Linking MOOC Courseware to Accommodate Diverse Learner Backgrounds

Shang-Wen Li and Victor Zue

MIT Computer Science and Artificial Intelligence Laboratory

{swli,zue}@mit.edu

Abstract

Massive Open Online Courses (MOOCs) brings great opportunities to millions of learners. However, the size of the learner population and the heterogeneity of the learners' backgrounds make conventional one-size-fits-all pedagogy insufficient. For example, learners lacking in prior knowledge may struggle with different concepts. In this paper, we propose a framework - educational content linking, to address the challenges. By linking and organizing scattered educational materials for a given MOOC into an easily accessible structure, this framework can provide guidance and recommendation of these contents, as well as improve navigation. Thus, learners can select appropriate supporting materials to suit their individualized needs and achieve self-exploring remediation. This paper describes an end-to-end case study, which found that learners, especially novices, can search learning materials faster without sacrificing accuracy, and can retain concepts more readily with our proposed approach. We have also obtained encouraging preliminary results that suggest that content linking can be achieved automatically using human language technology and stochastic modeling techniques.

Index Terms: MOOCs, learning at scale, automatic educational resource organization, hidden Markov model

1. Introduction

Since 2011, the revolution called MOOC (Massive Open Online Courses) has taken the world by storm [1, 2]. Today, there are thousands of courses, from science and engineering to humanities and law, being offered on the Internet using several platforms – Coursera, edX, etc. These platforms allow millions of learners around the world to take courses from top universities without the need for physical presence, thus potentially achieving the democratization of knowledge dissemination. At the same time, the openness of these platforms has also created a set of challenges. For example, the sheer size of the learner body, and the heterogeneity of their background – e.g., demographics, course preparedness, learning goals, motivation, etc., make it extremely difficult to meet everyone's learning needs [3, 4, 5].

Today's courseware on MOOC platforms typically consists of a myriad of high-quality materials that vary in type (e.g., lecture, slides, textbooks, discussion forum, problem sets), course level (e.g., college preparatory courses, graduate level courses), and pedagogy (e.g., active learning, mastery learning). These materials can potentially provide remediation for learners' heterogeneous needs. However, since these materials are conventionally made available to the students as disjoint entities, it is difficult to navigate them efficiently and find remediation. For example, a student interested in learning more about a specific topic described by the lecturer cannot easily look up relevant materials, such as from notes/slides to sections of the textbook, or from introductory materials to advanced ones, to broaden and reinforce his/her learning.

To address the challenges, we propose a framework - *educational content linking* [6]. By linking and organizing scattered educational materials for a given MOOC into an easily accessible structure, this framework can provide guidance and recommendation of these contents, as well as improving navigation. Thus, learners can potentially tailor the learning process to suit their background, and achieve self-exploring remediation for their heterogeneous learning needs (i.e., find appropriate supporting materials for their learning needs in a self-regulated way). To be more specific, one can imagine the linked courseware as a tree, in which the trunk corresponds to the curriculum that reflects the organization of concepts from instructors/experts, and the branches correspond to learning segments about the same topic but from various learning sources.

Fig. 1 below illustrates the two interfaces for navigating the course materials in our user studies ('*baseline*' *vs.* '*linked*'). These interfaces have an identical search module but different strategies for presenting retrieved result. The left panel of Fig. 1 illustrates how the search results are displayed to the user in '*baseline*' condition, i.e., the conventional way of delivering materials where each type of courseware is shown monolithically. By clicking the icon, the corresponding content will appear in a call-out box *independently*. In contrast, the right panel of Fig. 1 illustrates the '*linked*' interface. It is powered by our content linking result, which is described in the following. In this case, materials that are linked can be accessed together with one click, for learners to peruse at will.

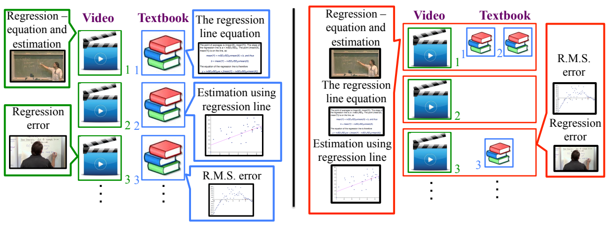

Figure 1: *An example of the 'baseline' and 'linked' interface used in our experiments.*

This paper is organized as follows. First, we summarize the results of a pilot study [6] comparing the two interfaces on "information search" tasks. We then substantially expand the previous experiment with more tasks and subjects, and extend the study to include a "concept retention" scenario for more evidence supporting our framework. In all the users studies, experts accomplished linking manually. With the assurance of the results that our framework is beneficial, we investigate an automatic linking method based on hidden Markov models

(HMM) to meet the scale issue. Lastly, we end this paper with a brief summary.

