the LIWC-derived function words as markers for accommodation. *BOW* refers to a standard bag of words feature set.

We test the performance on both in-domain and cross-domain cases using the datasets from the two different courses. While *BOW* performs significantly better than the other two coordination features-based methods for the in-domain cases, it does not generalize well for the cross-domain cases. This is because there are unique sets of technical vocabulary that are used in each respective course, which are often strong indicators of leadership or expertise in the domain. The proposed cluster-based method performs better than *LIWC* in both in-domain and cross-domain cases, showing that the proposed method better captures the leader's language influence on other students.

5 Conclusions

The main contributions of this paper are as follows: we have proposed that identifying student leaders from MOOC discussion forums is an important task that can potentially improve the quality of the courses by promoting a collaborative and engaging learning environment. We then proposed

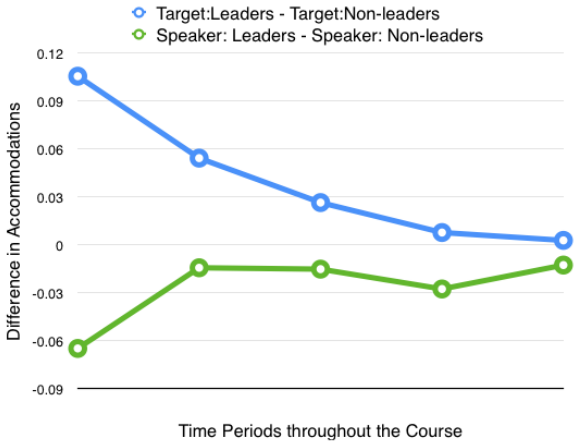


Figure 4: Language accommodation difference at each period throughout the Python course. The blue line (upper) refers to ΔC_{target}, whereas the green line (lower) refers to $\Delta C_{speaker}$. Higher absolute value of ΔC indicates that the method can distinguish leaders and non-leaders better.

	In-domain		Cross-domain	
Train:	Python	Psych	Python	Psych
Test:	Python	Psych	Psych	Python
Cluster	61.17	57.54	60.01	59.03
LIWC	58.34	55.10	58.52	57.92
BOW	73.12	69.23	53.26	54.07

Table 2: Classification accuracy of identifying a leader from a pair of students with different labeled roles. *Cluster* and *LIWC* refer to the coordination features using two different methods to track influence markers. The chance performance is 50 %.

a new method to measure language accommodation based on people's choices of words given a theme. We have shown that our proposed approach can better capture the language influence than previous literature on accommodation using the two MOOC datasets. We were also able to show that a community norm can be formed throughout the course, evidenced from the time-based analysis of language accommodation.

We plan to improve this research with respect to the way we measure language accommodation. Specifically, we would like to propose a new metric for measuring language accommodation by analyzing the propagation of influence, instead of looking at conversations locally. Suppose, for in-

stance, that during an online discussion a person b coordinates towards a with respect to a specific linguistic style marker m, and that within a short period of time, we find evidence that another person c coordinates towards b on the same marker m. We argue that c should be considered as pertaining to the influence graph of a, contributing to the evidence that a is a leader.

Acknowledgments

We would like to acknowledge the contributions made by Professor Carolyn Rosé for her valuable and constructive suggestions.

References

Sumit Basu, Chuck Jacobs, and Lucy Vanderwende. 2013. Powergrading: a clustering approach to amplify human effort for short answer grading. *NIPS Workshop on Data Driven Education*.

Leslie A. Baxter and Dawn O. Braithwaite. 2008. *Communication Accommodation Theory. Engaging theories in interpersonal communication: Multiple perspectives*.

Freimut Bodendorf and Carolin Kaiser. 2009. Detecting opinion leaders and trends in online social networks. In *Proceedings of the 2Nd ACM Workshop on Social Web Search and Mining*, SWSM '09, pages 65–68, New York, NY, USA. ACM.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialog. In *Proceedings of the ACL Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. *Proceedings of WWW 2012*.

Peter W. Foltz and Mark Rosenstein. 2013. Tracking student learning in a state-wide implementation of automated writing scoring. *NIPS Workshop on Data Driven Education*.

Amit Goyal, Francesco Bonchi, and Laks V. S. Lakshmanan. 2008. Discovering leaders from community actions. *CIKM '08*.

Phil Hall. 2008. Policespeak. *Dimensions of Forensic Linguistics*.

Yanyan Li, Shaoqian Ma, Yonghe Zhang, Ronghuai Huang, and Kinshuk. 2013. An improved mix framework for opinion leader identification in online learning communities. *Knowledge-Based Systems*, 43(0):43 – 51.

J MacQueen. 1967. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1: Statistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*.

K. G. Niederhoffer and J. W. Pennebaker. 2002. Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 21.

Aditya Pal and Joseph A. Konstan. 2010. Expert identification in community question answering: Exploring question selection bias. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 1505–1508, New York, NY, USA. ACM.

M. Zubair Shafiq, Muhammad U. Ilyas, Alex X. Liu, and Hayder Radha. 2013. Identifying leaders and followers in online social networks. *Selected Areas in Communications, IEEE Journal on (JSAC)*, 31.

Hossam Sharara, Lise Getoor, and Myra Norton. 2011. Active surveying: A probabilistic approach for identifying key opinion leaders. In *The 22nd International Joint Conference on Artificial Intelligence (IJCAI '11)*.