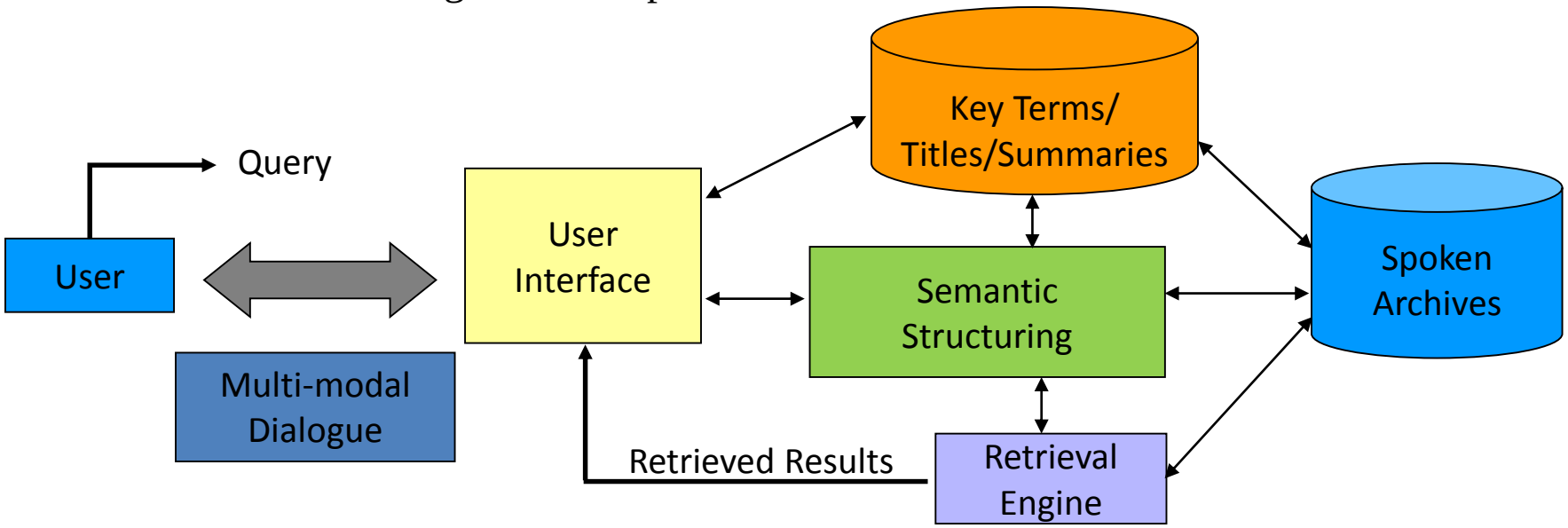
11.0 Spoken Document Understanding and Organization for User-content Interaction

References: 1. “Spoken Document Understanding and Organization”, IEEE Signal Processing Magazine, Sept. 2005, Special Issue on Speech Technology in Human-Machine Communication

2. “Multi-layered Summarization of Spoken Document Archives by Information Extraction and Semantic Structuring”, Interspeech 2006, Pittsburg, USA

User-Content Interaction for Spoken Content Retrieval

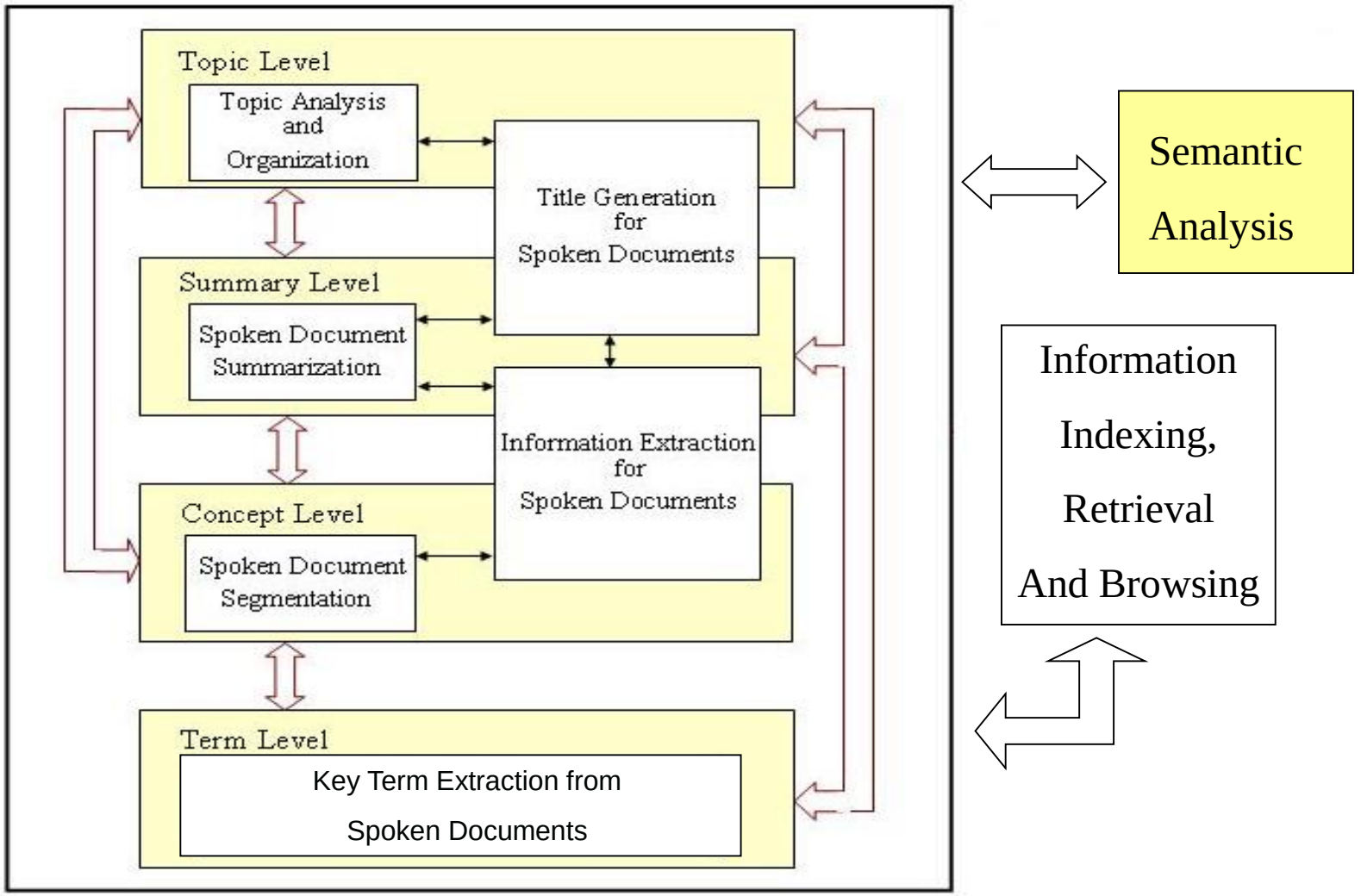
- **Problems**
  - Unlike text content, spoken content not easily summarized on screen, thus retrieved results difficult to scan and select
  - User-content interaction always important even for text content
- **Possible Approaches**
  - Automatic summary/title generation and key term extraction for spoken content
  - Semantic structuring for spoken content
  - Multi-modal dialogue with improved interaction



Multi-media/Spoken Document Understanding and Organization

- **Key Term/Named Entity Extraction from Multi-media/Spoken Documents**
  - **— personal names, organization names, location names, event names**
  - **— key phrase/keywords in the documents**
  - **— very often out-of-vocabulary (OOV) words, difficult for recognition**
- **Multi-media/Spoken Document Segmentation**
  - **— automatically segmenting a multi-media/spoken document into short paragraphs, each with a central topic**
- **Information Extraction for Multi-media/Spoken Documents**
  - **— extraction of key information such as who, when, where, what and how for the information described by multi-media/spoken documents.**
  - **— very often the relationships among the key terms/named entities**
- **Summarization for Multi-media/Spoken Documents**
  - **— automatically generating a summary (in text or speech form) for each short paragraph**
- **Title Generation for Multi-media/Spoken Documents**
  - **— automatically generating a title (in text or speech form) for each short paragraph**
  - **— very concise summary indicating the topic area**
- **Topic Analysis and Organization for Multi-media/Spoken Documents**
  - **— analyzing the subject topics for the short paragraphs**
  - **— clustering and organizing the subject topics of the short paragraphs, giving the relationships among them for easier access**

Integration Relationships among the Involved Technology Areas



Key Term Extraction from Spoken Content (1/2)

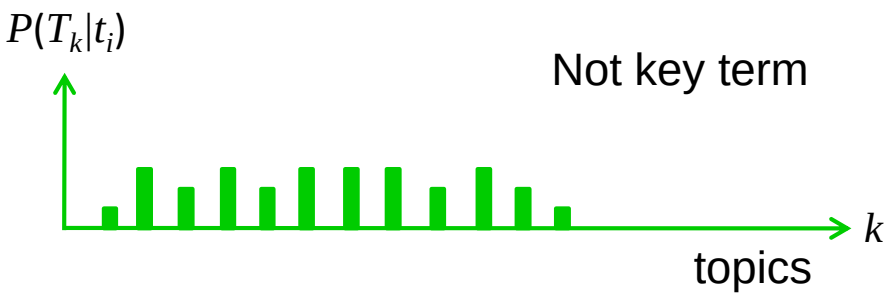
- Key Terms : key phrases and keywords
- Key Phrase Boundary Detection
- An Example

- "hidden" almost always followed by the same word
- "hidden Markov" almost always followed by the same word
- "hidden Markov model" is followed by many different words

