MIT develops lecture search engine to aid students

November 8 2007

The Lecture Server, as shown in this screenshot of MIT physics professor Walter Lewin, displays video and highlighted search terms.

Imagine you are taking an introductory biology course. You're studying for an exam and realize it would be helpful to revisit the professor's explanation of RNA interference. Fortunately for you, a digital recording of the lecture is online, but the 10-minute explanation you want is buried in a 90-minute lecture you don't have time to watch.

A new lecture search engine developed at MIT's Computer Science and

Artificial Intelligence Laboratory (CSAIL) could help with this dilemma. Created by a team of researchers and students led by MIT associate professor Regina Barzilay and principal research scientist James Glass, the web-based technology allows users to search hundreds of MIT lectures for key topics.

"Our goal is to develop a speech and language technology that will help educators provide structure to these video recordings, so it's easier for students to access the material," said Glass, who is head of CSAIL's Spoken Language Systems Group.

More than 200 MIT lectures are currently available on the site (web.sls.csail.mit.edu/lectures/). So far, most of the users are international students who access the lectures through MIT's OpenCourseWare (OCW) initiative, which makes curriculum materials for most MIT courses available to anyone with Internet access. Although the lecture-browsing system is still in the early development stages, a recent announcement in OCW's newsletter has drawn increased traffic to the site.

Barzilay and Glass expect the system will be most useful for OCW users and for MIT students who want to review lecture material. MIT World, a web site that provides video of significant MIT events such as lectures by speakers from MIT and around the world, is also participating in the project.

Many MIT professors record their lectures and post them online, but it's difficult to search them for specific topics. Because there is no way to easily scan audio, as you can with printed text, "you end up watching the whole thing, and it's hard to keep focused," said Barzilay, the Douglas T. Ross Career Development Associate Professor of Software Development in the Department of Electrical Engineering and Computer Science.

On the prototype web site, users can search lectures for any term they want and then play the relevant sections.

The lecture transcripts are created by speech recognition software. One major challenge is that the lectures usually contain many technical terms that might not be in the computer program's vocabulary, so the researchers use textbooks, lecture notes and abstracts to identify key terms and feed them into the computer.

"These lectures can have a very specialized vocabulary," said Glass. "For example, in an algebra class, the professor might talk about Eigenvalues."

When properly adapted to a speaker and topic, the lecture-based speech recognizer gets about four out of five words correct, however most of the errors occur in words that are not critical to the lecture topic, i.e., not the key vocabulary terms that people would use to search.

Once the transcript is complete, a language processing program divides the text into sections by topic. Chunks of text, about 100 words each, are compared with each other using a mathematical formula that calculates the number of overlapping words between the text blocks. Each word is weighted so that repetition of key terms has more weight than less important words, and chunks with the most similar words are grouped into sections.

In the future, Barzilay and Glass hope to add a lecture summarization feature to the language processing system. They also want to get users more involved in the project, by incorporating a Wikipedia-like function that would let users correct errors in lecture transcripts and allow them to add lecture notes.

The researchers presented their project at the Interspeech 2007

conference in Antwerp, Belgium, in August. The project was originally funded by Microsoft through the iCampus program and is now funded by the National Science Foundation.

Source: MIT