Automatic Generation and Insertion of Assessment Items in Online Video Courses

1st Author Name
Affiliation
Address
e-mail address
Optional phone number

2nd Author Name
Affiliation
Address
e-mail address
Optional phone number

3rd Author Name
Affiliation
Address
e-mail address
Optional phone number

ABSTRACT

In this paper, we propose a prototype system for automatic generation and insertion of assessment items in online video courses. The proposed system analyzes text transcript of a requested video lecture to suggest self-assessment items in runtime through automatic discourse segmentation and question generation. To deal with the problem of question genetation from noisy transcription, the system relies on semantically similar Wikipedia text segments. We base our study on a popular video lecture portal, namely, National Programme on Technology Enhanced Learning (NPTEL). However, it can be adapted to other portals as well.

Author Keywords

Online video courses, MOOCs, Question generation, Latent Semantic Analysis, Rhetorical Structure Theory

INTRODUCTION

Open online video course portals like MIT OpenCourseWare[1] and NPTEL[2] and others have gained enormous popularity as video lectures of high educational quality can be accessed from anywhere and anytime. This revolution in online education has further been boosted by the rise of Massive Online Open Courses (MOOCs). New generation MOOCs provide platforms for better interactivity through Web 2.0 to facilitate learner assessment, discussion forums and others. Even with this enhanced interactivity, current online educational video portals face following challenges:

- *Limited attention span:* Educational psychology and cognitive studies have provided evidences of attention lapse in learners in long lectures. According to different studies [5][2], a typical attention span varies between 10 to 15 minutes. Thus long video lectures may fail to keep students attentive. In traditional classroom setting, the lapse in student's attention is handled through various activities called "Change Ups" [12]. However, these "Change Ups" are not embedded in online video lectures.
- *Limited self-assessment:* MOOCs started including assessment items into video courses and they are presented in schedules prescribed by the course developers. However, a student may want to test his/her understanding at any point of time through self-assessment. Current video portals lack on-demand self-assessment.

One way to address the first challenge is to break a long video lectures into shorter 10-15 minute lectures and inserting "Change Ups" in between. But, adopting this strategy to existing video courses will spawn following issues:

- Shorter lecture video sequences will exhibit discourse level discontinuity with high possibility.
- Manual insertion of "Change Ups" demands huge effort from the course instructor.

Solving the second problem with human effort (where human instructor creates assessment items for every student request) is infeasible as a student's self-assessment request is unpredictable.

In this paper, we present a prototype system for automatic generation and insertion of "Change Ups" in the form of assessment items to address above mentioned challenges. The proposed system overcome the limitations by incorporating the following features:

- *Marking assessment timing:* The system marks temporal points in video to be the candidates for insertion of assessment items.
- *Assessment item generation:* Assessment items for a temporal point in video lecture is generated automatically.
- *On-demand self-assessment:* Upon student's request at anytime in video lecture session, the system presents assessment items.

Challenges in developing such a system are marking discourse boundaries in video lecture, automatic generation of question-answer pair and presenting the assessment items in video lecture interface. The originality of the proposed system are as follows:

- Text transcription corresponding to a video lecture is analyzed to automatically mark the discourse boundaries indicative of topical shifts.
- Text transcriptions are generally noisy due to several reasons. Firstly, transcriptions are conversational as observed in classroom discourse. Thus actual content is buried under

[1] `http://ocw.mit.edu/index.htm`
[2] `http://nptel.ac.in`

significant amount conversational cues. Secondly, for lectures that are not manually transcribed, automatic speech recognizers (ASRs) are used to generate transcriptions. As the ASRs are generally not perfect, the transcribed texts tend contain garbage texts. This work proposes to employ external knowledge sources (e.g., Wikipedia) to deal with this issue.

- The system proposes technique to automatically generate assessment items from text corresponding to a video lecture segment. The current work extends the state-of-art question generation research targeting assessment items capable of testing higher order skills like inferential, procedural etc.

- The proposed system adapts video lecture delivery interface to present the assessment items and provide feedback to students. This is achieved through APIs (e.g., YouTube API) exposed by portals hosting the video lectures.

