Automatic Identification of Questions in MOOC Forums and Association with Self-Regulated Learning

Fatima Harrak, François Bouchet, Vanda Luengo, Rémi Bachelet

Automatic identification of questions in MOOC forums and association with self-regulated learning

Fatima Harrak
Sorbonne Université
CNRS, Laboratoire
d'Informatique de Paris 6,
LIP6, F-75005 Paris, France
fatima.harrak@lip6.fr

François Bouchet
Sorbonne Université
CNRS, Laboratoire
d'Informatique de Paris 6,
LIP6, F-75005 Paris, France
francois.bouchet@lip6.fr

Vanda Luengo
Sorbonne Université
CNRS, Laboratoire
d'Informatique de Paris 6,
LIP6, F-75005 Paris, France
vanda.luengo@lip6.fr

Rémi Bachelet
Centrale Lille
PRES Université Lille Nord de
France
F-59650 Villeneuve d'Ascq
remi.bachelet@centralelille.fr

ABSTRACT
Discussion forums can be a rich source to analyze students' questions but it can be challenging to find relevant categories of questions. We considered here students' posts from the discussion forum of four editions of a same French MOOC on Project Management. We extended a coding scheme to annotate questions based on their content (course vs. non course) and trained 3 stages of an automatic annotation model. Then we studied the correlation between the nature of the questions asked and students' performance and self-regulation. The results are promising and reveal, for the minority of students active on forums, the possibility to use this feature to better estimate their performance and some of their self-regulation skills based on questions they ask.

Keywords
Student's question, discussion forum, coding scheme, self-regulation, student's performance, MOOC

1. INTRODUCTION
Students' questions play an important role in the learning process and are meaningful for both learning and teaching science [4]. The need for students to ask questions, or to point out errors in the course, are as salient in distance e-learning as they are in a classroom setting, thus emphasizing the importance of discussion forums in online learning and in MOOCs in particular [1]. Forums are not only a place for socialization, but also a place where learning happens, as learners post questions, opinions, and concerns, which are viewed, rated and answered by fellow learners and/or teaching staff. Therefore, we conducted analyses to explore the nature of questions asked by students in a MOOC forum and particularly tried to see the relationship between those questions and students' performance and self-regulation skills. More particularly, we wanted to answer to the following research questions:

(RQ1) Is it possible to reliably annotate questions extracted from MOOC forum posts according to a fine-grained multi-level coding scheme?

(RQ2) Is there a relationship between the nature of the questions asked on a MOOC and the students' performance and mastery of self-regulated skills?

2. STATE OF THE ART
Studying discussion in MOOC forums is still an ongoing topic of research. Zeng et al. [11] identified the confusion messages by using discussion forum posts derived from large open online courses. Sentiment analysis of MOOC forums discussions can also help in identifying the dropout behavior from students' posts [10].

Researchers have studied students' questions in a variety of educational settings, such as classroom [3], tutoring [7] and online learning environments [9]. Graesser and Person [7] developed a taxonomy of questions asked during tutoring sessions to be used for automatic question generation. Although their taxonomy could be relevant to our work, some categories included high quality 'deep-reasoning questions' and are associated to patterns of reasoning which are difficult to identify automatically.

We also investigated how self-regulated learning (SRL) was used to analyze students' interactions in online environments. Dettori et Persico [6] used a taxonomy of indicators of SRL to analyze directly what kind of students are self-regulated from their messages. Bouchet et al. [2] characterized students via clustering according to their interactions with an intelligent tutoring system fostering SRL.

Table 1: Descriptive statistics of the 4 MOOC sessions considered (registration, messages and success)

Session	GDP5	GDP6	GDP7	GDP8
Students reg.	17579	23315	19392	24603
Answered to quiz 1	4842	7537	5951	7998
Bas. certif. obtained	2282	3900	2393	4526
Adv. certif. obtained	503	697	559	589
Nb of posts	7655	10597	12224	14072
Nb of unique posters	2087	4717	3504	4760

3. CONTEXT AND DATASET

We consider in this paper four datasets made of forum messages posted in four different sessions (5 to 8) of the same biyearly French MOOC on project management called GDP (French acronym for project management) held in 2015 and 2016. The MOOC allows participants to obtain a basic certificate (15-25 hours estimated workload), and an advanced one (35-45 hours). Therefore for each participant we can determine two final grades and whether one, both or none of the certificates were obtained. The forum is organized around threads created by the pedagogical team to answer to technical or administrative issues, about homework or course content, *etc.* Table 1 provides some basic statistics on the forum usage and number of students registered.