2. Pilot Study

Before building a system that can automatically link various course materials, we must first validate our hypothesis that linking educational materials *will* lead to better learning for the students. This section will briefly summarize some of our previously published findings, which will help set the stage for our expanded experiments.

2.1. The Course Material

We have chosen to focus our investigation on materials around a single MOOC – Stat2.1x: Introduction to Statistics, offered by UC Berkeley on edX in 2013 [6, 7]. This course comes with three types of courseware common to many MOOCs – lecture videos, slides, and (electronic) textbook. There are 31 lectures totaling 7 hours of video, and 157 pages of slides. The suggested textbook [8] contains 77 sections, providing independent support to the lecture material.

2.2. Methodology for Content Linking

To create the '*linked'* interface, we must first delineate contents in each type of material into segments. These segments are subsequently organized into a proper curriculum. Finally, relevant topic segments from each type of material must then be linked, as shown in Fig. 1.

Proper segmentation will result in vignettes that are large enough to be self-contained as a learning unit, yet small enough to enable learners to search/browse effectively across material types. We start by segmenting the textbook into sections, and slides into pages. Since there are no clear structural breaks in videos, we recruited two researchers with expertise in statistics to manually align the video transcription to the deck of slides from the same lecture. Thus, the videos are segmented into vignettes, where each vignette corresponds to one aligned page of slides.

Then, these segments are organized into a proper curriculum and linked. First, we concatenate the 31 lectures together, and take the sequence of slides as the shared curriculum. The aligned video vignettes are linked to the slides accordingly. For each page of slides, the same two researchers also label the most relevant section in the textbook, and link the segment to the slide page. If there is a disagreement, the two researchers have to discuss until consensus is reached. With these steps, we can obtain a shared curriculum, and link the separate materials around the curriculum.

2.3. Pilot Experiment

The goal of our pilot experiment [6] is to provide early indication of whether linked content would lead to better learning. However, learning is a combination of mental processes such as attention, memory, problem solving, thinking, etc., it may be too elusive to ascertain in one set of experiments. Similar to [10], we thus adopted a specific set of learning-relevant activities – educational content navigation, as a proxy for learning. In this pilot experiment, we measured the subjects' performance on the task of 'information search', in which a learner in our experiment is given a question, and asked to retrieve a learning segment (in videos, slides, or Textbook) that can be used to solve the given question. This scenario attempts to emulate a situation where a learner is trying to review educational content and searching for useful information for problem solving.

For this pilot experiment, four questions (similar to the ones shown in Table 1) are sampled from the problem set in Stat2.1x, and 100 unique online workers on Amazon Mechanical Turk (AMT) are recruited for each question, resulting in 400 tasks. A total of 151 unique AMT workers participate in the experiment, instead of 400, since we allow a given worker to solve more than one task. The workers differ in their background – education level, exposure to MOOC, and familiarity with statistics. Such diversity allows us to understand the usefulness of our proposed model to a heterogeneous learner body. We randomly assign a worker in each task with either of the two interfaces shown in Fig. 1, and we measure the worker's performance in task completion time and the accuracy of the retrieved segment. By analyzing the difference in performance using each interface, we investigate whether our model benefits learners in navigating across educational contents.

Table 1. *Example tasks posed to the AMT workers.*

Instructions – "select a learning segment (a textbook section, a video chunk, or a page of slides) that helps you solve a given problem."	
Task 1	What is the formal definition for X^{th} percentile, where X is a general, real number between 0 and 100?
Task 2	Based on the given data, please plot a histogram of the distribution.

