Identifying relationships between students' questions type and their behavior

Fatima Harrak
Sorbonne Universités
UPMC Univ Paris 06,
CNRS, LIP6 UMR 7606,
4 place Jussieu, 75005 Paris, France
fatima.harrak@lip6.fr

François Bouchet
Sorbonne Universités
UPMC Univ Paris 06,
CNRS, LIP6 UMR 7606,
4 place Jussieu, 75005 Paris, France
francois.bouchet@lip6.fr

Vanda Luengo
Sorbonne Universités
UPMC Univ Paris 06,
CNRS, LIP6 UMR 7606,
4 place Jussieu, 75005 Paris, France
vanda.luengo@lip6.fr

ABSTRACT
We present the process of categorization of students' questions, and through a clustering on students, we show the relevance of this classification to identify different profiles of students. It opens perspectives in assisting teachers during Q&A sessions.

Keywords
Clustering, question taxonomy, students' behavior.

1. INTRODUCTION
Studying learners' questions while they learn is essential [1], not only to understand their level and eventually help them learn better [2] but to help teachers in addressing these questions. Analyzing students' questions can help for instance in distinguishing deep learning vs. shallow learning [3]. In this paper, we are interested in whether the type of questions asked by students on an online platform is characteristics of their classroom behavior. We investigate this question in the context of an hybrid curriculum (like [4]), where students have to ask questions before the class to help professors prepare their Q&A session. Our goal here is threefold: (RQ1) Can we define a taxonomy of questions relevant to analyze students' questions? (RQ2) Can we automatize the identification of these questions? (RQ3) Can annotated questions asked by a student inform us about their performance, attendance and questioning behavior?

2. RESEARCH METHODOLOGY
We addressed these research questions in 3 successive steps: (1) we conducted a manual process of categorization of students' questions, which allowed us to propose a taxonomy of questions, (2) we used this taxonomy for an automatic annotation of a corpus of students' questions, (3) to identify students' characteristics from the typology of questions they asked, we used clustering technique over two courses and then characterized the obtained clusters using a different set of features, as in [5].

The dataset used for this work is made of questions asked in 2012 by 1st year medicine/pharmacy students from a major public French university (Univ. Joseph Fourier). Each course is made of 4 to 6 4-week sequences on the PACES[1] platform. After a 1st week dedicated to learning from online material, during week 2 students must ask questions and vote for questions asked by other students on an online forum to help professors prepare their Q&A session in week 3. Therefore, for each of the 13 courses, we have 4 to 6 sets of questions asked by students (6457 questions overall) during the 2nd week of each sequence.

[1] paces.medatice-grenoble.fr

3. RESULTS

3.1 Categorization of questions
To answer to RQ1, we took a sample of 600 questions (around 10% of the corpus size) from two courses (biochemistry [BCH], histology & developmental biology [HBDD]), which are considered to be among the most difficult courses and had the highest number of questions asked. This sample was randomly divided in 3 sub-samples of 200 questions to apply 3 different categorization steps: a discovery step, a consolidation step and a validation step. Step 1 consisted in grouping sentences with similarities to extract significant concepts. Then we segmented the combined questions to standardize the previous annotation and we grouped the extracted categories into independent dimensions, where each dimension grouped similar concepts in sub-categories. Step 2 consisted in annotating the second sub-sample to validate the dimensions previously identified and to make sure they were indeed independent from each other. In step 3, we performed a double annotation to validate the generality of our categories on the remaining sub-sample of 200 sentences. Two human annotators used as a unique reference the taxonomy previously created. They annotated independently each dimension (average kappa = 0.70) – discussions to fix discrepancies led to a final refinement of the categories' description. Finally, a re-annotation was performed on the entire sample (600 sentences) to consider the changes and to provide a grounded truth for the automatic annotation. The final taxonomy is provided in Table 1.

Table 1. Final question taxonomy from manual annotation

Dim1	**Type questions**	**Description**
1	Re-explain / redefine	Ask for an explanation already done in the course material.
2	Deepen a concept	Broaden a knowledge, clarify an ambiguity or request for a better understanding
3	Validation / verification	Verify/validate a formulated hypothesis
Dim2	**Modality explanation**	**Description**
0	N/A	None – attributed when neither of the other values below applies
1	Example	Example application (course/exercise)
2	Schema	Schema application or an explanation about it
3	Correction	Correction of an exercise in course/exam
Dim3	**Type of explanation**	**Description**
0	N/A	None – attributed when neither of the other values below applies
1	Define	Define a concept or term
2	Manner (how?)	The manner how to proceed
3	Reason (why?)	Ask for the reason
4	Roles (utility?)	What's the use / function
5	Link between concepts	Verify a link between two concepts

Dim4	Optional: if question is a verification	Description
1	Mistake / contradiction	Detect mistake/contradiction in course or in teacher's explanation.
2	Knowledge in course	Verify knowledge
3	Exam	Check exam-related information

3.2 Automatic annotation

To answer to RQ2 and to annotate the whole corpus (and on the long term, to use it online to analyze the questions collected), we identified keywords representative of each value in each dimension (e.g. the word "detail" is representative of a "deepen a concept" question). Then we developed an automatic tagger which identifies for each question the main value associated to each dimension and tags the question as such. We validated the automatic annotator by comparing its results on the manually annotated subsample of 600 questions and obtained a kappa value of 0.74, enough to consider applying it to the full corpus.