Additionally, participants to the MOOC are invited to fill an optional (only 0.25% bonus points for filling it) research questionnaire after 2 weeks in the MOOC which included 21 psychometrically evaluated questions evaluating their SRL skills in an online setting [5] using 7-point Likert scales along 4 dimensions: (1) cognitive and metacognitive strategies, (2) procrastination, (3) context adaptation, and (4) peer support. We filtered out the participants who failed to answer to embedded attention check questions (*e.g.* "answer 5 here") to increase the reliability of the data considered.

4. QUESTION CODING SCHEME

To identify the nature of students questions in the forums, we considered a sample of 500 messages from the 4 sessions, randomly divided into 3 sub-samples (200/100/200). We applied 4 categorization steps to define a coding scheme as proposed in another context by Harrak et al. [8].

The raw corpus contains unstructured and noisy messages from students (*e.g.* a message can contain several questions, opinions, answers to issues not course related, *etc.*). We first filtered out the messages from the instructors, those that are a reply to other ones (*i.e.* not the root messages) and the explicitly non-course related topics (*e.g.* thread dedicated to technical issues). The messages were then segmented into several questions (using Python library NLTK) and annotated according to their content. The course-based questions were annotated with the coding scheme from [8] (summarized in upper Table 2) which consists in 4 independent dimensions: a mandatory main one (dimension 1), and 3 optional ones (dimensions 2 to 4). For instance, a question could be a request to re-explain the way something work by providing another example (tagged as Ree on dimension 1, Exa on dimension 2, Man on dimension 3, and nothing on dimension 4, *i.e.* vector [Ree,Exa,Man,0]). The non-course related questions were then annotated according to newly defined dimension 0 (*cf.* lower Table 2). Two human annotators made separate independent annotations on each dimension, and their agreement was evaluated using Cohen's Kappa. First the kappa was calculated for the agreement on whether a segment was a question or not ($\kappa = 0.85$) and then on explicit questions only ($\kappa = 0.96$). Then agreement was calculated for the topic of the question (course *vs.* non-course, $\kappa = 0.85$). Finally, kappas were calculated for dim1 to 4 ($\kappa_1 = 0.70$, $\kappa_2 = 0.61$, $\kappa_3 = 0.69$, $\kappa_4 = 0.57$) for course questions and for dim0 for non-course questions ($\kappa = 0.58$).

Table 2: Coding scheme for course-based students' questions (Dim 1 to 4, adjusted from [8]) and for non-course related questions (Dim0)

Code	Question category	Description
Dim1-4: Course-based questions		
Dim1: question type		
Ree	Re-explain / redefine	Ask for an explanation already done in the course material
Dee	Deepen a concept	Broaden a knowledge, clarify an ambiguity or request for a better understanding
Ver	Validation / verification	Verify or validate a formulated hypothesis
Dim2: explanation modality / question subject		
Exa	Example	Example application (course/exercise)
Sch	Schema	Schema application or an explanation about it
Cor	Correction	Correction of an exercise in course/exam
Dim3: explanation type		
Def	Define	Define a concept or term
Man	Manner (how?)	The manner how to proceed
Rea	Reason (why?)	Ask for the reason
Rol	Roles (utility?)	What's the use/function
Lin	Link between concepts	Verify a link between two concepts, define it
Dim4: verification type		
Mis	Mistake / contradiction	Found potential error in course/teacher's explanation
Kno	Knowledge in course	Verify knowledge
Exp	Expected knowledge	Verify expected information in exam or quiz (assessment)
Dim0: Non-course related questions		
Soc	Socialization	Social questions
Adm	Administrative issues	MOOC administration: registration, certificate, etc.
Exa	Exam/ quiz	Ask for assessment modality: notes, format, etc.
Tec	Technical issues	Detect a technical problem and ask for solution
Res	Ressources not found	Ask for not found ressources
Too	Tools	Ask for tools for a task
Pha	Phatic	Question that has no real value or information

5. AUTOMATIC ANNOTATION

To annotate the set of questions asked by the students, a semi-automatic tool based on rules and keywords manually weighted was used in prior work to annotate automatically the questions. Although effective (average kappa of 0.70), many questions were not annotated by this tool [8], which led us to develop another automatic tool using machine learning techniques trained on the corpus of questions.