- Left/right boundary of a key phrase detected by context statistics

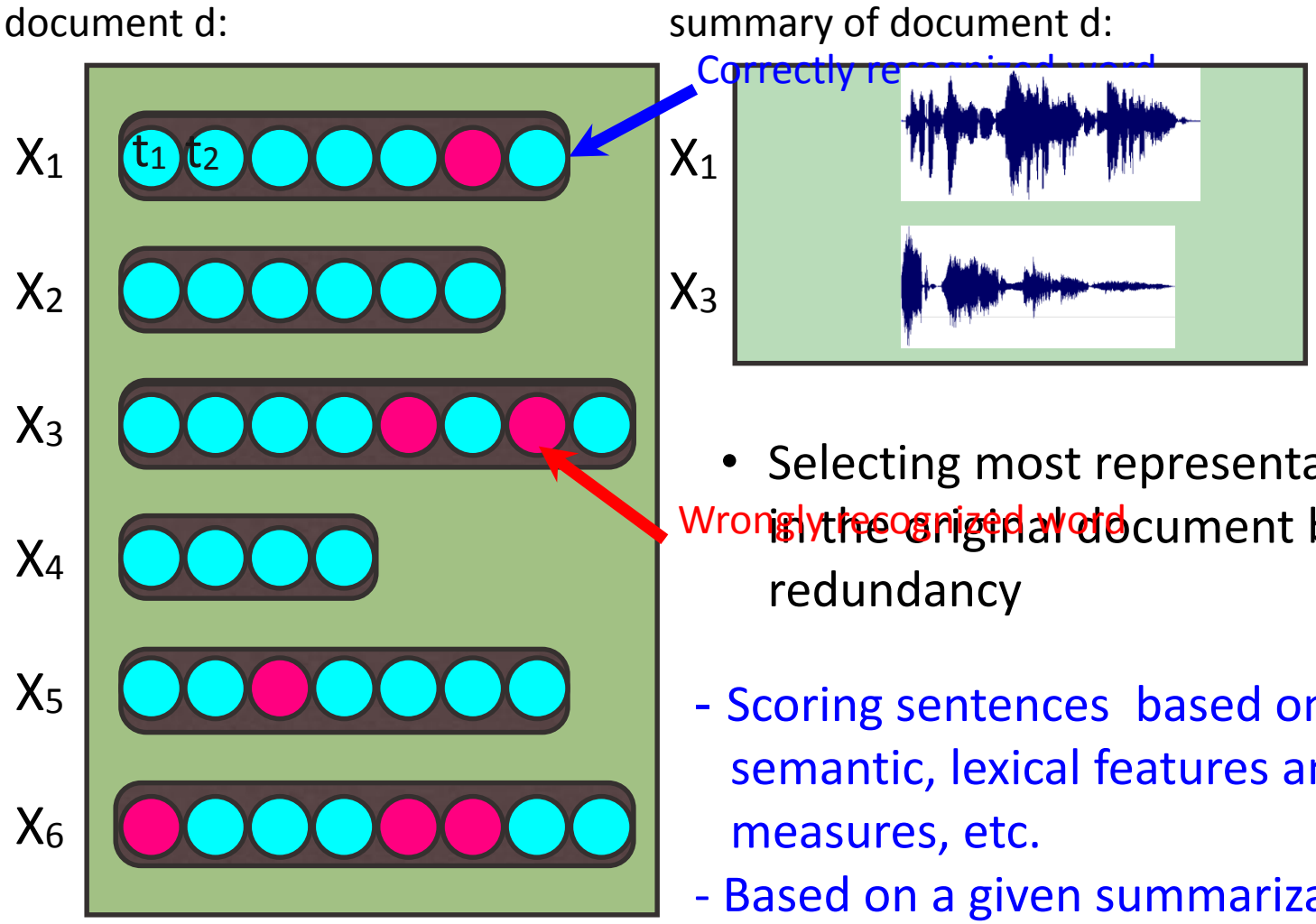
Key Term Extraction from Spoken Content (2/2)

- Prosodic Features
  - key terms probably produced with longer duration, wider pitch range and higher energy
- Semantic Features (e.g. PLSA)
  - key terms usually focused on smaller number of topics



- Lexical Features
  - TF/IDF, POS tag, etc.

Extractive Summarization of Spoken Documents

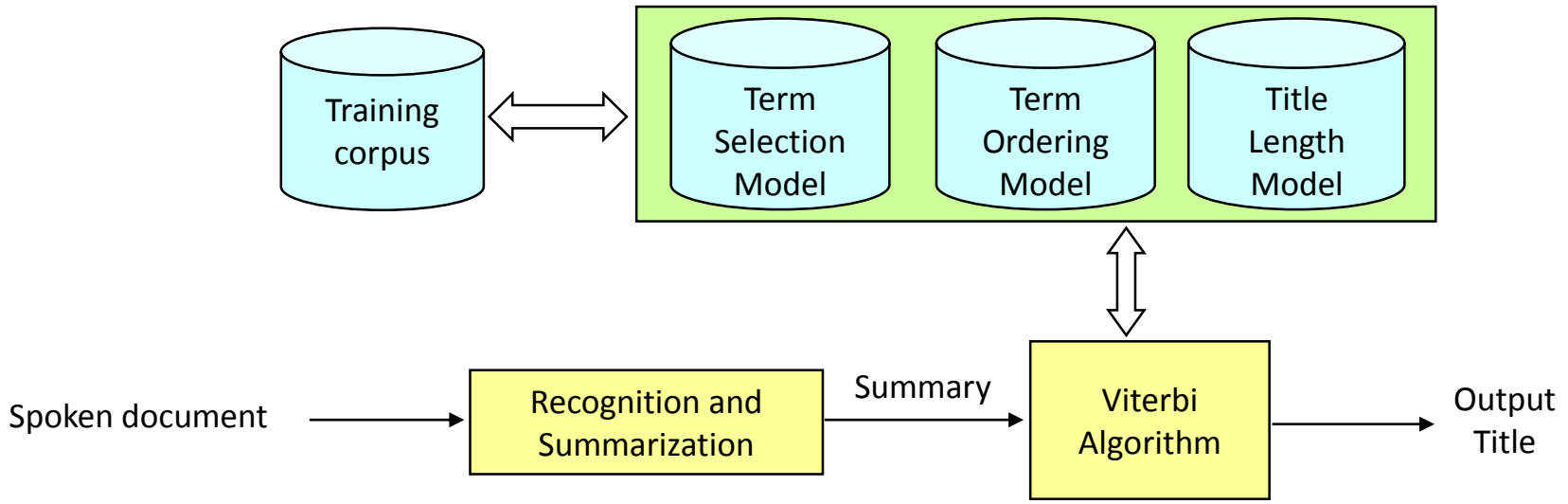


- Selecting most representative utterances in the original document but avoiding redundancy

- Scoring sentences based on prosodic, semantic, lexical features and confidence measures, etc.
- Based on a given summarization ratio

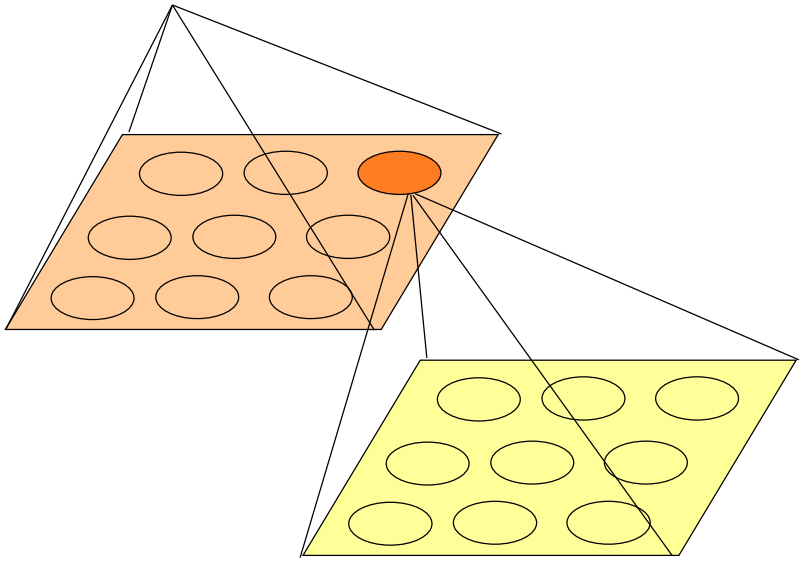
Title Generation for Spoken Documents

- **Titles for retrieved documents/segments helpful in browsing and selection of retrieved results**
- **Short, readable, telling what the document/segment is about**
- **One example: Scored Viterbi Search**



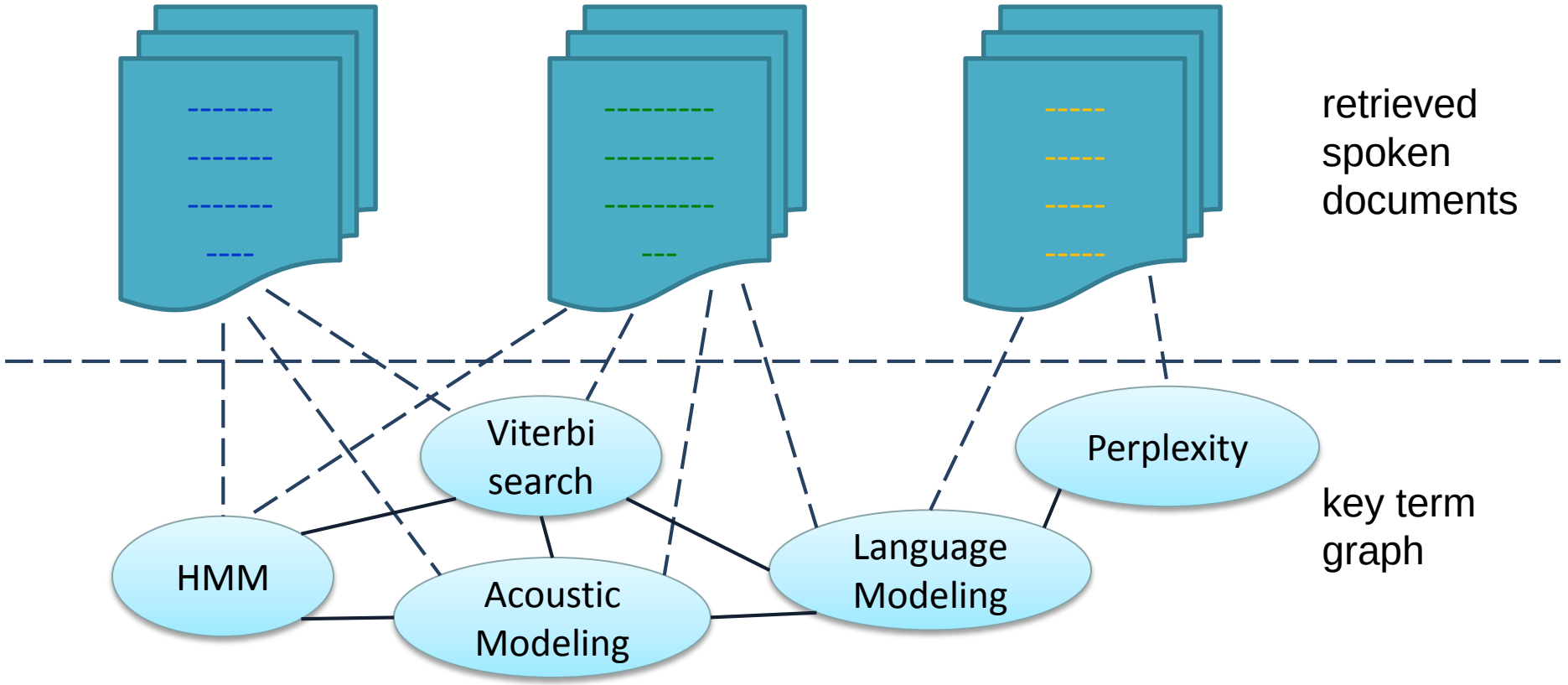
Semantic Structuring (1/2)

