Towards Spoken Knowledge Structuring and Organization

When Speech Processing Technology meets MOOCs

Speaker: Hung-yi Lee

Introduction

- 2012 is the year of the massive open online course (MOOC)
  - Instructors post recorded video/audio of their lectures on online lecture platforms.
  - Learners worldwide can easily access the curricula.
- More learning materials

Too much materials ……

videolectures●net

exchange ideas & share knowledge

Search: SVM - Matches: 212

I want to learn "SVM".

Learner

Tutorials:

Bernhard Schölkopf: Kernel Methods
The course will start with basic ideas of machine learning, followed by some elements of learning theory. It will also introduce po
kernels and their associated feature spaces, and show how to use them for kernel mean embeddings, SVMs, ...

Alexander J. Smola: Kernel methods and Support Vector Machines
The tutorial will introduce the main ideas of statistical learning theory, support vector machines, and kernel feature spaces. This i
derivation of the support vector optimization problem for classification and regression, the v-trick, various kernels and an overview

John Shawe-Taylor: Kernel Methods and Support Vector Machines
Kernel methods have become a standard tool for pattern analysis during the last fifteen years since the introduction of support ve
We will introduce the key ideas and indicate how this approach to pattern analysis enables a relatively easy ...

Fabio Ciravegna, Andrea Varga: Mining Complex Entities from Heterogeneous Information Networks
Most research on information mining has focused on classic Information Extraction (IE) tasks, from structured and unstructured
newspaper articles and web pages. In the last years however the staggering growth of social media as platform for sharing cont

Jerry (Xiaojin) Zhu: Semi-Supervised Learning
This tutorial covers classification approaches that utilize both labeled and unlabeled data. We will review self-training, Gaussian
models, co-training, multiview learning, graph-transduction and manifold regularization, transductive SVMs, and a PAC bound for
supervised learning. We then discuss some new ...

Chih-Jen Lin: Support Vector Machines
Support vector machines (SVM) and kernel methods are important machine learning techniques. In this short course, we will intr
basic concepts. We then focus on the training and optimization procedures of SVM. Examples demonstrating the practical use of

Manfred K. Warmuth, S.V.N. Vishwanathan: Survey of Boosting from an Optimization Perspective

A course is too much

機器學習基石
(by 林軒田)

I want to learn "SVM".

Learner

Learning From Data
(Yaser S. Abu-Mostafa)

- Lecture 14: Support Vector Machines
- Lecture 15: Kernel Methods

Machine Learning
(by Andrew Ng)

- XII. Support Vector Machines (Week 7)

A course is not enough

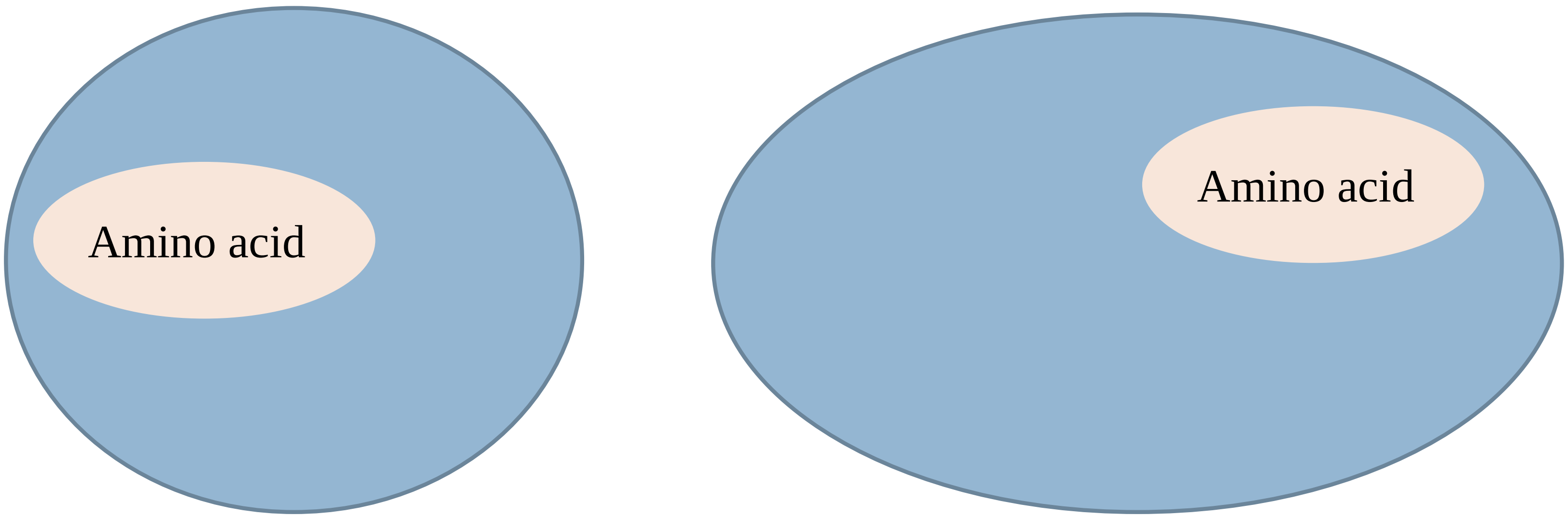
- Inter-discipline

I want to learn
“amino acid”.

Learner

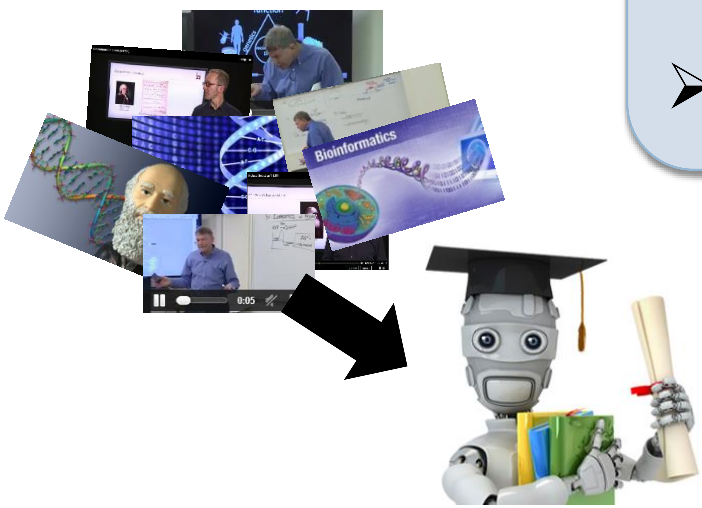
Introduction to Biology

Organic Chemistry



Vision: Personalized Courses

on-line learning material



- I want to learn “XXX”.
- I am a graduate student of computer science.
- I can spend 10 hours.

Learner

I open a course for you.

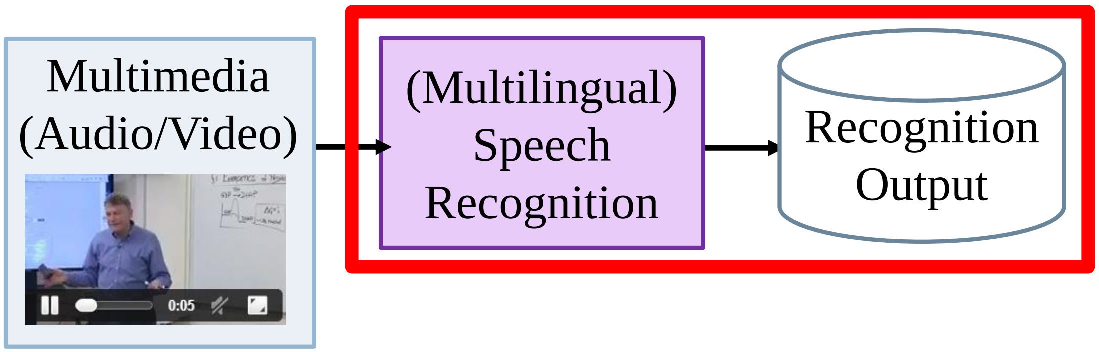
- Spoken Language Processing techniques can be very helpful.
- The spoken content in courses plays the most important role in conveying the knowledge.

Outline

- Part I: Overview each block in spoken knowledge structuring and organization
  - Speech Recognition
  - Temporal Structure
  - Spoken content retrieval
  - Linking related lectures
  - Speech summarization
  - Knowledge graph construction
  - Inferring prerequisite and advanced concepts
- Part II: Spoken Content Retrieval
- Part III: Speech Summarization
- Part IV: Demo

Part I: Overview

Speech Recognition



Speech Recognition

- Lectures on Coursera and edX has manual transcriptions
- Most lectures on the Internet do not have transcriptions
  - Speech recognition!

Speech Recognition

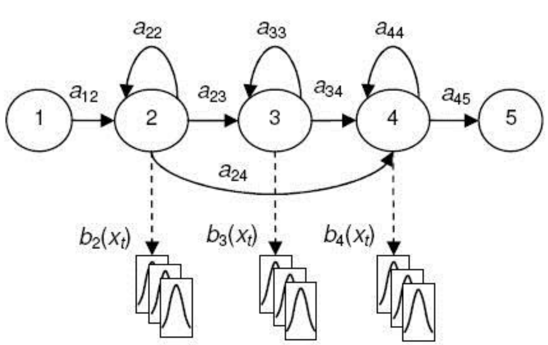
- Speech Recognition is the foundation of the following speech techniques

Spoken Lectures

Speech Recognition

DNA is the material of ……

Hidden Markov Model



N-gram Language Models

$$P(w_i|w_{i-(n-1)},\ldots,w_{i-1}) = \frac{count(w_{i-(n-1)},\ldots,w_{i-1},w_i)}{count(w_{i-(n-1)},\ldots,w_{i-1})}$$

Speech Recognition

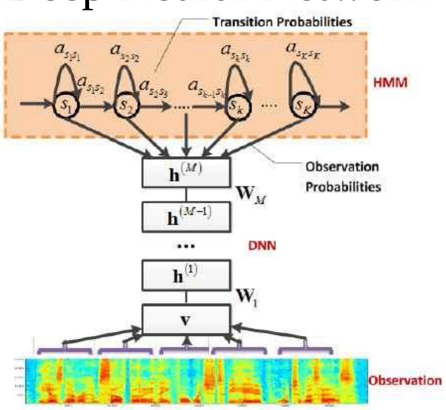
- Speech Recognition is the foundation of the following speech techniques

Spoken Lectures

Speech Recognition

DNA is the material of ……

Deep Neural Network



N-gram Language Models

$$P(w_i|w_{i-(n-1)},\ldots,w_{i-1}) = \frac{count(w_{i-(n-1)},\ldots,w_{i-1},w_i)}{count(w_{i-(n-1)},\ldots,w_{i-1})}$$

Speech Recognition

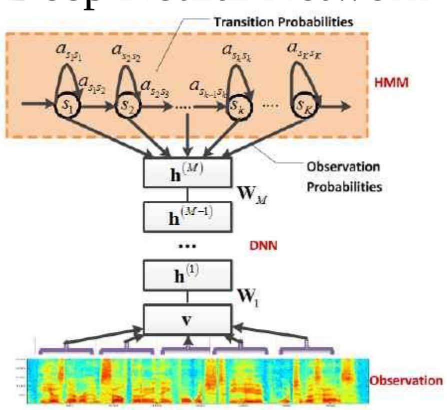
- Speech Recognition is the foundation of the following speech techniques

Spoken Lectures

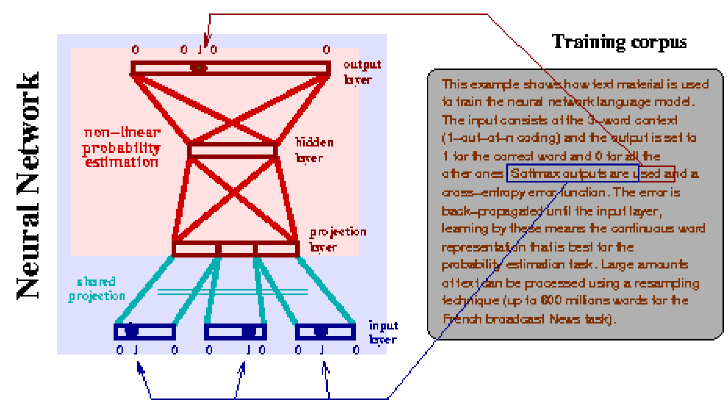
Speech Recognition

DNA is the material of ……

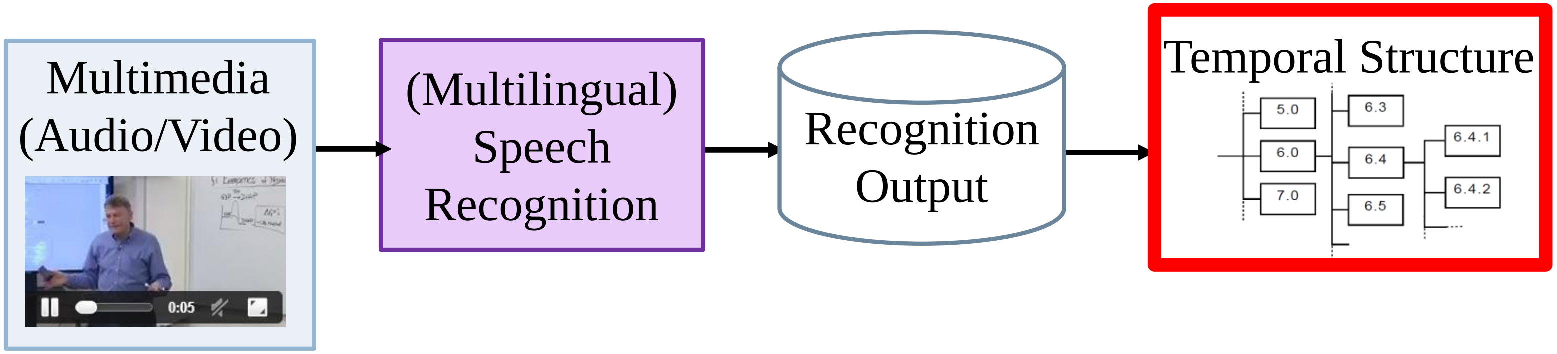
Deep Neural Network



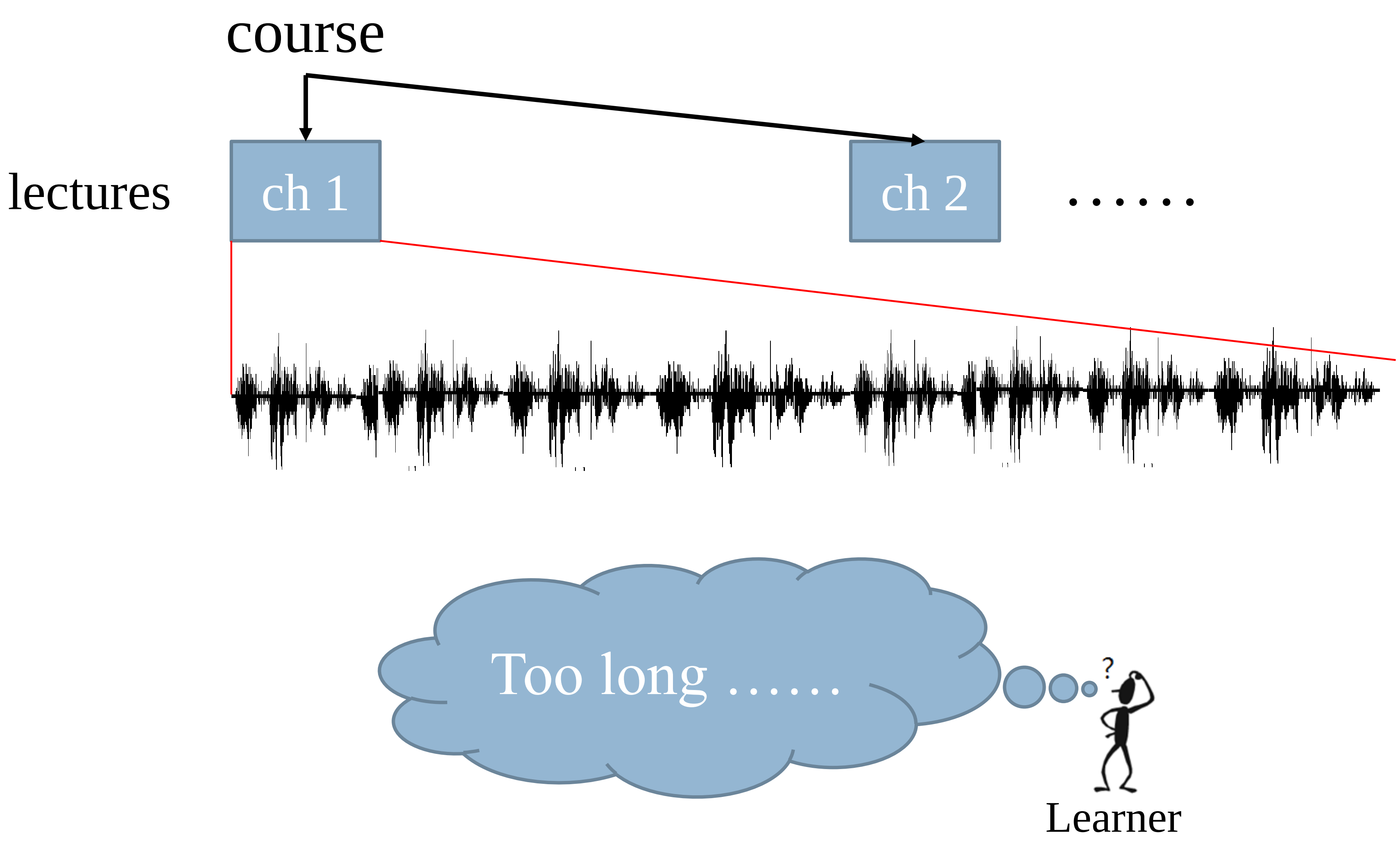
Continues Language Model



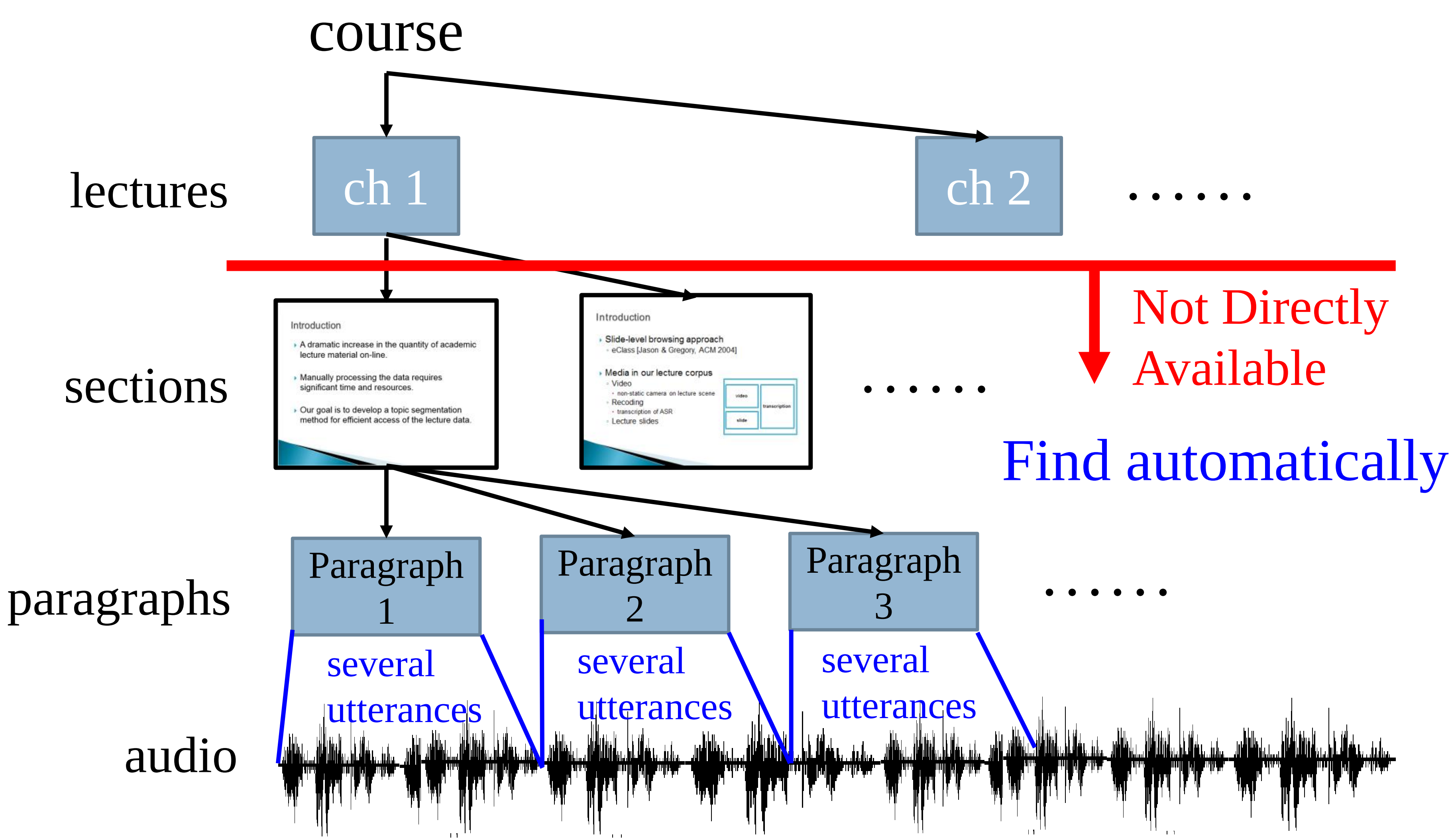
Multi-layer Temporal Structure



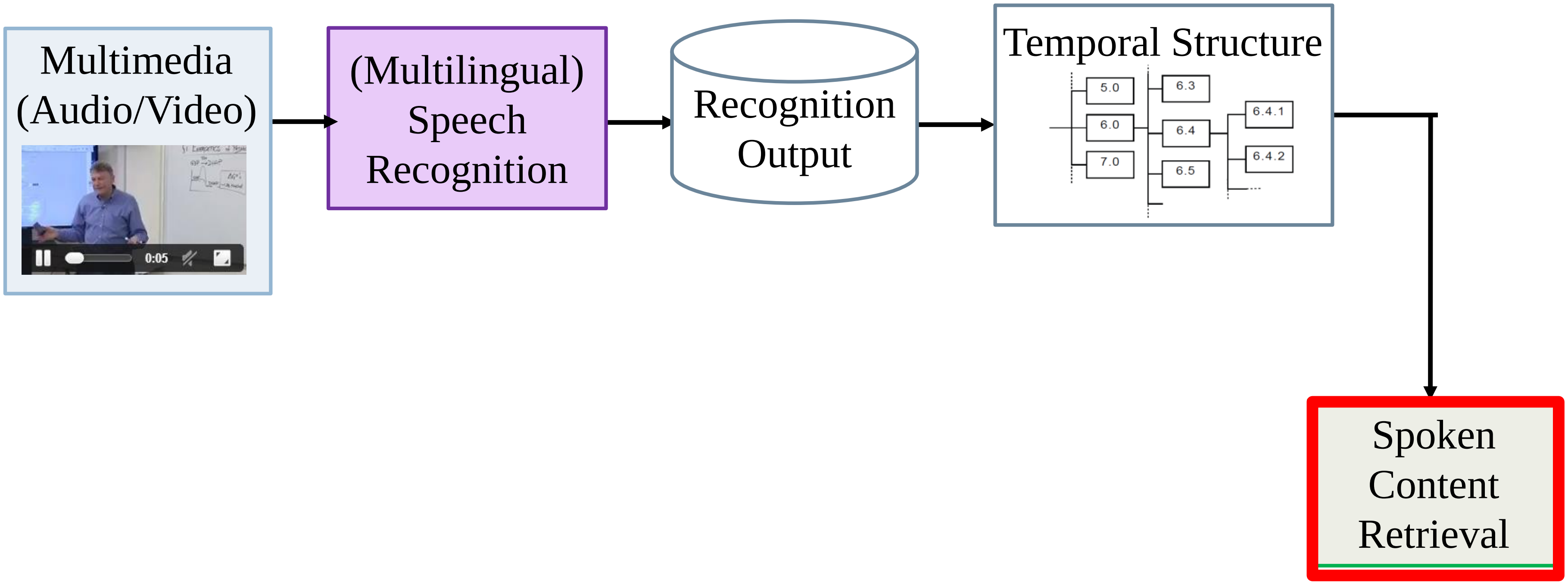
Multi-layer temporal structure



Multi-layer temporal structure

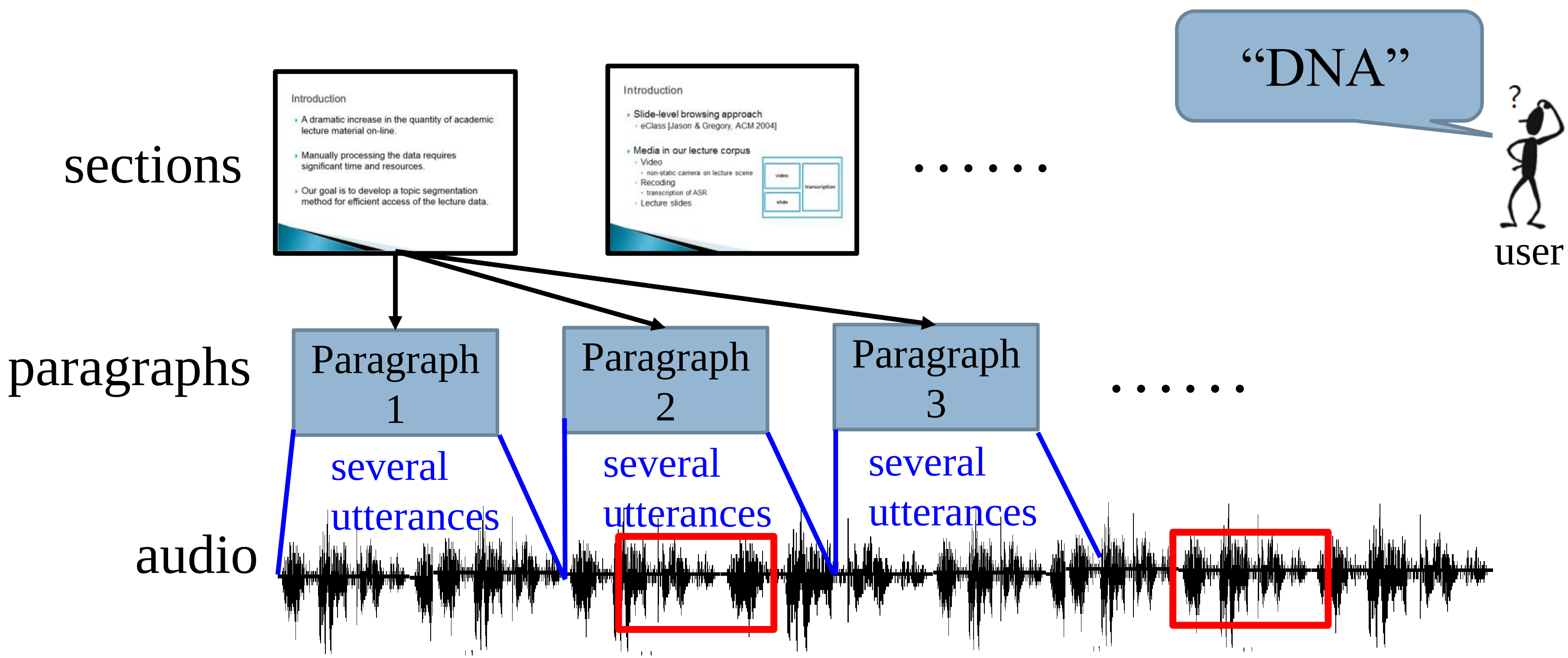


Spoken Content Retrieval



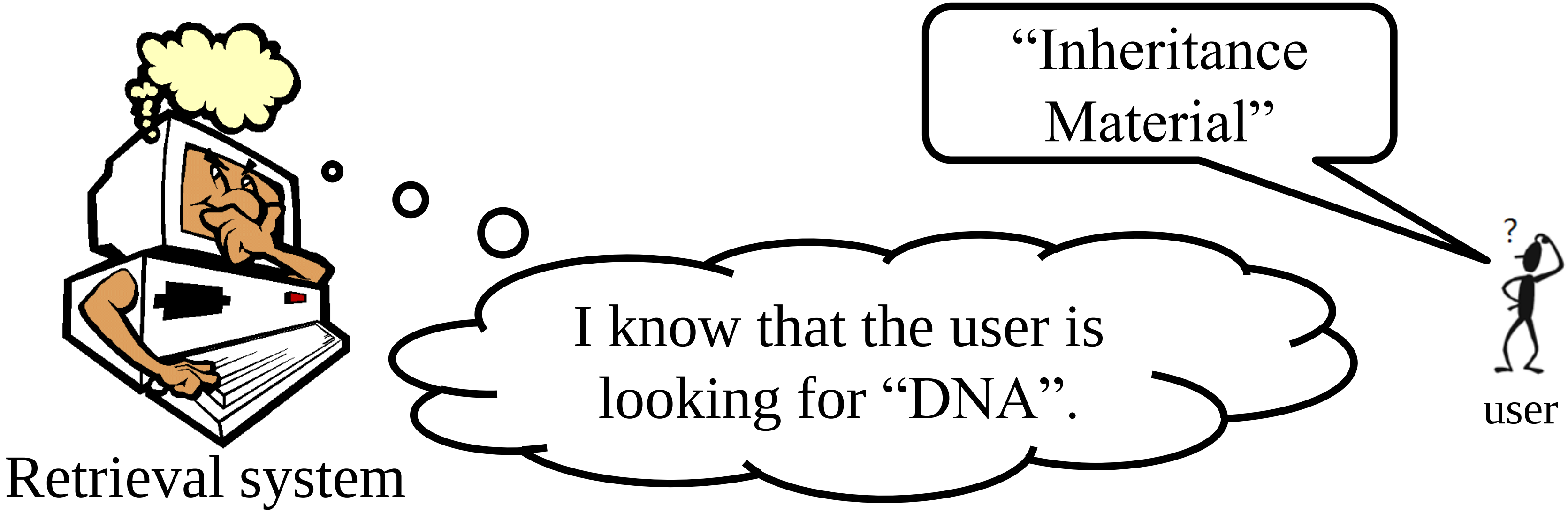
Spoken Content Retrieval – Goal

- Basic goal: return paragraphs or sections containing keywords
  - This is called “***Spoken Term Detection***”（口語詞彙偵測）