Table 3: Kappa values between automatic annotation and the reference manual annotation

Classifier	SVM	DT	GLM	GBT	K-NN	NB	RI
(1)	0.60	**0.91**	**0.91**	**0.91**	0.80	0.54	**0.91**
(2)	**0.66**	0.40	0.62	0.51	0.33	0.47	0.43
(3)	-	0.28	0.35	**0.37**	0.22	0.21	0.11
(4)	-	0.61	0.63	**0.68**	0.27	0.07	0
(5)	-	0.30	0.09	**0.39**	0.05	0.03	0.37
(6)	-	0.50	**0.56**	0.54	0.14	0	0.26
(7)	-	**0.48**	0.45	0.47	0.19	0.07	0.35

(-): Not suitable for non binary data

We performed the classical preprocessing steps on the training sample of 1307 segments (500 messages) manually annotated: tokenization, stemming, punctuation removal (except for '?') and stopwords (non-meaningful words) removal. We then extracted all the unigrams and bigrams and counted their occurrences in that sample. Each of the 1307 segments was represented by a binary word vector ('1' if the word is in the segment, '0' otherwise). We finally reduced the number of keywords extracted (high number of keywords compared to the number of segments) to keep the most important and significant ones using a feature selection technique (removing less frequent and correlated unigrams/bigrams).

We designed a 3-stage annotator to identify segments with questions, course *vs.* non-course related questions and the nature of those questions. Overall, 7 classifiers were trained to annotate a segment respectively: (1) into question/non-question; (2) into course/non-course related questions; (3) for non-course related questions, according to dim 0; (4-7) for course-based questions, according to dim 1 to 4. For each classifier we trained models using different machine learning techniques with a 10-fold cross-validation: Support Vector Machine (SVM), Generalized Linear Model (GLM), Gradient Boosted Trees (GBT), Decision Tree (DT), K-NN, Naive Bayes (NB) and Rule Induction (RI), each with various values of hyperparameters. For each classifier, the input was the words vectors representing the segments in terms of keywords, and the label to predict was the value associated to the segment in that dimension. Classifiers (1) and (2) took a binary values and each of the other classifiers took nominal values (varying from 3 to 7, according to the dimension).

We then calculated the Kappa values between the predictions from the classification models and the ground truth values from the manual annotation (*cf.* Table 3). The entire corpus of segments of messages was annotated by the techniques with the highest performance for each classification.

6. RELATIONSHIP BETWEEN QUESTIONS, SUCCESS AND SELF-REGULATION

6.1 Data coding

To study the relationship between question type and success and self-regulation, we coded for GDP8 students who posted on the forum the number of segments categorized as :

- an explicit question (NbQ)
- not a question or an implicit one (NbNQ)
- a non-course question (NbQ-NC)
- a non-course question corresponding to a socialization (NbQ-NC-Soc), an administrative issue (NbQNC-Adm), an exam (NbQ-NC-Exa), a technical issue (NbQ-NC-Tec), a ressource not found (NbQ-NC-Res), a tool issue (NbQ-NC-Too) or a phatic (NbQ-NC-Pha),
- a course question (NbQ-C),
- a course question about a reexplanation (NbQ-C1-Ree), a request to go deeper in a concept (NbQ-C1-Dee), or a verification (NbQ-C1-Ver)
- a course question requesting an example (NbQ-C2-Exa), a schema (NbQ-C2-Sch), or a correction (NbQ-C2-Cor),
- a course question asking for a definition (NbQ-C3-Def), the way to proceed (NbQ-C3-Man), the reason for something (NbQ-C3-Rea), the role of something (NbQ-C3-Rol), or the link between two concepts (NbQ-C3-Lin),
- a course question asking for a verification regarding an apparent mistake (NbQ-C4-Mis), knowledge from the course (NbQ-C4-Kno) or whether something is expected to be learned or not for the assessment (NbQ-C4-Exp).