Table 2. *Learner performance on 'information search' tasks using 'baseline' or 'linked' interface. We measure the performance by computing task completion time and accuracy.*

Learner background	Time consumed			Task accuracy		
	Seconds		P-value	% of correct tasks (# tasks)		P-value
	Baseline	Linked		Baseline	Linked	
$\geq$ Bachelor	306	284	0.16	52 (96)	65 (98)	0.03
$\leq$ Some college	322	257	**< 0.01**	66 (98)	55 (96)	0.94
MOOCs	277	286	0.63	70 (40)	63 (34)	0.74
No MOOCs	323	267	**< 0.01**	55 (154)	59 (160)	0.21
Statistics	294	268	0.10	60 (120)	61 (120)	0.45
No Statistics	346	276	0.01	57 (74)	58 (74)	0.44
Overall	315	271	**< 0.01**	58 (194)	60 (194)	0.38

2.4. The Results

Table 2 summarizes the learner performance in these tasks. The average task completion time and accuracy along with the significance test results are shown. We highlight, in boldface, the result where the differences are significant at P=0.01 level.

Focusing first on the last row of Table 2, we see that the overall performance suggests that the averaged search time using the linked interface is 14% less than using the baseline interface (cf. 315 *vs.* 271), and this improvement is statistically significant. In contrast, there is no significant difference in task accuracy for using the two interfaces. Looking over the top six rows of Table 2 for the individual results of the three demographic groups, we observe that the linked interface reduces search time in five of the six cases. The difference is statistically significant in two out of the three groups of novice learners (i.e., learners who are less educated and less experienced with MOOC), with the last one barely missing it (i.e., learners who are less familiar with the subject materials). Similarly, we compare the difference in the retrieval correctness, and no statistically significant degradation is observed. Thus, we conclude that, by having the educational content linked, we allow learners, especially novices, to find supporting learning segments more efficiently for solving problems, without sacrificing the searching correctness. This fact shows that our framework can potentially benefit educational content navigation.

Our results indicate that linking has little impact on task accuracy in most cases. This could be due to the fact that the difference between the two interfaces is about *how* the materials are presented, rather than the information itself. Therefore, learners can always find the correct learning segments with sufficient time and persistence.

3. New Experiments and Results

While the results of the pilot study were tantalizing, we were concerned about several shortcomings. We only measured one aspect of learning – the speed and accuracy of information search. As such, the number of tasks was relatively small. Allowing some workers to perform more tasks than others may have skewed the results. Therefore we expanded our experiment substantially in several dimensions.

3.1. Information Search Experiments

In this paper, we expand the previous 'information search' experiment and conduct user study on a larger scale. With more data, we expect to provide stronger evidence to show the benefit of linking. We increase the number of questions from four to ten, and recruit 200 *unique* AMT workers for each of the questions for a total of 2,000 tasks. In Table 3, we show the number of tasks completed by subjects with various backgrounds. In all, 497 distinct workers participate in the experiment. The experimental protocols remain the same as before.

Table 3. *Number of tasks completed by each subject group.*

Learner background	Number of tasks	
	Baseline	Linked
≥ Bachelor	573	522
≤ Some college	427	478
MOOCs	295	249
No MOOCs	705	751
Statistics	714	704
No Statistics	286	296
Overall	1,000	1,000

Table 4. *Learner performance on the expanded 'information search' tasks using 'baseline' or 'linked' interface.*

Learner background	Time consumed			Task accuracy		
	Seconds		P-value	% of correct tasks		P-value
	Baseline	Linked		Baseline	Linked	
≥ Bachelor	198	163	**< 0.01**	70.7	70.6	0.98
≤ Some college	208	136	**< 0.01**	67.5	68.5	0.79
MOOCs	166	139	0.06	72.0	70.6	0.76
No MOOCs	225	154	**< 0.01**	68.2	68.9	0.80
Statistics	166	147	0.05	71.1	70.5	0.81
No Statistics	295	160	**< 0.01**	64.9	67.1	0.64
Overall	206	152	**< 0.01**	69.2	69.5	0.89

Learner performance on the expanded 'information search' experiment is shown in Table 4. The average amount of time workers spend on completing the tasks with the '*baseline*' and '*linked*' interfaces is summarized in the first two columns of the table, respectively, and the significance test result of the time difference between the two interfaces is listed in the third column. Columns 4 and 5 tabulate the percentage of tasks where the retrieved segments are correct for the two interfaces, respectively. Column 6 shows the significance test results.