SYSTEM OVERVIEW

Overview of the proposed system is presented in Figure 1. Text transcription corresponding to a NPTEL video lecture hosted in YouTube[3] is extracted using YouTube API. A long video lecture generally discusses about several topics and an instructor distributes a lecture duration among them. Topical changes are marked by discourse boundaries. Through *Text Segmentation* module, several discourse segments are generated from the text transcript with each segment signifying a discourse distinct from neighboring segments. In order to generate items relevant to a discourse segment, semantically similar Wikipedia text segments are retrieved through *Segment Similarity* module. The system relies on Wikipedia segments as they present more well-formed and formal discourse than transcription segments. The *Item Generator* module extracts discourse relations inspired by Rhetorical Structure Theory (RST) [11] from Wikipedia segment. Question-answer pairs are generated by applying relation specific rules to text span covering the relation. The generated assessment items are then inserted in marked discourse boundaries in video lecture.

SYSTEM COMPONENTS

In this section, we provide technical details of important modules of the proposed system.

Topic Boundary Detection

Discourse boundaries marking the changes of topic are potential places for inserting assessment items. Discourse boundaries are detected in text transcription using TextTiling algorithm [7]. Transcribed text is first tokenized, stop words are removed and the remaining tokens are stemmed. Pseudo sentences are formed by grouping 20 consecutive stemmed words. For each gap between pseudo sentences *lexical cohesion score* signifying the average similarity between words in pseudo sentences before and after the gap is computed. The cohesion scores for all the gaps if plotted graphically will show peaks and valleys. *Depth score* at each gap is the distance from a valley to two peaks corresponding to the similarity valley. The gaps having depth scores greater than a threshold are marked as topic boundaries. This step outputs a sequence $T_S = \{T_1, T_2, \ldots T_n\}$ of transcript segments.

[3] `https://www.youtube.com/user/nptelhrd/channels?view=60`

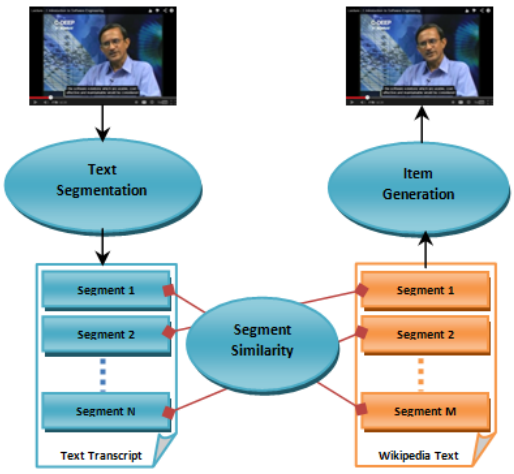

Figure 1. Overview of system components

Wikipedia Segment Retrieval

As mentioned earlier, text transcriptions tend to be noisy and generating valid questions from them is not feasible. To handle this issue, we rely on Wikipedia articles that are semantically similar to transcriptions. Wikipedia segment semantically similar to a text transcription segment (T_i) is retrieved through following processing steps:

1. *Concept linking*: At the first step, it is required to decide the set of Wikipedia articles that are relevant to an input text transcription segment. This is achieved through linking concepts from a text transcription to Wikipedia articles using WikipediaMiner [13] web service. Output of this step is a set of $\langle C_j, URL_j \rangle$ pairs where C_j is the j^{th} concept in text transcription and URL_j is the linked Wikipedia article.

2. *Wikipedia article segmentation*: Text content of sections of each article (URL_j) from the article set corresponding to T_i are retrieved. This step generates $W = \{W_1, W_2, \ldots, W_m\}$ Wikipedia segments corresponding to T_i.

3. *Similarity computation*: From a set of Wikipedia segments generated for a text transcription in the previous step, the task here is to find the most similar Wikipedia segment. Traditional keyword-based similarity measures like TF-IDF (term frequency-inverse document frequency) fail to judge similarity of text spans in semantic space. Distributional semantics based models are potential candidates for computing semantic similarity and we have adopted Latent Semantic Analysis (LSA) [9] based model for this purpose. First, an LSA model is trained using a corpus consisting of all the text transcript segments in NPTEL video lectures and corresponding Wikipedia segments. Similarity score

is computed for each text pair $\langle T_i, W_k \rangle$ by converting corresponding texts into LSA vectors ($\vec{T_i}$ and $\vec{W_k}$) followed by cosine-based similarity score computation as given in Equation 1.