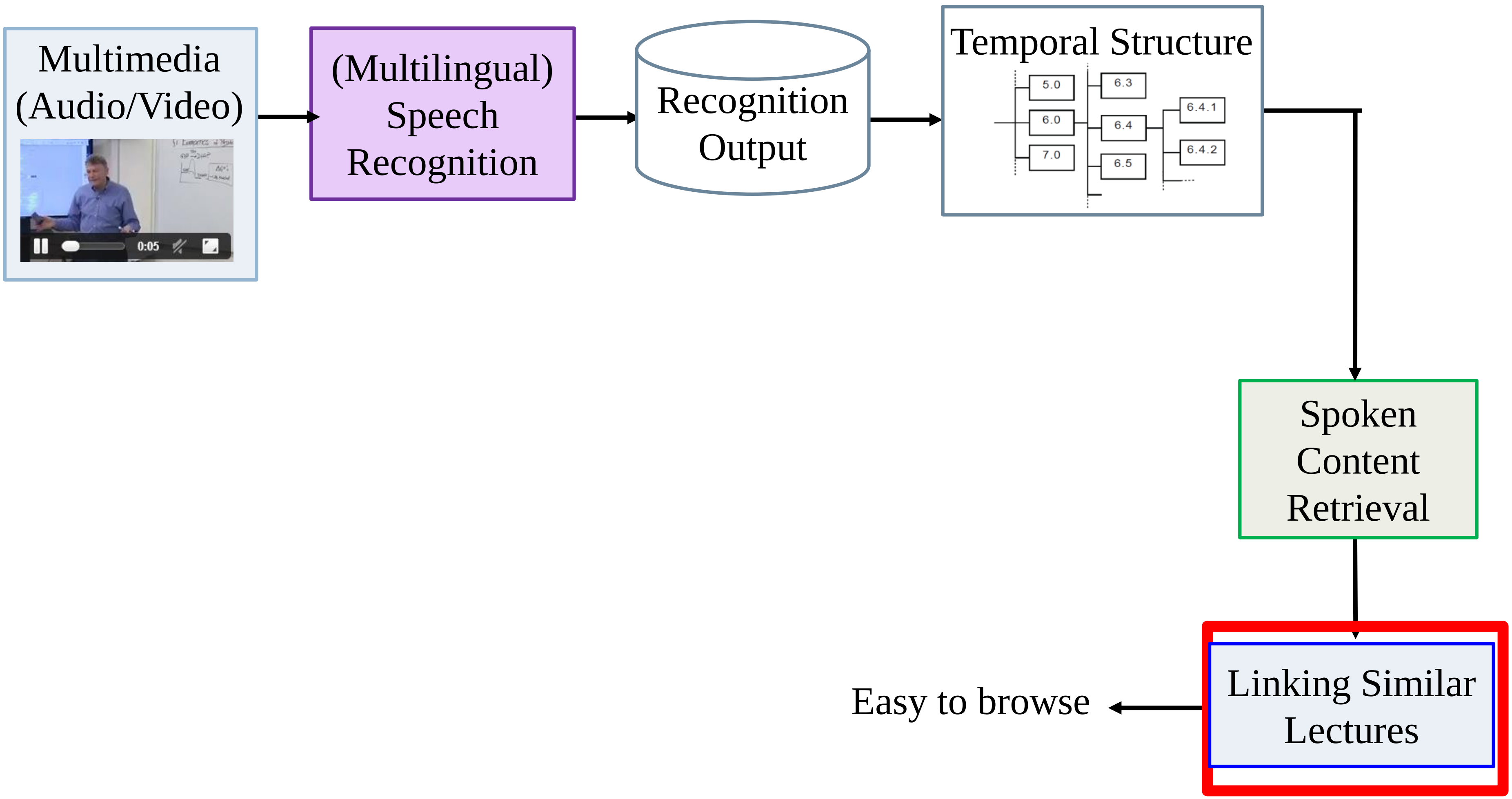


Spoken Content Retrieval – Goal

- Basic goal: return paragraphs or sections containing keywords
  - This is called "***Spoken Term Detection***" (口語詞彙偵測)
- Advanced goal: Semantic retrieval of spoken content

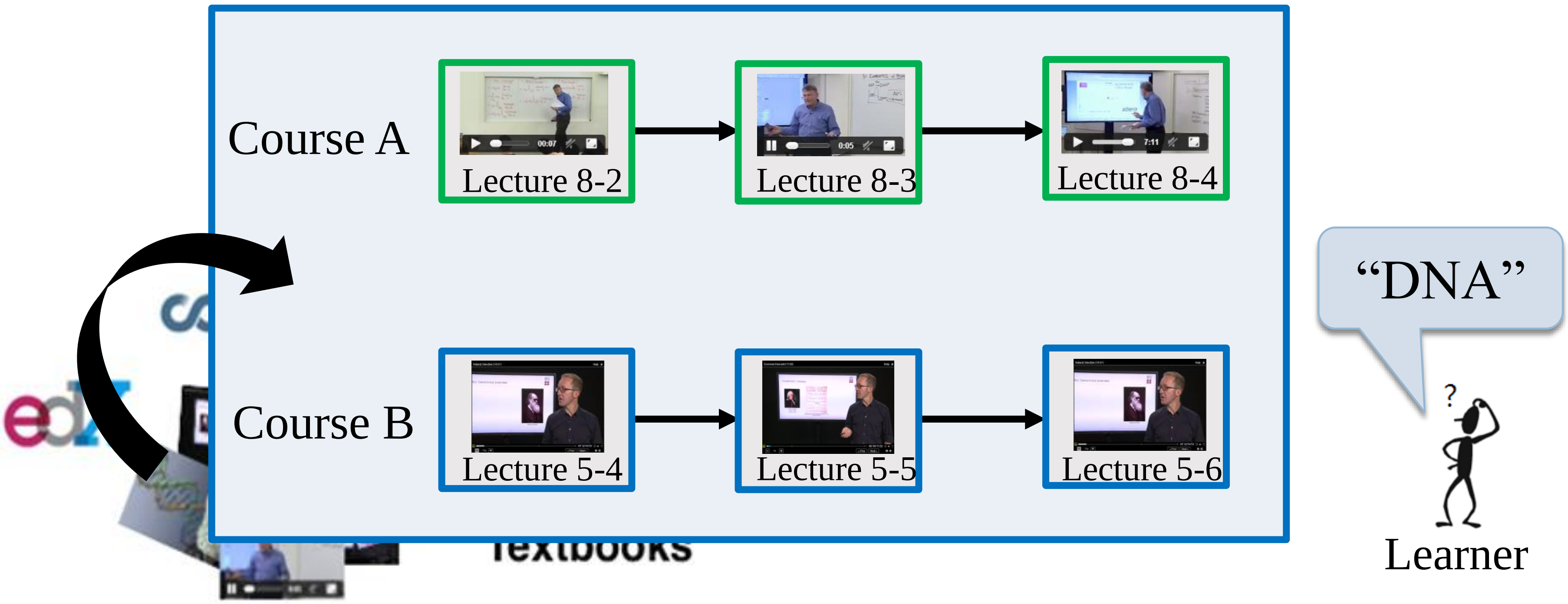


Visualizing Search Results



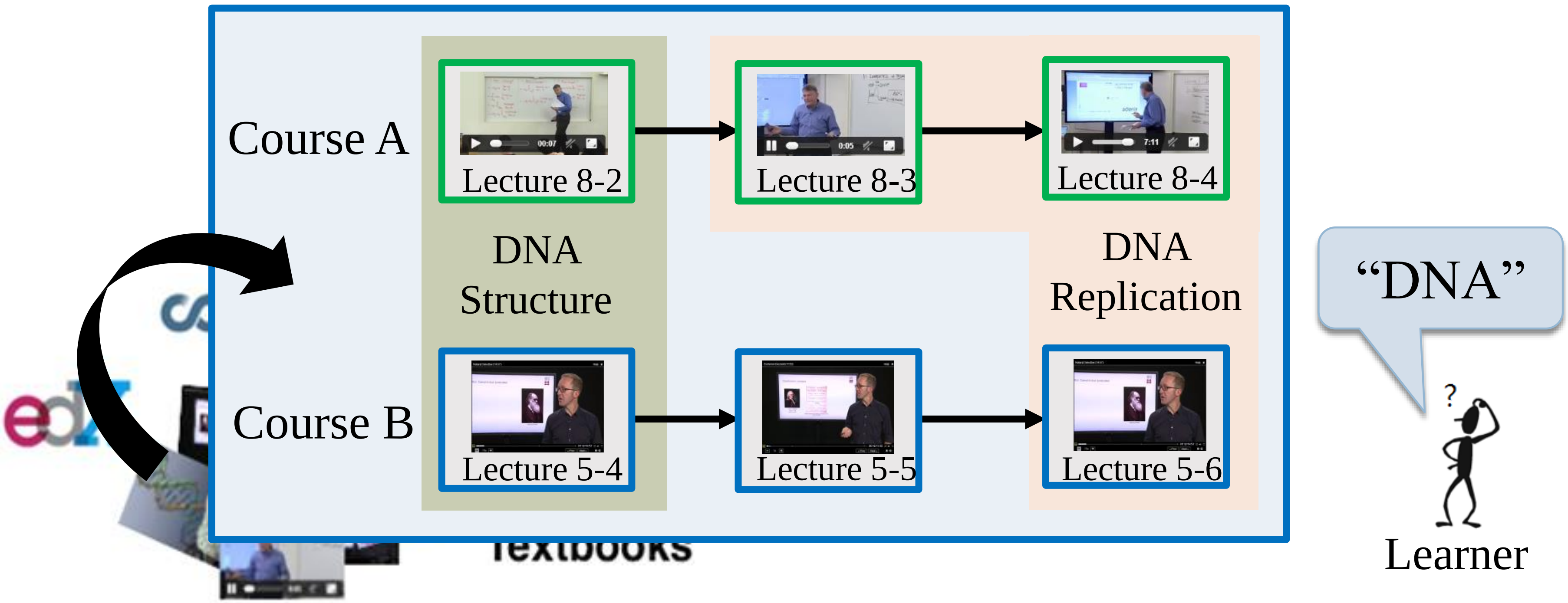
Search

- With spoken content retrieval, we can use keywords to search related lectures

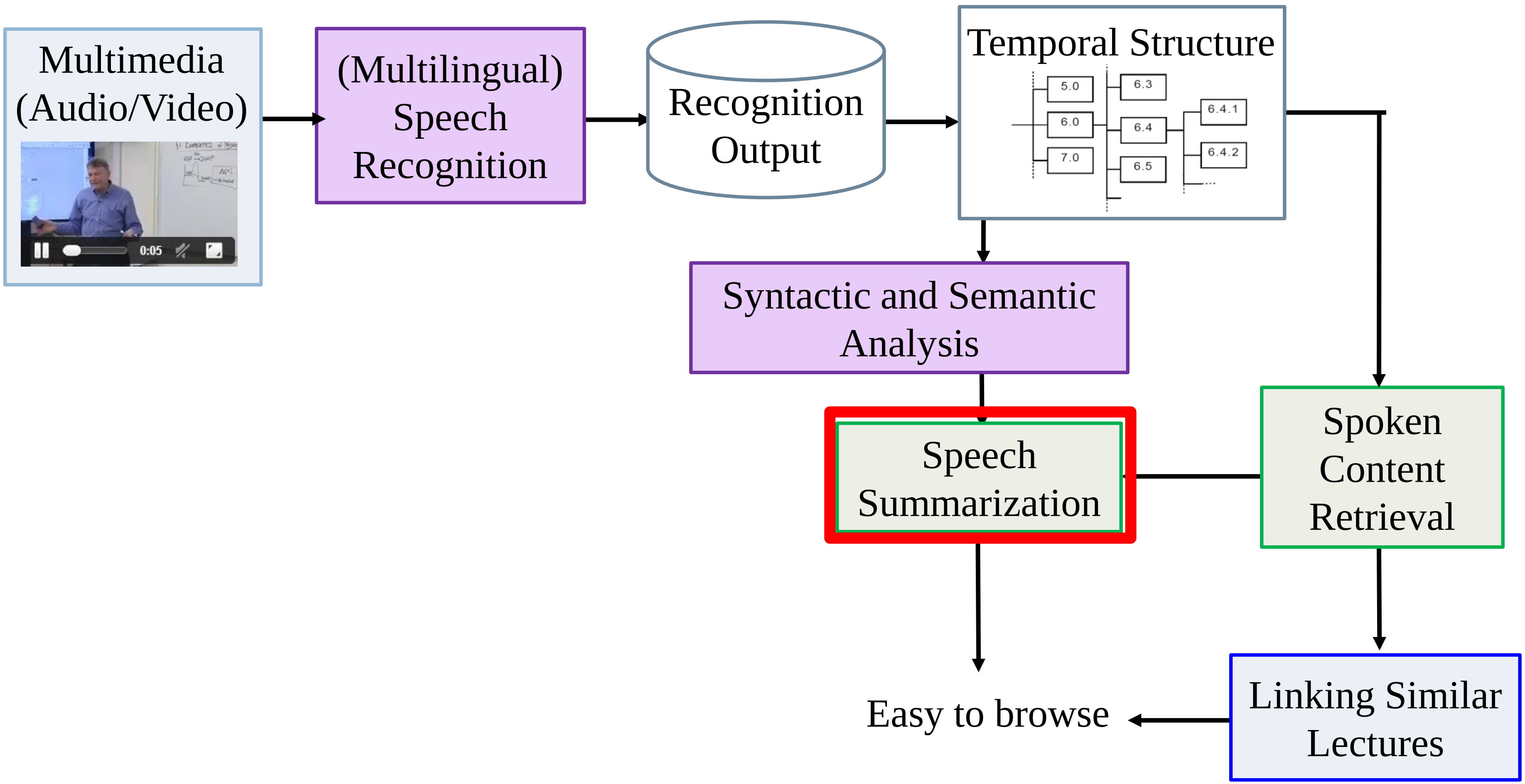


Linking

- Linking lectures with similar content
  - Compute similarity between lectures in courses and sections in textbooks
  - Merge the materials with high cosine similarity

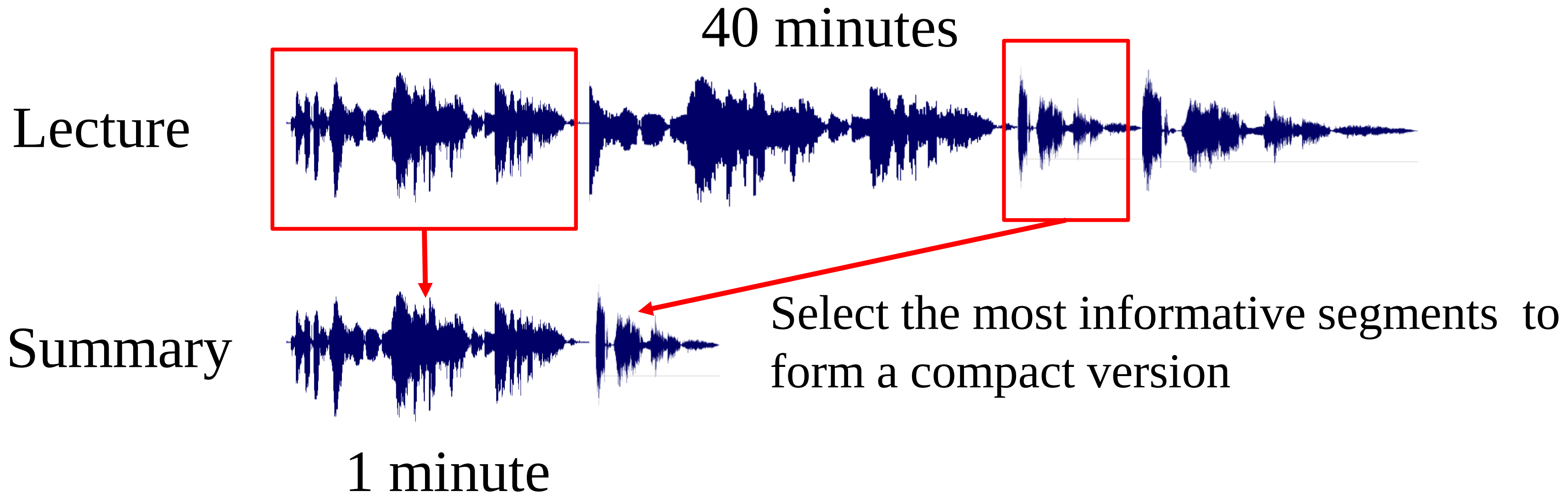


Speech Summarization

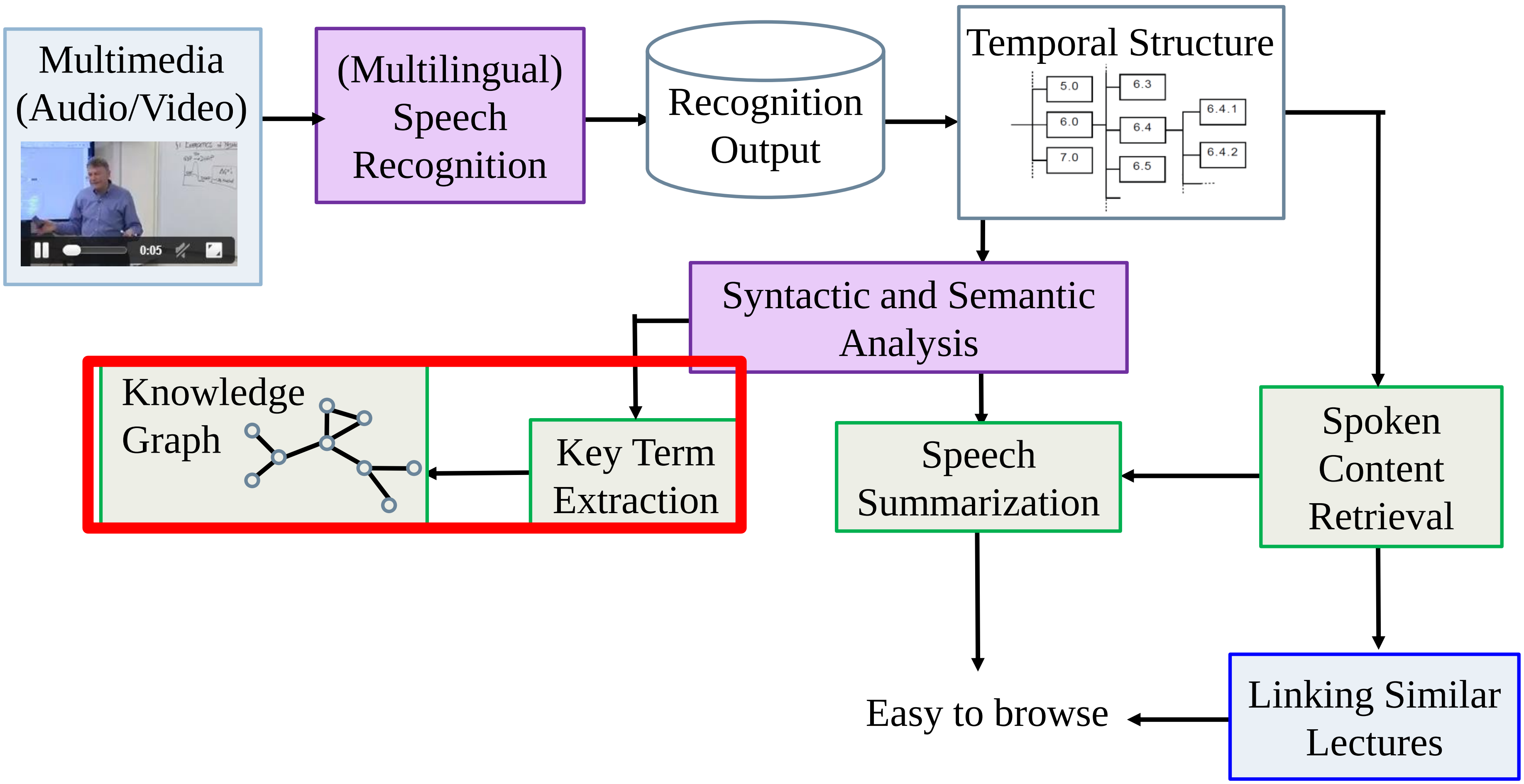


Speech Summarization

- Audio is hard to browse



Knowledge Graph Construction



Knowledge Graph Construction - Keyterm Extraction

- Knowledge graph construction
  - Keyterm extraction

Audio transcriptions of video clips

Keyterm Extraction

DNA
Inheritance
RNA
Adenosine triphosphate
……

Knowledge Graph Construction - Relation Extraction

- Knowledge graph construction
  - Keyterm extraction
  - Find relation between keyterms

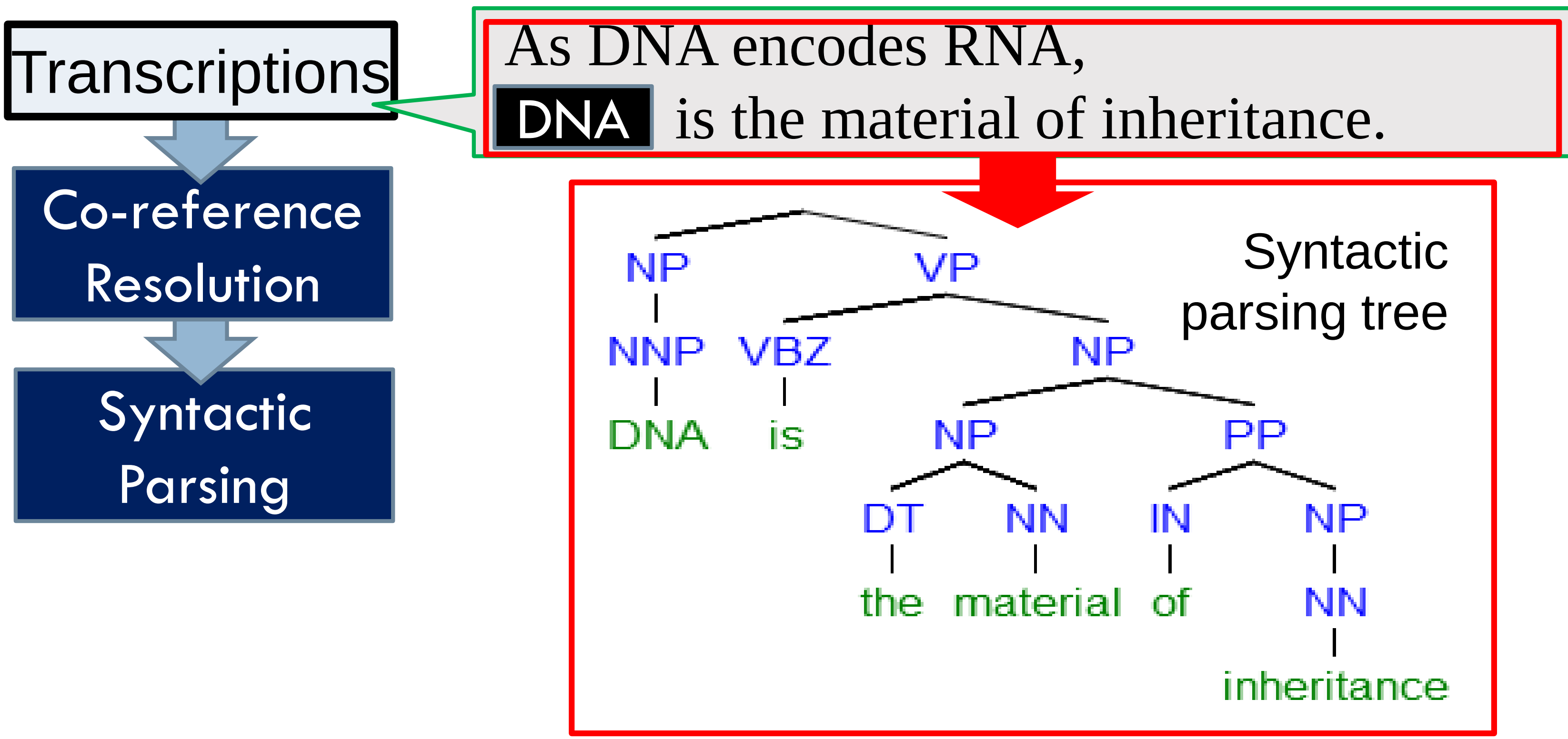
Transcriptions

Co-reference Resolution

As DNA encodes RNA,
it is the material of inheritance.

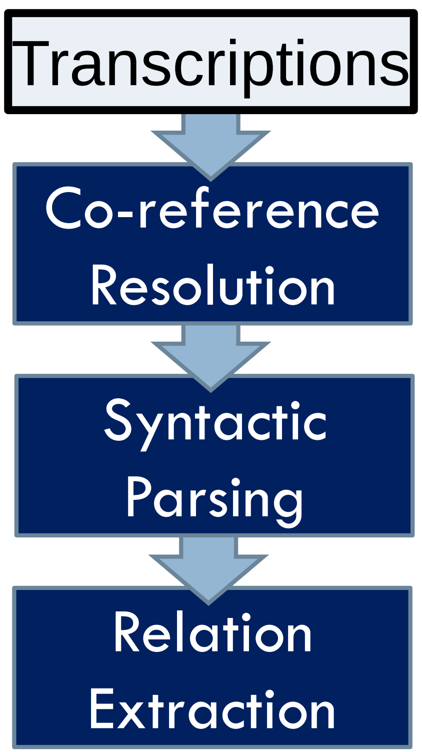
Knowledge Graph Construction - Relation Extraction

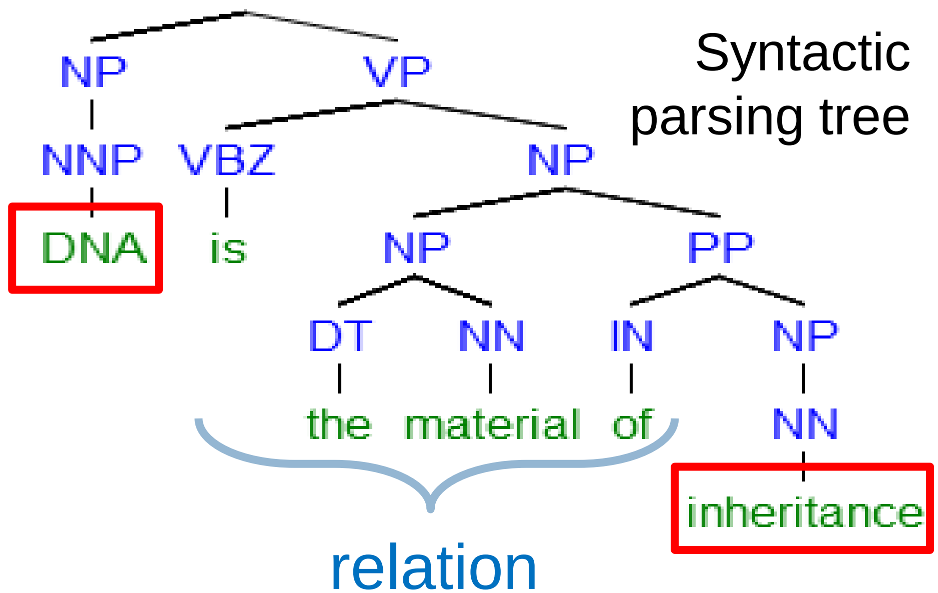
- Knowledge graph construction
  - Keyterm extraction
  - Find relation between keyterms



Knowledge Graph Construction - Relation Extraction

- Knowledge graph construction
  - Keyterm extraction
  - Find relation between keyterms





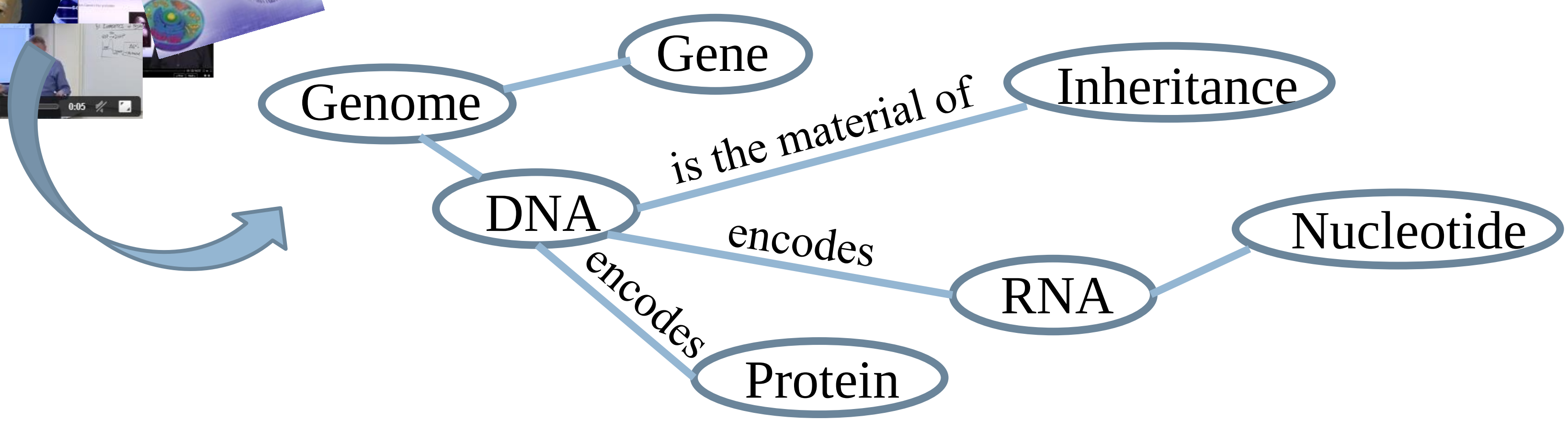
Syntactic parsing tree

[Mausam, EMNLP’12].