In addition to those variables relative to the questions asked, we also have for each student :

- **four SRL scores**, measured by a questionnaire [5]. Although the authors average the 4 individual scores into an overall SRL score (with procrastination coded in a reverse manner), we believed they captured different facets of SRL which could individually be associated to different question asking behavior. Therefore we defined for each student their lack of procrastination score (Sco_{NPr}), their context score (Sco_{Ctx}), their strategy score (Sco_{Str}) and their peer support score (Sco_{Pee}). Each score is an average of 5 to 6 questions, between 1 and 7,
- **two performance scores** for the basic/advanced MOOC track (Sco_{Bas} and Sco_{Adv}), a value between 0 and 100.

6.2 Correlation analysis

Method: We calculated for each question variable (NbQ-*) its Pearson correlation coefficient (r) with each of the self-regulation and performance variables (Sco_*).

Results: 286 students posted at least one message with a segment containing an explicit question. The results (not all detailed here) reveal that asking explicit questions (on the basic $[p = .012, r = .148]$ and the advanced tracks $[p = .000, r = .237]$), and questions on a topic relevant to the course (on the basic $[p = .010, r = .153]$ and the advanced tracks $[p = .000, r = .253]$), is a behavior positively correlated with the performance. The questions the most strongly positively correlated to performance are the ones to check one's understanding (Ver) regarding a theme of the course ($p = .000, r = .292$) or a skill one is expected to master for the final exam ($p = .000, r = .282$).

123 students among those who posted at least one message with an explicit question had also filled the SRL questionnaire. The results (summarized in Table 4) reveal that procrastination is not correlated with any particular type of question. It is however logically negatively correlated with the score in the basic ($r = -.349$) and advanced ($r = -.372$) tracks of the MOOC, *i.e.* students who procrastinate have lower scores overall. The context facet of SRL is positively correlated only with the number of messages not containing a question (NQ). The two other facets are more interesting, as the students who self-report being good at using cognitive and metacognitive strategies (such as note-taking) while

Cat.	Sco_{NPr}		Sco_{Ctx}		Sco_{Str}		Sco_{Pee}	
	r	p	r	p	r	p	r	p
NQ	−.064	.480	.178	**.049***	−.137	.131	.259	**.004***
Q	−.064	.485	.056	.541	.169	.061	.247	**.006***
NC	−.104	.254	.123	.174	−.161	.076	.212	**.019***
NC-Soc	−.095	.295	.081	.376	−.107	.239	.198	**.028***
NC-Adm	−.058	.520	.162	.074	.018	.840	.258	**.004***
NC-Exa	−.090	.322	−.006	.951	−.216	**.016***	.078	.392
NC-Tec	−.060	.513	.023	.802	.001	.995	.155	.087
NC-Res	.067	.459	.138	.128	.039	.665	.166	.067
NC-Too	.007	.939	.103	.257	.016	.858	−.050	.584
NC-Pha	−.018	.844	.079	.385	−.226	**.012***	.093	.307
C	−.028	.761	.005	.958	−.143	.115	.222	**.014***
C1-Ree	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
C1-Dee	−.040	.660	.075	.407	−.001	.987	.219	**.015***
C1-Ver	−.020	.660	−.022	.805	−.179	**.048***	.196	**.030***
C2-Exa	.005	.957	.123	.174	−.123	.174	.068	.487
C2-Sch	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
C2-Cor	−.105	.247	.038	.675	−.045	.619	.118	.193
C3-Def	.005	.957	−.130	.151	.135	.136	.055	.547
C3-Man	−.089	.329	.111	.221	−.002	.984	.187	**.038***
C3-Rea	−.068	.457	.140	.123	.024	.796	.222	**.014***
C3-Rol	n/a	n/a	n/a	n/a	.n/a	n/a	n/a	n/a
C3-Lin	−.009	.918	.117	.197	−.067	.459	.076	.406
C4-Mis	−.112	.216	−.063	.491	−.067	.460	−.048	.599
C4-Kno	−.009	.919	−.048	.597	−.176	.052	.191	**.034***
C4-Exp	−.024	.795	.097	.288	−.136	.135	.169	.062

n/a: no segment annotated with this code

Table 4: Correlation between the question types and the four SRL scores

learning online are asking less question about the organization of the final exam, less phatic questions and less verification questions. Thus it seems that being more organized, maybe when they watch the video or go through pages of contents, they have a lesser need to verify information probably already mentioned somewhere. As for students who self-report being good at interacting with others to learn in a more efficient manner, logically they post more messages (both questions and non-question), which can be related to the course or not. When analyzing the nature of the questions they ask, they socialize more with others and ask more administrative questions. They also tend to ask very practical questions about the course about how to perform a task or the reason some concept is working that way.