$$similarity(\vec{T_i}, \vec{W_k}) = cos(\theta) = \frac{\vec{T_i}.\vec{W_k}}{\|\vec{T_i}\|\|\vec{W_k}\|} \quad (1)$$

Wikipedia segment having the highest similarity score is retrieved following Equation 2.

$$W^* = \underset{W_k}{\operatorname{argmax}}\, similarity(\vec{T_i}, \vec{W_k}) \quad (2)$$

Assessment Item Generation

Assessment item is defined as a complex object consisting of a question, model answer, feedback strategy, score, completion time etc. As the assessment items are primarily used as "Change Ups" in the proposed system, we restrict to the following definition of assessment items:

Assessment Item *An assessment item is a tuple* $AI = \langle Q, A, F \rangle$*, where* Q *is a question,* A *is model answer and* F *is a feedback strategy. For MCQs,* F *takes value from* $\{correct, incorrect\}$ *and for text-based answer it takes a value from* $\mathbb{R}$*, representing similarity of learner's answer with* A*.*

Question Generation and Answer Extraction

Retrieved Wikipedia segment (W^*) corresponding to input transcription segment (T_i) is analyzed to generate questions that demands varying cognitive abilities in the answerer's part. For example, factual questions test 'recall' skill whereas explanation type questions test 'inference' ability. We focus on generating questions of two modalities: Multiple Choice Questions (MCQs) and Short Answer Questions (SAQs).

Factual question generation works in sentence level. Following the work by Heilman [8], simplified factual sentences are extracted by performing different operations on an input sentence. Operations include removing appositive phrases, relative clauses and others. This operations are performed over syntactic parse trees with help of two tree query and manipulation tools: *Tregex* and *Tsurgeon* [10]. Questions from the simplified sentences were generated by marking an answer phrase, moving the answer phrase to appropriate position and replacing it with relevant WH-word.

This approach towards question generation though effective for factual or yes/no questions, is not readily applicable in generating questions of complex type (for question category see [6]). Apart from factual questions, we intend to generate questions belonging to categories like *causal antecedent, causal consequence* and *explanation*. For the mentioned question categories, the question-answer pairs span multiple sentences as contrast to factual questions where question-answer cues reside in same sentence. Thus generation of non-factual questions demands text to be treated as span of multiple sentences bound together with inter-sentential relations. Rhetorical Structure Theory (RST) [11] provides a formal way of representing inter-sentential bonding with a set of rhetorical relations. These relations in general designate one sentence or clause as Nucleus (N) and other as Satellite (S) where nucleus conveys the primary message and satellite provides additional information about nucleus. There are several rhetorical relations reported in literature. However, we restrict to a subset: *evidence, justify, condition, explanation, manner-means, Non-volitional cause, non-volitional result, purpose, volitional cause, volitional result*.

RST style discourse parser [4][4] is used to mark discourse relations between different non-overlapping text spans in W^*. The generated discourse parse tree is then traversed to extract discourse relations and corresponding text spans. We adopt a rule-based approach towards generation of questions by defining relation specific rules represented through relation template (Table 1).

Relation Name	Name of the relation	
Nucleus	Nucleus span of the relation (N)	
Satellite	Satellite span of the relation (S)	
AI Block 1	**A1.Question_Span**	Either N or S
	A1.Answer_Span	Either N or S
	A1.WH_Phrase	WH-phrase to be used in question
...	...	...
AI Block N	...	...

Table 1. Relation template for question generation (AI=Assessment Item).

In a relation template, *nucleus* and *satellite* slots are filled with respective text spans for the relation. The template contains a set of *Assessment Item Blocks*. Each block contains cues for generating questions semantically different from that generated from other blocks. `Ai.Question_Span` slot represents the text span from which question will be generated and is filled with either nucleus or satellite. `Ai.Answer_Span` stores text span that may contain answer. `Ai.WH_Phrase` stores a set of Wh-phrases that will be used in surface representations of syntactically different but semantically similar questions. Instantiations of template for different relations are presented in Table 2[5]. After selecting question span and WH-phrase, syntactic parse tree corresponding to satellite or nucleus is manipulated using Tsurgeon to insert WH-phrases in proper place in the tree. Question sentence is then generated from the modified parse tree. Generation of answer is done through modification of parse tree corresponding to the answer span with removal of cue phrases indicative of discourse relations.