Knowledge Graph Construction - Relation Extraction

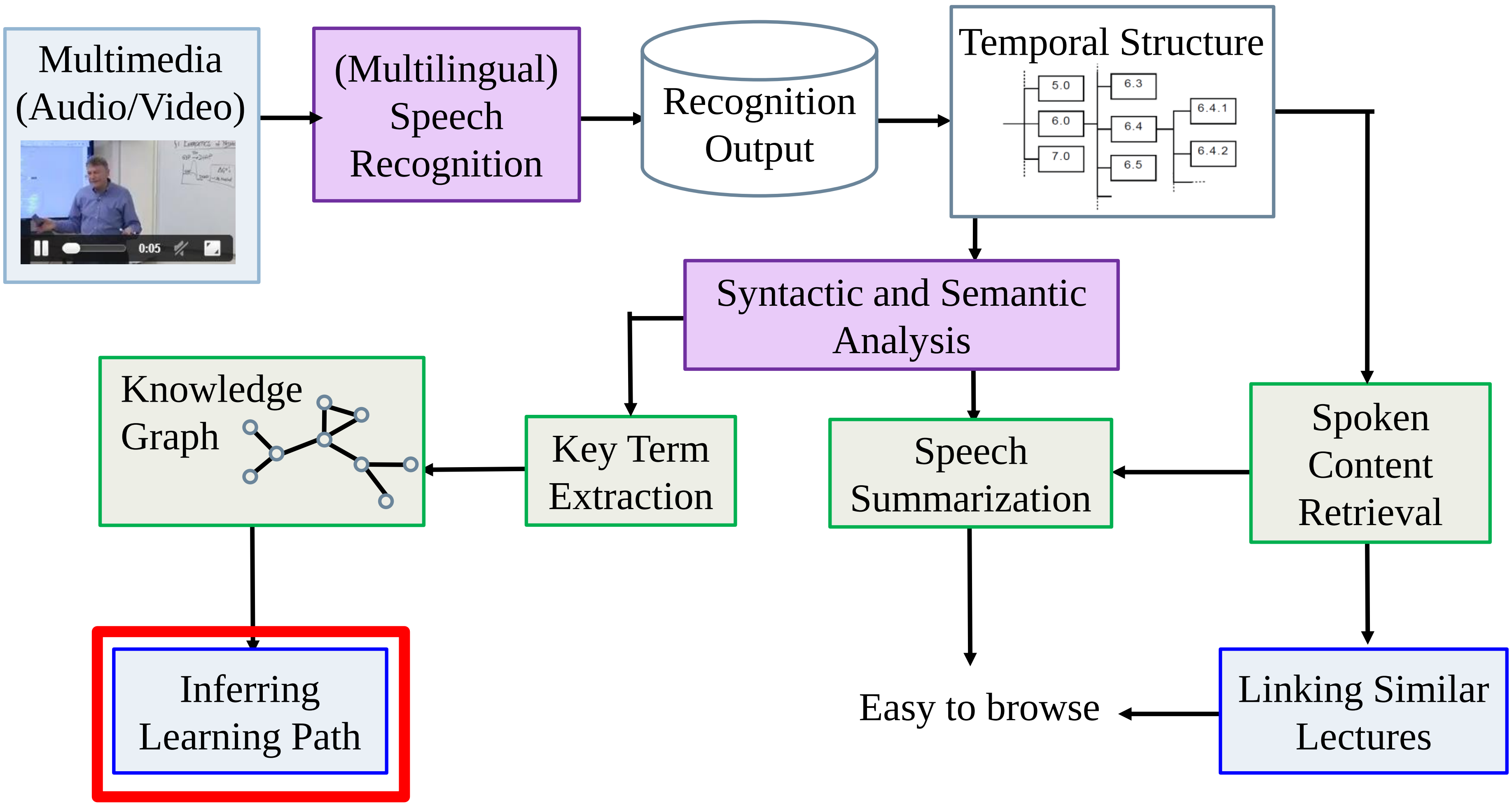
- Knowledge graph construction
  - Keyterm extraction
  - Find relation between keyterms

Knowledge Graph



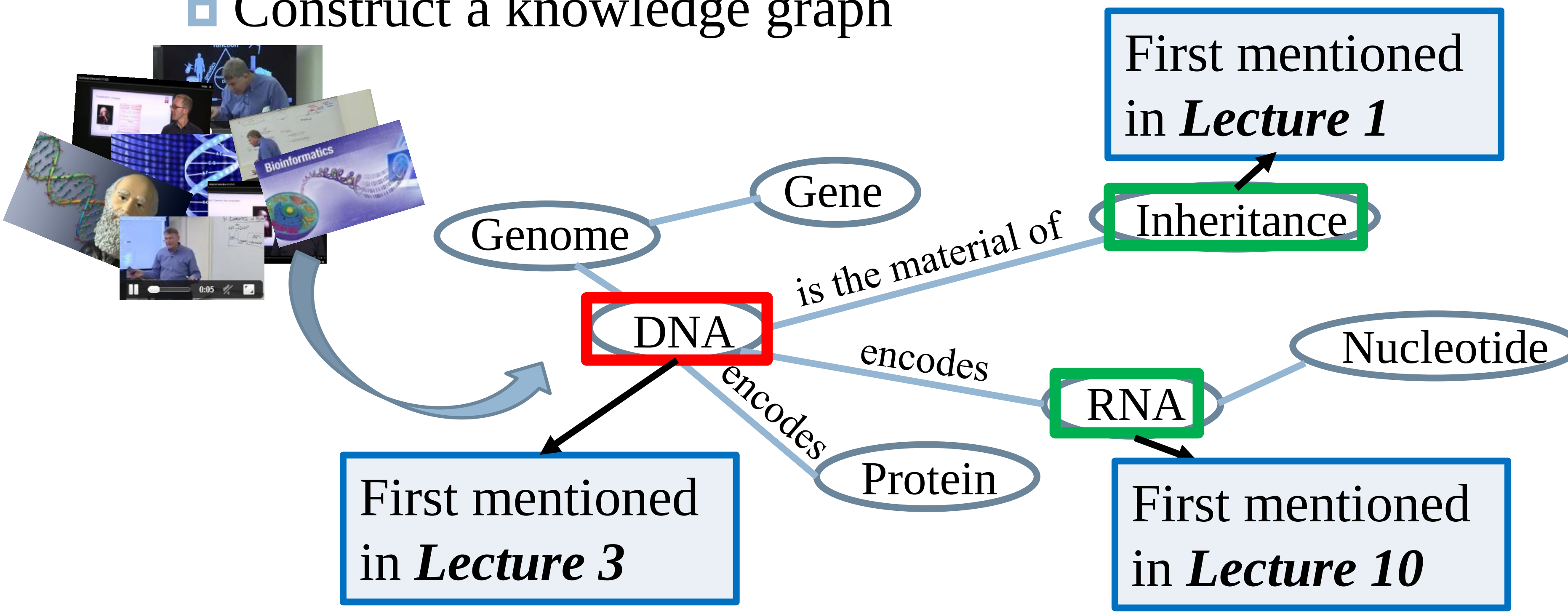
Knowledge Graph

Inferring Learning Path



Inferring Learning Path

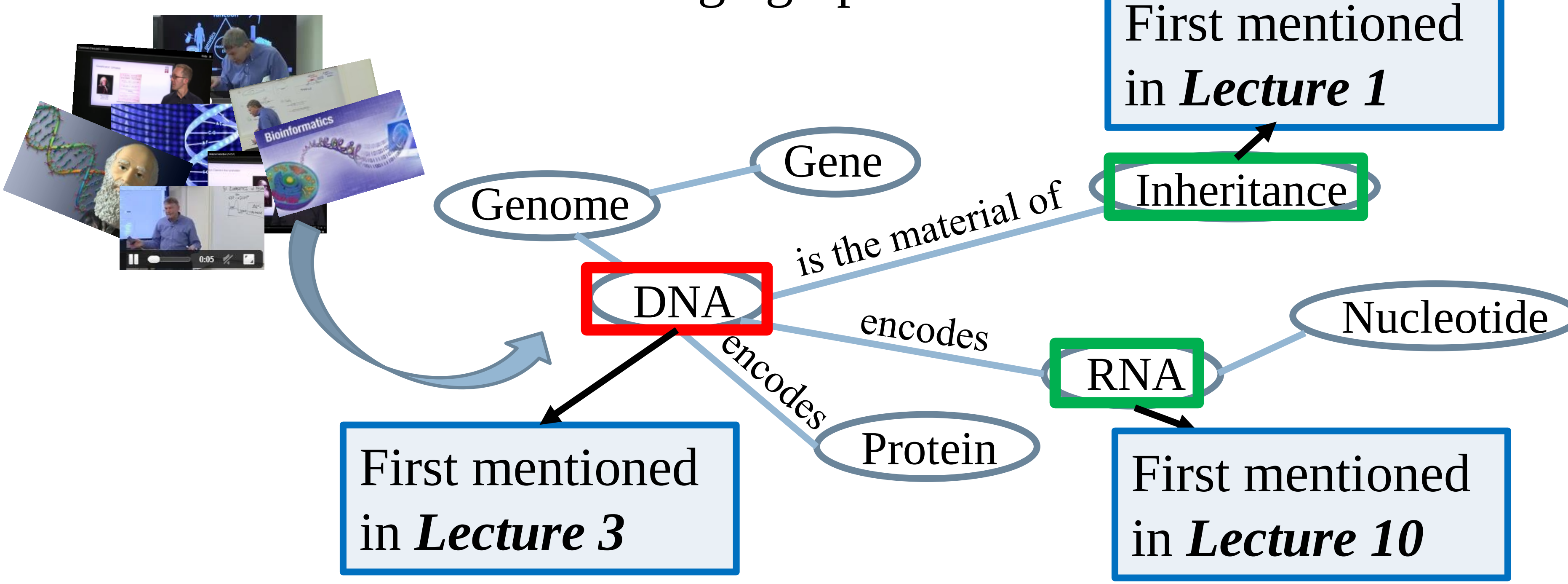
- Inferring prerequisite and advanced concepts
  - Construct a knowledge graph



➢ Analyze the positions where the concepts are mentioned the first time in a course

Inferring Learning Path

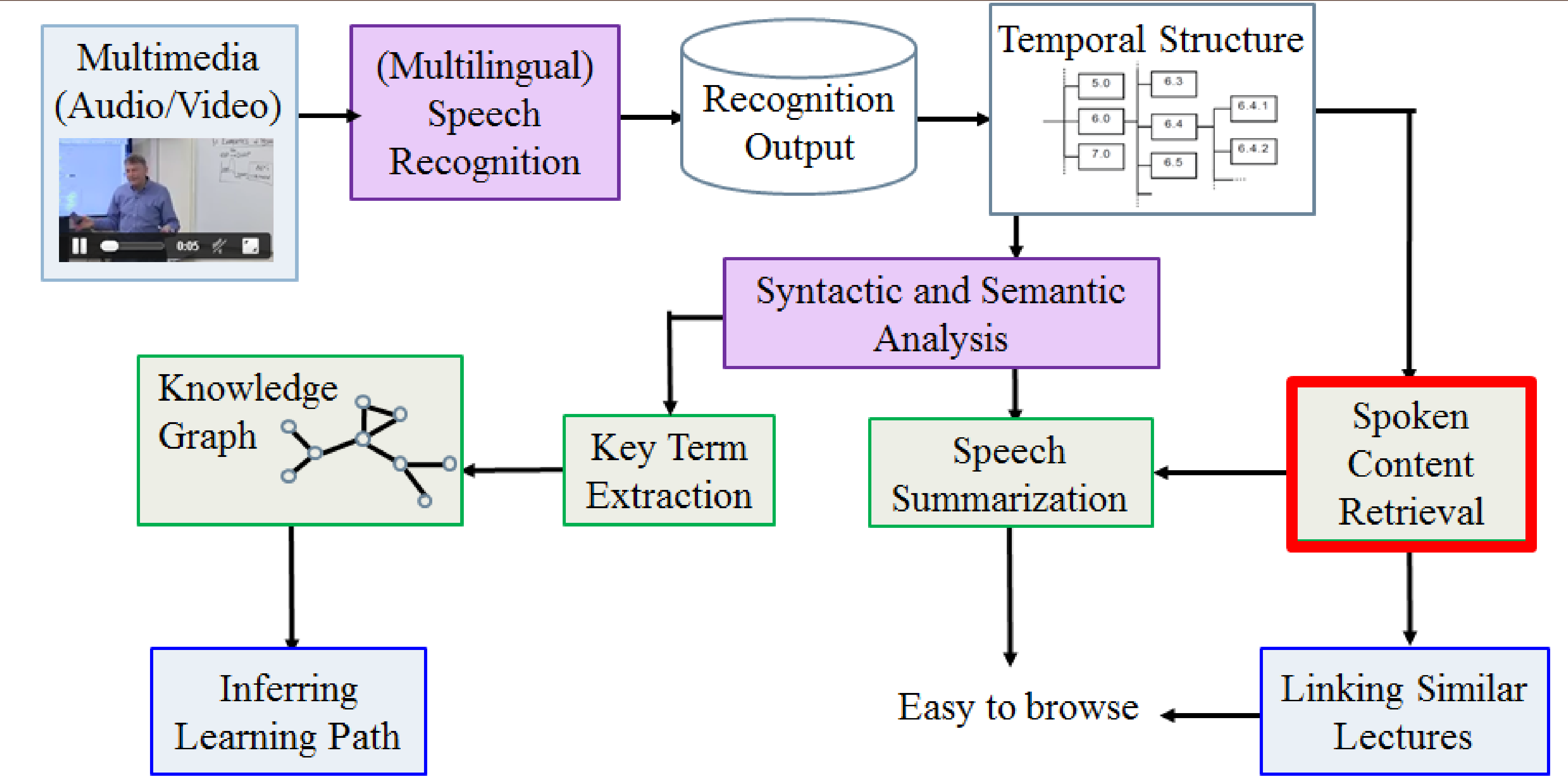
- Inferring prerequisite and advanced concepts
  - Construct a knowledge graph



“Inheritance” is the prerequisite concept of “DNA”

“RNA” is the advanced concept of “DNA”

Part II:
Spoken Content Retrieval



People think ……

Spoken Content Retrieval

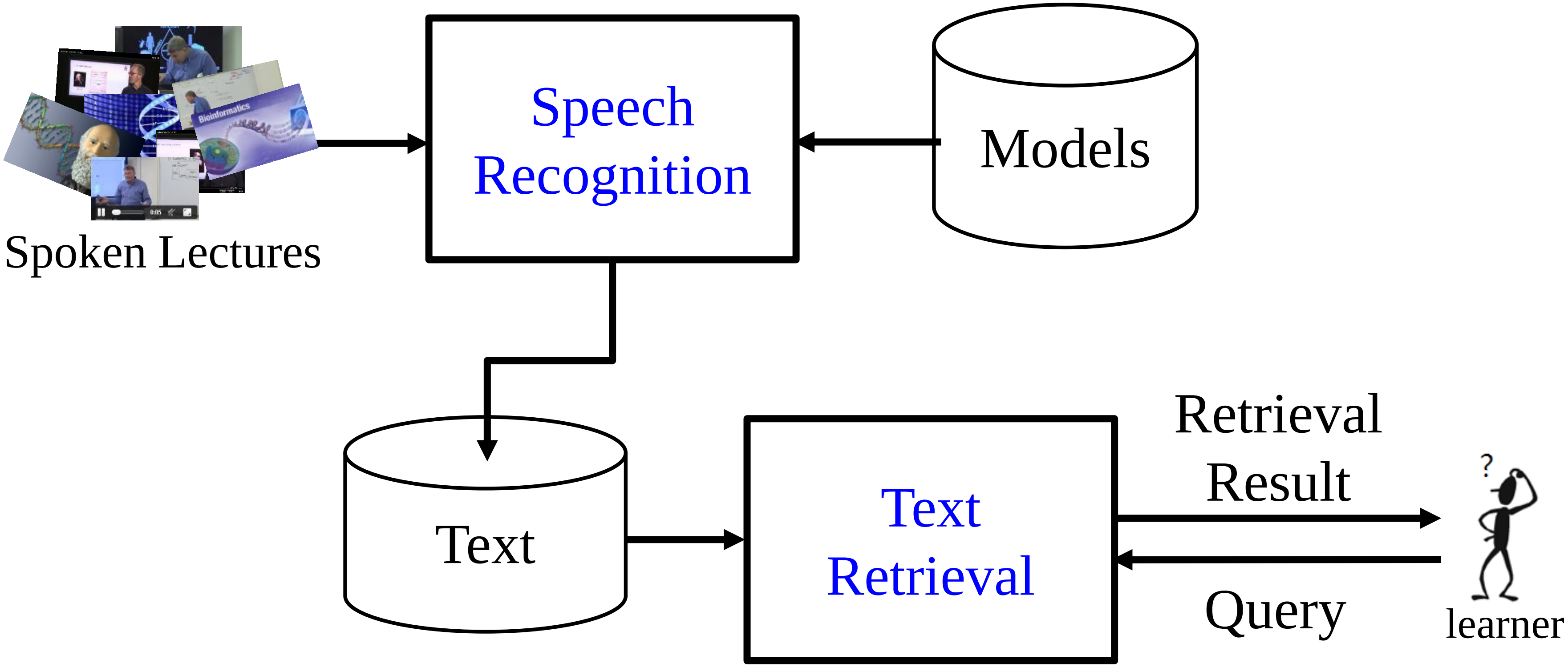
=

Speech Recognition

+

Text Retrieval

Speech Recognition + Text Retrieval

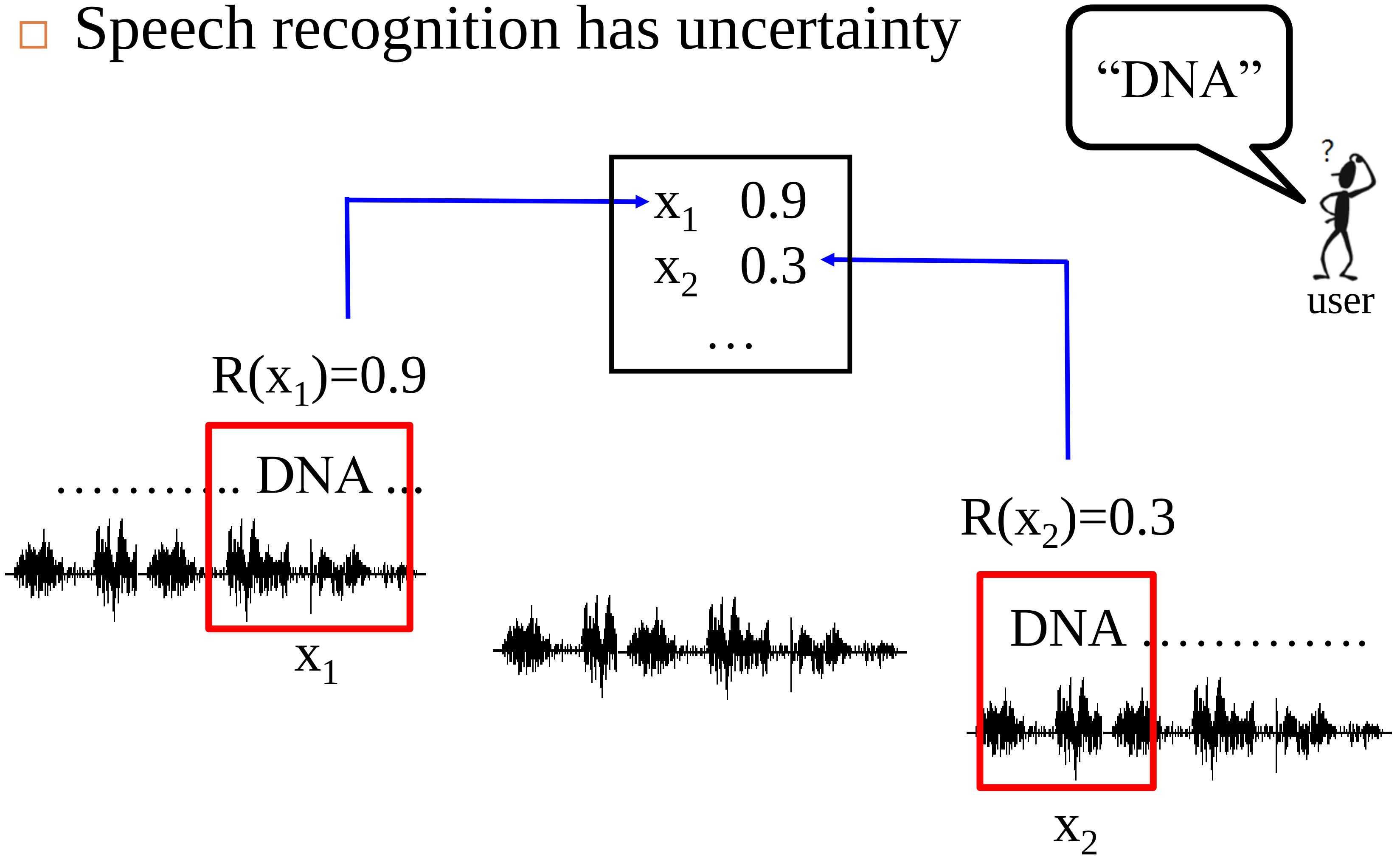


Spoken Content Retrieval

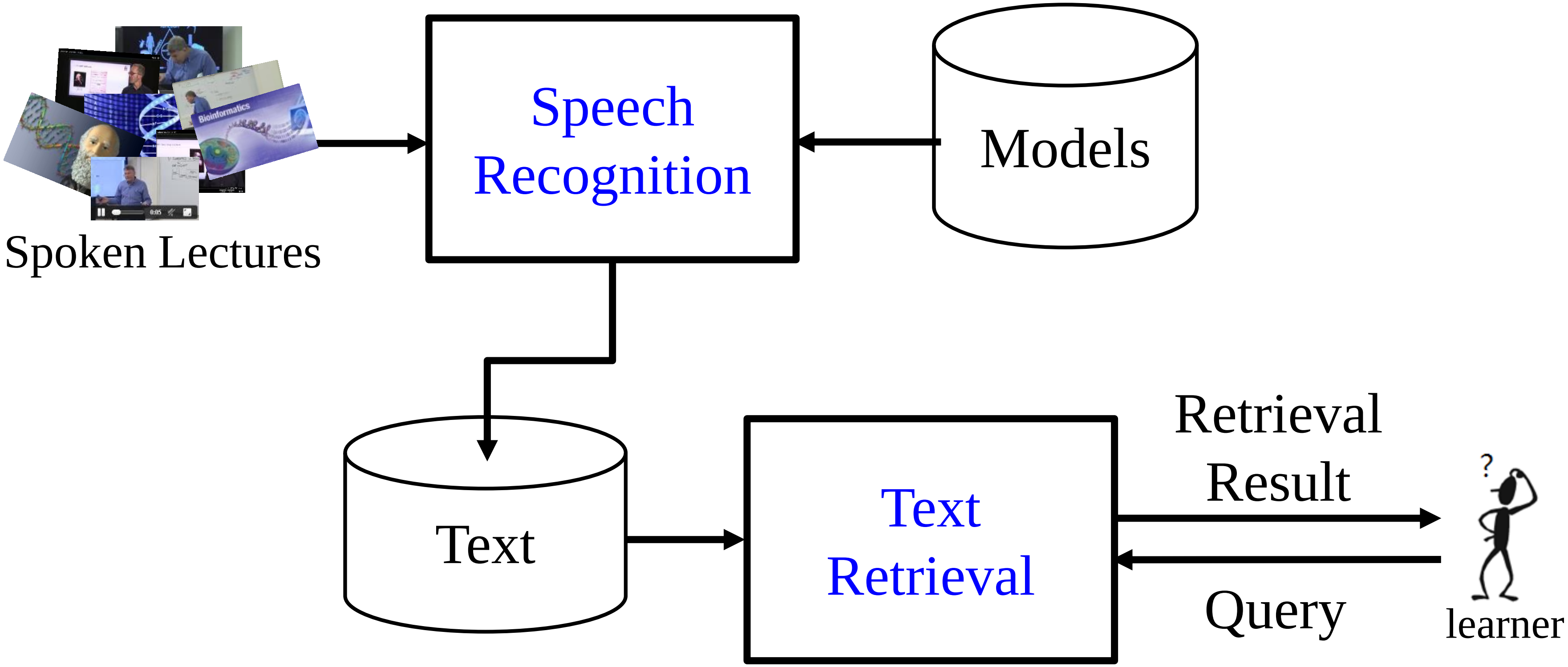
= Speech Recognition + Text Retrieval

Speech Recognition + Text Retrieval

- Speech recognition has uncertainty

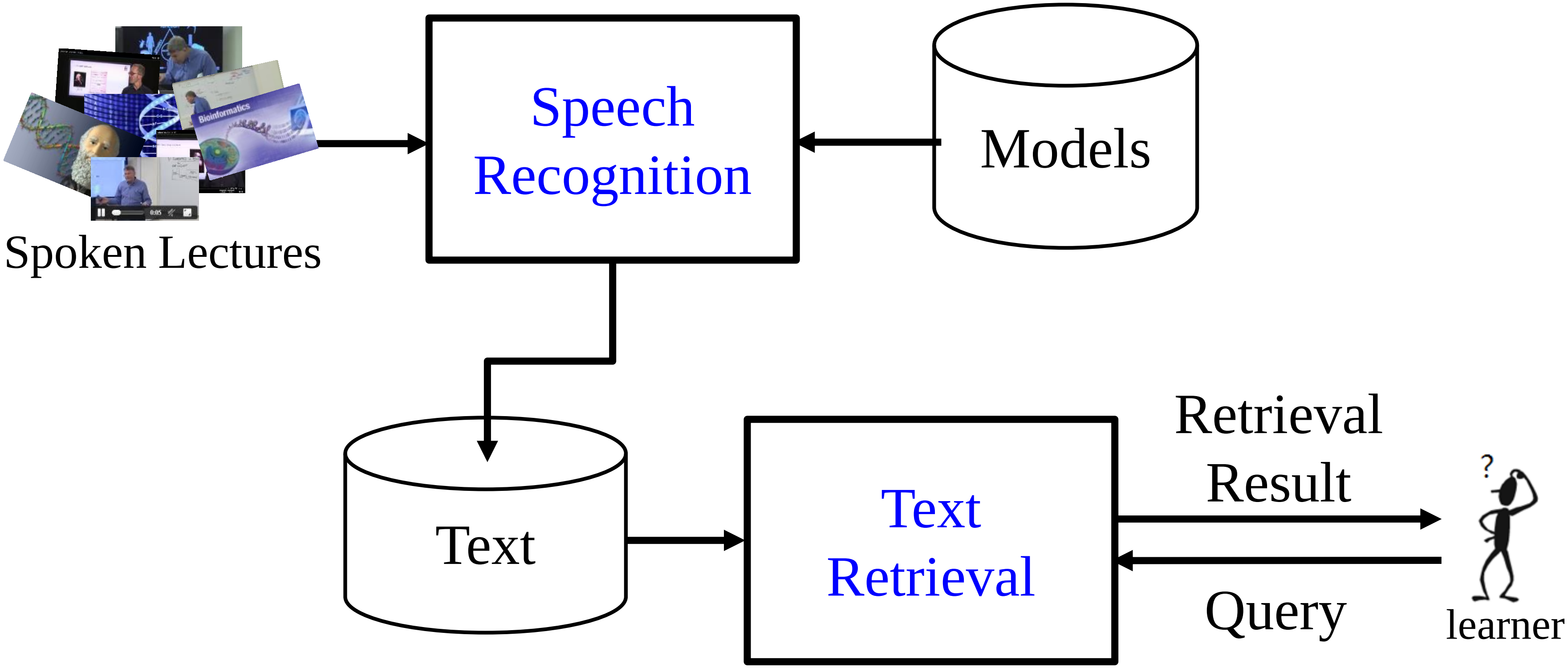


Is the problem solved?



- The retrieval performance seriously degrades with inevitable recognition errors.
- In real application, speech recognition accuracy can be low.

Is the problem solved?



- To make retrieval performance less limited by recognition errors
- We need new ideas beyond cascading speech recognition and text retrieval.