Compared to the previous results in Table 2, we observe in Table 4 a stronger evidence of the usefulness of linking. The overall search time is reduced by 36% (cf. 206 *vs*. 152), and this trend holds for all sub-categories of workers. The difference is statistically significant for five of the seven groups, including all three novice groups. As for task accuracy, the '*linked*' interface yields some improvement (ranging from +0.7% to +2.2%) in the three novice groups, and a smaller difference (ranging from -0.1% to -1.4%) in the experienced counterparts. Similar to the results in our previous study, none of the differences are statistically significant.

Thus, we reach the same conclusion as before, except with greater confidence that learners can search desired information more efficiently without sacrificing the accuracy when learning materials are linked. Since our experiments are conducted remotely, it is inevitable that studies based on crowdsourcing have to deal with outliers and spammers in the collected data. By increasing the scale of user study, we can mitigate the noise from spammers and achieve more reliable conclusions.

3.2. Concept Retention Experiments

With similar experimental setup [6], we also design another set of experiments – 'concept retention,' to explore whether the '*linked*' interface can benefit learning from a different aspect. In this scenario, we attempt to measure how efficiently a learner could peruse the materials to capture, memorize, and recall key concepts. Specifically, each subject is first assigned a topic. Then, the subject has ten minutes to learn about the topic with either of the two interfaces shown in Fig. 1. After the learning session, he/she is asked to write a short essay to summarize what he/she learned about the topic. We then evaluate the learner performance in this scenario by counting the number of unique key concepts mentioned in the essay, where key concepts are extracted from the textbook glossary and defined as the entries belonging to the corresponding topic.

For this scenario, we also sample ten topics, and recruit 200 unique AMT workers for each topic. Several examples are shown in Table 5. A total of 751 different AMT workers are recruited.

Table 5. *Example tasks for the 'concept retention' scenario and the task hint given to experimental subjects.*

Instruction – "learn the given topic based on the content you can find in the interface, and write an essay to summarize what you have learned and remembered about the topic"			
Topic 1	Regression	Topic 2	Standard deviation
Topic 3	Correlation	Topic 4	Mean

Table 6 summarizes learner performance on the 'concept retention' scenario. The first two columns of the table contain the average number of unique concepts in the subjects' essay for the two interfaces, respectively, for each subject group. We also list the P-values in column 3, and the number of tasks for each group of subjects in columns 4 and 5.

Table 6. *Learner performance on 'concept retention' tasks using 'baseline' or 'linked' interface.*

Learner background	# unique key concepts			# tasks	
	Baseline	Linked	P-value	Baseline	Linked
≥ Bachelor	4.73	5.23	**< 0.01**	549	519
≤ Some college	3.98	4.60	**< 0.01**	451	481
MOOCs	4.83	5.14	0.27	205	287
No MOOCs	4.27	4.77	**< 0.01**	795	713
Statistics	4.71	5.11	0.02	594	597
No Statistics	3.98	4.60	**< 0.01**	406	403
Overall	4.39	4.91	**< 0.01**	1,000	1,000

Focusing first on the last row of Table 6, we see that, overall, subjects are able to produce a greater number (~12%) of key concepts while using the '*linked*' interface (cf. 4.39 vs. 4.91), and the difference is statistically significant. Looking over the top six rows of Table 6, we observe that there is a similar trend to that in the 'information search' scenario, where the '*linked*' interface yields consistent improvement over each group of subjects, and the novice learners benefit more than their experienced counterparts. In four of the six cases (including all the three novice groups), the improvement passes the statistical significance test. These findings reveal another benefit of linking in navigating content and learning.

From the results in the two sets of experiments, we notice that the '*linked*' interface yields better performance, especially for novice subjects. This fact is perhaps not surprising. Because of the shortcomings of these subjects - less education, less experience with MOOC, and less familiarity with the subject matter, they may lack a broad perspective to explore the various resources on their own. By organizing the learning materials in a '*linked*' interface, which is easy to visualize and manipulate, we can potentially enhance their ability to navigate through the knowledge space more effectively, which could lead to improved knowledge acquisition. This is consistent with previous study that shows "guidance" is particularly crucial for learners who are likely to struggle [11].

In conclusion, with *educational content linking*, learners can find supporting learning segments faster with no degradation in searching accuracy. They can also review materials and capture concepts in a topic more efficiently. Furthermore, the improvement from linking is more significant in novices. This fact shows the potential of our linking framework in reducing the knowledge gap among the heterogeneous learners in MOOCs. We interpret these findings as evidence that the proposed framework benefits learners in navigating content and exploring remediation.