7. DISCUSSION AND CONCLUSION

We have shown it is possible to annotate not only messages from a MOOC forums, but individual questions within sometimes long messages. Segmenting messages allows to distinguish finer-grain intent of the student, using an adapted coding scheme for both course and non-course related questions. This result opens the way to automatically tagging MOOC posts, for instance to help the pedagogical team to quickly know the intent of the messages that have not received a reply yet. Another interesting aspect is the fact that the nature of the questions asked within the messages provides information on some aspects of students' level of self-regulation (their tendency to interact with others for learning and their use of cognitive and metacognitive strategies). It is also worth noting that some of the patterns found here, such as the fact that students who ask verification questions tend to succeed overall better than others, are consistent with previous results in a different context [8].

Some limits include: the topic of the MOOC which hindered the classifiers performance with its low technical vocabulary (words overlap between the content and context of the course) and the average kappa values obtained for the classifiers which can reduce the impact of some correlations observed, correlation values which are themselves never extremely high even when $p < .05$. Finally, as always with results relative to MOOCs forum, they are only used by a minority of active learners. Future directions involve considering some messages excluded here (messages that are not root in the thread, technical or socialization threads which could fit in the non-course coding scheme), and considering forums from MOOCs on different themes to build up a larger corpus of messages, to try to improve the annotator performance.

8. REFERENCES

[1] M. A. Andresen. Asynchronous discussion forums: success factors, outcomes, assessments, and limitations. *J. of Educ. Technology & Society*, 12(1):249–257, 2009.

[2] F. Bouchet, J. M. Harley, G. J. Trevors, and R. Azevedo. Clustering and Profiling Students According to their Interactions with an Intelligent Tutoring System Fostering Self-Regulated Learning. *J. of Educ. Data Mining*, 5(1):104–146, 2013.

[3] C. Chin and G. Kayalvizhi. Posing Problems for Open Investigations: What questions do pupils ask? *Research in Science & Technological Education*, 20(2):269–287, 2002.

[4] C. Chin and J. Osborne. Students' questions: a potential resource for teaching and learning science. *Studies in science education*, 44(1):1–39, 2008.

[5] L. Cosnefroy, F. Fenouillet, and J. Heutte. Développement et validation d'une échelle d'apprentissage autorégulé en ligne. In *2e Colloque international e-Formation des Adultes et Jeunes Adultes*, Lille, France, 2018.

[6] G. Dettori and D. Persico. Detecting Self-Regulated Learning in Online Communities by Means of Interaction Analysis. *IEEE Transactions on Learning Technologies*, 1(1):11–19, 2008.

[7] A. C. Graesser and N. K. Person. Question asking during tutoring. *American educational research journal*, 31(1):104–137, 1994.

[8] F. Harrak, F. Bouchet, V. Luengo, and P. Gillois. Profiling Students from Their Questions in a Blended Learning Environment. In *Proc. of the 8th Int. Conf. on Learning Analytics and Knowledge*, LAK '18, 102–110, New York, NY, USA, 2018. ACM.

[9] H. Li, Y. Duan, D. N. Clewley, B. Morgan, A. C. Graesser, D. W. Shaffer, and J. Saucerman. Question Asking During Collaborative Problem Solving in an Online Game Environment. In *Intelligent Tutoring Systems*, LNCS, 617–618. Springer, Cham, 2014.

[10] M. Wen, D. Yang, and C. P. Rosé. Sentiment Analysis in MOOC Discussion Forums: What does it tell us? In *Educational Data Mining* 130–137, 2014.

[11] Z. Zeng, S. Chaturvedi, and S. Bhat. Learner Affect Through the Looking Glass: Characterization and Detection of Confusion in Online Courses. In *Int. Conf. on Educational Data Mining*, 272–277, 2017.