My point

Spoken Content Retrieval

$\cancel{=}$

Speech Recognition

+

Text Retrieval

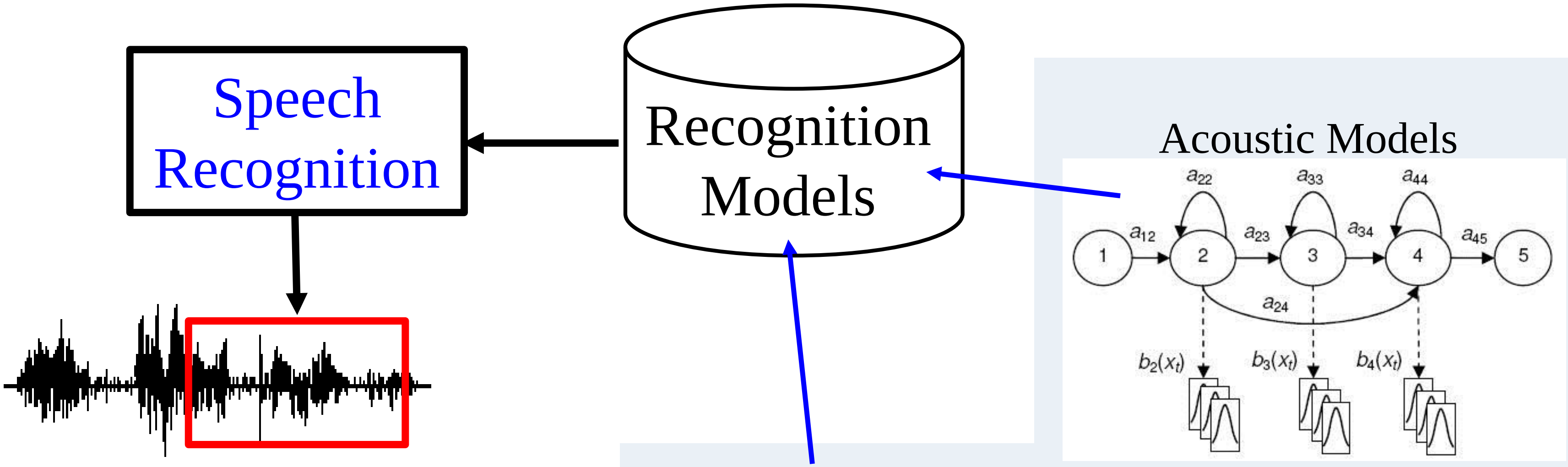
Beyond Cascading Speech Recognition and Text Retrieval

- Incorporating Information Lost in Standard Speech Recognition
- Improving Recognition Models by User Relevance feedback
- Query Expansion with Speech Signals
- Spoken Content Retrieval without Speech Recognition
- Interactive Retrieval

Beyond Cascading Speech Recognition and Text Retrieval

- Incorporating Information Lost in Standard Speech Recognition
- Improving Recognition Models by User Relevance feedback
- Query Expansion with Speech Signals
- Spoken Content Retrieval without Speech Recognition
- Interactive Retrieval

Similarity



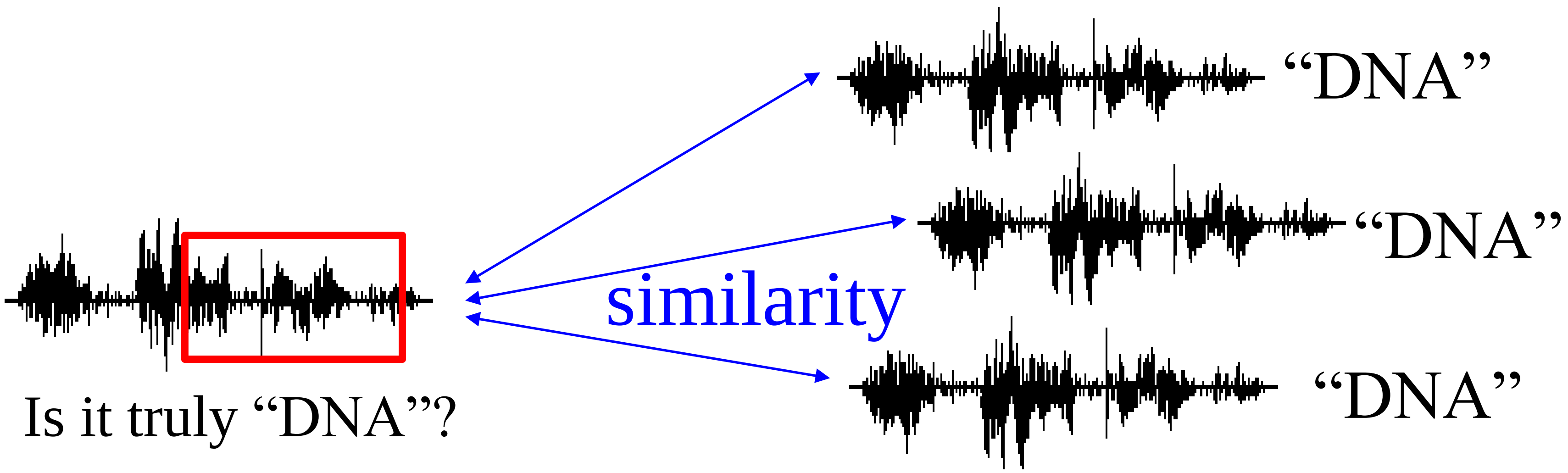
Is it truly “DNA”?

Language Model

$$P(w_i|w_{i-(n-1)},\ldots,w_{i-1}) = \frac{count(w_{i-(n-1)},\ldots,w_{i-1},w_i)}{count(w_{i-(n-1)},\ldots,w_{i-1})}$$

Lots of inaccurate assumption

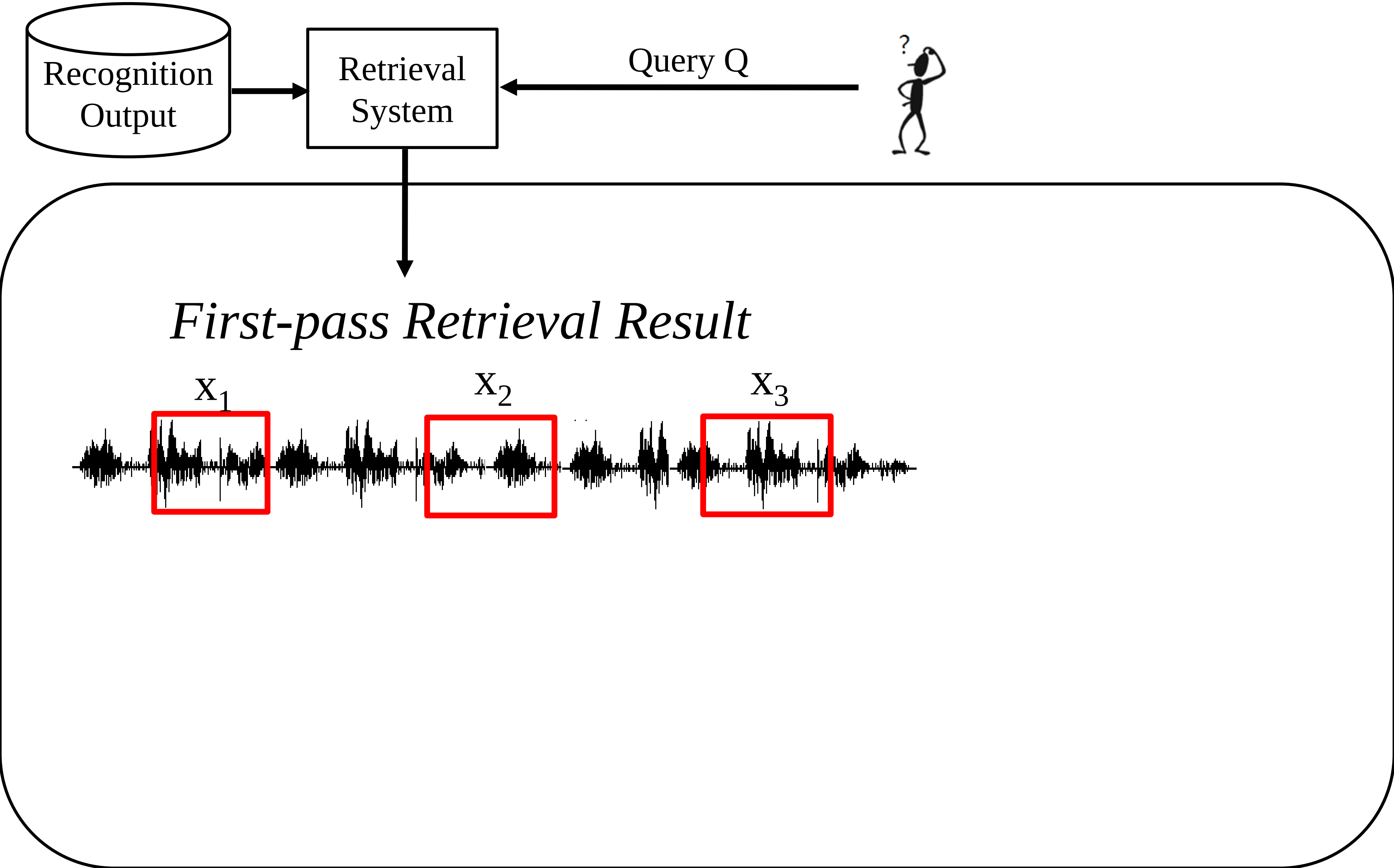
Similarity



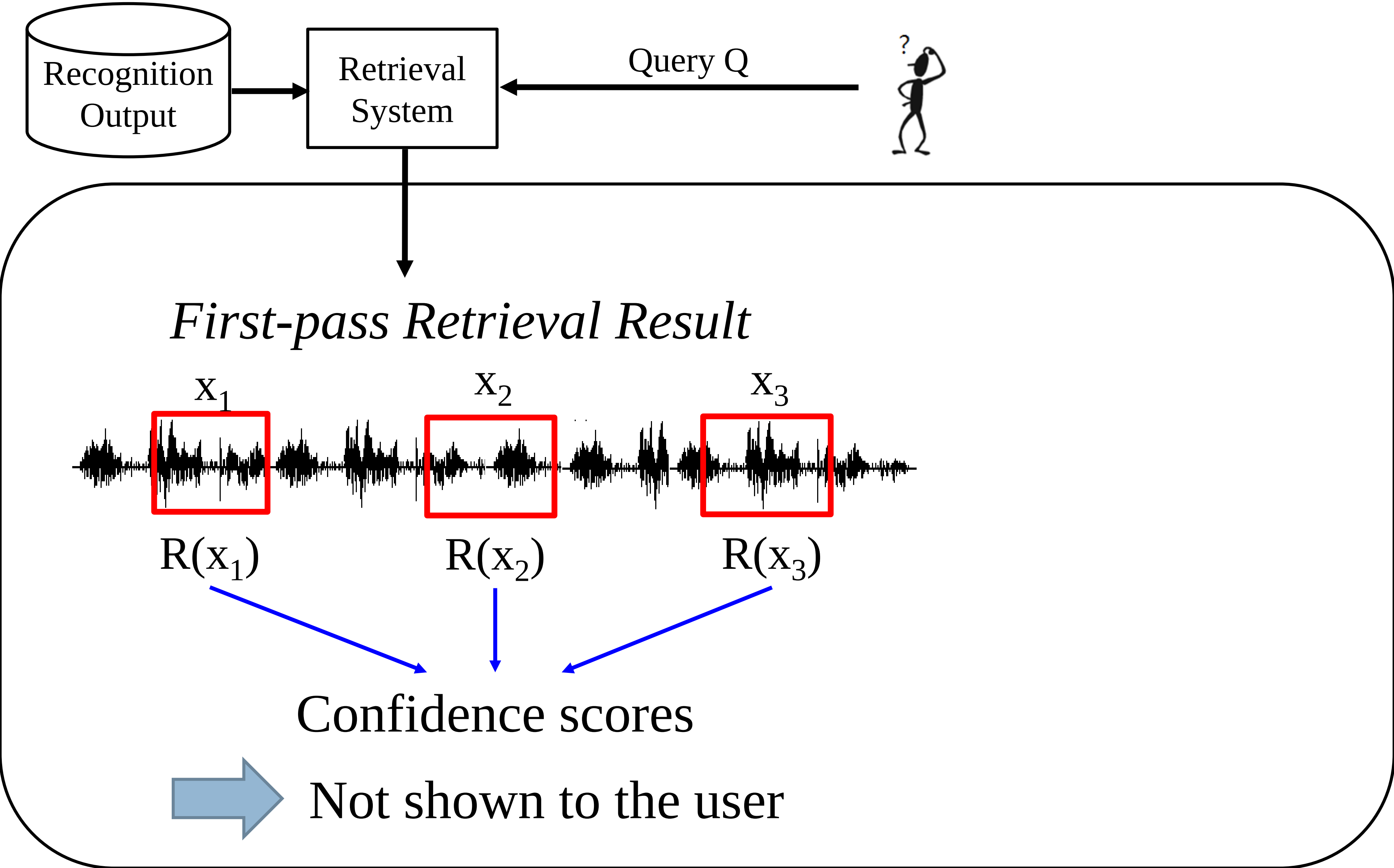
It is not realistic to find examples for all queries.

Use Pseudo-relevance Feedback (PRF)

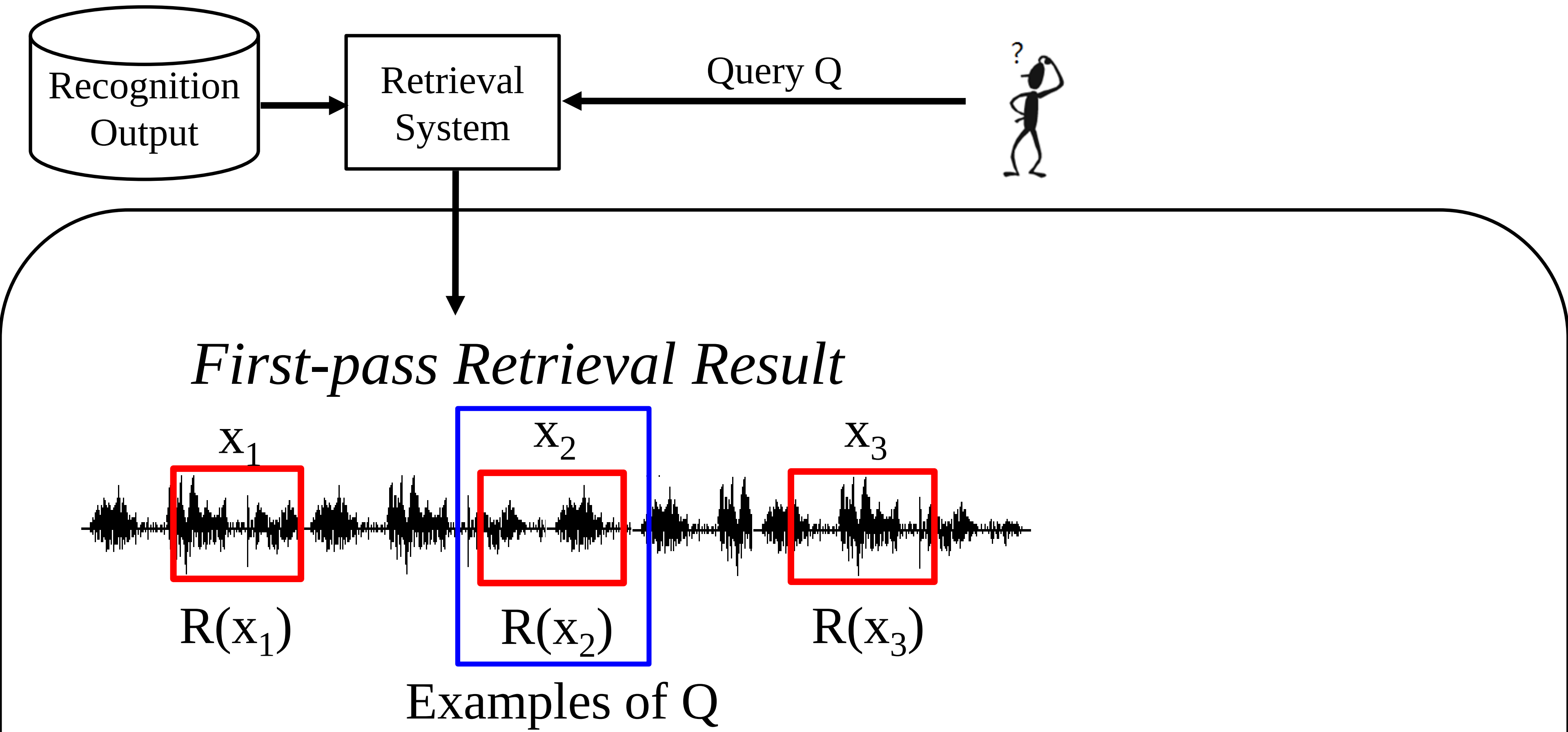
Pseudo Relevance Feedback (PRF)



Pseudo Relevance Feedback (PRF)



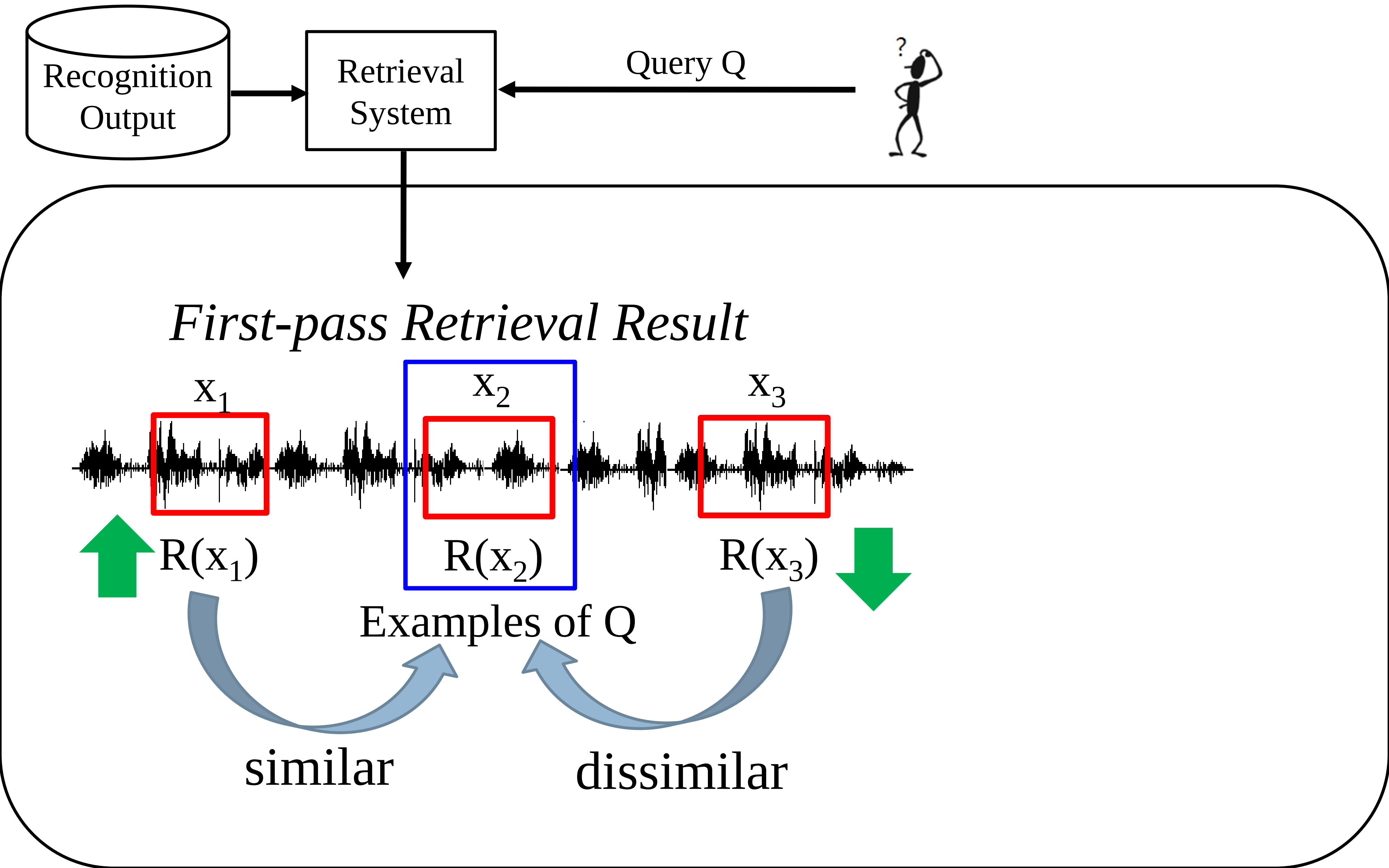
Pseudo Relevance Feedback (PRF)



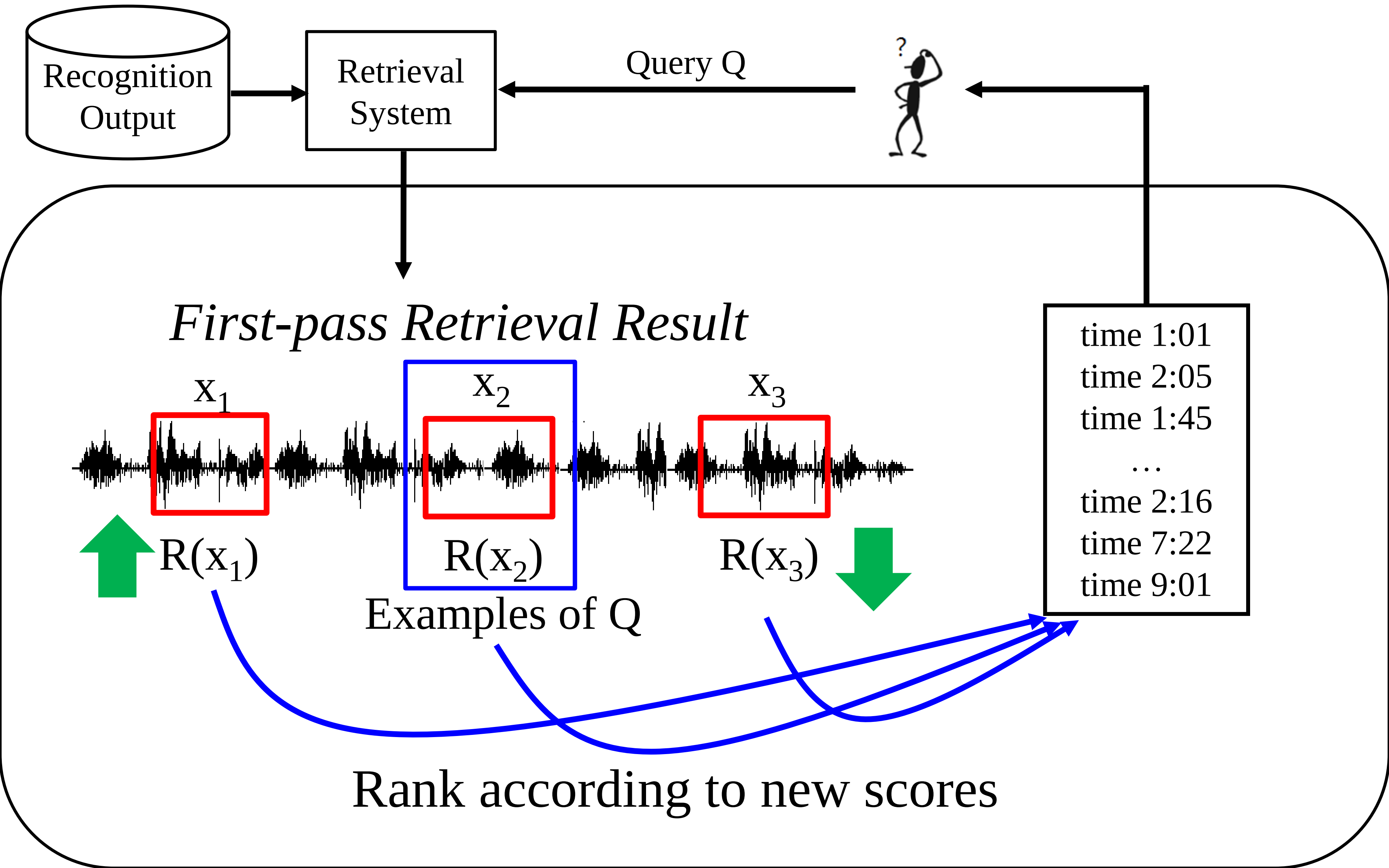
Assume the result with high confidence scores as correct

Considered as examples of Q

Pseudo Relevance Feedback (PRF)

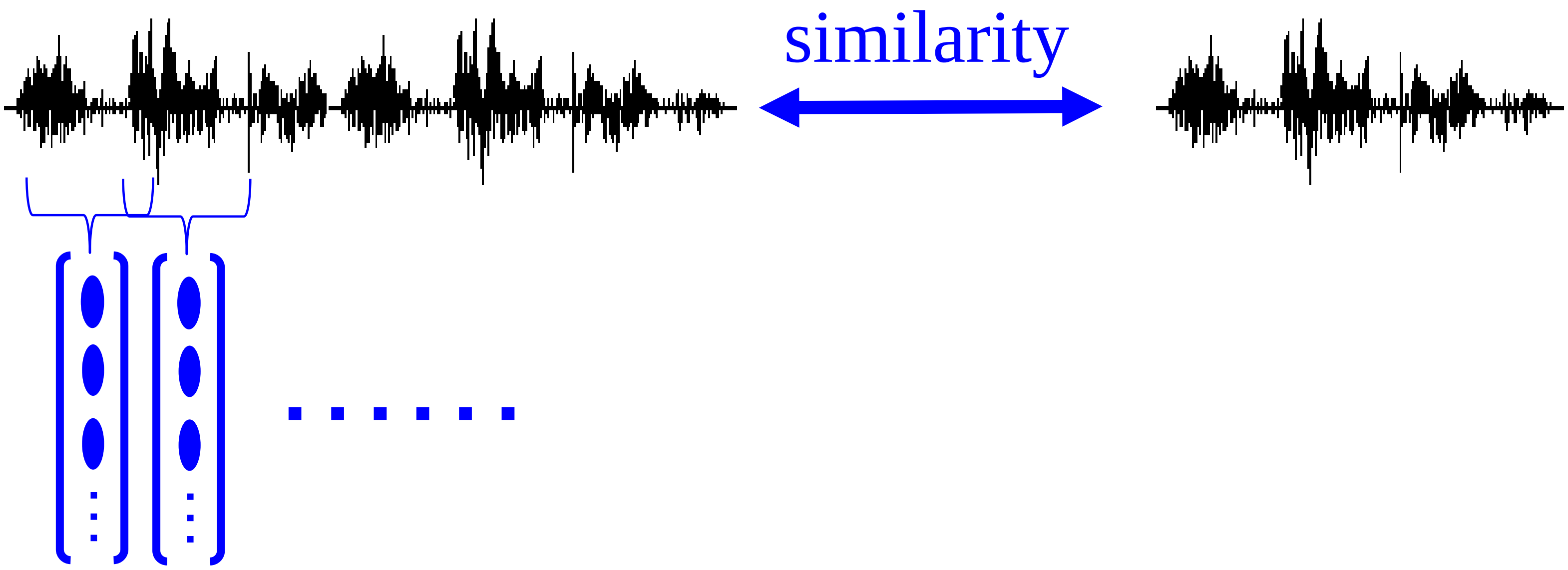


Pseudo Relevance Feedback (PRF)



Similarity between Audio Segments

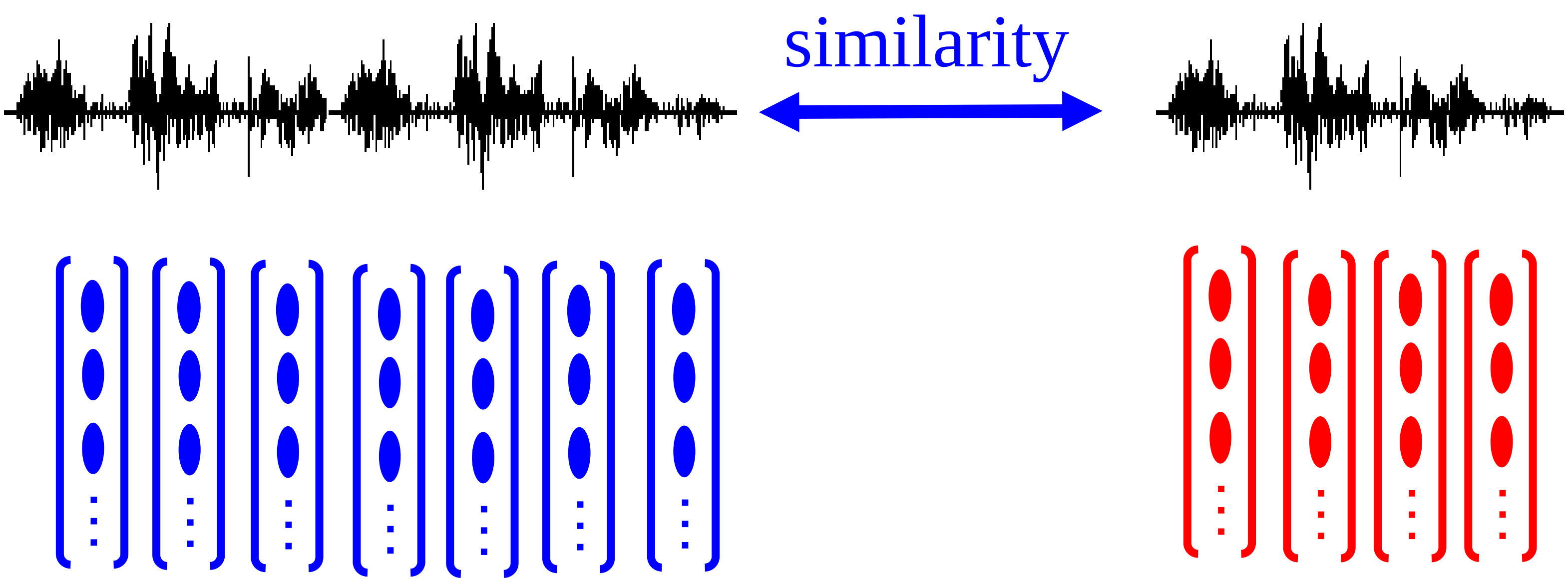
- How to compute the similarity of two audio segments?



Use a feature vector to present a short time span.

Similarity between Audio Segments

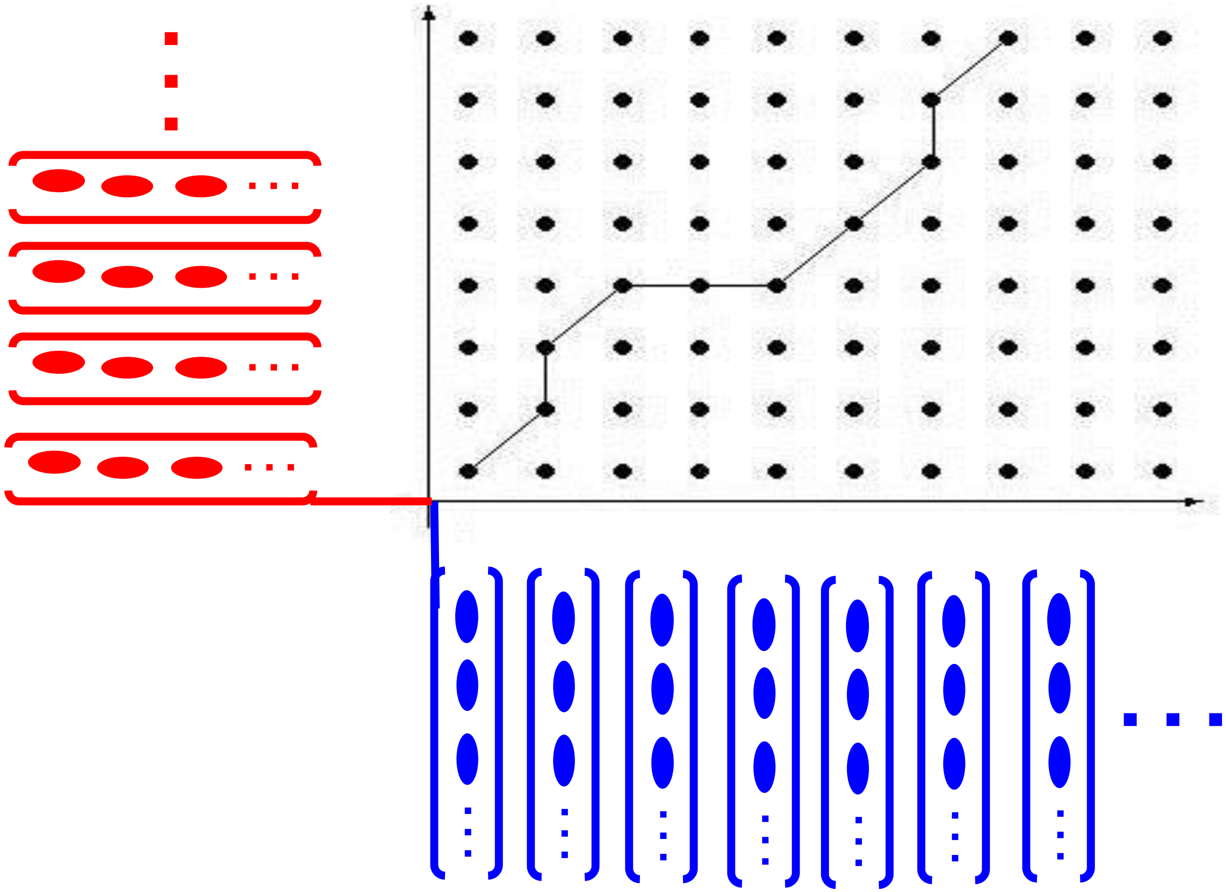
- How to compute the similarity of two audio segments?