- **Example 1: retrieved results clustered by Latent Topics and organized in a two-dimensional tree structure (multi-layered map)**
  - each cluster labeled by a set of key terms representing a group of retrieved documents/segments
  - each cluster expanded into a map in the next layer



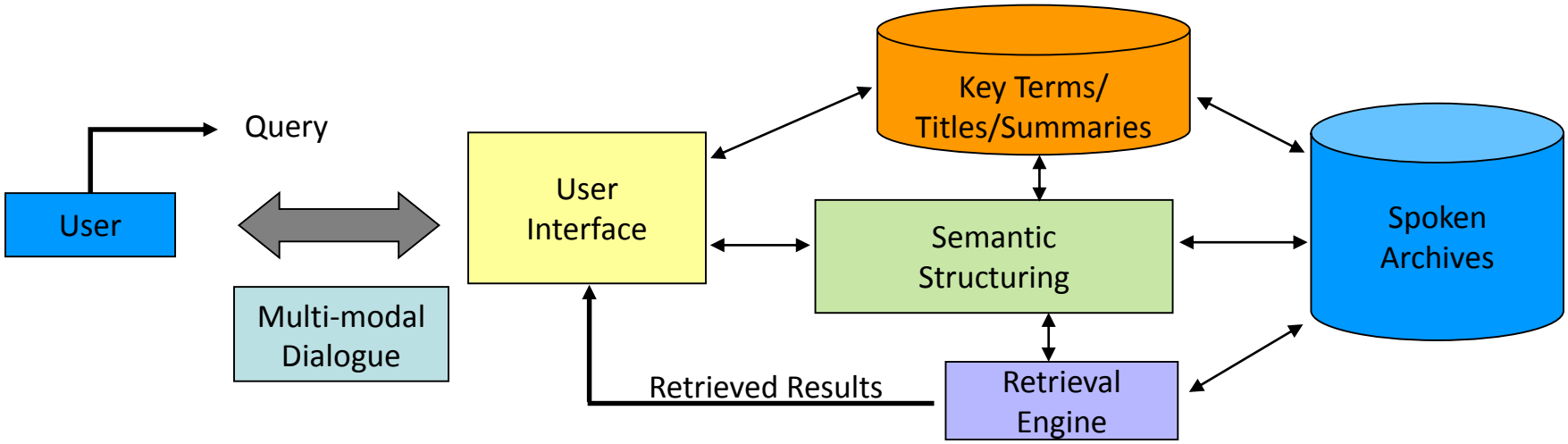
Semantic Structuring (2/2)

- **Example 2: Key-term Graph**
  - each retrieved spoken document/segment labeled by a set of key terms
  - relationships between key terms represented by a graph



Multi-modal Dialogue

- **An example: user-system interaction modeled as a Markov Decision Process (MDP)**



- **Example goals**
  - small average number of dialogue turns (average number of user actions taken) for successful tasks (success: user's information need satisfied)
  - less effort for user, better retrieval quality

Spoken Document Summarization

- **Why summarization?**
  - Huge quantities of information
  - Spoken content difficult to be shown on the screen and difficult to browse

Mails

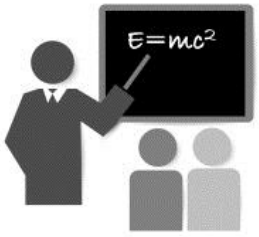
News articles

Meeting

Books

Broadcast News

Lecture

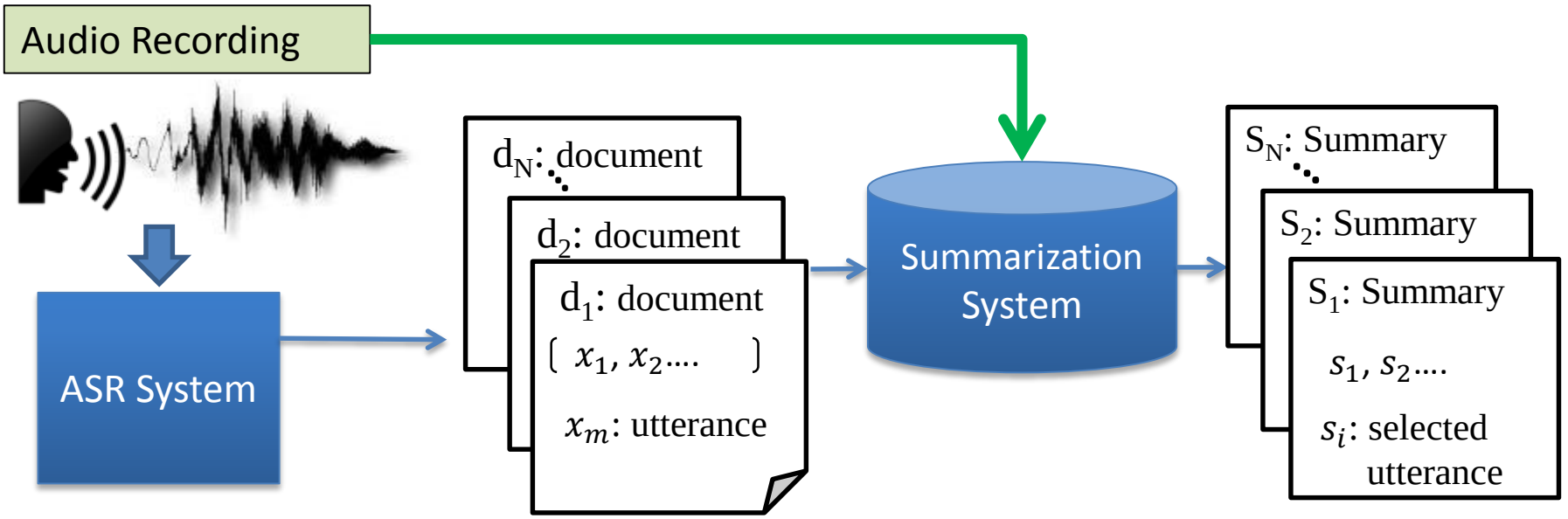


Websites

Social Media

Spoken Document Summarization

- **More difficult than text summarization**
  - Recognition errors, Disfluency, etc.
- **Extra information not in text**
  - Prosody, speaker identity, emotion, etc.



Unsupervised Approach: Maximum Margin Relevance (MMR)

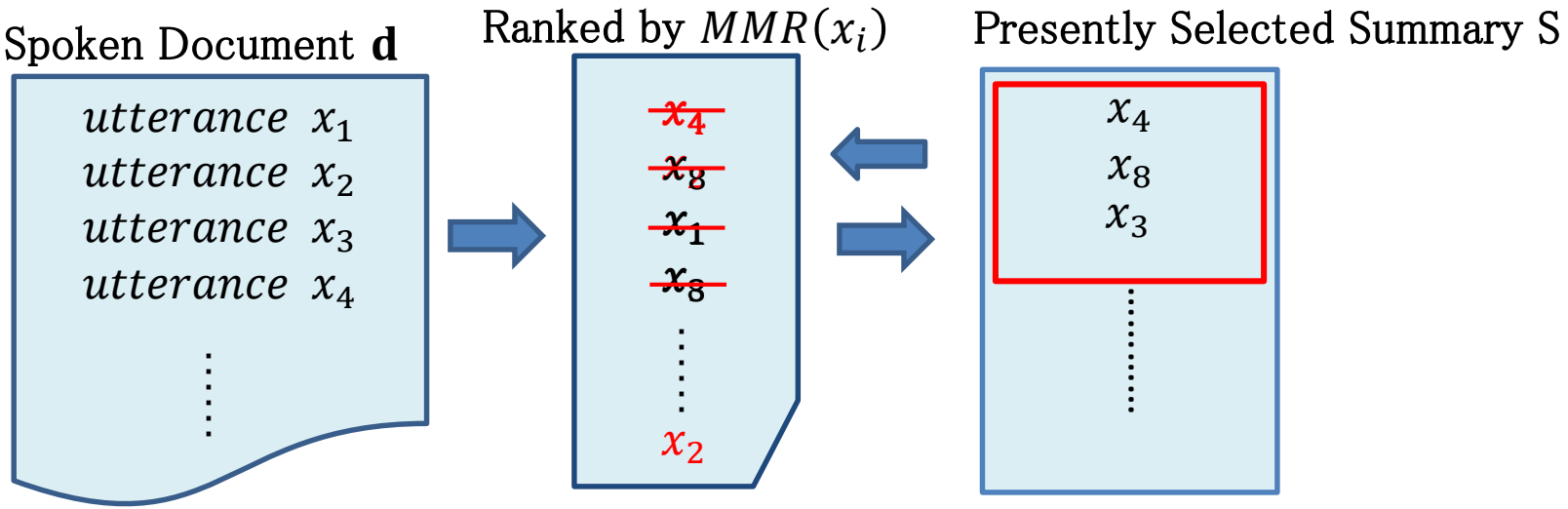
- **Select relevant and non-redundant sentences**