4. Automatic Linking Using HLT

In this section, we investigate methods to link courseware automatically. Due to the heterogeneous learner body, students have various prior knowledge and learning needs (e.g., they can struggle for a myriad of different reasons and require various remediation). We show linking can potentially help learners navigate course materials in the previous section. However, it is cost-prohibitive and not scalable to manually link all available course contents for covering every possible learning need for remediation (e.g., link the tens of thousands of forum discussions). Thus, we propose a human language technology (HLT) based method to generate linking automatically and at scale. HLT is a major focus here, since human language is an integral part in education for knowledge transferring, and HLT has been proved successful in many applications of information retrieval [12-17]. Thus, we believe methods based on HLT will be more generalizable to different courses, and more likely to provide high-quality automatic linking.

4.1. Hidden Markov Model

We employ HMM to link various types of course materials automatically. HMM is a special case of graphical model [18]. Conventional information retrieval methods in HLT, e.g., cosine similarity, infer the relation among a repository of documents based on lexical cues of the content [12]. As compared to that, a graphical model can additionally express the ontology and global structure behind the repository. This characteristic allows us to understand the curriculum and extract global information for more accurate linking prediction among learning materials. Thus, we adopt HMM to model the sequential structure of the curriculum.

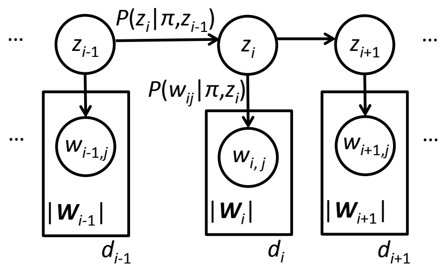

Figure 2: *The graphical representation of HMM.*

Fig. 3 is an illustration of HMMs. First, we denote an ordered sequence of learning segments extracted from a type of course materials (e.g., a deck of slides) as $\boldsymbol{D} = \{d_i\}$, where $i = \{1, 2, ..., |\boldsymbol{D}|\}$. For this sequence, we assume there is a corresponding sequence of hidden state variables, $\boldsymbol{Z} = \{z_i\}$, mapped to these segments. Given a list of hidden states, the value of variable z_i is the list index of the hidden state generating d_i. A hidden state s can be interpreted as the red rectangle in the right panel of Fig. 1. If z_i = s, we link d_i (can be interpreted as the video/textbook icons in the red rectangle) to the state. Therefore, to generate the linking automatically, all we need is to infer the values of the hidden variable sequence $\boldsymbol{Z}$, based on the observation of segments $\boldsymbol{D}$.

We model the inference problem with maximization of the probability $P(\boldsymbol{D}, \boldsymbol{Z} | \pi)$, where the parameter set π is learned from a training corpus. To expand the probability, we first notice that, with the first-order Markov assumption made in HMM, each hidden variable z_i depends only on the hidden variable for its preceding segments d_{i-1}, and the observation of d_i is independent of other segments given z_i. In addition, we represent each observed segment d_i with its word occurrence $\boldsymbol{W}_i = \{w_{ij}\}$, where $j = \{1, 2, ..., |\boldsymbol{W}_i|\}$, and w_{ij} is the word index of the j^{th} word in d_i. With this representation, $P(d_i | \pi, z_i)$ can be expressed as $\prod_{j=1}^{|\boldsymbol{W}_i|} P(w_{ij}|\pi, z_i)$. Thus, we can rewrite the probability $P(\boldsymbol{D}, \boldsymbol{Z} | \pi)$ to:

$$P(z_1|\pi) \prod_{j=1}^{|\boldsymbol{W}_1|} P(w_{1j}|\pi, z_1) \prod_{i=2}^{|\boldsymbol{D}|} \prod_{j=1}^{|\boldsymbol{W}_i|} P(w_{ij}|\pi, z_i) P(z_i|\pi, z_{i-1}) \quad (1).$$

The two types of dependence, $P(z_i | \pi, z_{i-1})$ and $P(w_{ij} | \pi, z_i)$, are used to model the sequential structure of the curriculum, and the lexical information in each hidden state variable (e.g., the word distribution) respectively.