3.3 Links between questions and behavior

To identify whether the type of questions asked can inform us on students' characteristics, first we performed two clustering analyses using K-Means algorithm (with k varying between 2 and 10) over two datasets: students who asked questions in the BCH course (1227 questions asked by N_1=244 students) and in the HBDD course (979 questions asked by N_2=201 students). We performed the clustering using as features for each student the proportion of each question asked in each dimension (*e.g.* the proportion of questions with value 1 in dimension 1) asked (a) overall, (b) during the first half of the course, and (c) during the second half of the course (44 features overall). Distinguishing (b) and (c) in addition to (a) allowed us to take into account whether it was a change in questions asked that could be meaningful, more than the overall distribution. We obtained 4 clusters in both cases.

The second step consisted in characterizing the clusters by considering attributes not used for the clustering: students' grade on the final exam on this course (out of 20), attendance ratio (from 0 [never there] to 1 [always there]), the number of questions asked in this course, and the number of votes from other students on their questions in this course. Students for whom this data was not available were excluded from the datasets, leading to two smaller sample sizes (N'_1=173 and N'_2=161). We performed two one-way ANOVA for grades on these two clusterings and found statistically significant differences (p<0.001 and p<0.001). For the other variables, the distribution did not follow a normal law and we therefore performed a Kruskal-Wallis H test on ranks associated to each variable. The test showed that there was a statistically significant difference for attendance (p=0.04 and p=0.02), number of questions asked (p<0.001 and p<0.001) and number of votes received (p=0.04 and p<0.001) for BCH and HBDD respectively. Results are summarized in Table 2.

Table 2. Differences between the 4 BCH and HBDD clusters

Course	Cluster	*N*	Grade (/20)	Attendance	# quest.	# votes
BCH	A	53	7.97	0.86	2.83	3.06
	B	63	8.54	0.90	2.92	2.69
	C	86	9.38	0.93	6.23	2.61
	D	42	11.2	0.93	11.74	1.22
HBDD	A	59	7.43	0.89	3.53	5.57
	B	34	9.78	0.92	2.44	2.47
	C	72	10.11	0.92	6.54	3.69
	D	36	11.78	0.95	7.00	1.71

4. DISCUSSION AND CONCLUSION

Overall, when considering the results presented in Table 2, we see two similar clusters in both cases: A and D. Cluster A is made of around 28-41% of the students with grades lower than average, attending less to classes, asking less questions than average but which are particularly popular (probably because of votes from similar students, but that information was unfortunately not available). In terms of questions asked, they had a higher number of "how to" questions (*cf.* dim3-2 in Table 1) than any other cluster. On the other end of the spectrum, cluster D is made of around 21% of the students with grades above average, high attendance, who ask more questions than average that are fairly unpopular – we can assume these must be very precise questions that already require a good understanding of the content of the course, and are thus not deemed as important by other students. Interestingly, when comparing the proportion of questions asked in the first vs. second half of the class, cluster D students are the only ones who asked more questions in the 2nd half of the 4-6 sequences than in the 1st half, presumably because the concepts presented at the beginning were simpler and easier for them to understand. In between, clusters B and C represent more average students who differ mostly in terms of number of questions asked.

Therefore, to answer to RQ3 we have shown that although the clustering was performed exclusively on semantic features (*cf.* taxonomy in Table 1), it correlates with information relative to students' performance, attendance and questioning/voting behavior. Our work has some limits: we have applied it only to 2 courses (because a minimum number of questions is required) and we have not considered if it would be possible to classify students in clusters online or even if the same clusters could be found in the same courses on different years. Furthermore, not all questions could be automatically annotated, which reduced the dataset size and is particularly problematic for students who asked few questions. However, this work demonstrates the validity and the usefulness of our taxonomy, and shows the relevance of this classification to identify different students' profiles. It also suggests the taxonomy could be useful for our long-term goal which is to assist teachers in choosing questions to be explained in Q&A sessions. We also intend to apply this taxonomy to different datasets (*e.g.* questions asked in a MOOC) to see if it can also be useful in these contexts and if similar patterns appear.

5. REFERENCES

[1] A. C. Graesser and N. K. Person, "Question Asking During Tutoring," *Am. Educ. Res. J.*, vol. 31, no. 1, pp. 104–137, 1994.

[2] J. Sullins *et al.*, "Are You Asking the Right Questions: The Use of Animated Agents to Teach Learners to Become Better Question Askers.," in *FLAIR*, 2015, pp. 479–482.

[3] C. Chin and J. Osborne, "Students' questions: a potential resource for teaching and learning science," *Stud. Sci. Educ.*, vol. 44, no. 1, pp. 1–39, Mar. 2008.

[4] Q. Liu, W. Peng, F. Zhang, R. Hu, Y. Li, and W. Yan, "The Effectiveness of Blended Learning in Health Professions: Systematic Review and Meta-Analysis," *J. Med. Internet Res.*, vol. 18, no. 1, Jan. 2016.

[5] 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. Educ. Data Min.*, vol. 5, no. 1, pp. 104–146, May 2013.