$$MMR(x_i) = Rel(x_i) - \lambda\, Red(x_i, S)$$

Relevance : $Rel(x_i) = Sim(x_i, d)$

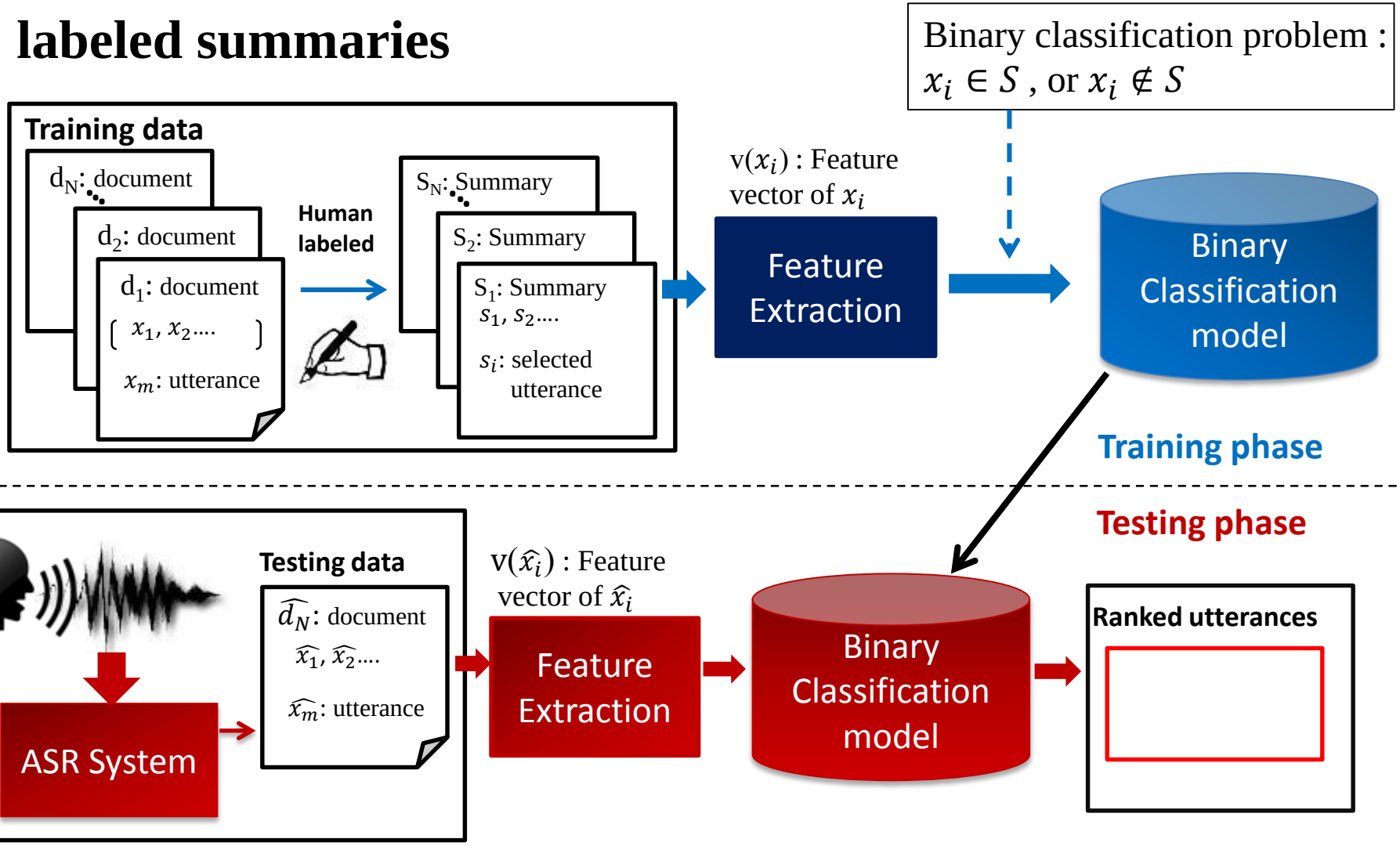
Redundancy : $Red(x_i, S) = Sim(x_i, S)$

Sim $(x_i, \bullet)$: Similarity measure



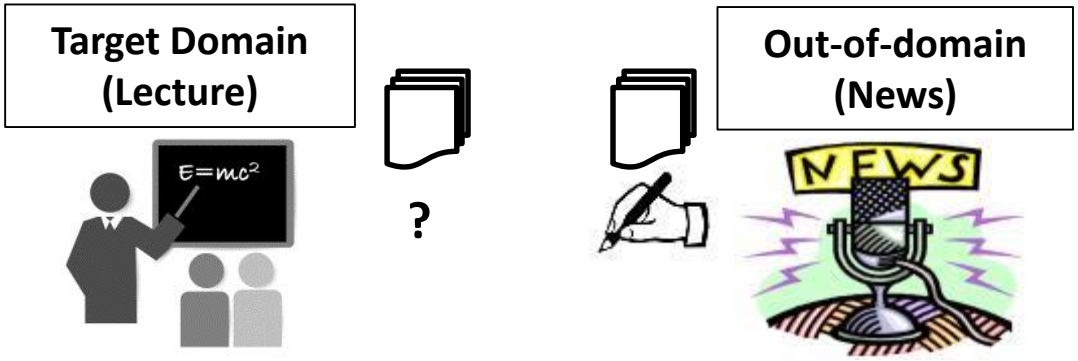
Supervised Approach: SVM or Similar

- **Trained with documents with human labeled summaries**



Domain Adaptation of Supervised Approach

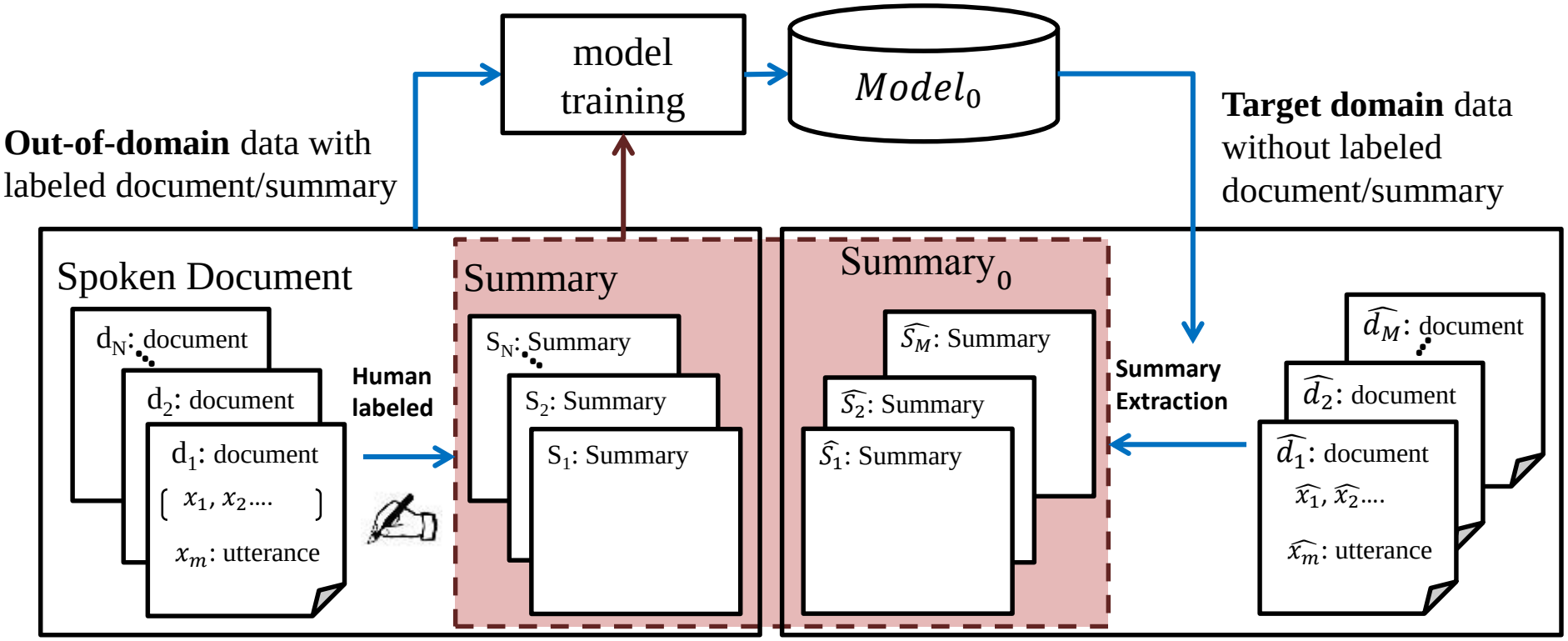
- **Problem**
  - Hard to get high quality training data
  - In most cases, we have labeled **out-of-domain references** but not labeled **target domain references**



- **Goal**
  - Taking advantage of **out-of-domain data**

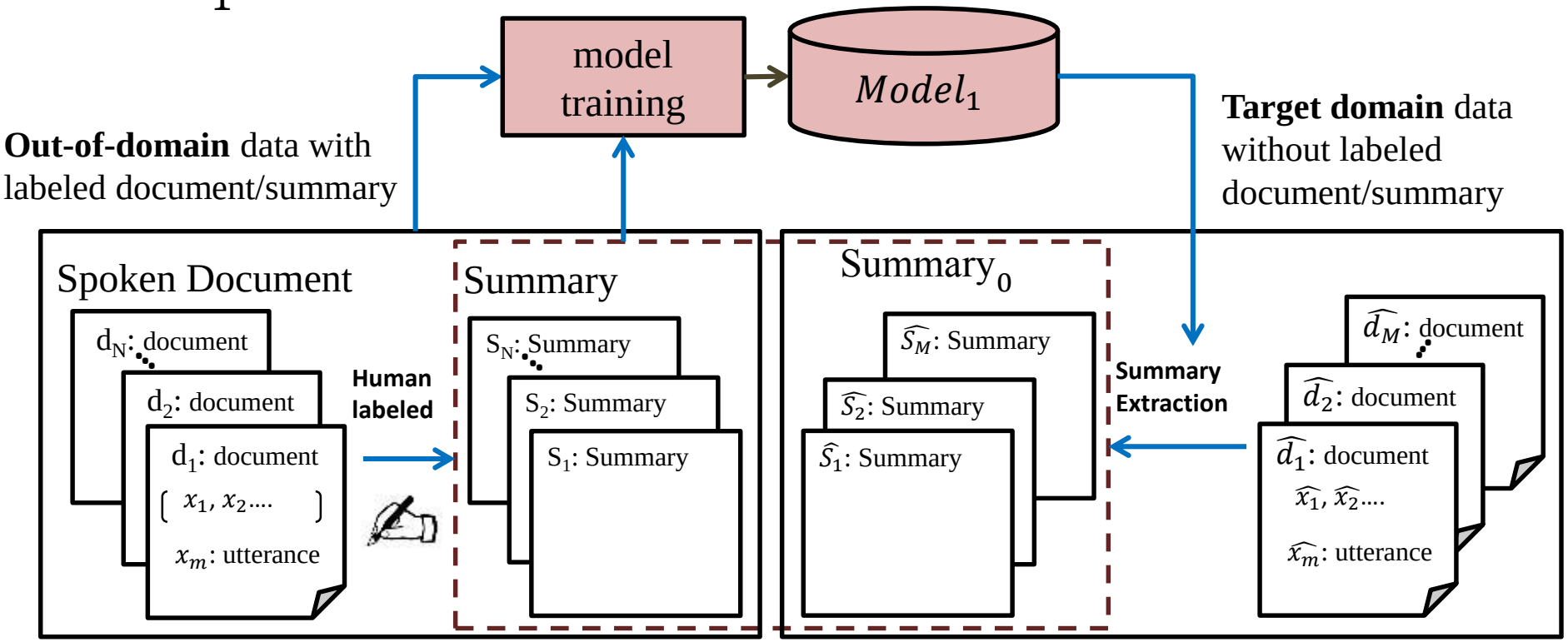
Domain Adaptation of Supervised Approach