To solve the inference problem and predict linking with HMM, we first have to train the model to find the parameter setting, $\hat{\pi}$, which maximizes the probability $P(\boldsymbol{D}, \boldsymbol{Z} | \pi)$ over all possible settings π on the training corpus $\boldsymbol{D}$. The EM algorithm is adopted for estimating $\hat{\pi}$. Then, we predict linking with the learned setting $\hat{\pi}$. For a sequence of testing segments $\boldsymbol{D'}$, we find the value assignment $\hat{Z}$, which maximizes the probability $P(\boldsymbol{D'}, \boldsymbol{Z} | \hat{\pi})$ over $\boldsymbol{Z}$. With $\hat{Z}$, we can link each segment d_i' to the hidden state according to the value $\hat{z_i}$, and thus to the corresponding part in the curriculum.

4.2. Experimental Results and Findings

We then evaluate our automatic linking generation method on Stat2.1x. Two linking tasks are studied - video-to-slide linking and video-to-textbook linking. In the video-to-slide task, we adopt the 157 pages of slides as the training set due to the clear structural breaks in slides, and select the seven-hour video transcription as the test set. Then, we define a hidden state as a page of slides. We train the HMM parameter setting with the lexical information (e.g., the word occurrence counts in each page of slides) and the material structure (e.g., $P(z_i = s_i | \pi, z_{i-1} = s_{i-1}) = 1$ if s_i and s_{i-1} are adjacent pages of slides, otherwise the probability is 0) in the training set. With the learned parameter setting, we predict the hidden state for every sentence in the test set. Each sentence is then linked to the page of slides represented by the predicted states. As for the video-to-textbook task, we conduct an experiment with similar design, except that we replace the slides with the 77 textbook sections as the training set and define a section as a hidden state. These tasks allow us to study our model performance when the materials to be linked are matched (i.e., video and slides) or mismatched (i.e., video and textbook).

Table 7. *The sentence accuracy (%) of the predicted linking from video transcription to slides or textbook.*

	Video-to-slide Linking		Video-to-textbook Linking	
Models \ Features	Word frequency	TFIDF	Word frequency	TFIDF
Cosine similarity	73.3	75.5	19.1	25.7
HMM	80.6	84.1	31.8	33.0

We take the expert-labeled linking described in section 2.2 as ground truth for evaluation, and compute two performance metrics for the two tasks respectively - the percentage of sentences that are linked to the correct 1) page of slides, or 2) textbook section. Table 7 summarizes the results. As a reference, we also implement a baseline method – cosine similarity, in which we link each sentence to the page/section with the most similar bag-of-word representation measured by cosine similarity.

In the video-to-slide tasks (first two columns), with the additional information from material structure, our method yields a 7.3% absolute linking accuracy improvement over the baseline method. After normalizing the times a word appears in a learning segment with the frequency of the word in the corpus (i.e., TF-IDF, or term frequency-inverse document frequency), we obtain a feature that can better discriminate between function and topic words. The TF-IDF feature further improves the accuracy by 3.5%. As for the video-to-textbook tasks (last two columns), similar trends can be observed - our method yields a 12.7% improvement over the baseline with additional modeling of the structure information; by using a more discriminating feature (the TF-IDF), we further improve the performance by 1.2%. However, the accuracy here is significantly lower than that of the video-to-slide task with comparable experimental settings.

In summary, by modeling structure and lexical information simultaneously, our HMM-based method yields a significant improvement over the baseline in generating linking automatically. The performance can be further improved with the TF-IDF feature. We believe that, with refined models and features, the proposed method is likely to achieve comparable performance to manual linking. Thus, this method is a promising solution to link MOOC contents automatically and support learners finding remediation at scale.

5. Summary and Future Work

This paper describes a continuation of our effort to provide students with diverse background the ability to enhance their learning through a '*linked*' interface. We extend our previous study [6] and provide more evidence to validate the benefit of the proposed framework in learning. With the assurance of the results, an automatic linking method based on HMM is further investigated for our framework to scale well at MOOC setting.

Our results suggest that learners, especially novices, can be more efficient in reviewing materials and capturing concepts in a topic with the '*linked*' interface. Combined with our previous findings in [6], these results provide evidence that the proposed framework is beneficial in educational content navigation. Thus, our framework can potentially help learners find materials for remediating confusion or broadening their learning. We believe our linking framework is well suited to MOOC, in which there is a high demand for providing multiple alternatives of materials in order to accommodate the diverse background of learners. It is the novice learners who will need the most help and who stand the most to benefit [11].