A audio segment is a sequence of feature vectors.

Similarity between Audio Segments

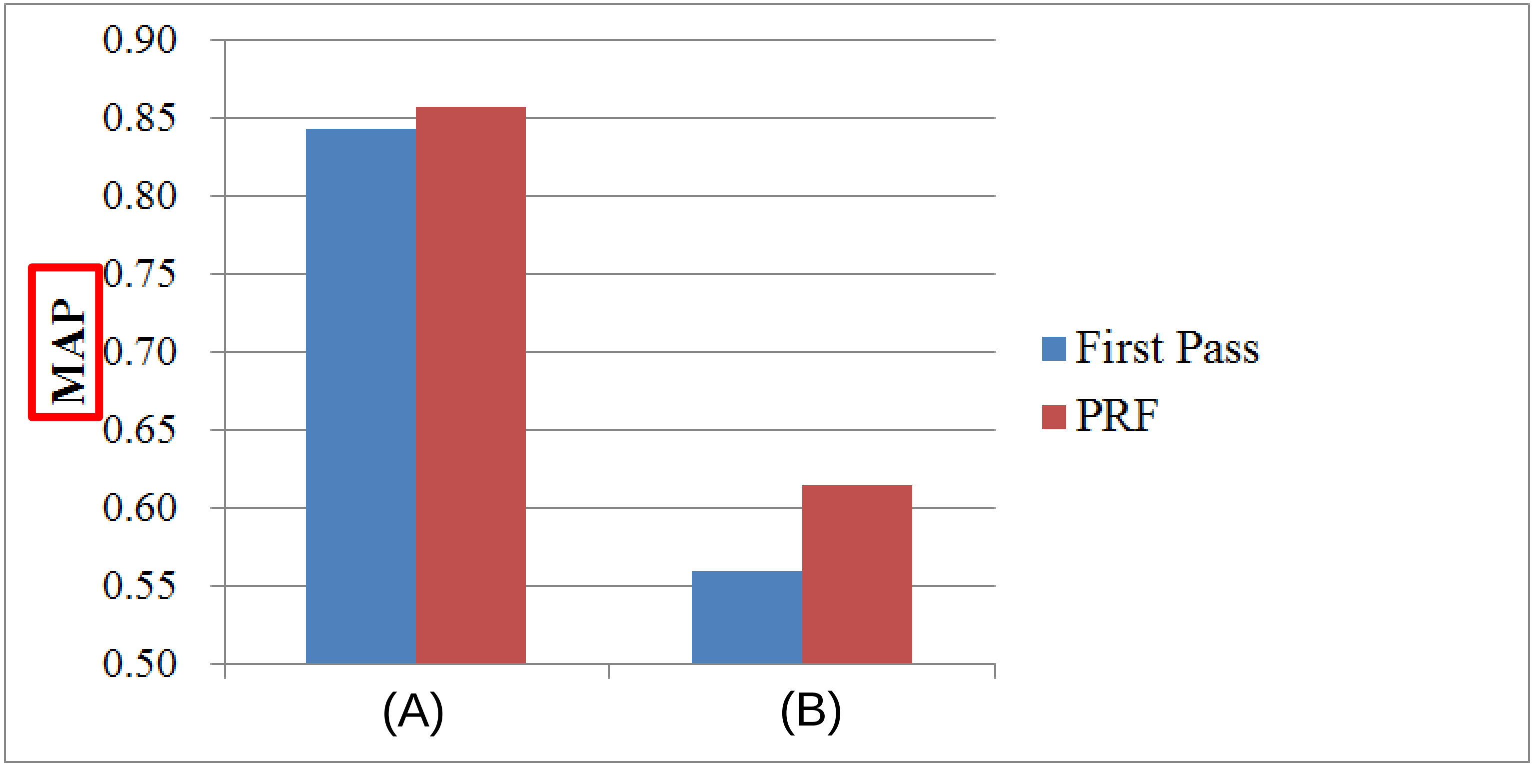
Dynamic Time Warping (DTW)



Pseudo Relevance Feedback (PRF) - Experiments

- Digital Speech Processing (DSP) of NTU based on lattices

Evaluation Measure: MAP (Mean Average Precision)

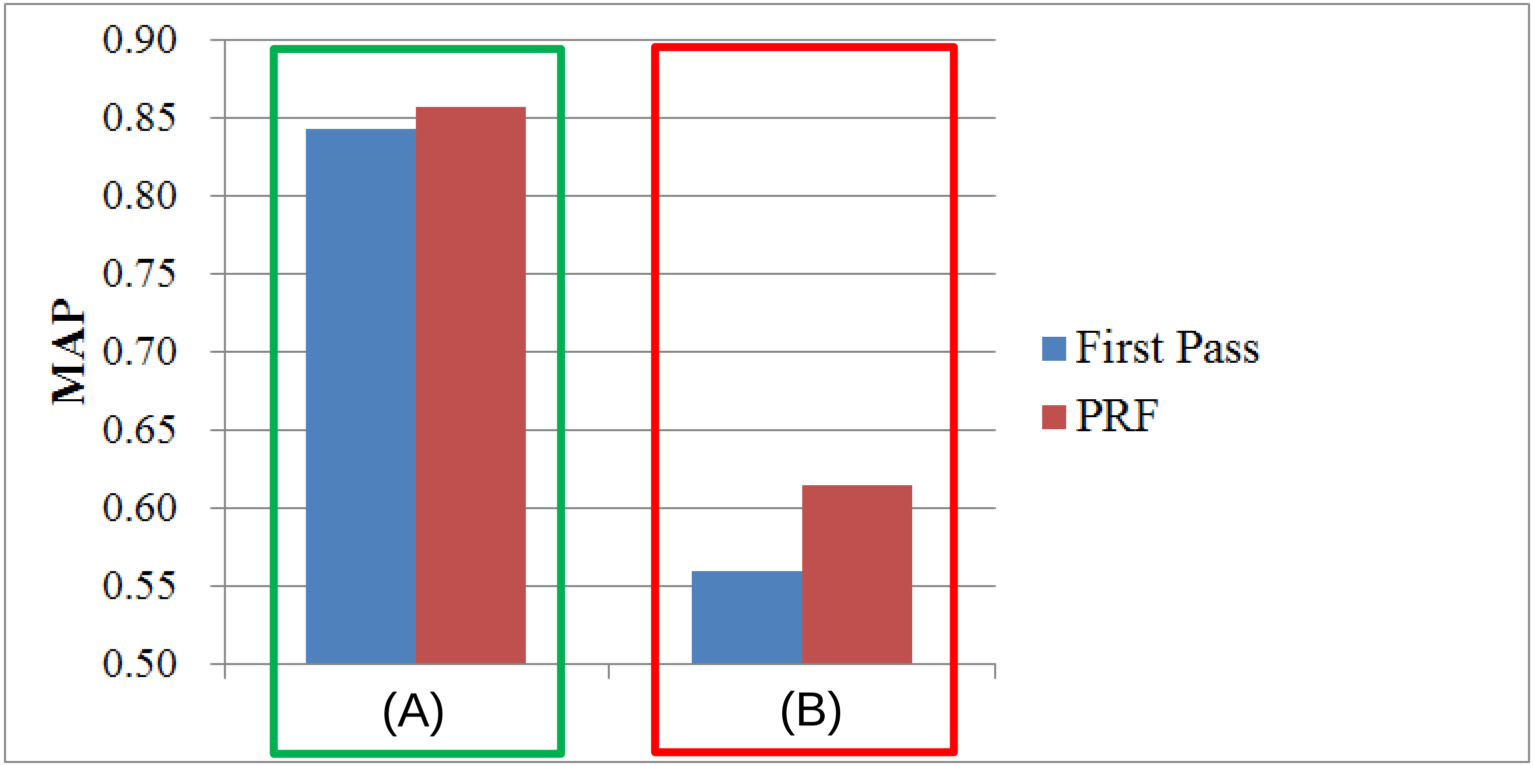


Pseudo Relevance Feedback (PRF) - Experiments

(A) and (B) use different speech recognition systems

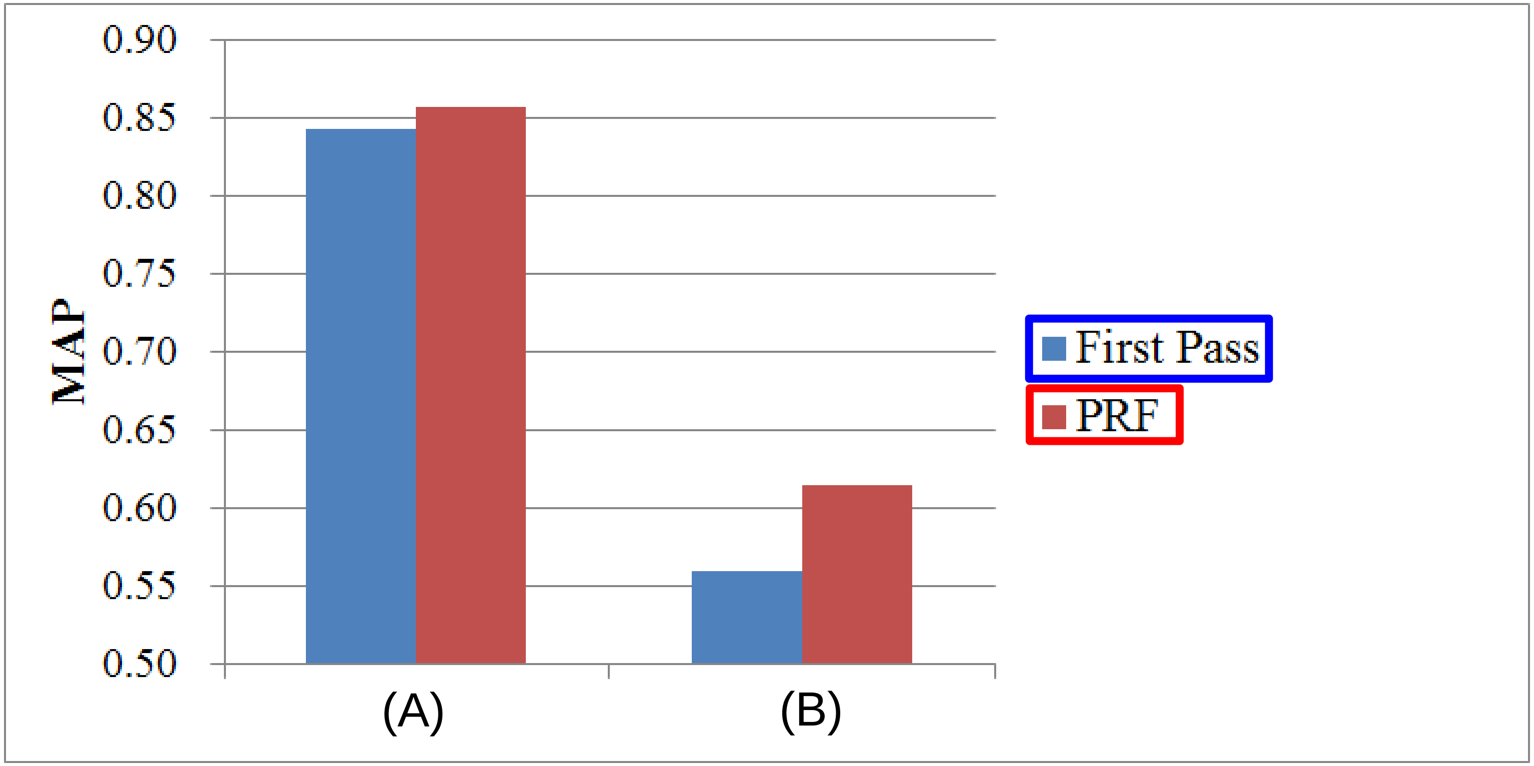
(A): speaker dependent (84% recognition accuracy)

(B): speaker independent (50% recognition accuracy)



Pseudo Relevance Feedback (PRF) - Experiments

- PRF (red bars) improved the first-pass retrieval results with lattices (blue bars)

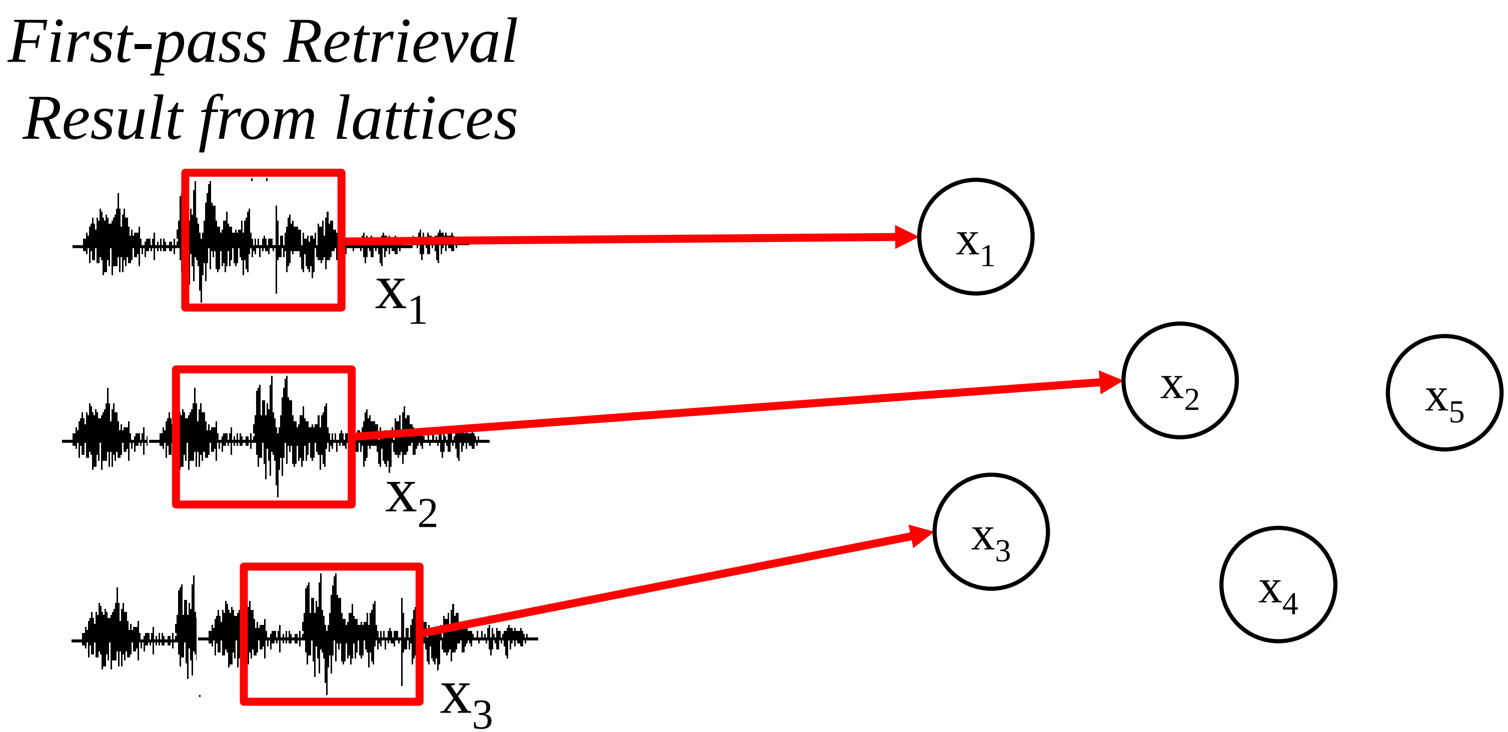


Graph-based Approach

- In PRF, each result considers the similarity to some examples
- Consider the similarity between all results
- Formulated as a problem on graph

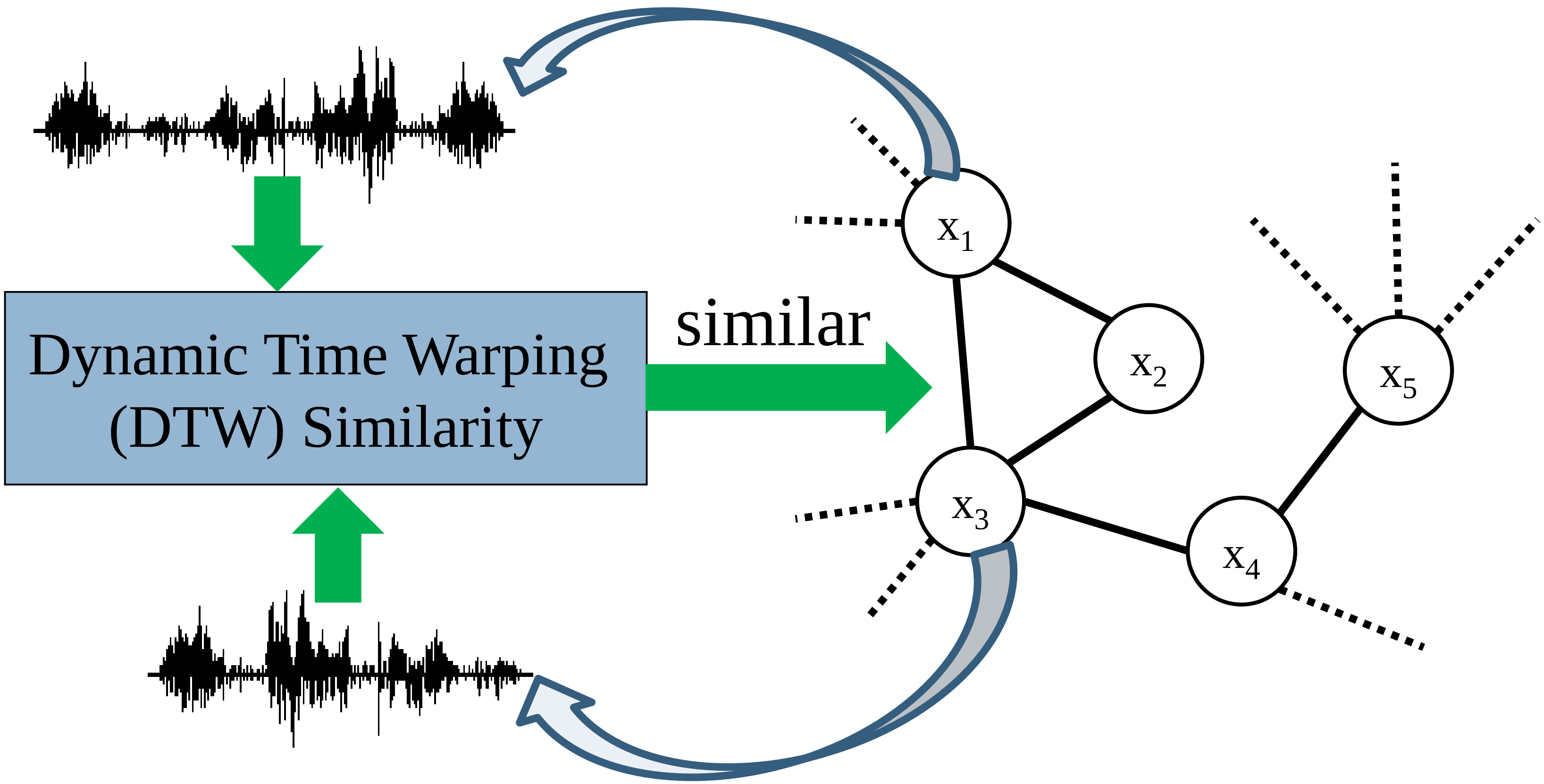
Graph Construction

- The first-pass results is considered as a graph.
  - Each retrieval result is a node



Graph Construction

- The first-pass results is considered as a graph.
  - Nodes are connected if their retrieval results are similar.

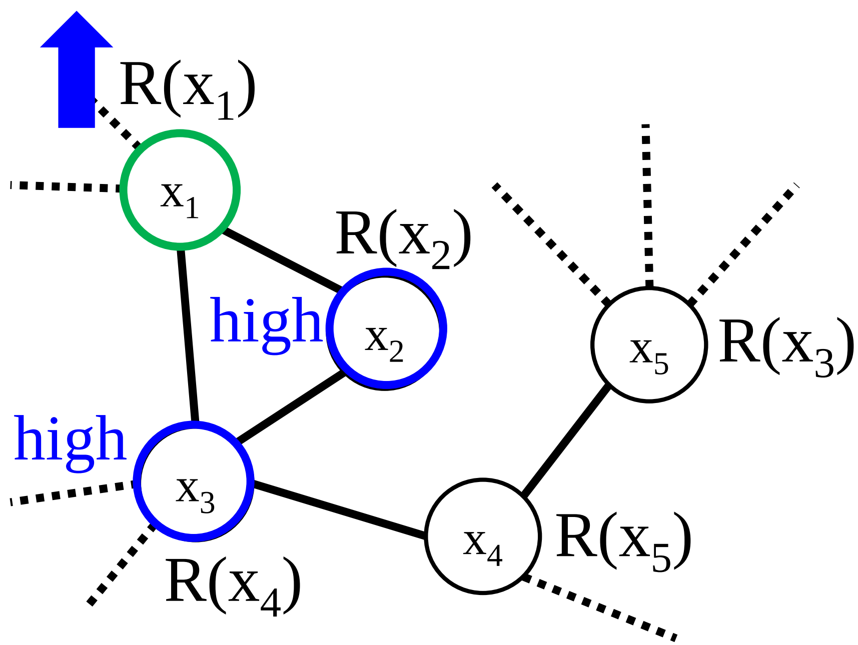


Changing Confidence Scores by Graph

- The score of each node depends on its neighbors.

近朱者赤

近墨者黑

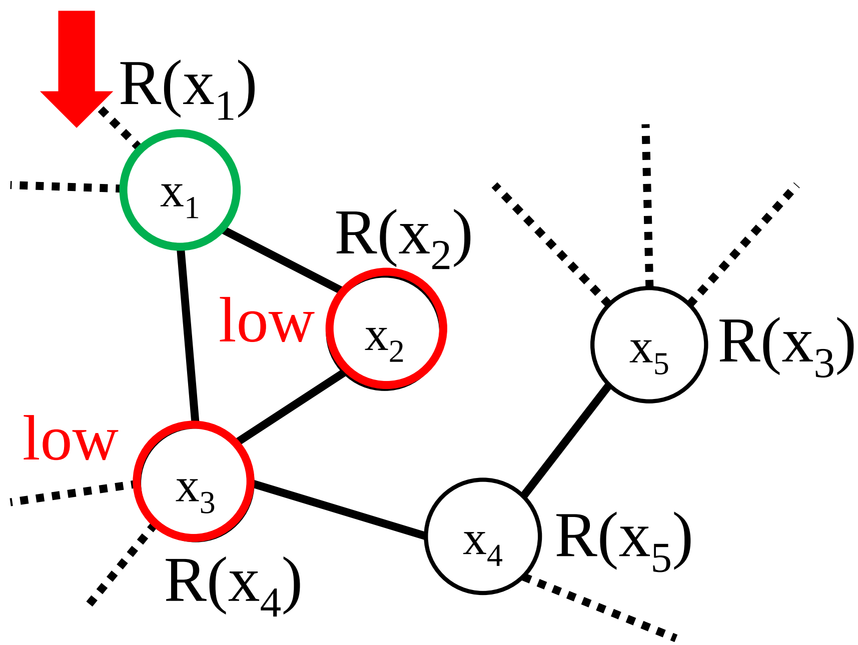


Changing Confidence Scores by Graph

- The score of each node depends on its neighbors.

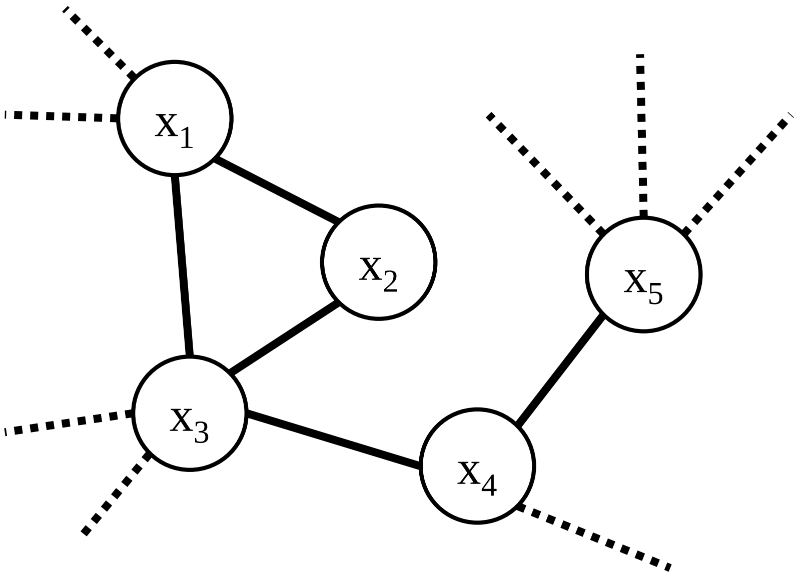
近朱者赤

近墨者黑



Changing Confidence Scores by Graph

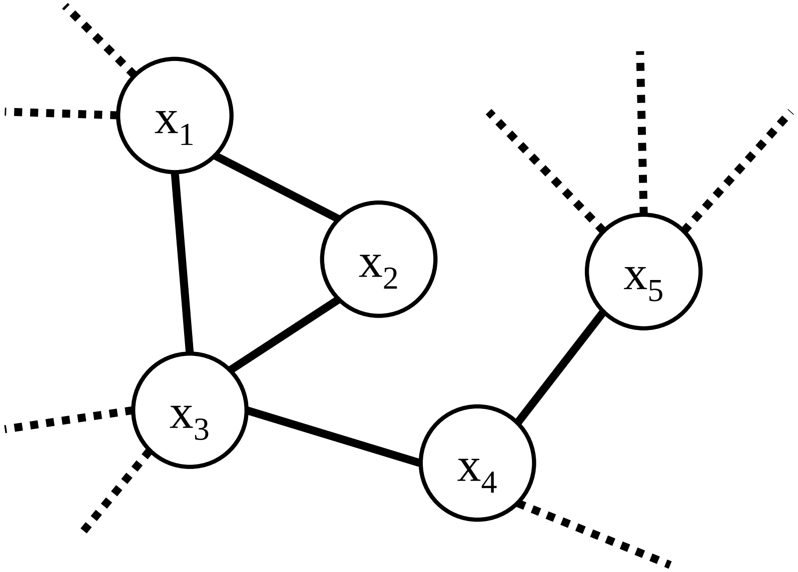
- The score of each node depends on its connected nodes.



- Score of x_1 depends on the scores of x_2 and x_3

Changing Confidence Scores by Graph

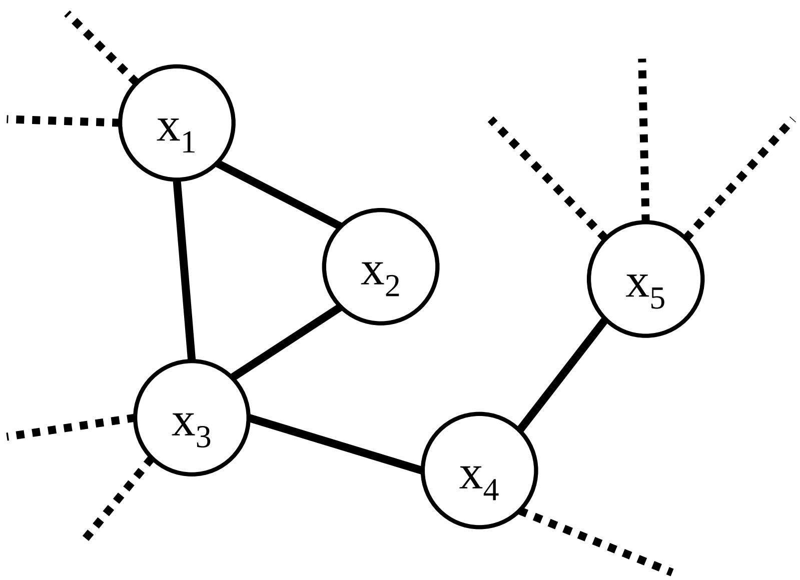
- The score of each node depends on its connected nodes.



- Score of x_1 depends on the scores of x_2 and x_3
- Score of x_2 depends on the scores of x_1 and x_3

Changing Confidence Scores by Graph

- The score of each node depends on its connected nodes.