- $Model_0$ **trined by out-of-domain data, used to obtain** $summary_0$ **for target domain**



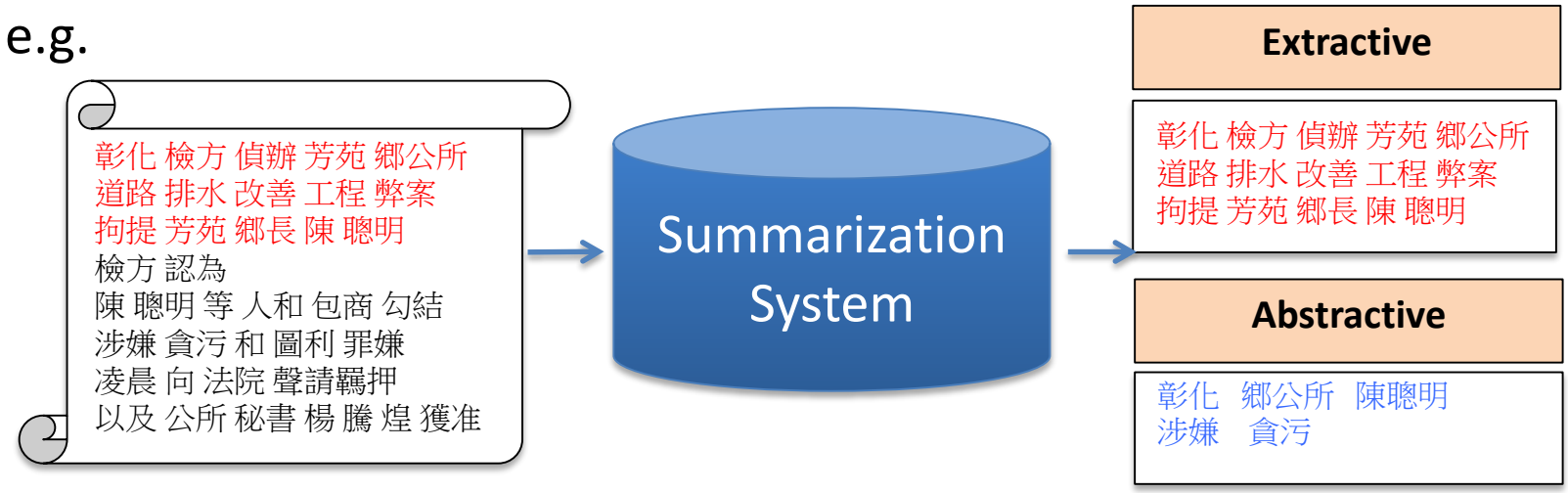
Domain Adaptation of Supervised Approach

- $Model_0$ **trined by out-of-domain data, used to obtain** $summary_0$ **for target domain**
- $summary_0$ **together with out-of-domain data jointly used to train** $Model_1$



Document Summarization

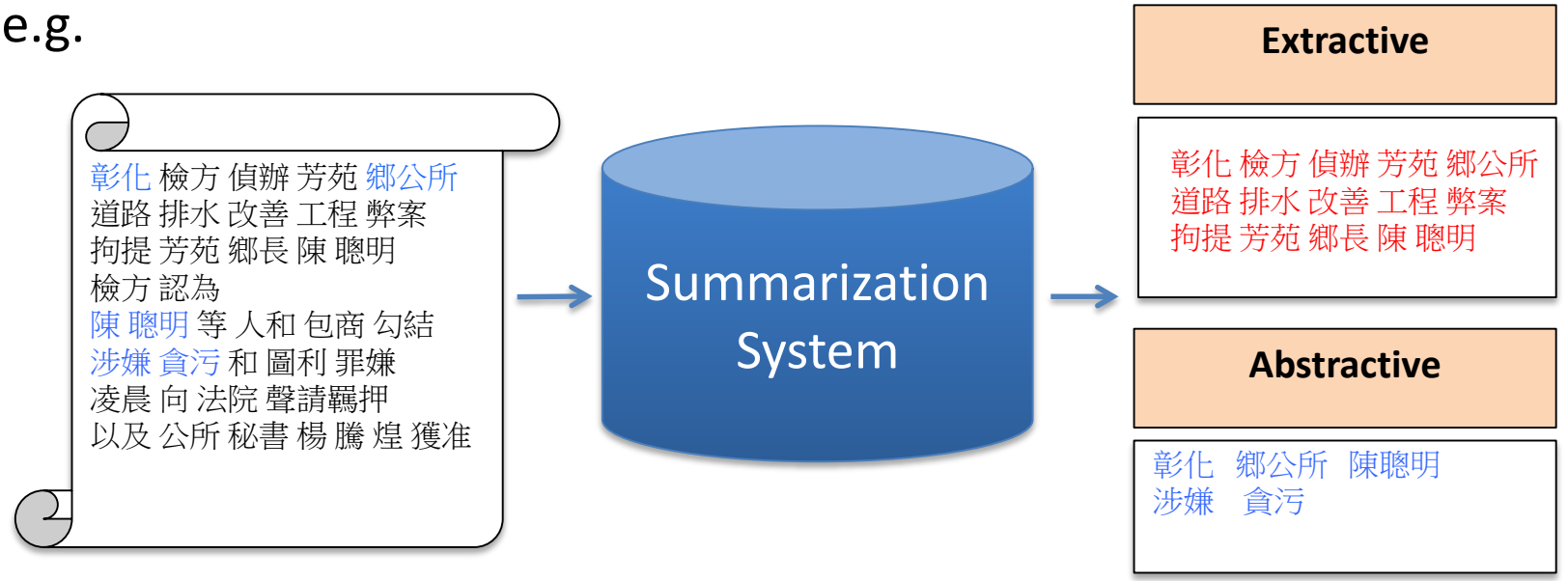
- **Extractive Summarization**
  - select **sentences** in the document
- **Abstractive Summarization**
  - Generate sentences describing the content of the document



Document Summarization

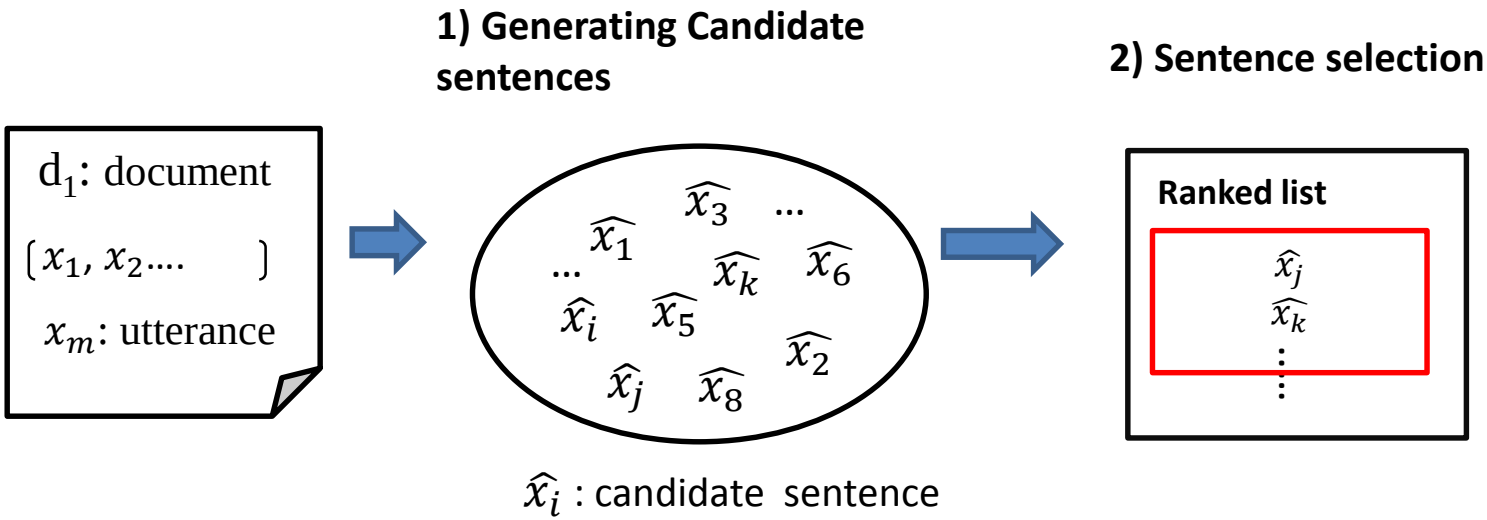
- **Extractive Summarization**
  - select **sentences** in the document
- **Abstractive Summarization**
  - Generate sentences describing the content of the document

e.g.



Abstractive Summarization (1/4)

- **An Example Approach**
  (1) Generating candidate sentences by a graph
  (2) Selecting sentences by topic models, language models of words, parts-of-speech(POS), length constraint, etc.



