Textbook Construction from Lecture Transcripts

Aliabbas Petiwala   Kannan Moudgalya   Pushpak Bhattacharya
IIT Bombay , Mumbai India
aliabbas@iitb.ac.in,kannan@iitb.ac.in,pb@cse.iitb.ac.in

ABSTRACT
This paper presents ongoing work on the proposal for Techniques for Construction of Pedagogically Sound Syllabus Guided Textbooks through Quality Courses and Collaboration using collaborative platforms like Wikis, customized Open Educational Resources (OER) and natural language processing technologies.The course videos from NPTEL at IITs are quality courses presented by leading faculties of IITs. The problem with such rich video courses is the question of usability of the courses, it would be nice if there would exist a textbook companion for these courses customized to the syllabus of the student's home university. This will save the student's time by enabling the student to study only what is required to be studied and thus provide a customized textbook catering to the respective university syllabus or even allowing the student to generate ones own personalized textbook companion for the video courses.A suitable platform needs to be developed which would act as a central repository for collaboratively organizing, indexing and proper interfacing of these Instructional material. A major goal of this platform should be to generate customized textbooks according to the syllabus.

KEYWORDS: Textbooks, lecture transcript, NLP, Authoring.

Proceedings of the Workshop on Speech and Language Processing Tools in Education, pages 43–50,
COLING 2012, Mumbai, December 2012.

1 Introduction

In today's world of Open Online Universities, video lectures and OER from eminent professors around the world, high quality lecture video has become easily available to the student. The major problem in creation of OER is how useful it is to the user community. Different universities have substantially different syllabus for the same course. The videos available for each course even have surprisingly different content. Its a common case when students require certain parts of the course which is taught at the user's home university is not available in the course lectures and similarly it is possible that a student who is not interested in some topics and yet he inadvertently goes through the non topics which are not taught at the home university.

Also,What if a student is interested only in parts of the video lectures. If a topic required in his exam is missing in the course, how does he find it? Is there a textbook that goes with the video lecture? To answer these questions and to make the learning material widely useful, we propose a textbook project.

2 Related Work

None of the previous work on textbook generation discussed explores the possibility of customized automatic textbook generation based on syllabus by an authoritative instructor.An important distinction from the work of ebook generation from web (Chen et al., 2005) and connexions project (Henry et al., 2003) is that we require to generate the textbook from the Wiki course content repository which is closely moderated and compiled by the instructor.

3 The Textbook Authoring Platform

The proposed solution to the mentioned problem can be depicted by the data flow diagram in Fig.1.Although a good portion of the textbook is expected to be generated automatically, a substantial amount of manual intervention is also expected, at least at the initial stages to ensure quality. It is likely that some material required for the syllabus of a university may not be present in the wiki. Through the moderation-collaboration route mentioned earlier, the missing information can be added.

A brief Outline of the proposed methodology the TextBook Project are:

1. Convert course lecture transcripts + reference materials of IIT professors to a semantic format like wiki, cnxml, eLML etc
2. Make this available to Subject matter experts over a Wiki. The experts will add\mix \match new information going through a thorough review process.
3. The content for each course is stored in a semantic repository that would grow phenomenally over time.The contents of the course in this repository will be the union of all similar named courses taught at different universities across India.
4. Use a syllabus tool to force the Instructors to submit a detailed syllabus for their course.
5. The syllabus tool would allow instructors to specify detailed meta info about the course like keywords, non topics,beyond scope topics, chapter difficulty,length etc.
6. Extract the information guided by the syllabus to generate a book for that syllabus performing extractive and logical summarization on the fly. The syllabus keywords or used to recommend the most promising text segment from the lecture transcript for the given topic. Many of the existing available techniques(Chandrasekar et al., 1996)(Fujii et al., 2008) (Das & Martins, 2007). can be used to simplify and reorganize the text.The techniques described by (Siddharthan, 2006) can be used for syntactic simplification

and retaining discourse cohesion of the rewritten text. Linguistic problems attacked in the techniques would be simplifying sentences, deciding determiners, deciding sentence order and preserving rhetorical and anaphoric structure.