MCQ Generation

Challenges in MCQ generation include identifying the key, generating stem and effective distractors. In this work, we have focused on generating MCQs for factual questions. Generation of stem and key are done through factual question generation described earlier. There have been several approaches towards automatic MCQ generation [14][17]. We adopt ontology-based strategy [15][1] towards distractor generation. As video lectures may belong to varying disciplines, developing ontology for all is not feasible. We rely on Wikipedia category taxonomy as a replacement for ontology

[4] `http://www.cs.toronto.edu/~weifeng/software.html`

[5] Due to space limitation instantiations for all the relation have not been provided

Relation	Cue-phrase	Q-Span	A-span	Example
Manner-means	by	Why<N>	S	[Abstraction is a conceptual process]$_N$ [by which general rules and concepts are derived from the usage and classification of specific examples, literal ("real" or "concrete") signifiers, first principles or other methods]$_S$
				Question: *Why is abstraction a conceptual process?*
				Answer: General rules and concepts are derived from the usage and classification of specific examples, literal ("real" or "concrete") signifiers, first principles or other methods
Explanation	therefore	Why<S>	N	[All risks can never be fully avoided or mitigated simply because of financial and practical limitations.]$_N$ [Therefore all organizations have to accept some level of residual risks.]$_S$
				Question: W*hy all organizations have to accept some level of residual risks?*
				Answer: All risks can never be fully avoided or mitigated simply because of financial and practical limitations.
Condition	if	What happens if <S>	N	[A logical ER model does not require a conceptual ER model]$_N$ [especially if the scope of the logical ER model includes only the development of a distinct information system.]$_S$
				Question: *What happens if the scope of the logical ER model includes only the development of a distinct information system?*
				Answer: A logical ER model does not require a conceptual ER model.

Table 2. Instantiations of relation template for different RST relations.

sacrificing the availability of finer ontological relationships. As the key is a domain concept, it will have a Wikipedia entry. First Wikipedia category for the article corresponding to the key is determined. The siblings of the extracted category in Wikipedia taxonomy are used as distactors. Generated MCQ for a factual sentence "bounds-checking elimination is a compiler optimization" is presented in Figure 2.

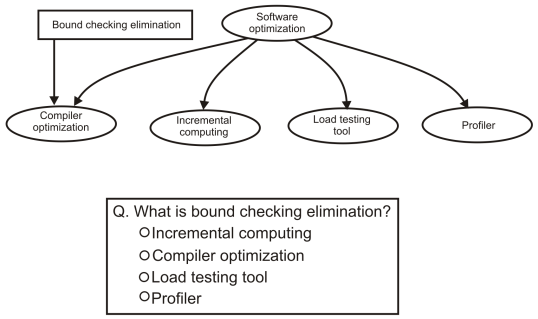

Figure 2. MCQ distractor generation from Wikipedia taxonomy.

Feedback Design

For MCQs, the system provides binary feedback (e.g., correct or incorrect) depending on learner's response. For other questions, the system expects the learner to input text-based answers. As the video lectures are used for self-learning purpose, formative assessment is more relevant than summative assessment. Accordingly, our system provides qualitative feedback to the learner ignoring grading aspect.

The qualitative feedback is generated by computing semantic similarity between learner's answer and automatically extracted model answer. We make use of two types of similarity measures: Wordnet-based [3] and LSA-based. The similarity scores are computed through SEMILAR toolkit [16] with LSA trained on Wikipedia corpus. Weighted average of similarity values obtained from two measures is used to assign final similarity score. Depending on the range of similarity score, qualitative feedback value is assigned with either of high similarity, medium similarity or low similarity.

USER INTERFACE

The back-end subsystem analyzes text transcript to generate assessment items corresponding to each identified discourse boundary and stores extracted information in an XML file. As a learner inputs an YouTube video lecture to the system, the user interface renders the video stream. Consulting the XML, discourse boundaries are identified and the learner is notified about the availability of assessment items. Upon learner's confirmation, the streaming is paused and the system takes the learner to in-citu assessment interface; otherwise streaming is continued. The system provides feedback on the learner's responses depending upon the question category. A representative snapshot of the user interface is shown in Figure 3.