Abstractive Summarization (2/4)

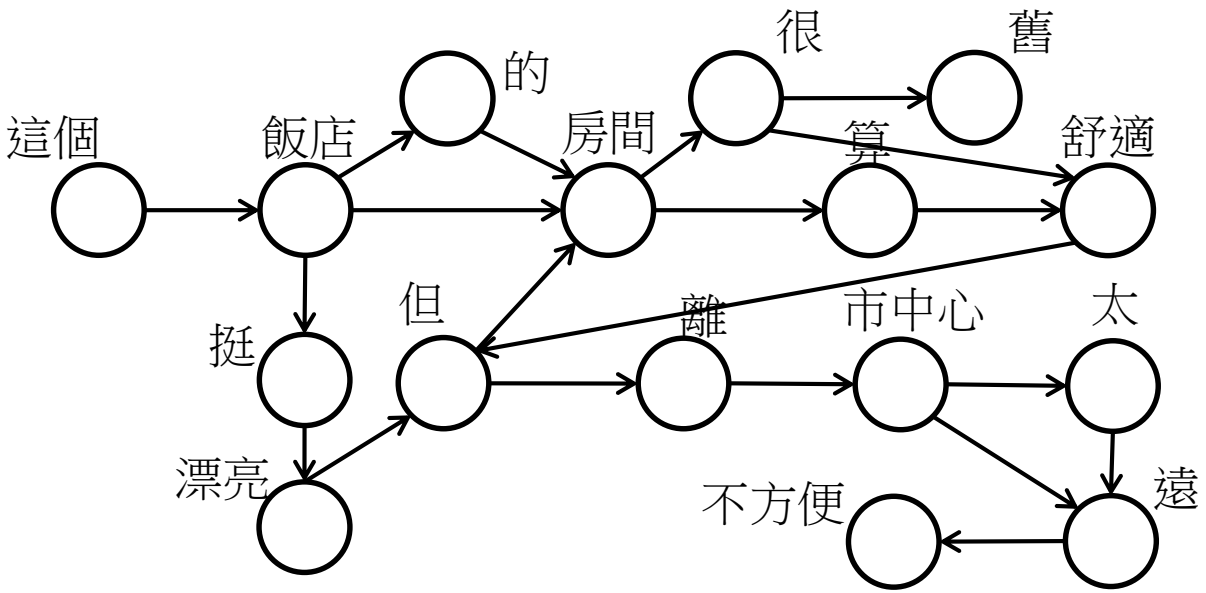
- 1) Generating Candidate sentences **Graph construction** + search on graph
  - Node : “word” in the sentence
  - Edge : word ordering in the sentence

- X1 : 這個 飯店 房間 算 舒適.
- X2 : 這個 飯店 的 房間 很 舒適 但 離 市中心 太遠 不方便
- X3 : 飯店 挺 漂亮 但 房間 很 舊
- X4 : 離 市中心 遠

Abstractive Summarization (3/4)

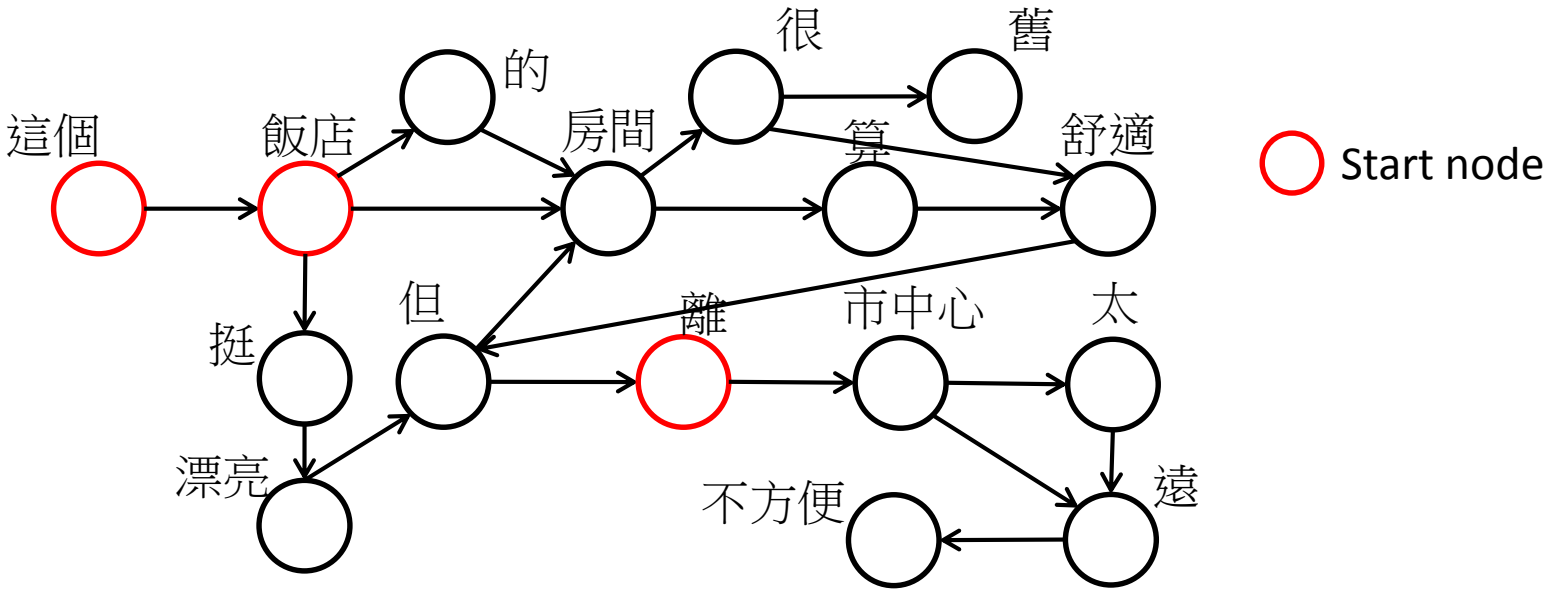
- 1) Generating Candidate sentences **Graph construction** + search on graph

  - X1 : 這個 飯店 房間 算 舒適
  - X2 : 這個 飯店 的 房間 很 舒適 但 離 市中心 太遠 不方便
  - X3 : 飯店 挺 漂亮 但 房間 很 舊
  - X4 : 離 市中心 遠



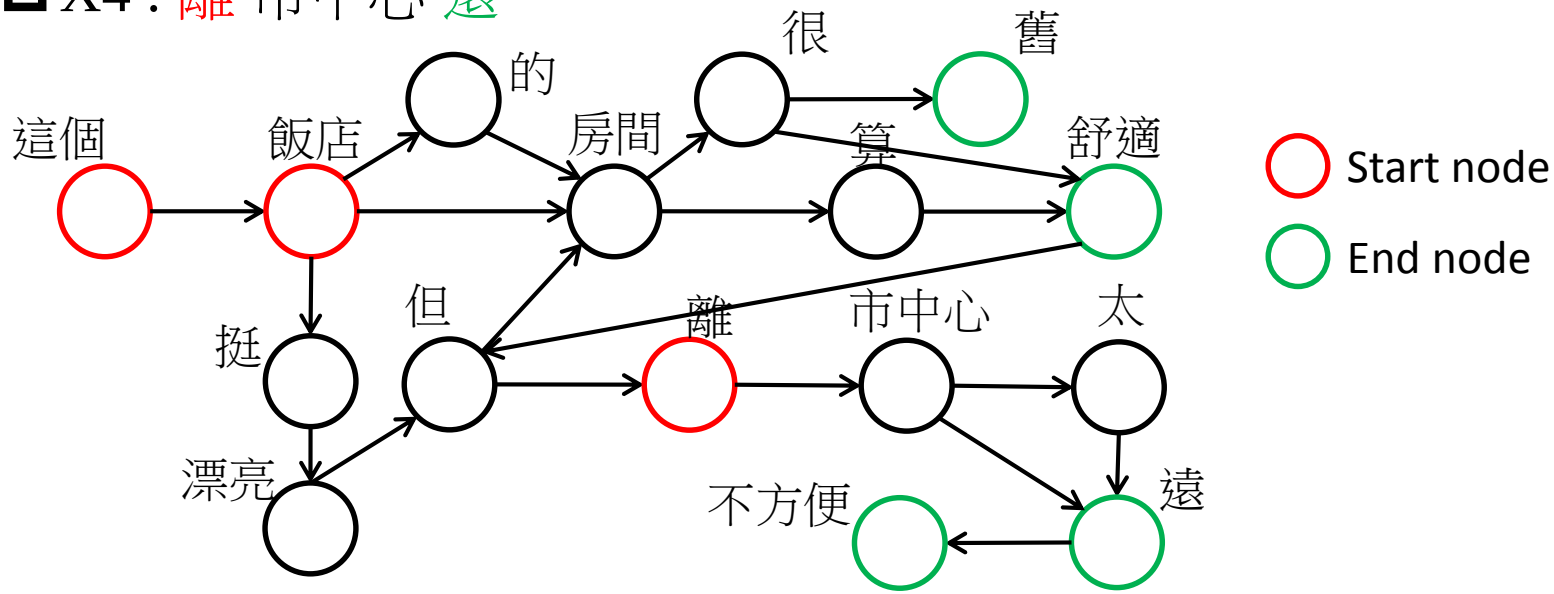
Abstractive Summarization (3/4)

- **1) Generating Candidate sentences** **Graph construction**
  + search on graph
  - X1 : 這個 飯店 房間 算 舒適
  - X2 : 這個 飯店 的 房間 很 舒適 但 離 市中心 太遠 不方便
  - X3 : 飯店 挺 漂亮 但 房間 很 舊
  - X4 : 離 市中心 遠



Abstractive Summarization (3/4)

- **1) Generating Candidate sentences** **Graph construction + search on graph**
  - X1 : 這個 飯店 房間 算 舒適
  - X2 : 這個 飯店 的 房間 很 舒適 但 離 市中心 太遠 不方便
  - X3 : 飯店 挺 漂亮 但 房間 很 舊
  - X4 : 離 市中心 遠