7. This expanded course will help generate several textbooks corresponding to the syllabi of various Indian universities.
8. The repositories are open to collaboration under strict control of the Wiki Moderator\Teaching Asst. and Instructors.
9. This will help resulting in better transparency , standardization and openness of what is taught across different universities and ease the student in finding the right content according to custom requirements easily.

4 Results and Contributions

This section describes the ongoing work and the contributions made which are on going . The results of applying NLP techniques to the lecture transcripts and textbooks has been discussed in the following sections.

4.1 Applying NLP to Lecture Transcripts Analysis

A corpus of lecture transcripts was constructed containing 40 IIT Bombay lecture transcripts from a basic Electrical Engineering course, 20 MIT lecture transcripts from Introductory AI course and 34 MIT lecture transcripts from physics course. The lecture transcripts were transcribed by a human transcriber. True to our expectation we found that the lecture transcripts seem to incorporate a lot of features which characterize informal active speech which we call it as “lecture speech” .A statistical frequency analysis of the lecture transcripts corpus was done using the NLTK toolkit, and GATE(Cunningham et al., 2011). Following are the prominent features of these lecture transcripts examples are quoted in the parenthesis:

- Frequent use of first and second person ("I", "you")
- Personal references(“I will”)
- Active voice (“Let’s see”)
- Simple words and sentences
- Frequent Short sentences
- Interruptions, frequent topic changes and returning back to topics
- Frequent Questioning followed by answer to the posed question by the lecturer himself(“What”,”How”)
- Contractions (“wont”,”lets”...)
- Abbreviations(“IIT”,”Btech”)
- Frequent use of demonstrative pronouns to present ideas, slides and examples (’this’, ’that’, ’there’, ’refer’, ’see’, ’these’)

Similarly a textbook corpus was constructed consisting of four famous textbooks in traditional engineering curriculum, each with pages ranging from 300-1500 pages of text. For a textbook we observed that a typical textbook is written in an academically structured and formal style incorporating sound pedagogical principles. A similar statistical frequency analysis on a engineering textbook corpus consisting four authoritative English textbooks in the Engineering domain found the following features:

- Use of Impersonal style (third person:It, that)
- Passive voice(It was found that ...)