- Score of x_1 depends on the scores of x_2 and x_3
- Score of x_2 depends on the scores of x_1 and x_3
- ……

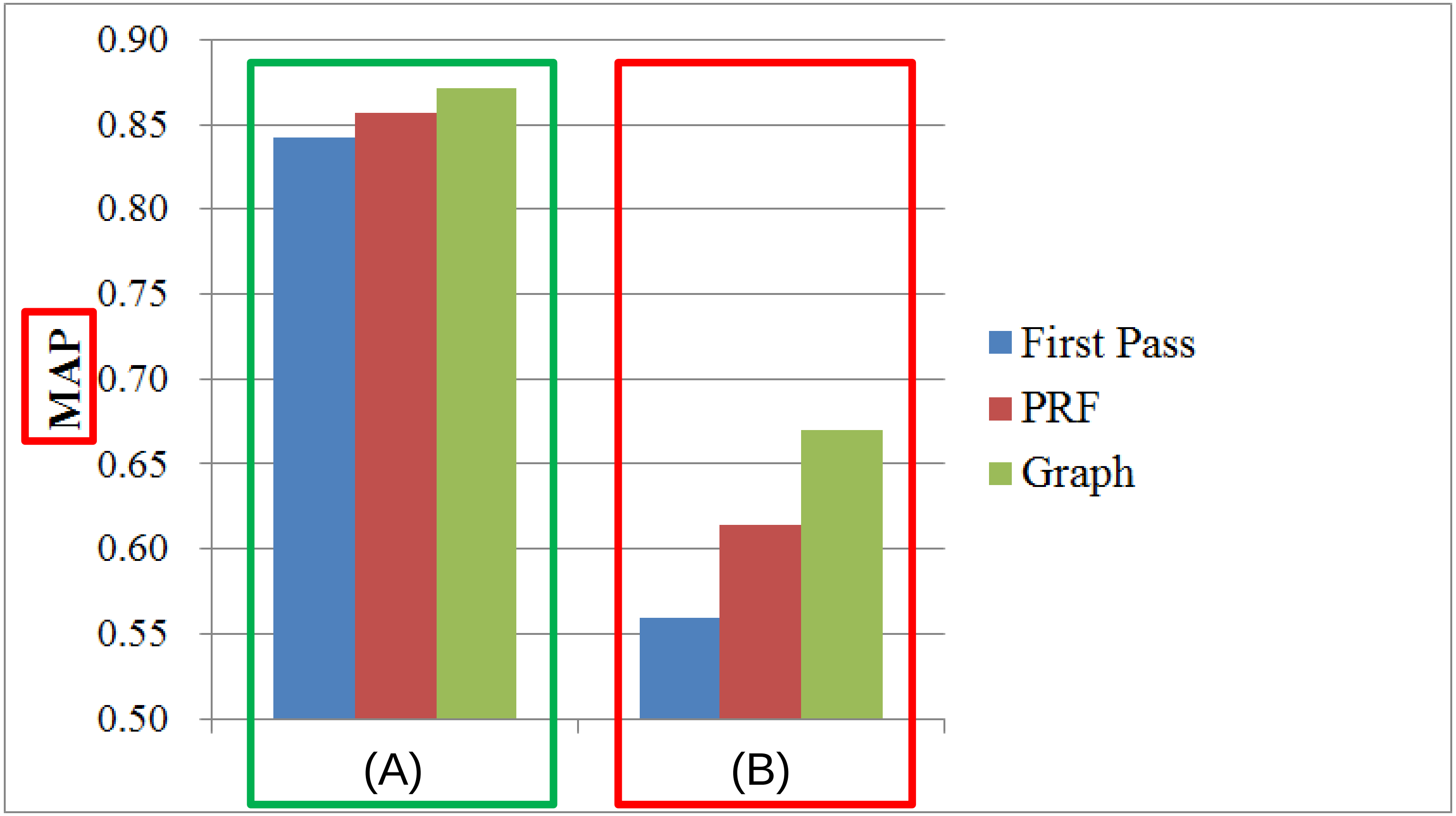
The scores are found by ***random walk*** algorithm.

Graph-based Approach - Experiments

- Digital Speech Processing (DSP) of NTU based on lattices

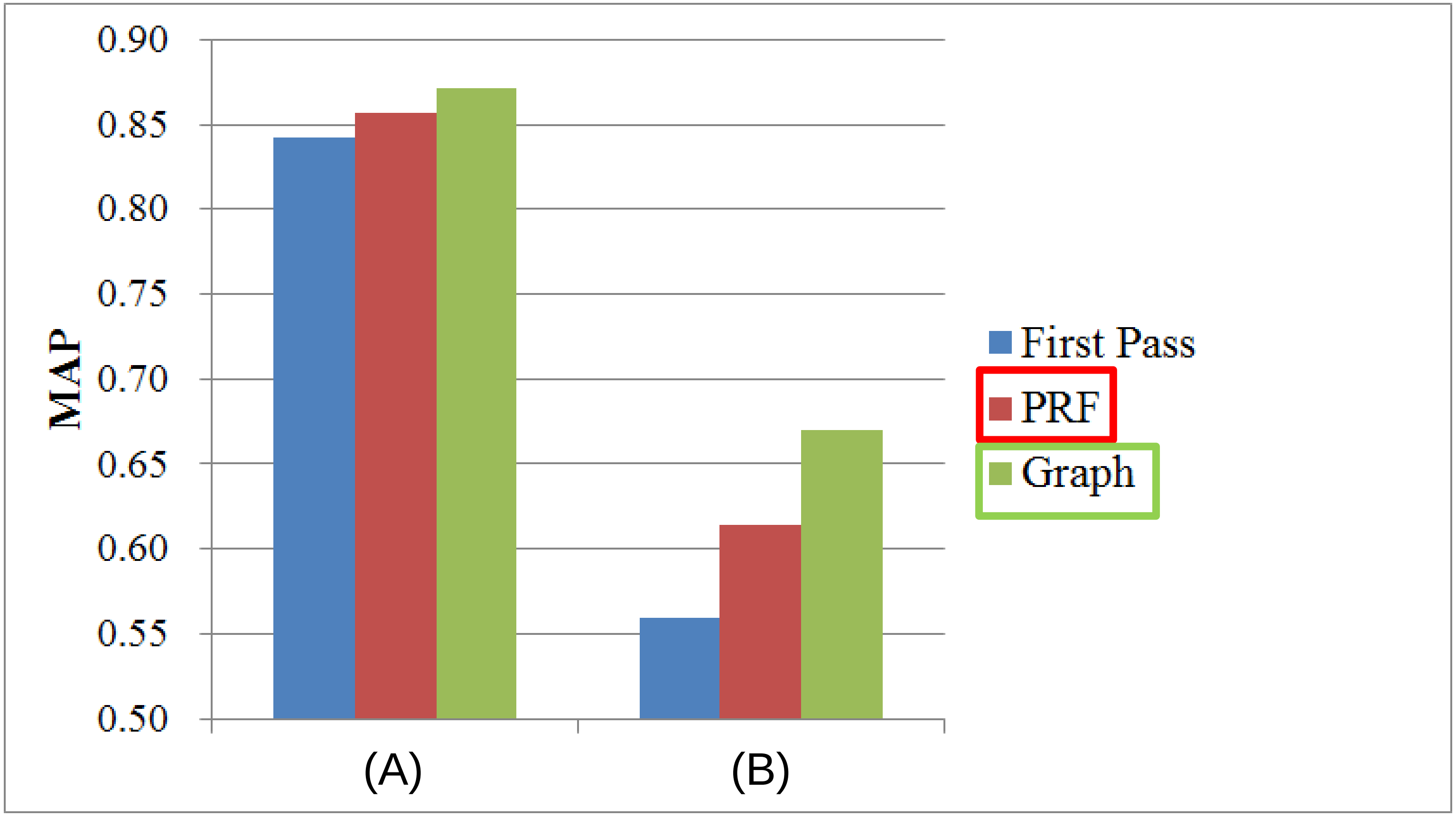
  (A): speaker dependent (high recognition accuracy)

  (B): speaker independent (low recognition accuracy)



Graph-based Approach - Experiments

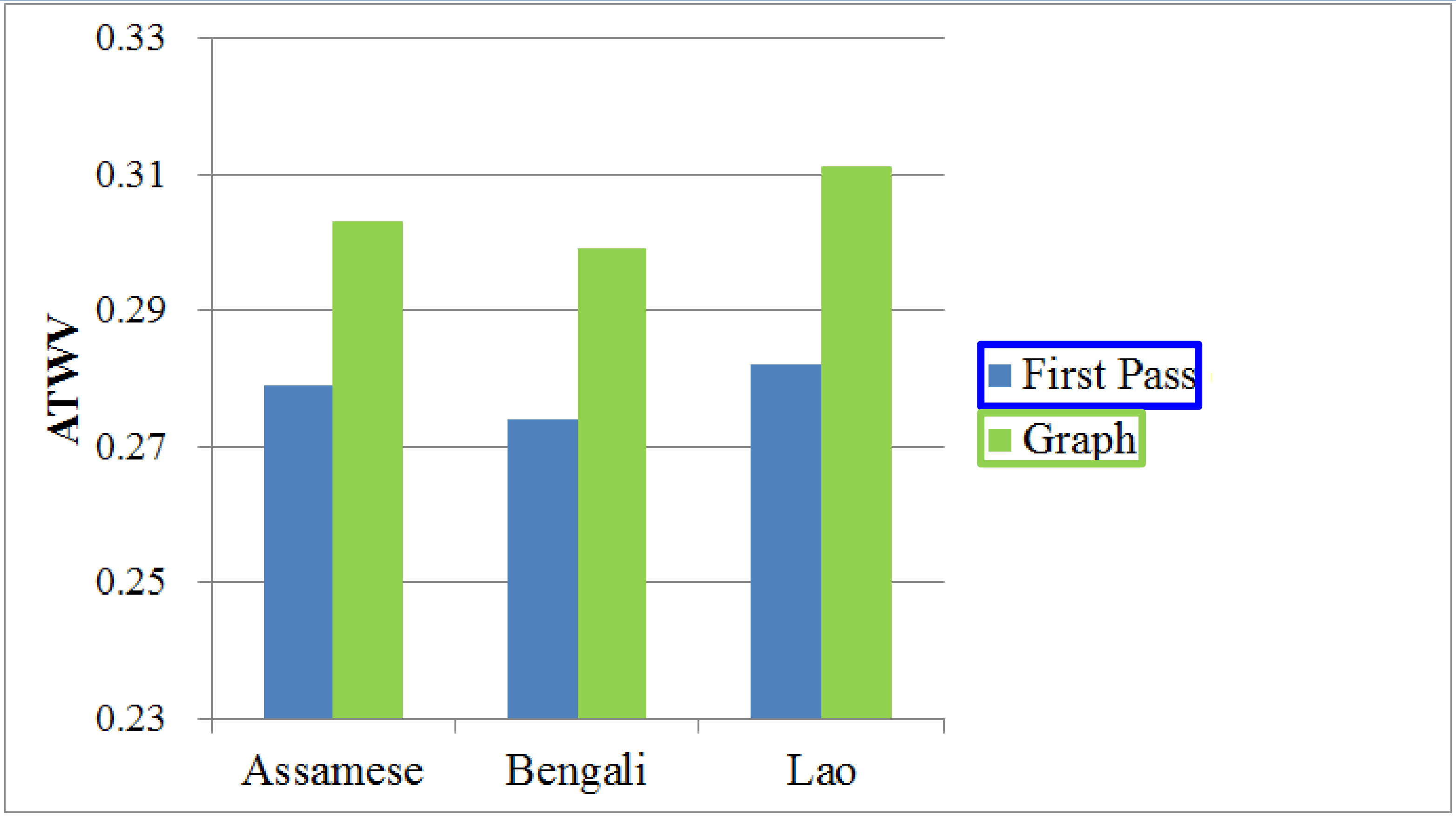
- Graph-based re-ranking (green bars) outperformed PRF (red bars)



Graph-based Approach – Experiments on Babel Program

- Join Babel program（巴別塔計畫）at MIT
- Evaluation program of spoken term detection
  - More than 30 research groups divided into 4 teams
  - Spoken content to be retrieved are in special languages

Graph-based Approach – Experiments on Babel Program



Speech recognition system is based on deep neutral networks

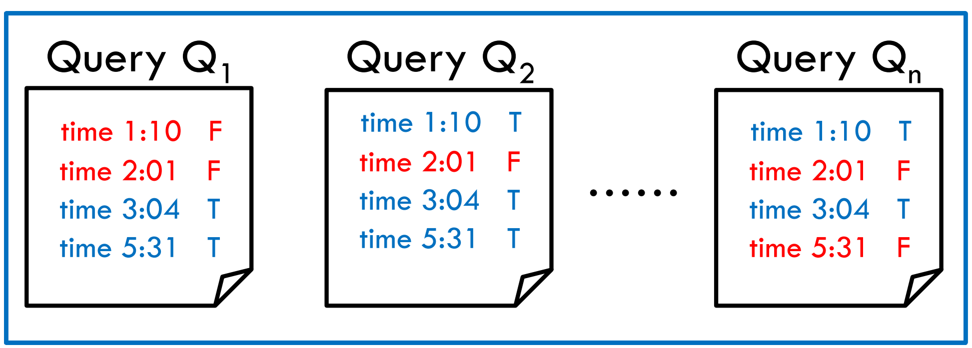
3 out of 4 teams used this approach

New Directions for Spoken Content Retrieval

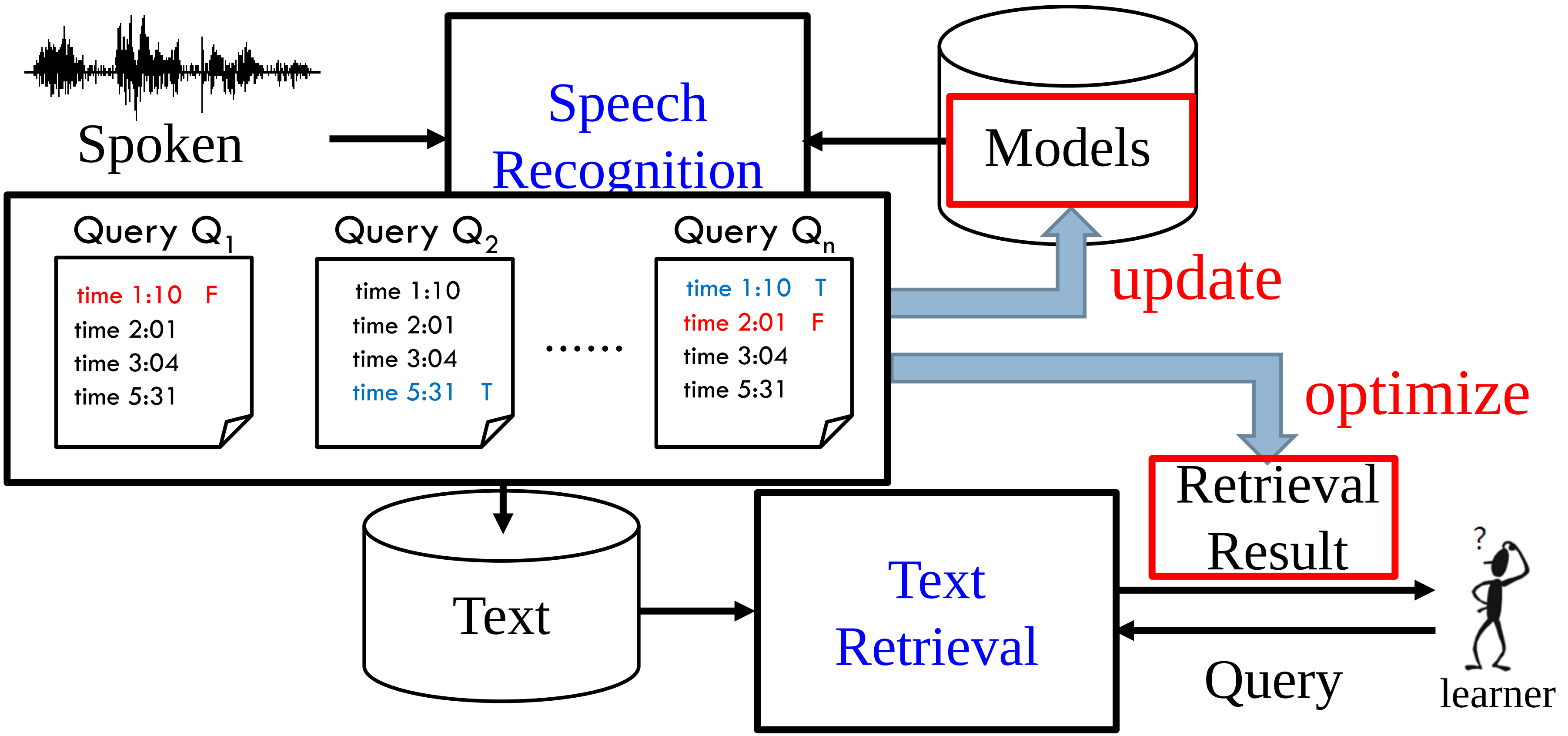
- Incorporating Information Lost in Standard Speech Recognition
- Improving Recognition Models by User Relevance feedback
- Query Expansion with Speech Signals
- Spoken Content Retrieval without Speech Recognition
- Interactive Retrieval

User Relevance Feedback

- Online search engine optimizes performance by user relevance feedback
  - E.g. click-through data [T Joachims, SIGKDD 02]



Update Recognition Models



New Directions for Spoken Content Retrieval

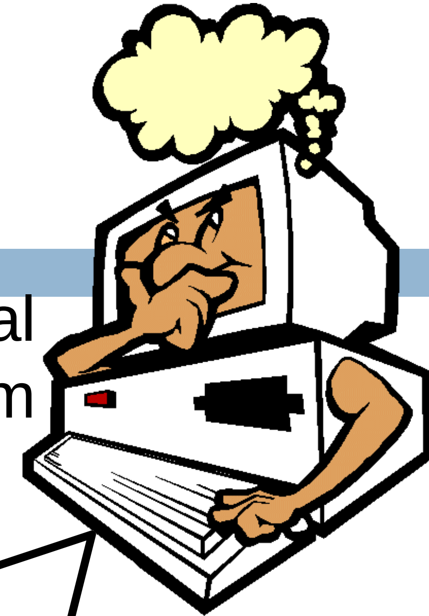
- Incorporating Information Lost in Standard Speech Recognition
- Improving Recognition Models by User Relevance feedback
- Query Expansion with Speech Signals
- Spoken Content Retrieval without Speech Recognition
- Interactive Retrieval

Query Expansion for Text Retrieval

learner

“Inheritance material”

Retrieval system



Search “Inheritance material” and “DNA”

To handle the problem of semantic retrieval, retrieval system will expand the user query.

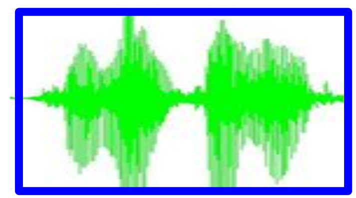
Query Expansion for Spoken Content Retrieval

Retrieval system

learner

"Inheritance material"

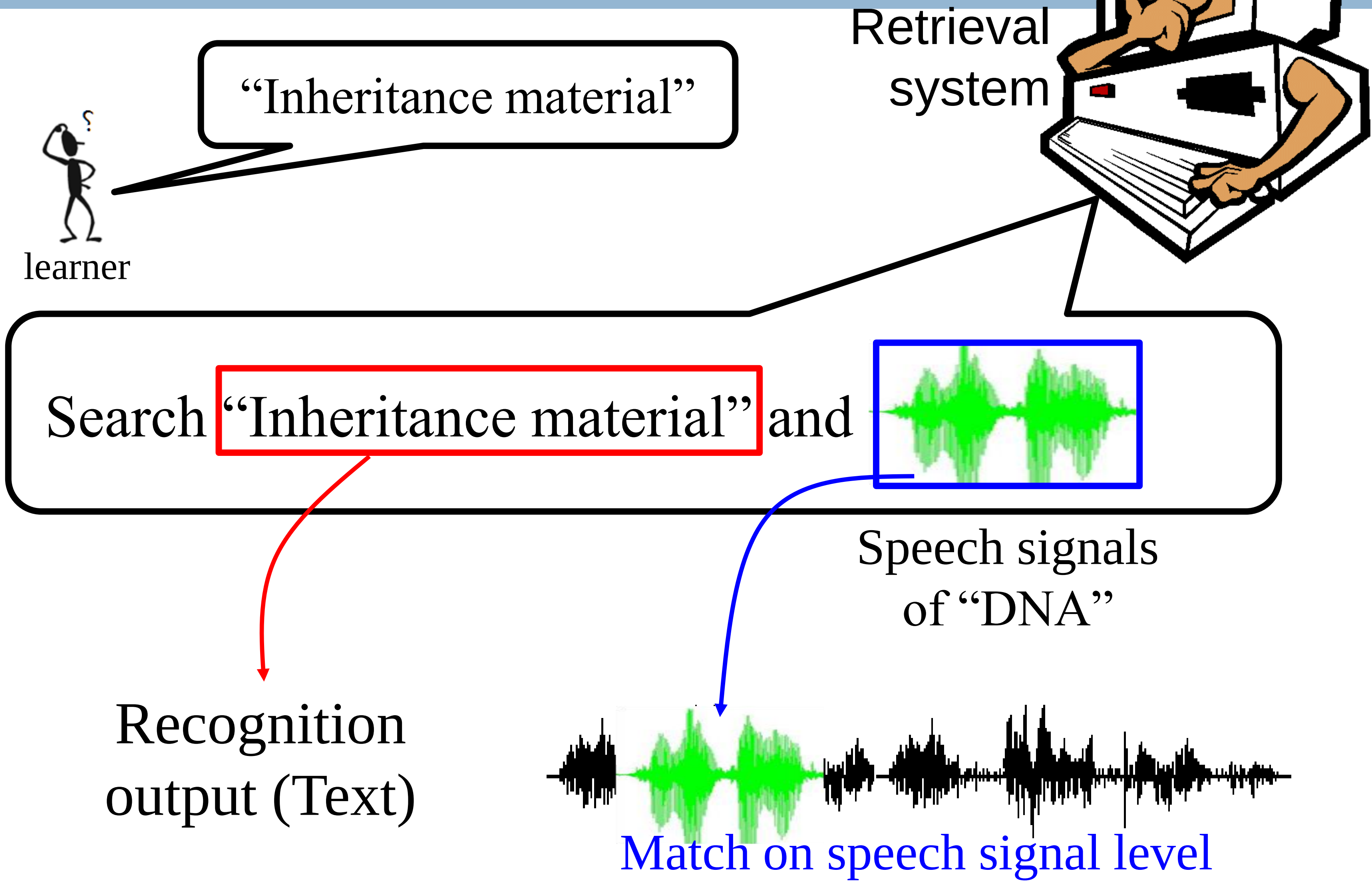
Search "Inheritance material" and



Speech signals of "DNA"

Expand the queries by speech signals

Query Expansion for Spoken Content Retrieval

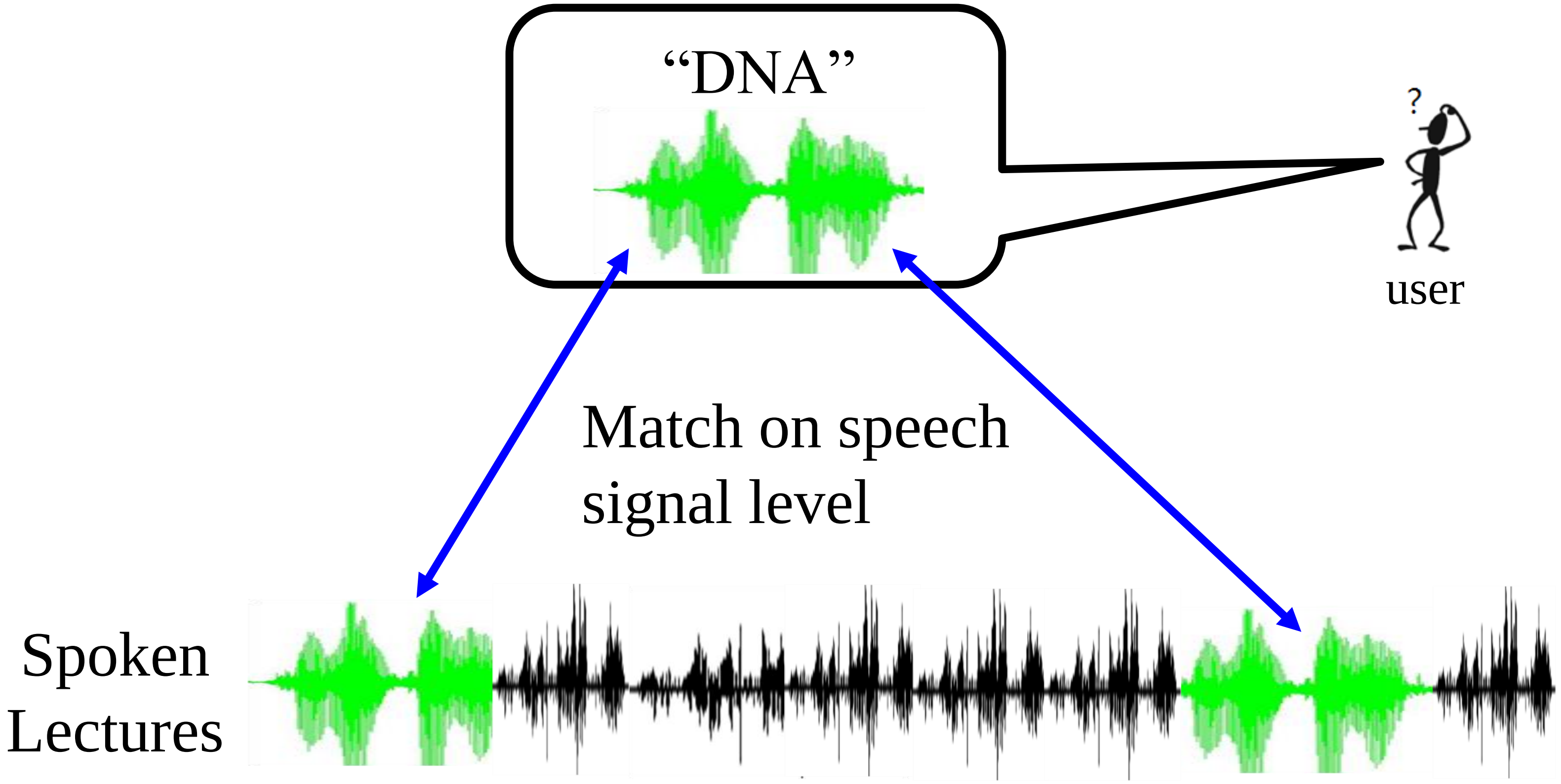


New Directions for Spoken Content Retrieval

- Incorporating Information Lost in Standard Speech Recognition
- Improving Recognition Models by User Relevance feedback
- Query Expansion with Speech Signals
- Spoken Content Retrieval without Speech Recognition
- Interactive Retrieval

Spoken Content Retrieval without Speech Recognition

- Spoken Queries

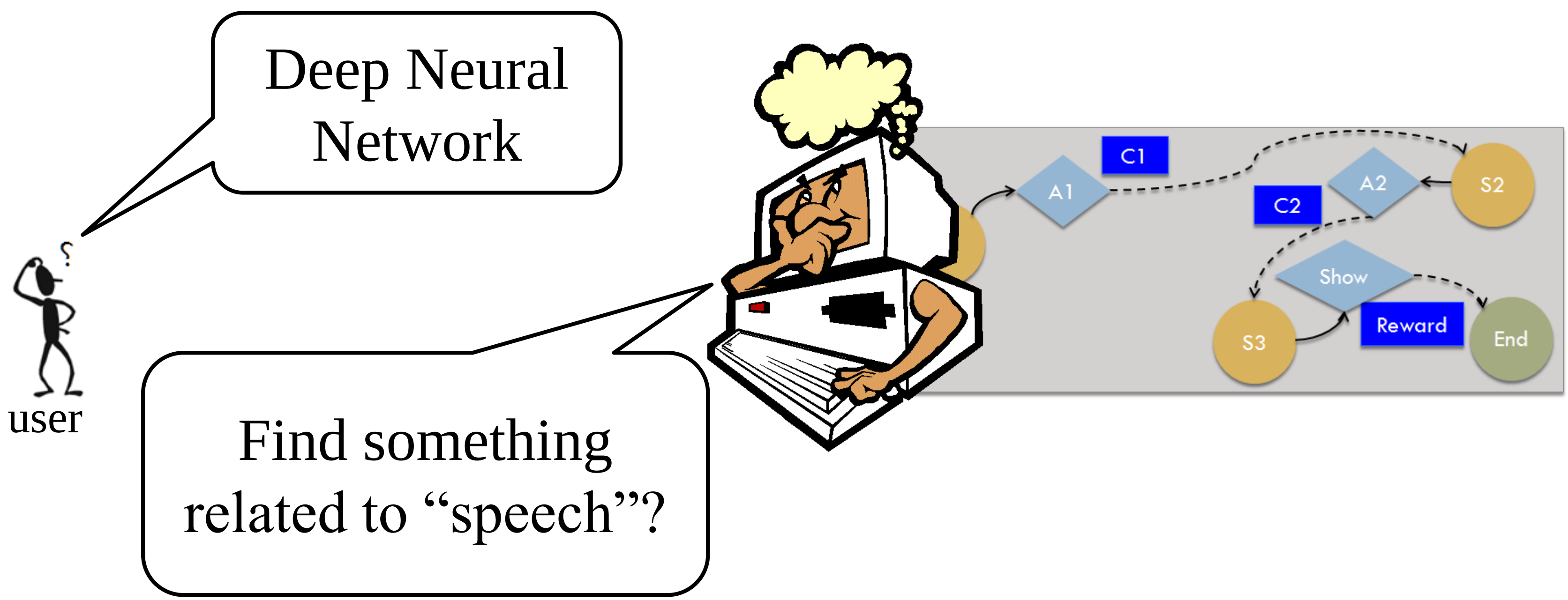


New Directions for Spoken Content Retrieval

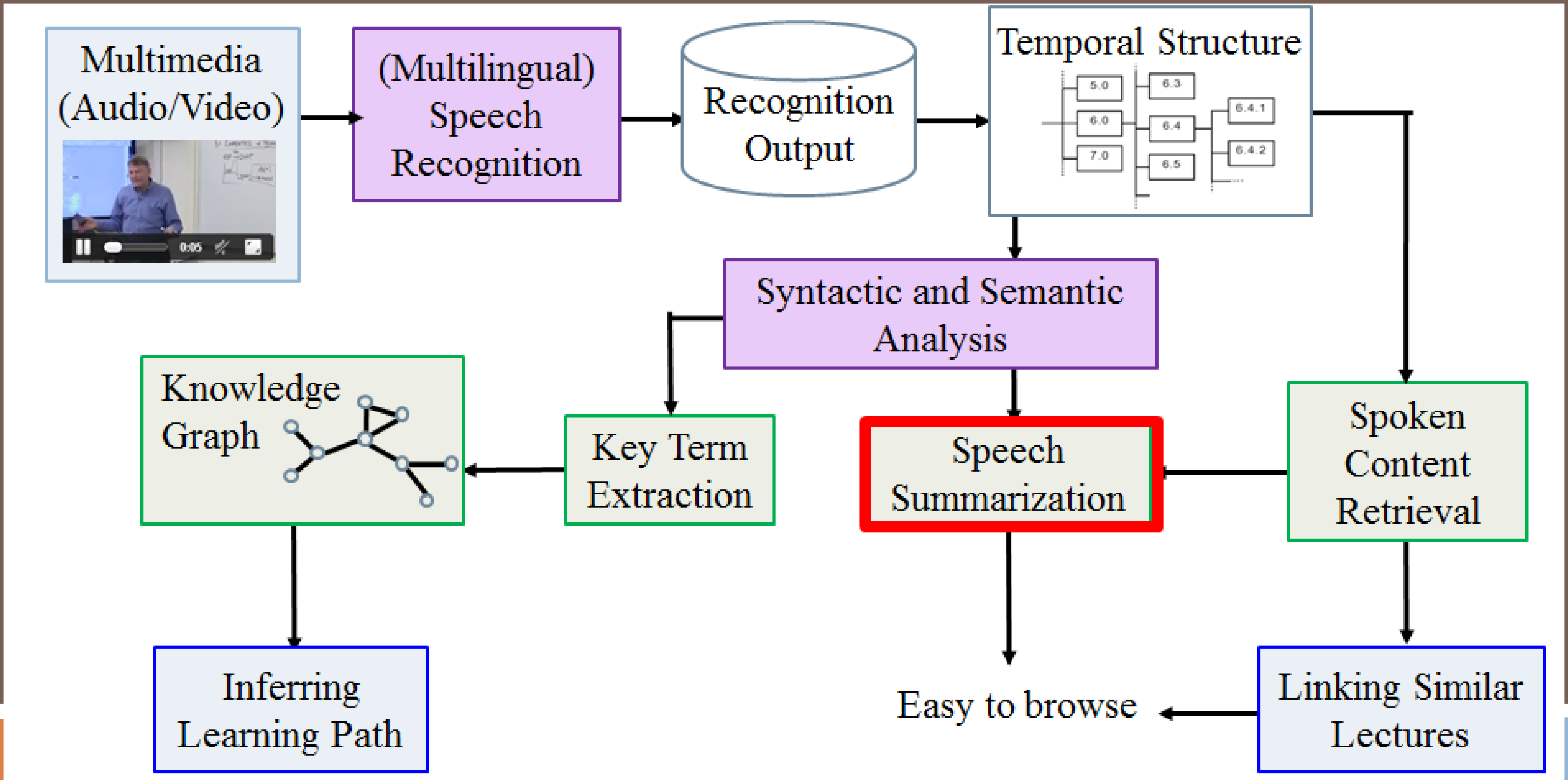
- Incorporating Information Lost in Standard Speech Recognition
- Improving Recognition Models by User Relevance feedback
- Query Expansion with Speech Signals
- Spoken Content Retrieval without Speech Recognition
- Interactive Retrieval