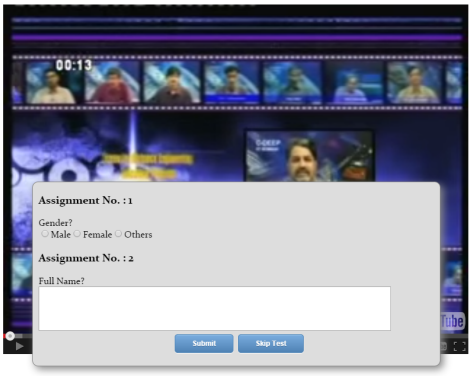

Figure 3. Interface showing inserted assessment items in a discourse boundary.

CONCLUSION

This paper presents a prototype system for inserting assessment items automatically in online video lectures. The challenges in this task include topic boundary detection, Wikipedia segment retrieval, automatic generation of assessment items and feedback design. The authors envisage further scopes for improvements in tasks involved specifically question generation, model answer extraction and formative feedback design for text-based answer with deeper natural language understanding. Apart from assessment items, inclusion of other "Change Ups" like trivia, animations will make the system more appealing.

REFERENCES

1. Al-Yahya, M. M. Ontoque: A question generation engine for educational assesment based on domain ontologies. In *ICALT*, IEEE Computer Society (2011), 393–395.

2. Benjamin, L. Lecturing. In *The Teaching of Psychology: Essays in Honor of Wilbert J. McKeachie and Charles L. Brewer*, S. F. Davis and B. W., Eds. Lawrence Erlbaum Associates Inc. Publishers, 57–67.

3. Corley, C., and Mihalcea, R. Measuring the semantic similarity of texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, EMSEE '05, Association for Computational Linguistics (Stroudsburg, PA, USA, 2005), 13–18.

4. Feng, V. W., and Hirst, G. A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers* (2014), 511–521.

5. Goss Lucas, S., and Sandra Bernstein, D. A. *Teaching Psychology: A Step By Step Guide*. Lawrence Erlbaum Associates Inc. Publishers.

6. Graesser, A., R. V., and Cai, Z. Question classification schemes. In *Proceedings of the Workshop on Question Generation* (2008).

7. Hearst, M. A. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Comput. Linguist. 23*, 1 (Mar. 1997), 33–64.

8. Heilman, M. Automatic factual question generation from text.

9. Landauer, T. K., Foltz, P. W., and Laham, D. An Introduction to Latent Semantic Analysis. *Discourse Processes*, 25 (1998), 259–284.

10. Levy, R., and Andrew, G. Tregex and tsurgeon: tools for querying and manipulating tree data structures. In *Proceedings of the fifth international conference on Language Resources and Evaluation* (2006).

11. Mann, W. C., and Thompson, S. A. Rhetorical structure theory: Toward a functional theory of text organization. *Text 8*, 3 (1988), 243–281.

12. Middendorf, J., and Kalish, A. The "change-up" in lectures. *The National Teaching & Learning Forum 5*, 2 (1996).

13. Milne, D. N., and Witten, I. H. An open-source toolkit for mining wikipedia. *Artificial Intelligence 194* (2013), 222–239.

14. Mitkov, R., An Ha, L., and Karamanis, N. A computer-aided environment for generating multiple-choice test items. *Nat. Lang. Eng. 12*, 2 (June 2006), 177–194.

15. Papasalouros, A., Kanaris, K., and Kotis, K. Automatic generation of multiple choice questions from domain ontologies. In *IADIS International Conference e-Learning 2008, Amsterdam, The Netherlands, July 22-25, 2008. Proceedings* (2008), 427–434.

16. Rus, V., Lintean, M. C., Banjade, R., Niraula, N. B., and Stefanescu, D. SEMILAR: the semantic similarity toolkit. In *51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, Proceedings of the Conference System Demonstrations, 4-9 August 2013, Sofia, Bulgaria* (2013), 163–168.

17. Traynor, D., and Gibson, J. P. Synthesis and analysis of automatic assessment methods in cs1: Generating intelligent mcqs. In *Proceedings of the 36th SIGCSE Technical Symposium on Computer Science Education*, SIGCSE '05, ACM (New York, NY, USA, 2005), 495–499.