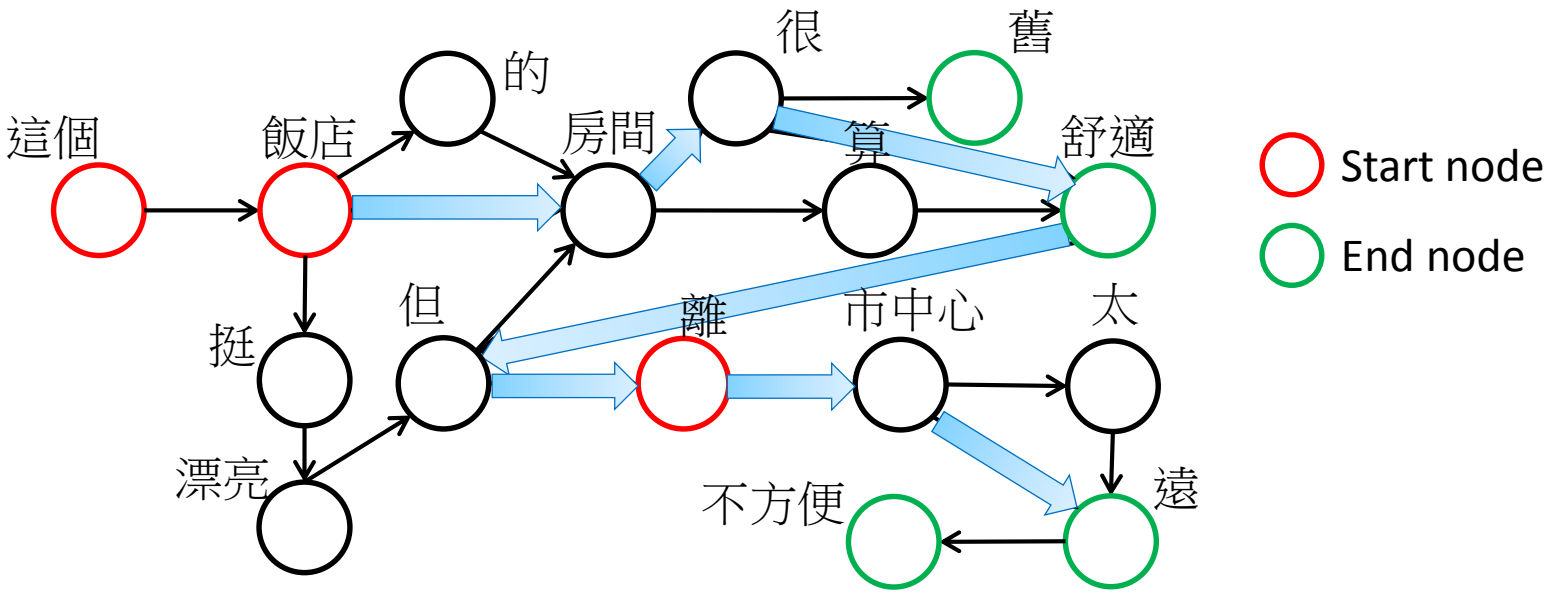


Abstractive Summarization (4/4)

- **1) Generate Candidate sentences** **Graph construction + search on graph**
  - Search : find Valid path on graph
  - Valid path : path from start node to end node

e.g. 飯店 房間 很 舒適 但 離 市中心 遠

- X1 : 這個 飯店 房間 算 舒適
- X2 : 這個 飯店 的 房間 很 舒適 但 離 市中心 太遠 不方便
- X3 : 飯店 挺 漂亮 但 房間 很 舊
- X4 : 離 市中心 遠

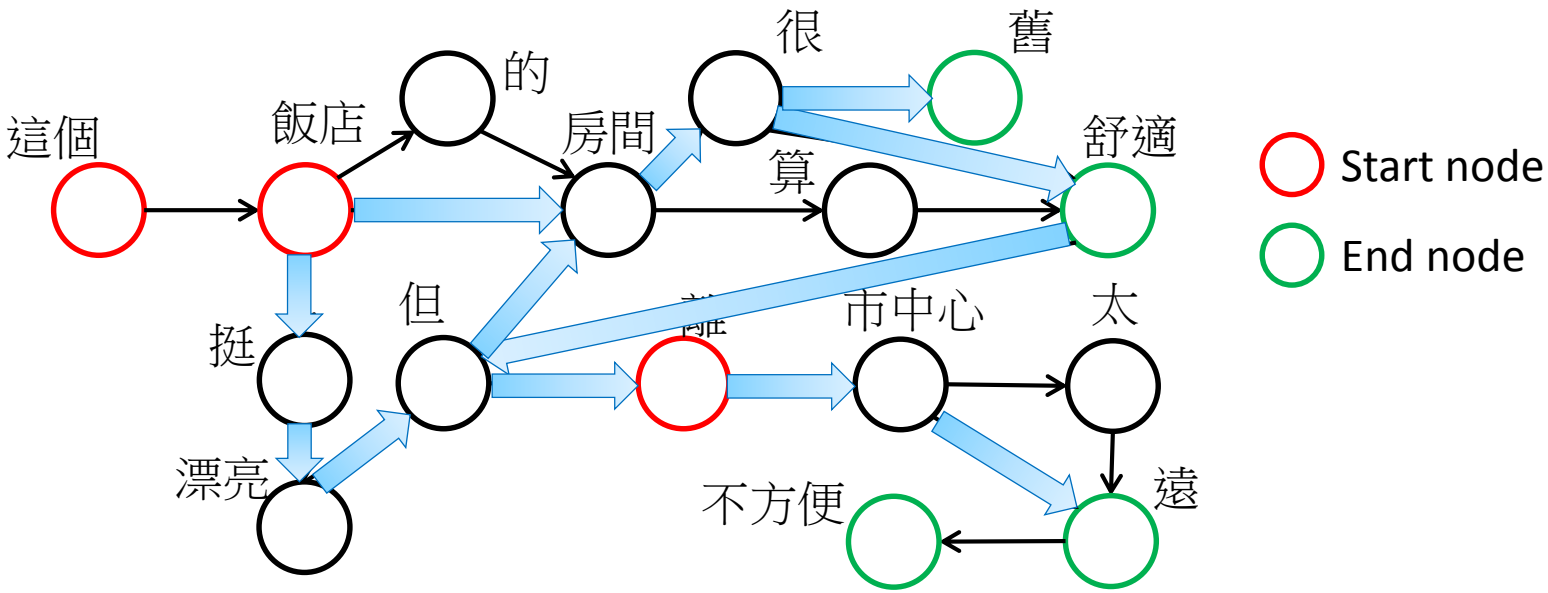


Abstractive Summarization (4/4)

- **1) Generating Candidate sentences** **Graph construction**
  + search on graph
  - Search : find Valid path on graph
  - Valid path : path from start node to end node

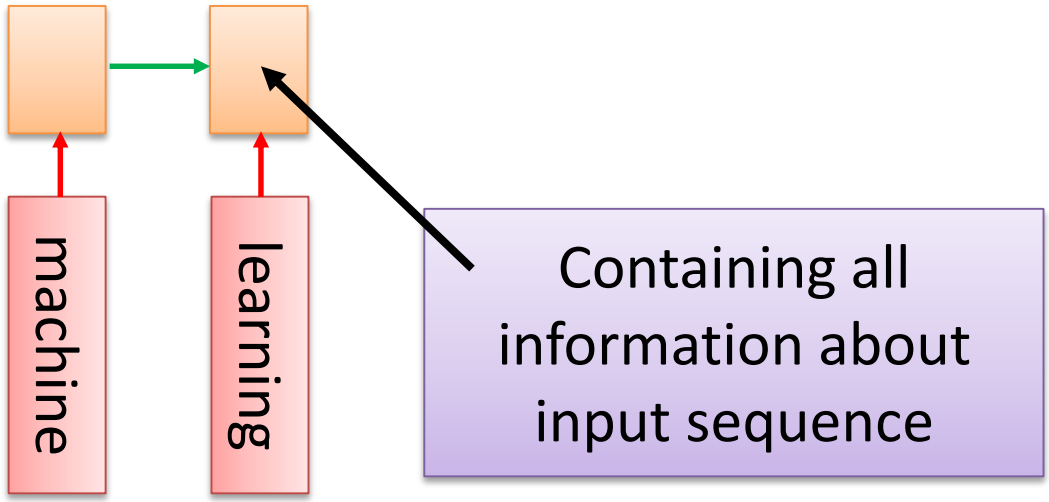
- X1 : 這個 飯店 房間 算 舒適
- X2 : 這個 飯店 的 房間 很 舒適 但 離 市中心 太遠 不方便
- X3 : 飯店 挺 漂亮 但 房間 很 舊
- X4 : 離 市中心 遠

e.g. 飯店 房間 很 舒適 但 離 市中心 遠
飯店 挺 漂亮 但 房間 很舊



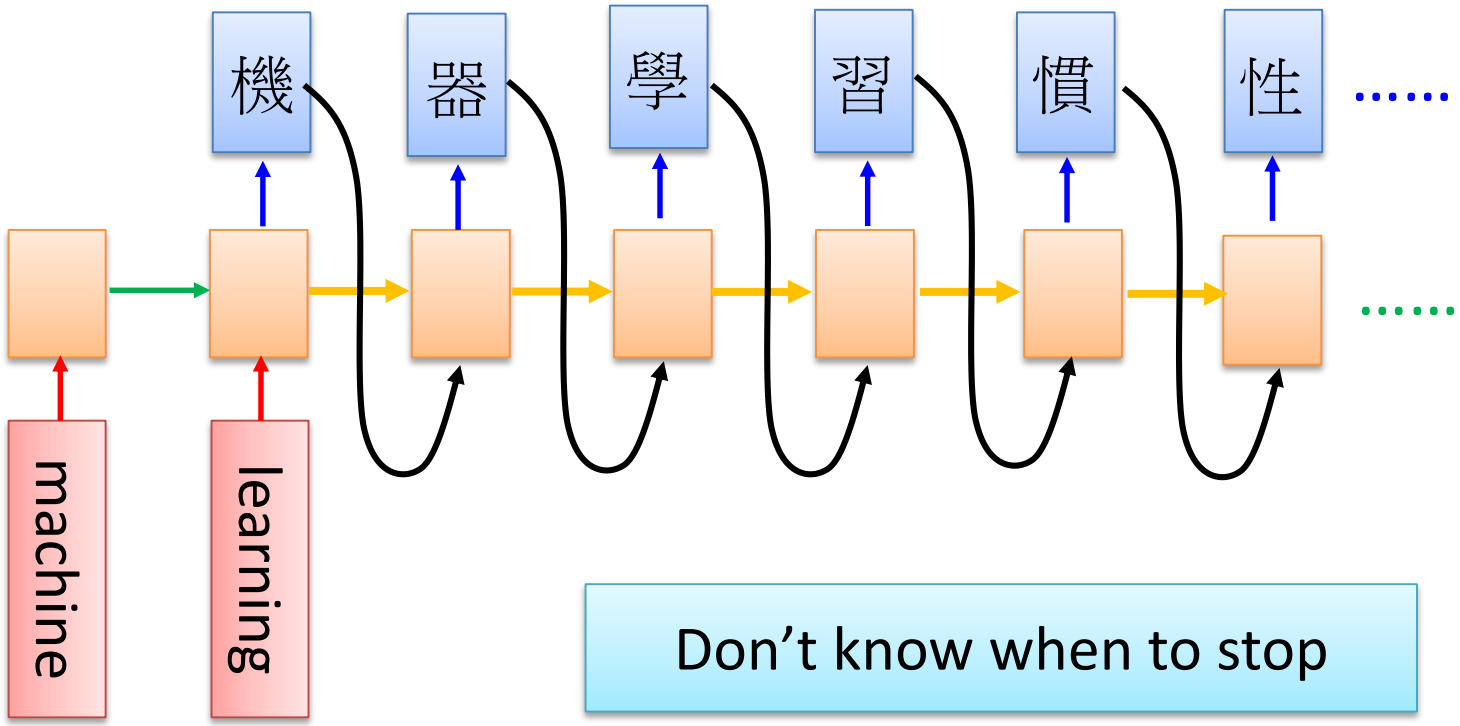
Sequence-to-Sequence Learning (1/3)