Interactive Retrieval

- Model the interactive retrieval process as Markov Decision Process (MDP)



Part III: Speech Summarization



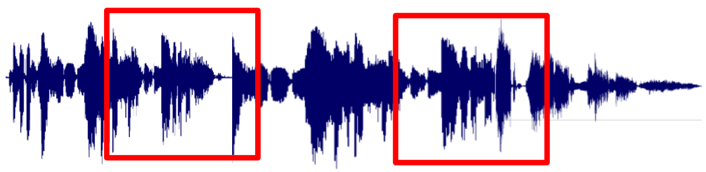
MMR approach

- Maximum marginal relevance (MMR) approach
  - Unsupervised approach: Use heuristic rules to select utterances
    - Select utterances whose content are similar to the whole lectures
    - Minimize redundancy in summary at the same time

Supervised Approach

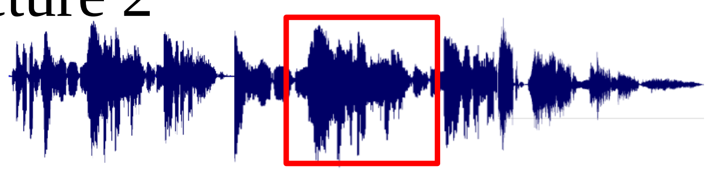
- Training data

Lecture 1



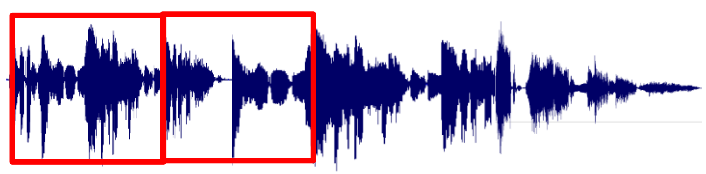
2^{nd} and 4^{th} utterances form the summary

Lecture 2



3^{rd} utterances form the summary

Lecture 3

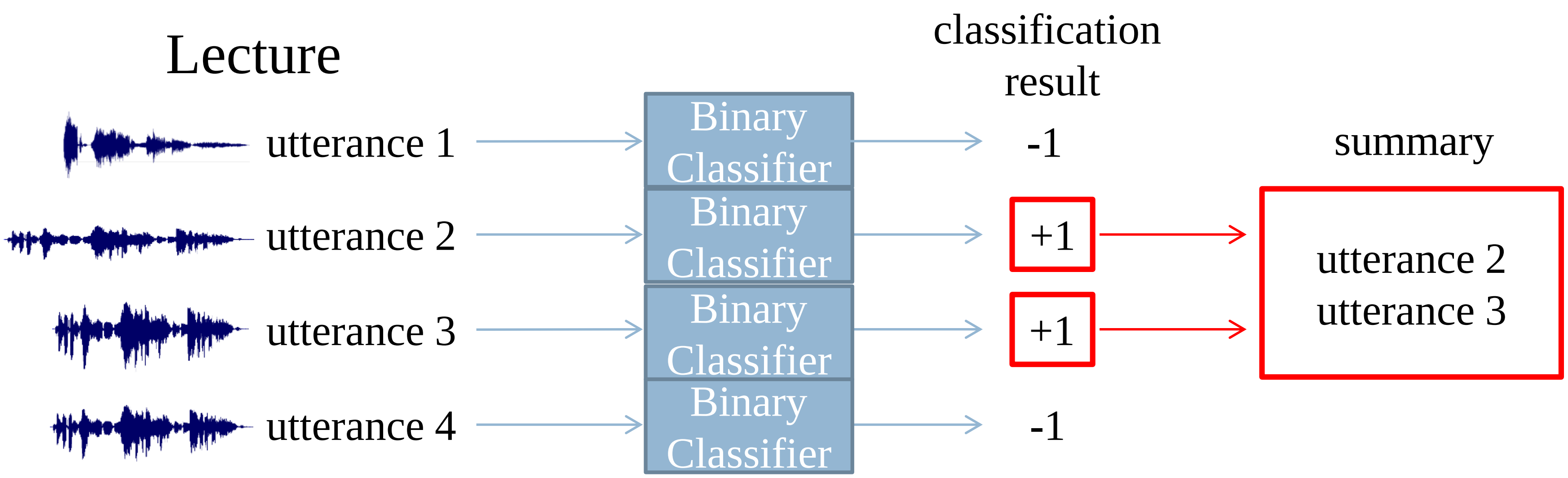


1^{st} and 2^{nd} utterances form the summary

Use the training data to learn model for summarization

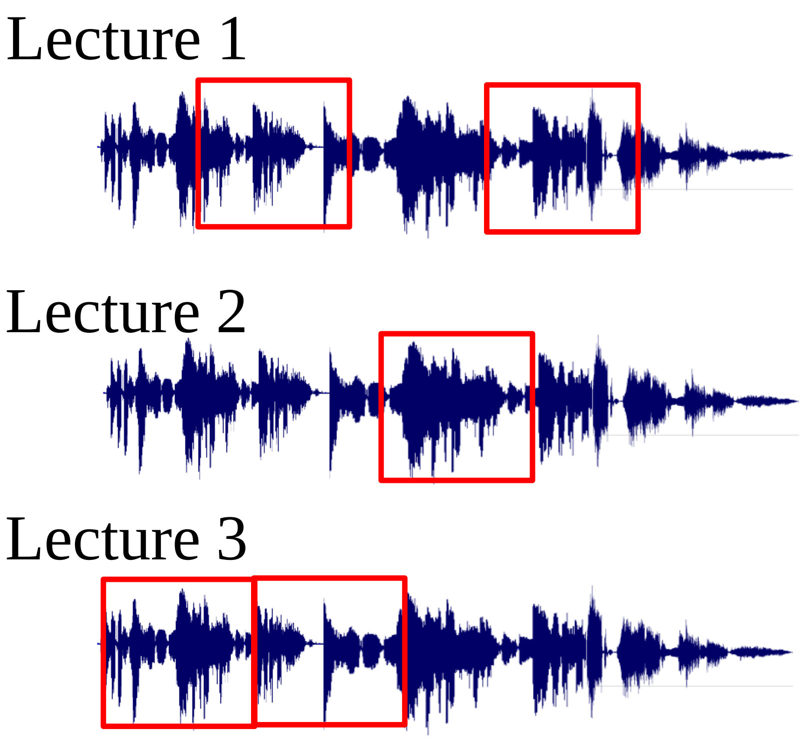
Supervised Approach – Binary Classification

- Summarization problem can be formulated as a binary classification program
  - Included in the summary or not



Supervised Approach – Binary Classification

- Training data

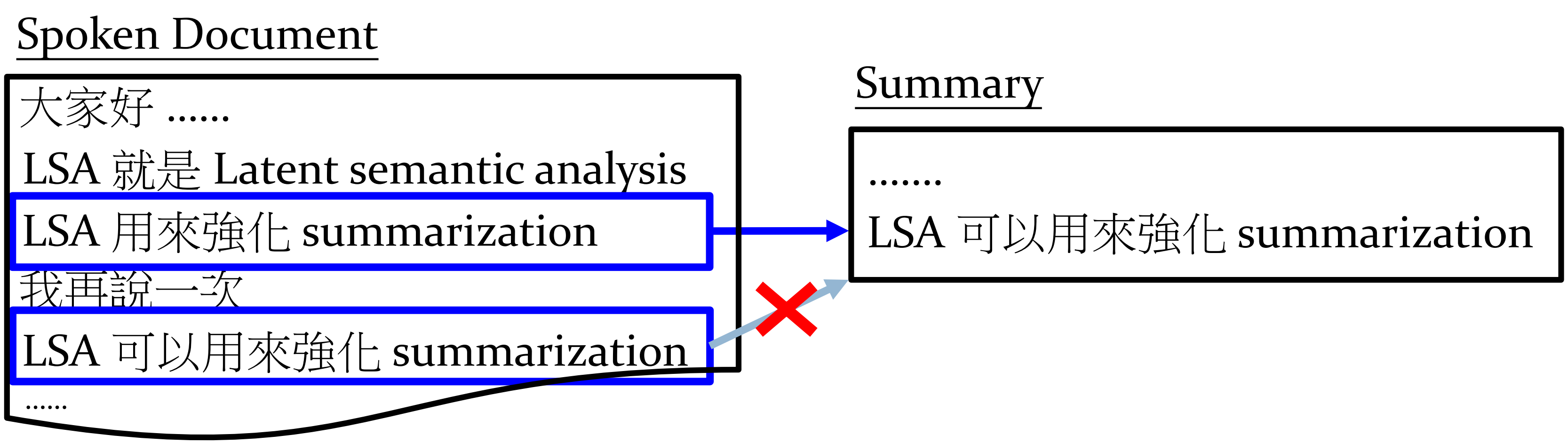


- The utterances in the summary are positive examples.
- Otherwise, negative examples

Train a binary classifier

Supervised Approach – Binary Classification

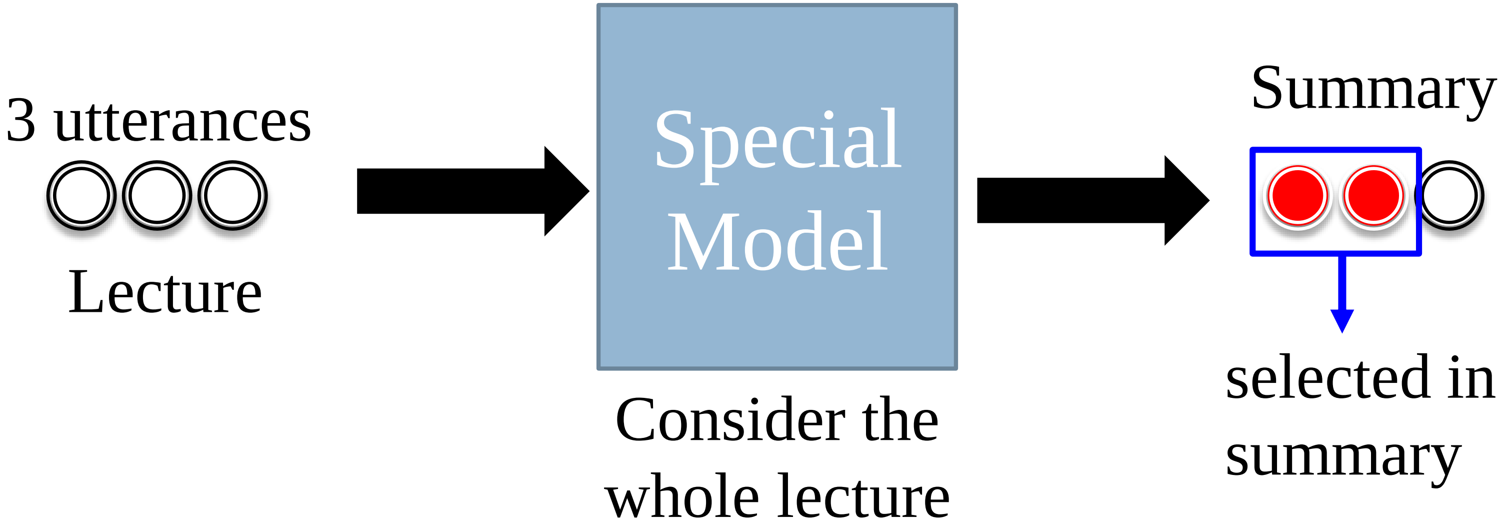
- Binary classifier individually considers each utterance
- To generate a good summary, “global information” should be considered
- Example: summary should be concise



More advanced machine learning techniques

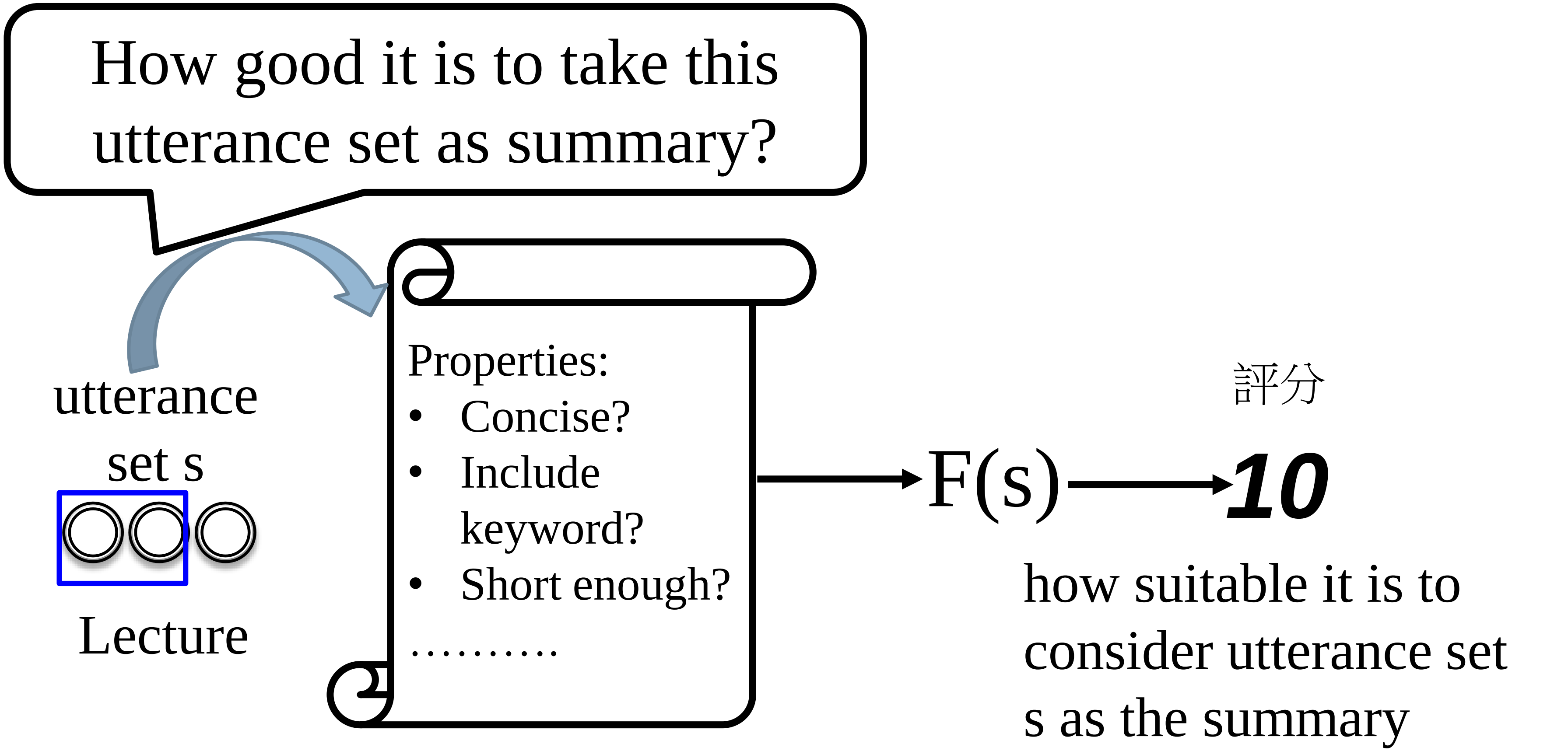
Globally considering the whole spoken lectures

- Learn a special model by structured learning techniques
  - Input: whole lecture
  - Output: summary



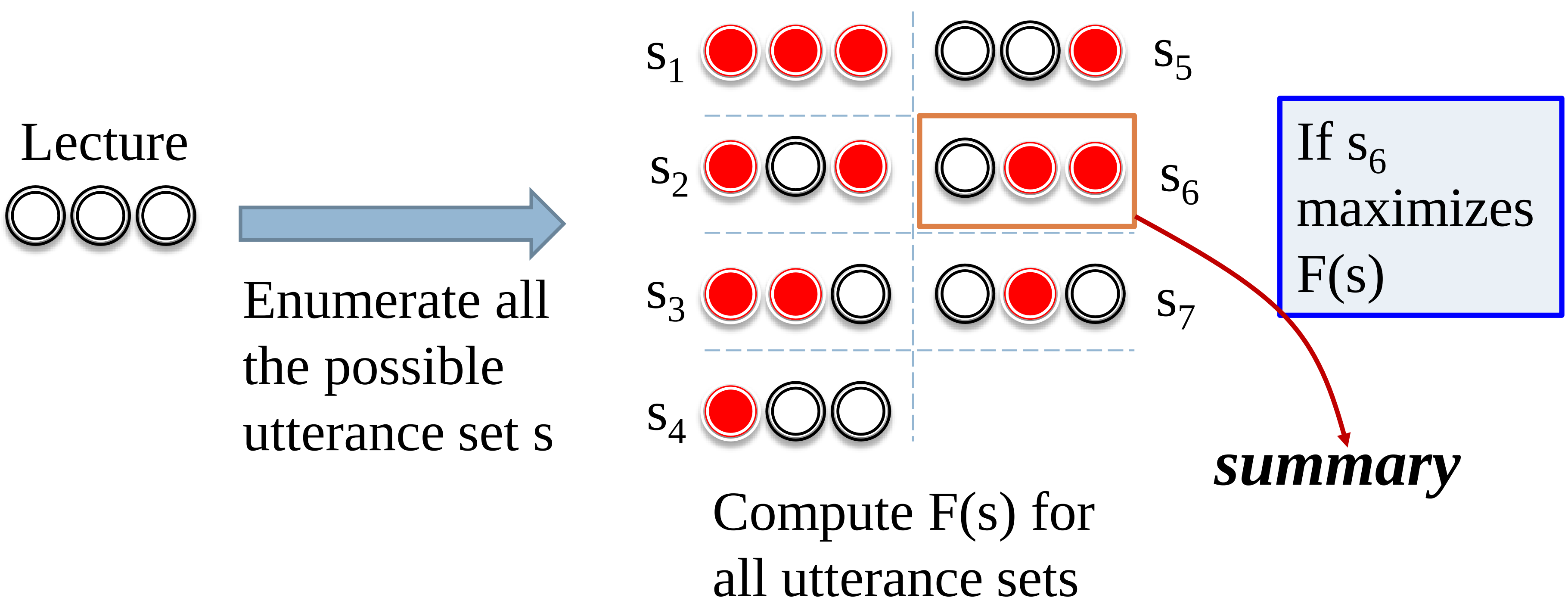
Evaluation Function

- Evaluation function of utterance set F(s)
  - s: utterance set in a lecture



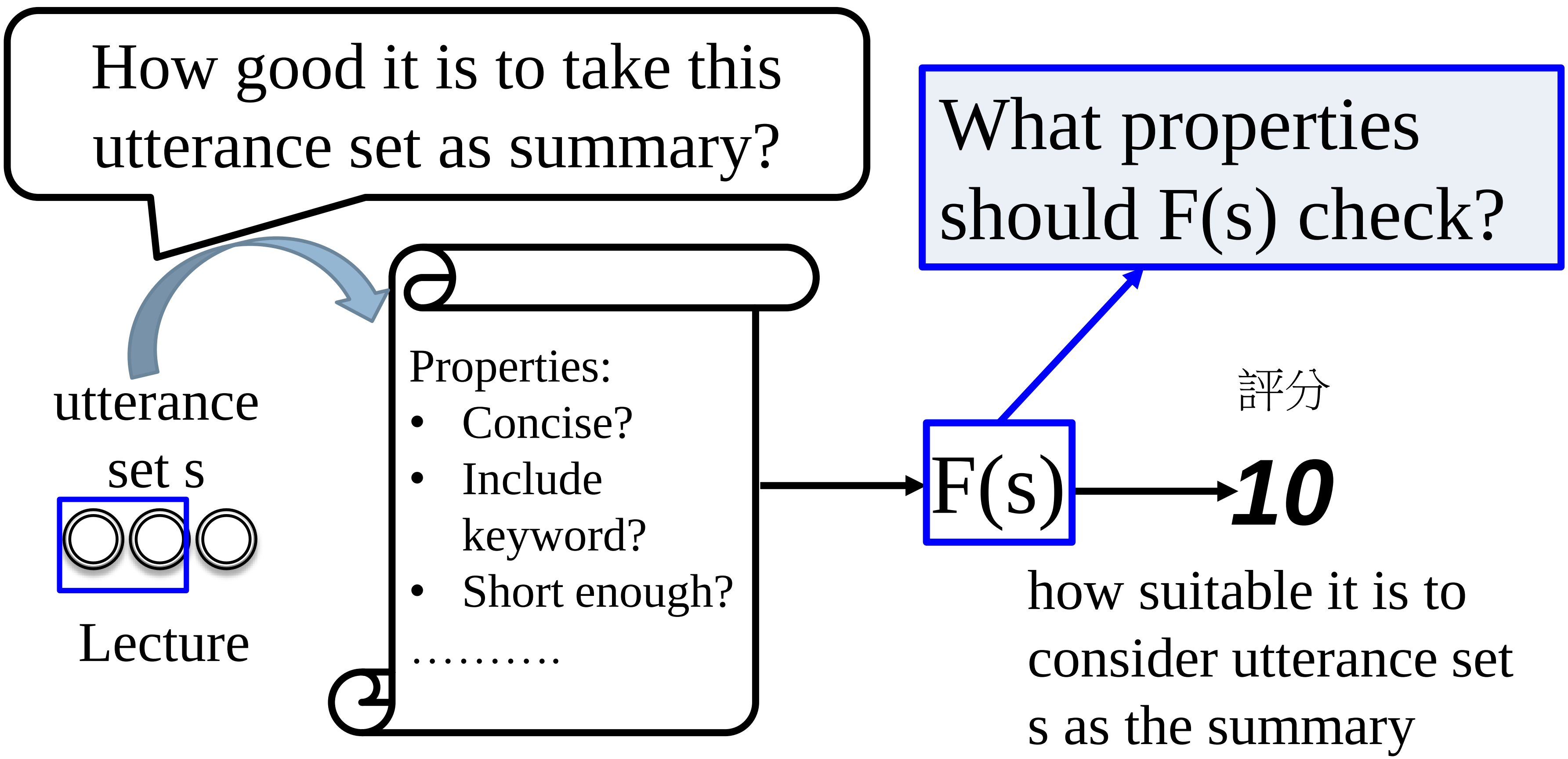
Evaluation Function – How to summary

- With F(s), we can do summarization on new lectures now



Evaluation Function

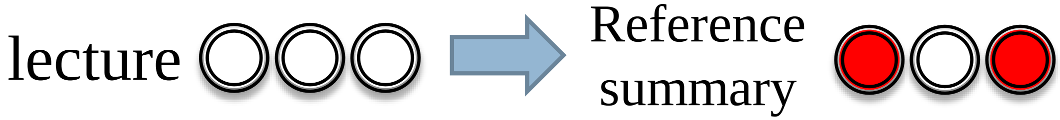
- Evaluation function of utterance set F(s)
  - s: utterance set in a lecture



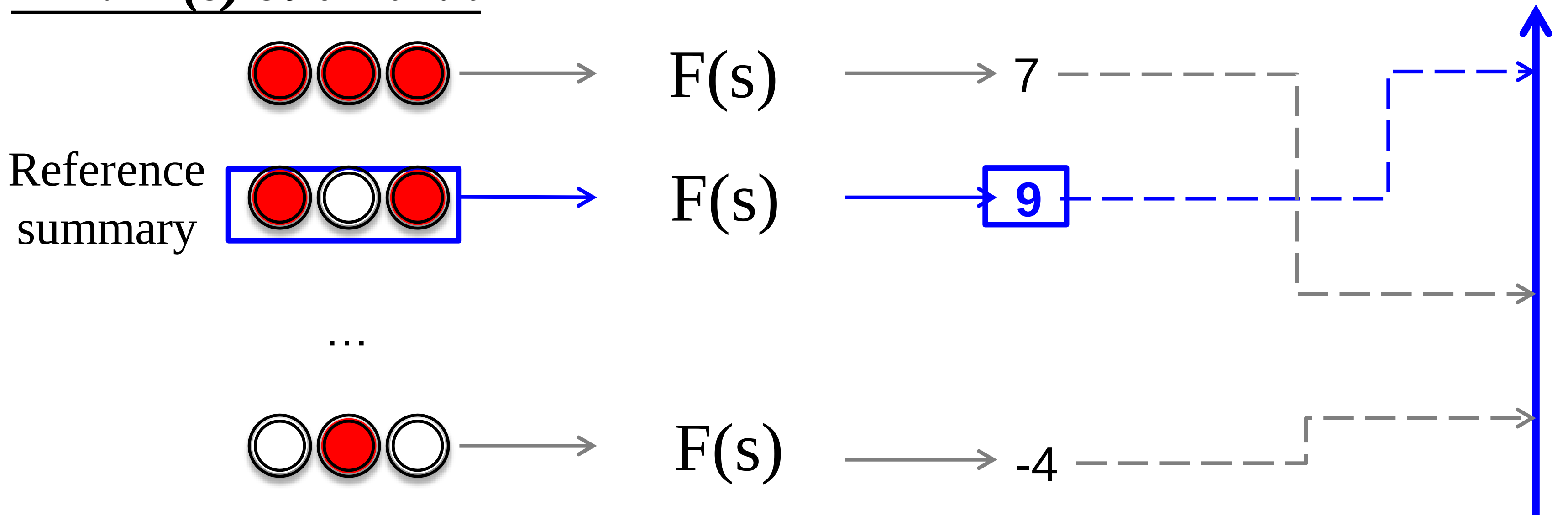
Evaluation Function - Training

- Learn F(s) from training data

<u>Training data</u>



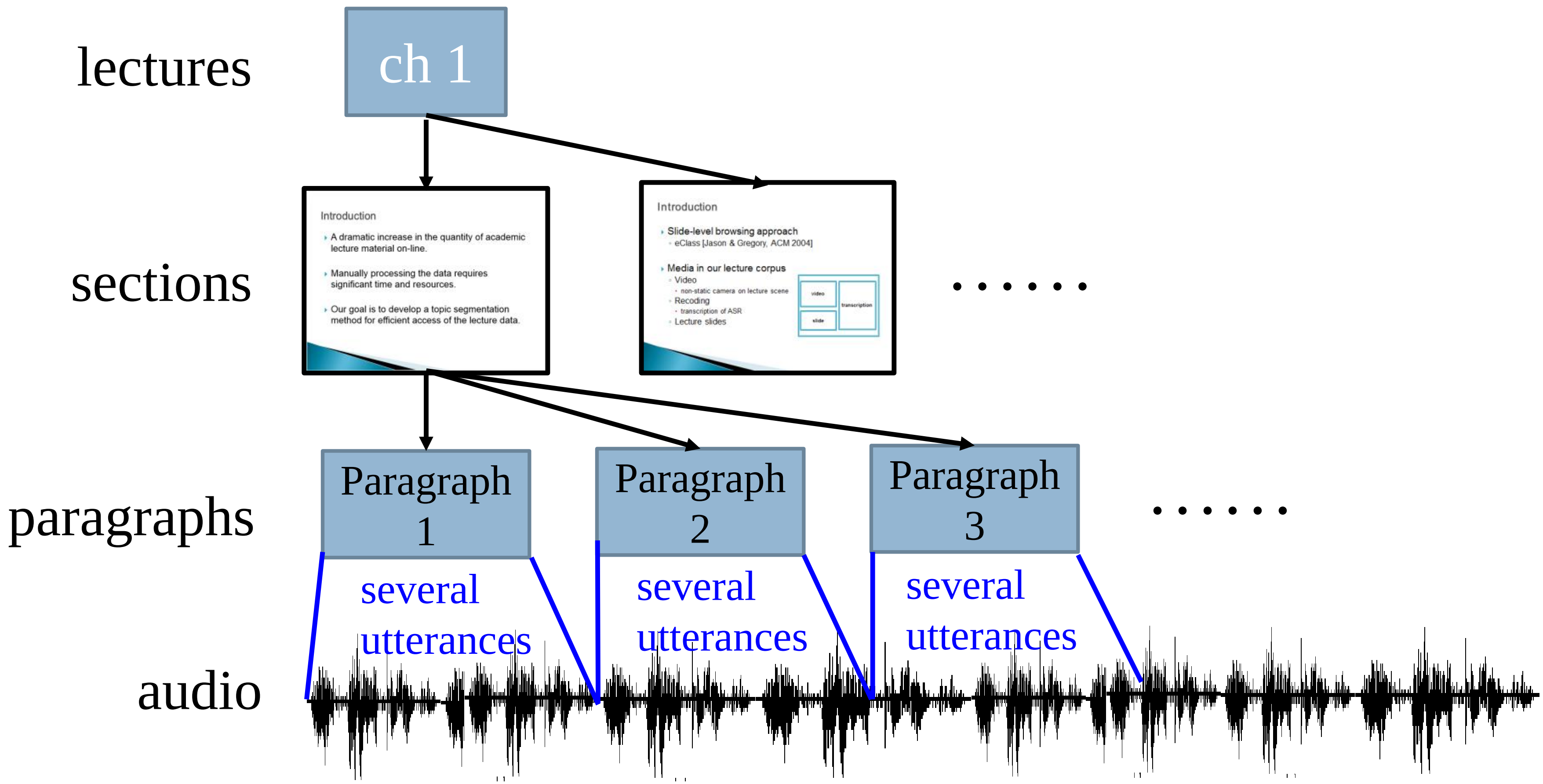
<u>Find F(s) such that</u>



Structured SVM: I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support Vector Learning for Interdependent and Structured Output Spaces, ICML, 2004.