Furthermore, we observe that the proposed HMM method outperforms the baseline in generating linking automatically. Structure information and more discriminating features are two key factors yielding the improvement. This encouraging result suggests that linking can be achieved at scale with such automatic methodology.

Future work for our research will follow several directions. First, we plan to refine our experimental procedure and expand our repertoire of learning tasks. Similar experiments will be conducted with various educational materials/modalities (e.g., speech, text, and video) and on other MOOCs, to further validate our findings and investigate the generalizability. Second, we will explore advanced features (e.g., click-through information) and models to refine our automatic linking generation. We will strive to replace human with a machine in linking, and help learners in MOOCs find remediation at scale.

6. Acknowledgements

The authors would like to thank Hung-Yi Lee and Chengjie Sun for insightful discussions and assistance in developing the interface. The work is sponsored by Quanta Computer, Inc. under the Qmulus Project.

7. References

[1] L. Breslow, D. E. Pritchard, J. DeBoer, G. S. Stump, A. D. Ho, and D. T. Seaton, "Studying learning in the worldwide classroom research into edX's first MOOC," *Research and Practice in Assessment*, vol. 8, pp. 13 – 25, 2013.

[2] L. Pappano, "The year of the MOOC," *The New York Times*, 2012.

[3] J. DeBoer, A. D. Ho, G. S. Stump, and L. Breslow, Changing "course": Reconceptualizing educational variables for massive open online courses. *Educational Researcher*, 2014.

[4] R. F. Kizilcec, C. Piech, and E. Schneider, "Deconstructing disengagement: analyzing learner subpopulations in massive open online courses," in *Learning Analytics and Knowledge*, 2013.

[5] D. T. Seaton, Y. Bergner, I. Chuang, P. Mitros, and D. E. Pritchard, "Who does what in a massive open online course?" *Communications of the ACM*, vol. 57 No. 4, pp 58-65, 2014.

[6] S.-W. Li, and V. Zue, "Would linked MOOC courseware enhance information search?" *IEEE Conf. on Adv. Learning Technologies* (accepted 2015).

[7] Introduction to Statistics: Descriptive Statistics "https://www.edx.org/course/introduction-statistics-descriptive-uc-berkeleyx-stat2-1x".

[8] SticiGui, "http://www.stat.berkeley.edu/~stark/SticiGui/index.htm".

[9] Amazon Mechanical Turk, "https://www.mturk.com".

[10] J. Janssen, C. Tattersall, W. Waterink, B. van den Berg, R. van Es, C. Bolman, and R. Koper, "Self-organising navigational support in lifelong learning: How predecessors can lead the way," *Computers and Education*, 49(3), pp. 781-793, 2007.

[11] P. A. Kirschner, J. Sweller, and R. E. Clark, "Why minimal guidance during instruction does not work: an analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching," *Educational Psychologist*, 41(2), 75-86, 2006.

[12] D. Shahaf, C. Guestrin, and E. Horvitz, "Trains of thought: Generating information maps," in *World Wide Web*, 2012.

[13] R. Rose, A. Norouzian, A. Reddy, A. Coy, V. Gupta, and M. Karafiat, "Subword-based spoken term detection in audio course lectures," in *ICASSP*, 2010.

[14] Y. Fujii, K. Yamamoto, N. Kitaoka, and S. Nakagawa, "Class lecture summarization taking into account consecutiveness of important sentences," in *Proc. INTERSPEECH*, 2008.

[15] I. Szoke, J. Cernocky, M. Fapso, and J. Zizka, "Speech@FIT lecture browser," in *Spoken Language Technology Workshop (SLT), 2010 IEEE*, 2010.

[16] J. Glass, T. J. Hazen, D. S. Cyphers, I. Malioutov, D. Huynh, and R. Barzilay, "Recent progress in the MIT spoken lecture processing project," in *Proc. INTERSPEECH*, 2007.

[17] K. Riedhammer, M. Gropp, and E. Noth, "The FAU video lecture browser system," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*, 2012.

[18] P. Smyth, D. Heckerman, and M. Jordan, "Probabilistic independence networks for hidden markov probability models," *Neural Computation*, vol. 9, pp. 227–269, 1996.