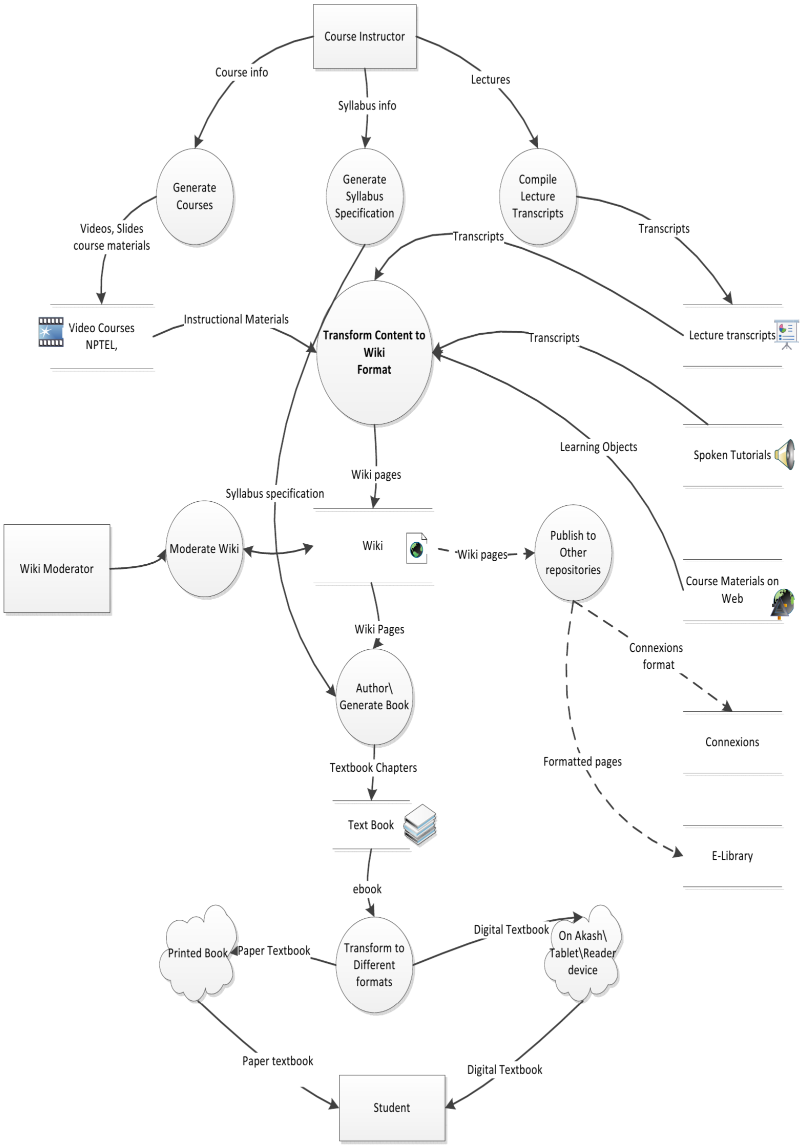

Figure 1: Data flow diagram for TextBook Project

- Reported Speech
- Complex words and sentences to express complex points
- Equations
- Technical jargon
- No contractions
- Use of proper English style , precise education technical vocabulary and several other best practices in literature writing
- Use of objective style, using facts and references to support an argument
- Absence of vague expressions and slang words
- Citations to other books
- Bibliography

The identification of above features of lecture transcripts and textbook content will allow us to define a mapping between the concrete linguistic features of the lecture transcripts and concrete linguistic features of textbooks. It is thus hypothesized that this mapping would help in transforming lecture transcripts to formal textbooks.

A typical lecture transcript contains many unique topics as well as alternating between various topics , need arises to segment the text into relevant cohesive text units and find similarities among them which would help in merging similar sections in the transcript to generate a well structured and coherent text that can serve to achieve the goal of generating a textbook. The following sections elaborate the the techniques and results to find topical changes in the transcripts and its cluster analysis.

4.1.1 Text Segmentation

The text tiling algorithm(Hearst, 1997) was applied to the lecture transcripts and the number of topical segments found in the transcripts is portrayed in Fig.2. The number of segments in the lecture transcripts roughly correspond to the number of topic changes in the lecture. A mean of 14 segments was found per transcript for the EE111 course as shown in the figure.

4.1.2 Clustering to Find and Merge Intra Segment Similarities

The lecture transcripts and for that matter any instructional material contains a lot of redundancy across the entire course. To generate a textbook from the transcripts it is also required to find both inter transcript and intra transcript similarities for the course. The lecture transcripts for each course were clustered to find clusters of similar lecture transcripts that can be probably linked together for better information extraction for textbook generation. A bottom up hierarchical clustering of unsupervised learning was used for cluster analysis. Initially, each transcript is represented as vector of tf-idf features.A tf-idf feature matrix was constructed for all the transcripts assuming each transcript as a document. The bag of words approach was used to compute features of the transcripts. Using a Euclidean distance metrics distances between these vectors of word counts is computed, the closest two texts are grouped into a cluster. The distance from this new cluster to all other transcripts is then recalculated in a recursive fashion. transcripts and clusters are thus compared for similarity, and clustering continues until all transcripts belong to a single top level cluster. The resulting tree can be visualized as a dendrogram. The number of features was empirically fixed at top 50 features ordered by term frequency across the corpus after experimenting with different values for the number of features and fixing a value which optimizes the number of clusters in the hierarchical clustering . The results of this hierarchical clustering was visualized as a dendrogram as shown in Fig.3. The y axis represents the similarity distance between the lecture transcripts at the leaves. It is evident from the dendrogram that the lectures transcripts for each of the lectures are highly successively ordered corresponding to the natural style of the delivered lectures as new vocabulary is introduced as lecture proceeds. These clusters can help us finding similar lectures allowing us to merge similar content. Similarly, to find inter transcript similarities the segments found in the segmentation of the lecture transcripts were clustered to find similar segments across the lecture transcripts for the same course. A tf-idf feature matrix was constructed for all the segments assuming each segment as a document. The number of features was empirically fixed at 50 after experimenting with different values for the number of features and fixing a value which optimizes the number of clusters in the hierarchical clustering . The results of this hierarchical clustering was visualized as a dendrogram as shown in Fig.4. The distinct color bunches of hair in the dendrogram correspond to the clusters of the segments. A zoomed analysis showed that most of the lecture segments are contiguous in natural sequence. This cluster analysis is envisioned for merging, discarding or augmentation of different segments within the transcript for enhancing the coherency and cohesiveness of the target textbook.