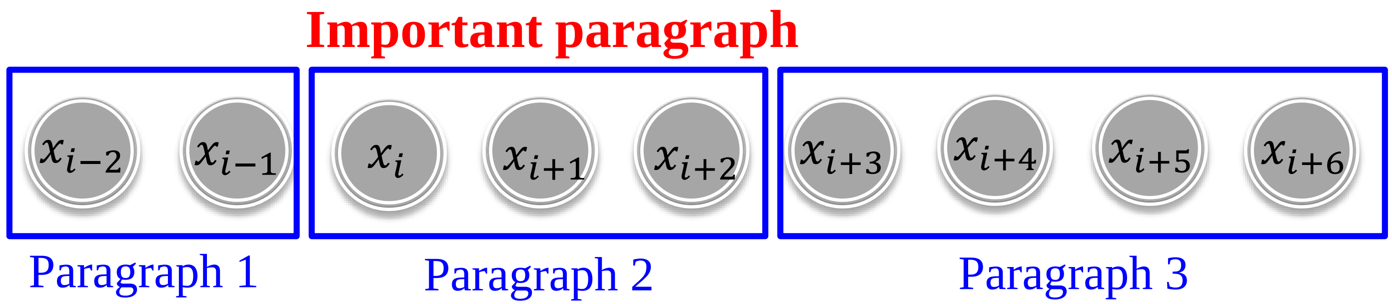
Speech Summarization - Structure

- Temporal structure helps summarization

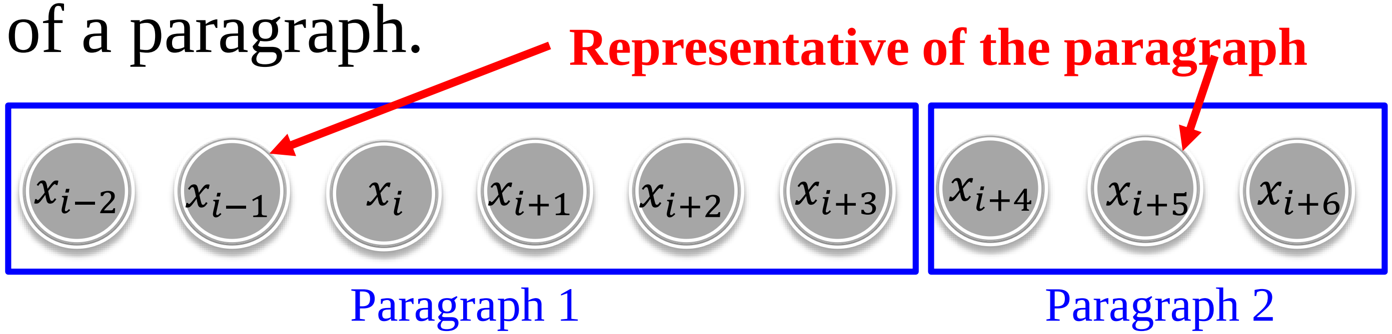


Speech Summarization - Structure

- Temporal structure helps summarization
  - Long summary: consecutive utterances in a paragraph are more likely to be

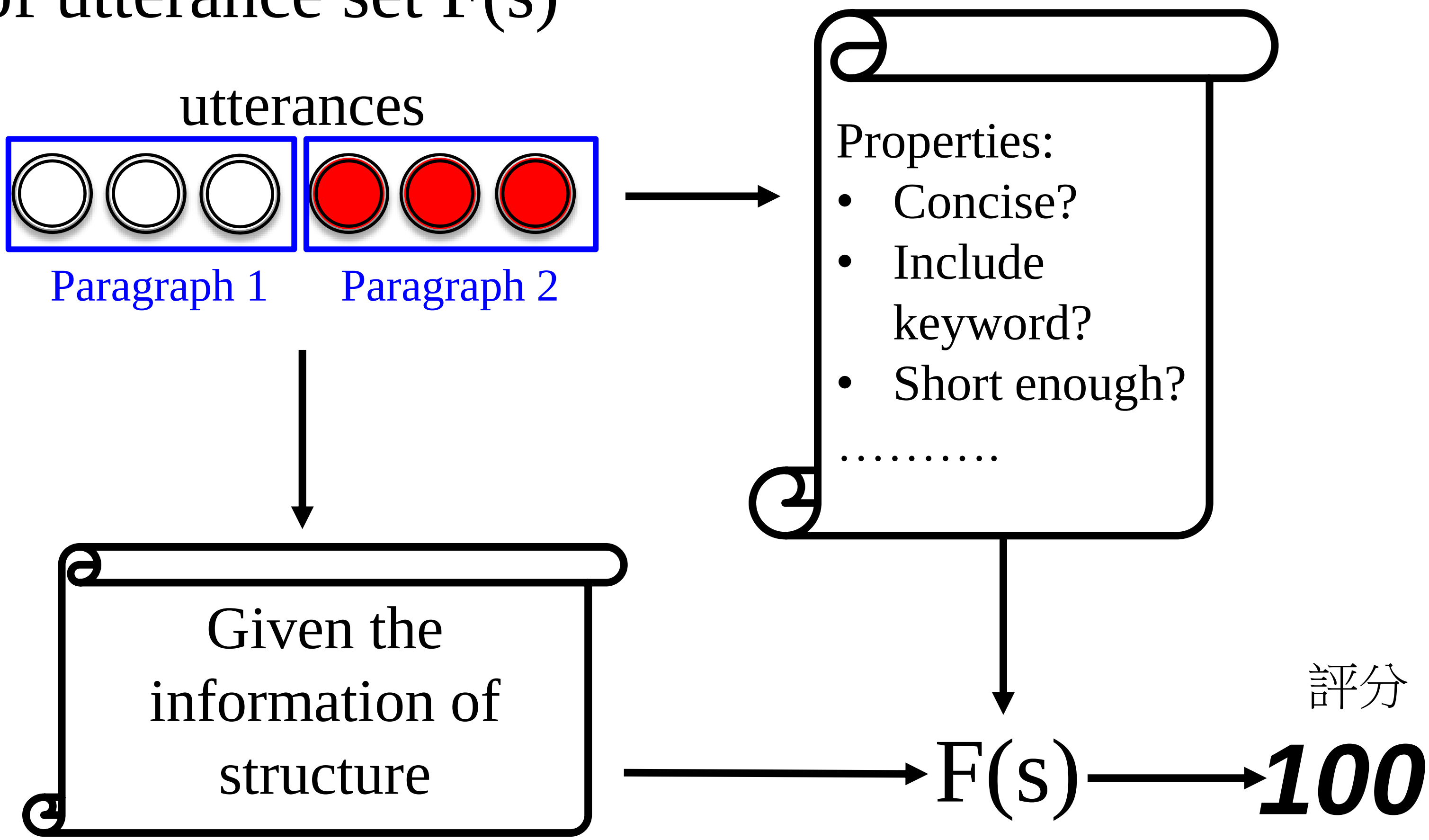


  - Short summary: one utterance is selected on behalf of a paragraph.



Evaluation Function - Structure

- Add structure information into evaluation function of utterance set F(s)

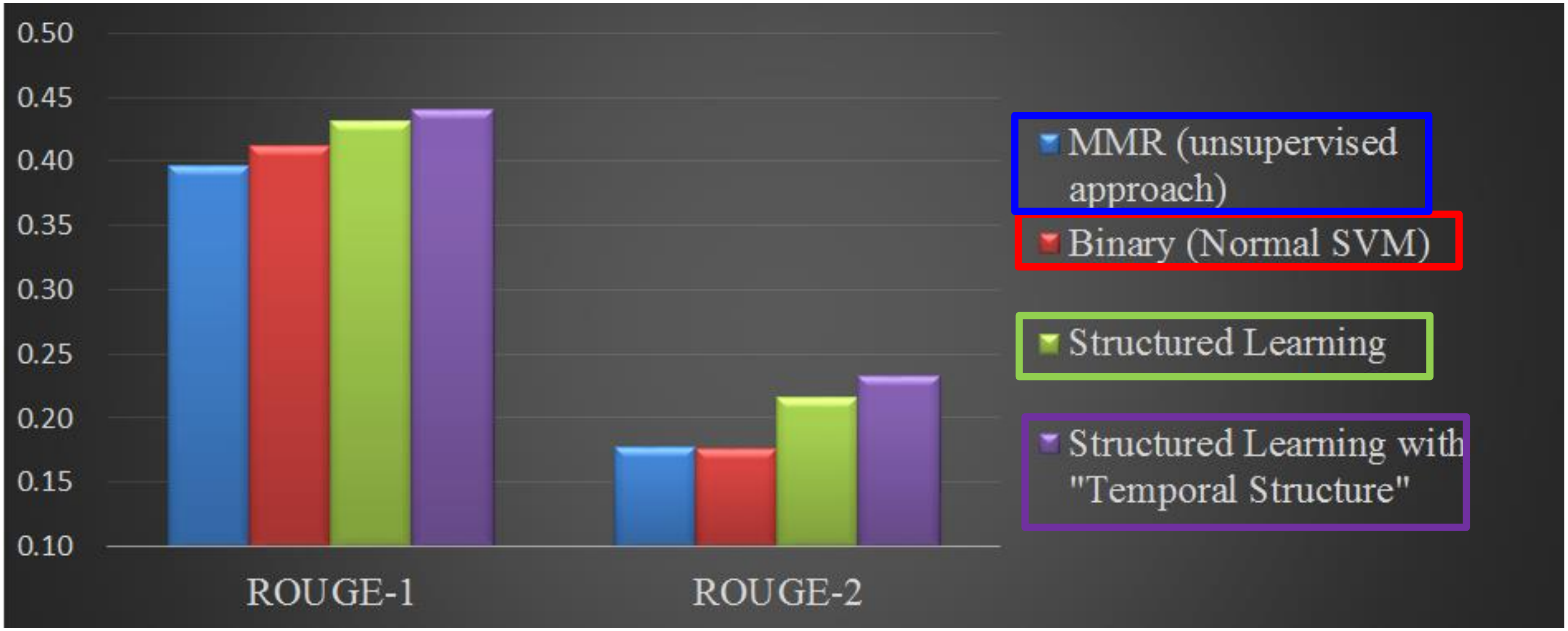


Speech Summarization - Structure

- Structure in text are clear
  - Paragraph boundaries are directly known
- For spoken content, there is no obvious structure
  - Here the structure are considered as “hidden variables”
  - Structured learning with hidden variables

Speech Summarization - Experiments

- Evaluation Measure: ROUGE-1 and ROUGE-2
  - Larger scores means the machine-generated summaries is more similar to human-generated summaries.

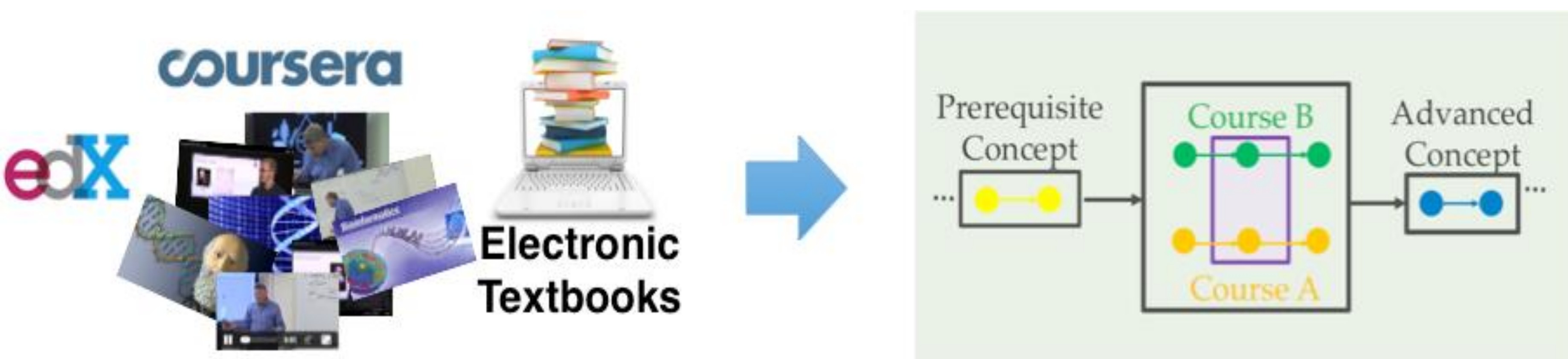


Part IV:

Demo

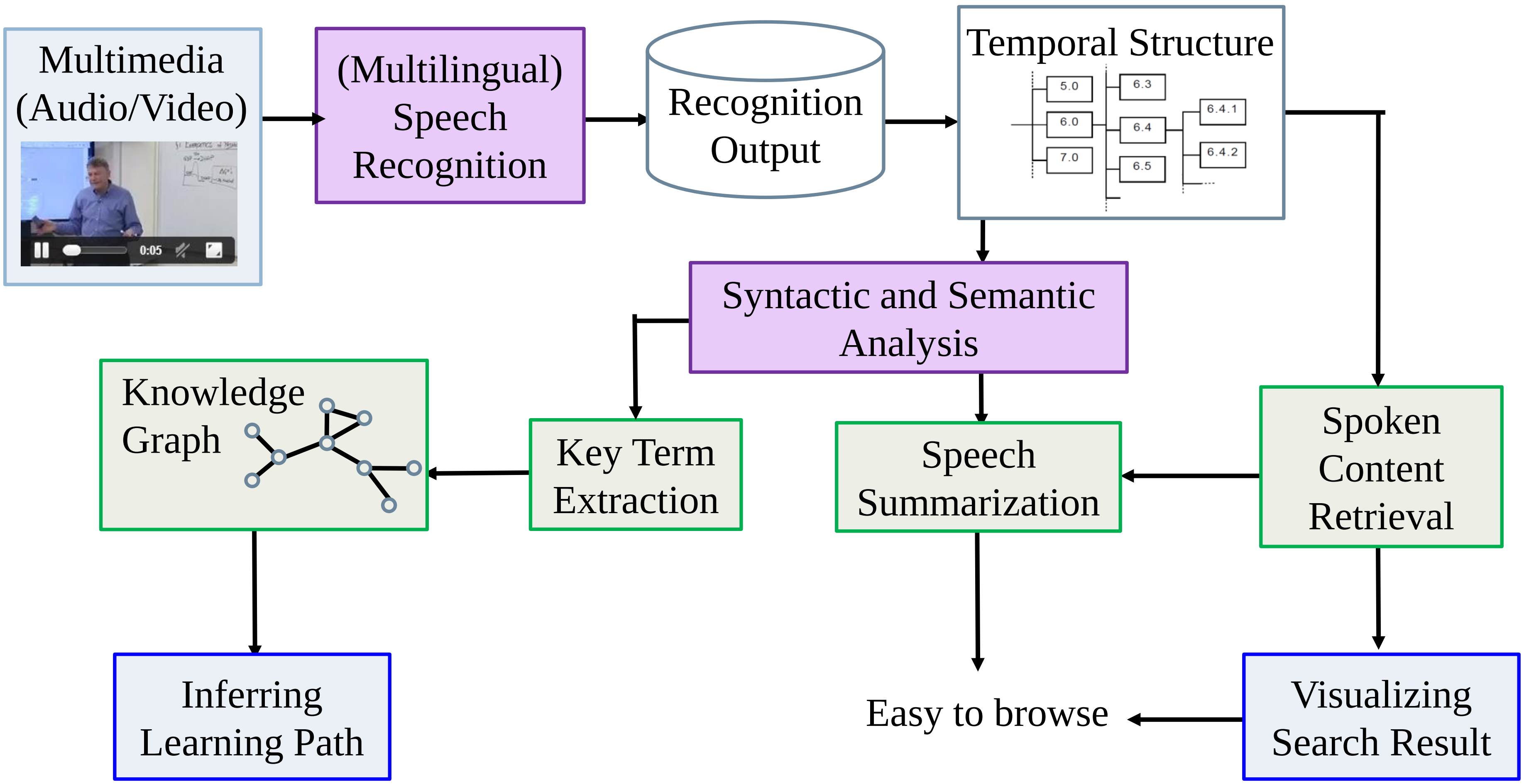
On-line lecture platforms (MIT)

- “Cang-Jie (倉頡)”:
  - Search lecture recording and textbook
  - Linking video clips or textbook sections with similar content
  - Inferring prerequisite and advanced concepts
  - http://people.csail.mit.edu/tlkagk/Cangjie/



Concluding Remarks

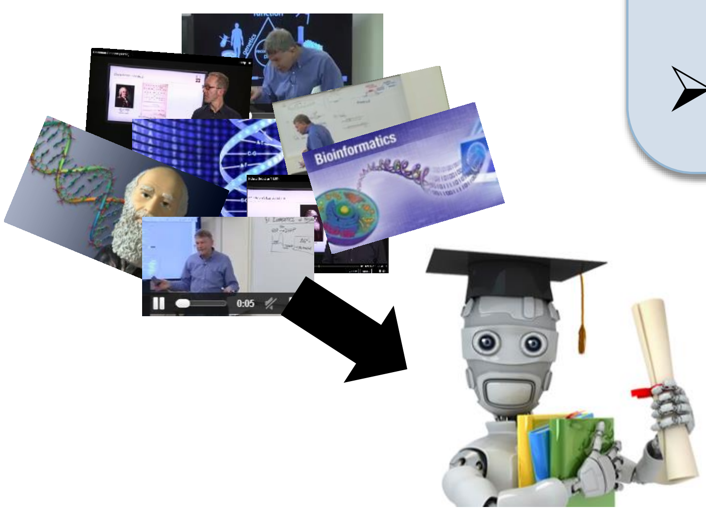
Towards Spoken Knowledge Structuring and Organization



Ultimate Goal

- Personalized course for each learner

on-line learning material



- I want to learn “XXX”.
- I am a graduate student of computer science.
- I can spend 10 hours.

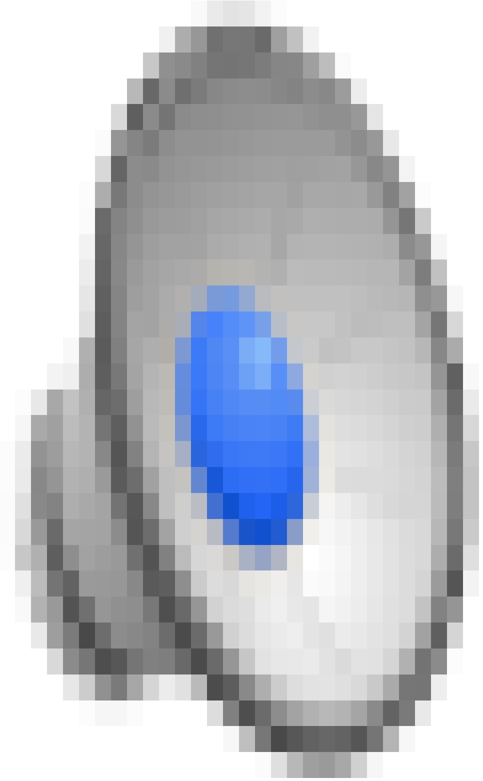
Learner

I open a course for you.

Thank You for Your Attention

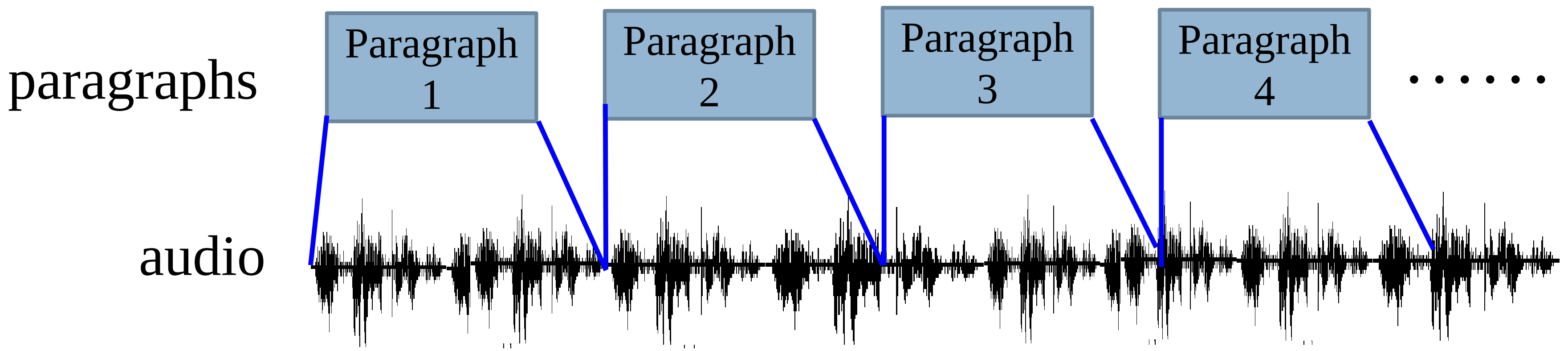
Appendix

Video Demonstration



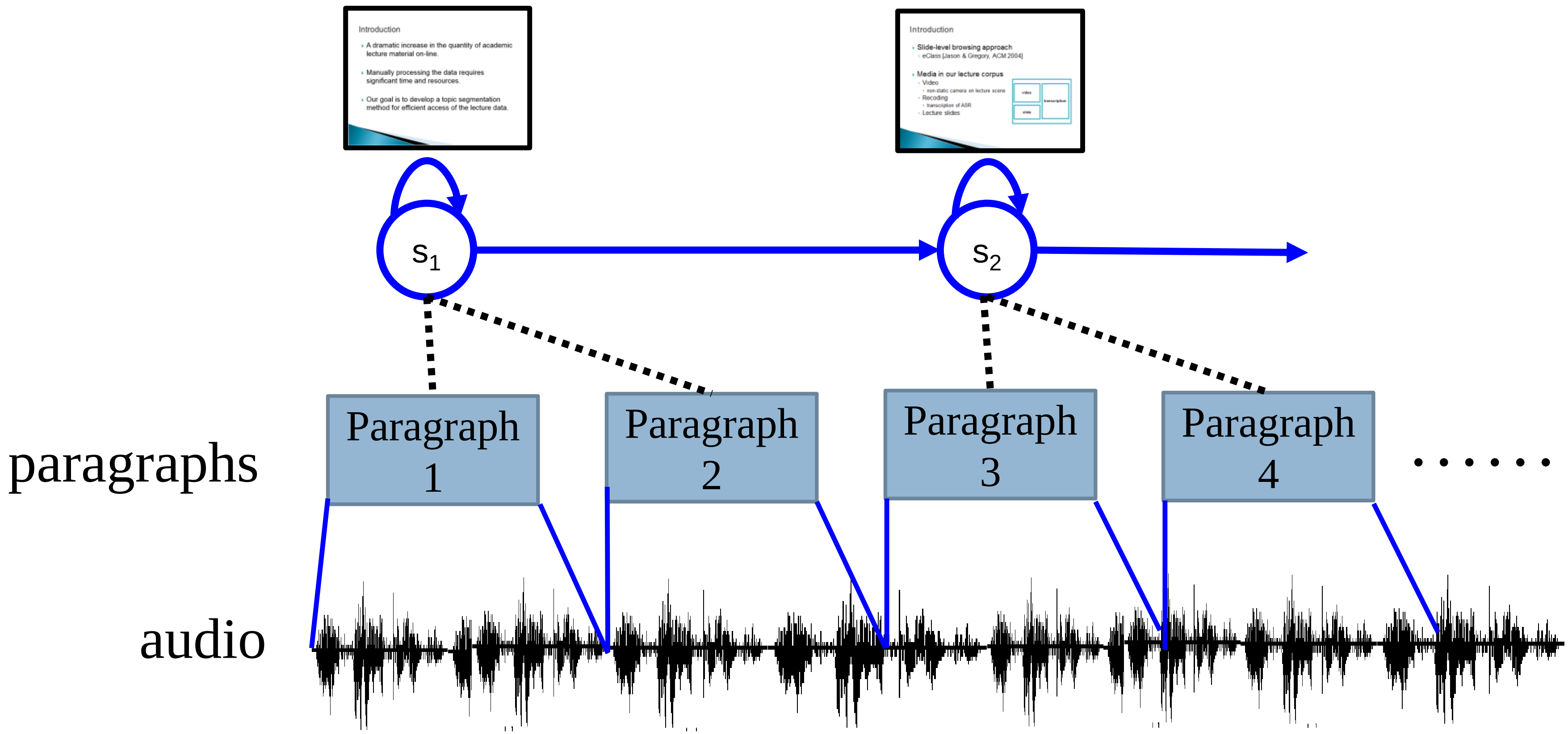
Paragraph Boundaries

- With speech recognition, we know the content of each utterances
  - Compute their similarities
- Find the boundary of paragraph such that
  - The content of the utterances in a paragraph is similar



Slide Boundaries

- The slides are modeled as HMMs
  - Align the slides with paragraphs



Evaluation function F(s)

- A good summary should
  - 1. *include the most important utterance*
  - 2. but *minimize the redundancy* at the same time
  - 3. not too long

Utterance set s fulfill the above requirement should have large F(s)

Evaluation function F(s)

- A good summary should
  - 1. *include the most important utterance*
  - 2. but *minimize the redundancy* at the same time
  - 3. not too long

$$F(s) = \sum_{x_i \in s} I(x_i)$$

$I(x_i)$: importance of utterance x_i

$$I(x_i) = w \cdot \boxed{f(x_i)} \rightarrow$$ $f(x_i)$: feature of sentence x_i

Evaluatio

- A good su
  - 1. *include*
  - 2. but *min*
  - 3. not too

$$F(s) =$$

- Lexical feature: use speech recognition to transcribe each utterance into text
  - similarity to the transcriptions of whole lectures
  - latent topic distribution
  - how many keywords
  - ……
- Prosodic feature:
  - Energy, pitch, syllable duration, pause duration

I(x_i): importance of utterance x_i

$$I(x_i) = \boxed{w} \cdot \boxed{f(x_i)}$$

f(x_i): feature of sentence x_i

weights for each feature

Evaluation function F(s)

- A good summary should
  - 1. *include the most important utterance*
  - 2. but *minimize the redundancy* at the same time
  - 3. not too long

$$F(s) = \sum_{x_i \in s} I(x_i) - \lambda \sum_{x_i, x_j \in s} Sim(x_i, x_j)$$

$Sim(x_i, x_j)$: similarity between utterances x_i and x_j

λ is a parameter to be determined.

Evaluation function F(s)

- A good summary should
  - 1. *include the most important utterance*
  - 2. but *minimize the redundancy* at the same time
  - 3. not too long

$$F(s) = \sum_{x_i \in s} I(x_i) - \lambda \sum_{x_i, x_j \in s} Sim(x_i, x_j)$$

$$\sum_{x_i \in s} L(x_i) < K \quad \text{(constraint)}$$

$L(x_i)$: length of utterance x_i

K: length constraint of summary

Evaluation function F(s)

- A good summary should
  - 1. *include the most important utterance*
  - 2. but *minimize the redundancy* at the same time
  - 3. not too long

$$F(s) = \sum_{x_i \in s} I(x_i) - \boxed{\lambda} \sum_{x_i, x_j \in s} Sim(x_i, x_j)$$

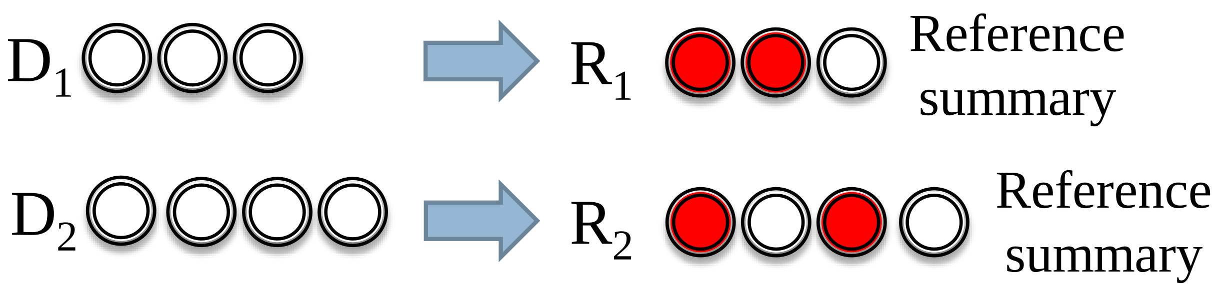
$$\sum_{x_i \in s} L(x_i) < K$$ (constraint)

$$I(x_i) = \boxed{w} \cdot f(x_i)$$

Jointly learn from training data

Idea of Training

- Training data



……

Find w and λ in F(s) such that

$F(R_1) > F(\boxed{s_{D1}})$ s_{D1} is all utterance set in D_1, except R_1

$F(R_2) > F(\boxed{s_{D2}})$ s_{D2} is all utterance set in D_2, except R_2

……

I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support Vector Learning for Interdependent and Structured Output Spaces, ICML, 2004.