Conclusion

We presented the architecture automatic authoring of textbooks from lecture transcripts. This is proposed to be achieved through identifying the features of the transcripts and the textbook and creating a mapping between the two and finally using this mapping to achieve the goal of generating a textbook out of the lecture transcript. Similar segments and cluster analysis results of the transcripts would enable to merge similar topical sections.

References

Chandrasekar, R., Doran, C., & Srinivas, B. (1996). Motivations and methods for text simplification. In *Proceedings of the 16th conference on Computational linguistics-Volume 2* (pp. 1041–1044).

Chen, J., Li, Q., & Jia, W. (2005). Automatically generating an e-textbook on the web. *World Wide Web*, 8, 377–394.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Roberts, I., Gorrell, G., Funk, A., Roberts, A., Damljanovic, D., Heitz, T., Greenwood, M. A., Saggion, H., Petrak, J., Li, Y., & Peters, W. (2011). *Text Processing with GATE (Version 6)*.

Das, D. & Martins, A. F. T. (2007). A survey on automatic text summarization. *Literature Survey for the Language and Statistics II course at CMU*, 4, 192–195.

Fujii, Y., Yamamoto, K., Kitaoka, N., & Nakagawa, S. (2008). Class lecture summarization taking into account consecutiveness of important sentences. In *Ninth Annual Conference of the International Speech Communication Association*.

Hearst, M. A. (1997). TextTiling: segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23(1), 33–64.

Henry, G., Baraniuk, R., & Kelty, C. (2003). The connexions project: Promoting open sharing of knowledge for education. *Syllabus, Technology for Higher Education*.

Siddharthan, A. (2006). *Syntactic simplification and text cohesion*. PhD thesis.

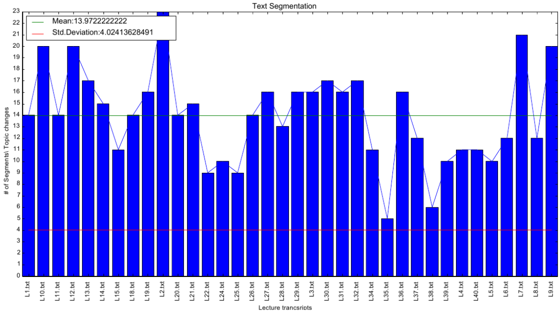

Figure 2: Segmentation Statistics of Lecture Transcripts

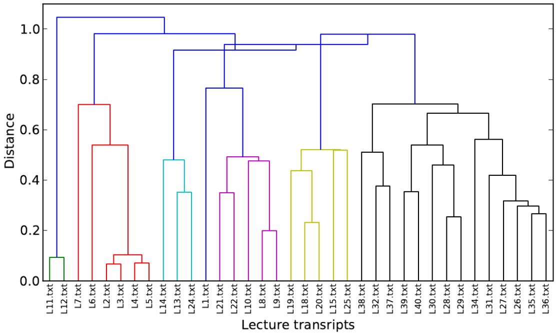

Figure 3: Dendrogram for Hierarchical Clustering of Lecture Transcripts

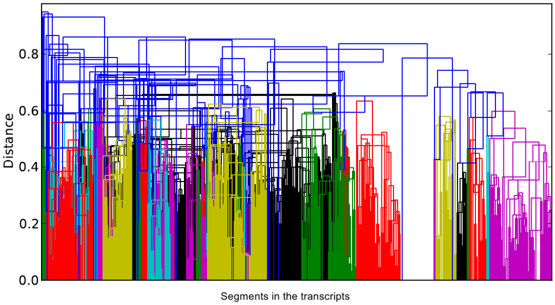

Figure 4: Dendrogram for Text Segment Clustering in the EE lecture Transcripts