- **Both input and output are sequences *with different lengths*.**
  - machine translation (machine learning→機器學習)
  - summarization, title generation
  - spoken dialogues
  - speech recognition



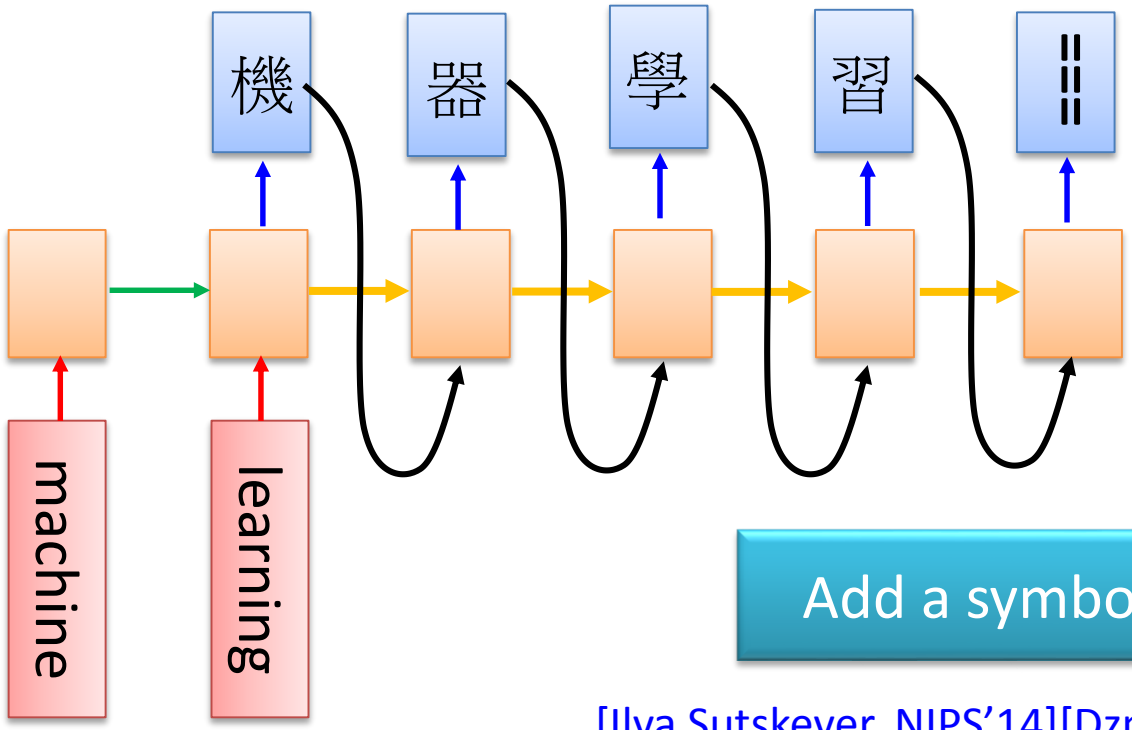
Sequence-to-Sequence Learning (2/3)

- **Both input and output are sequences *with different lengths*.**
  - machine translation (machine learning→機器學習)
  - summarization, title generation
  - spoken dialogues
  - speech recognition



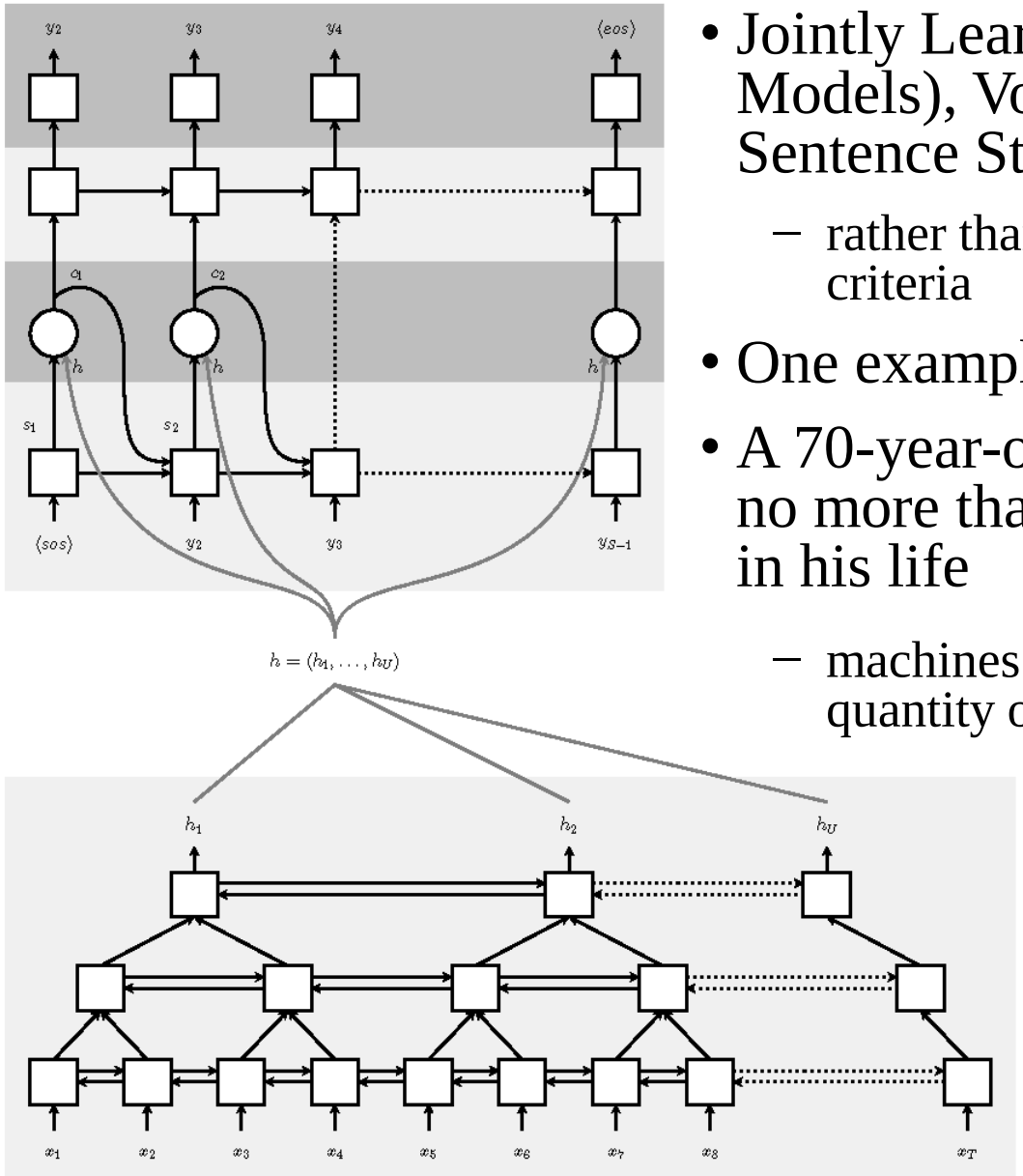
Sequence-to-Sequence Learning (3/3)

- **Both input and output are sequences *with different lengths*.**
  - machine translation (machine learning→機器學習)
  - summarization, title generation
  - spoken dialogues
  - speech recognition



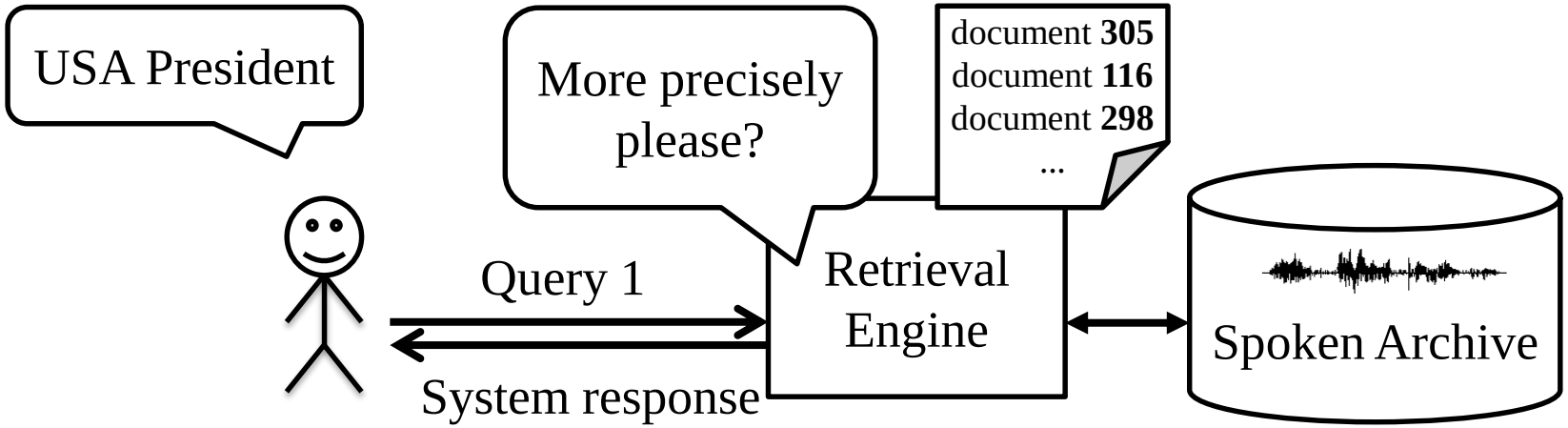
[Ilya Sutskever, NIPS'14][Dzmitry Bahdanau, arXiv'15]

End-to-end Deep Learning for Speech Recognition



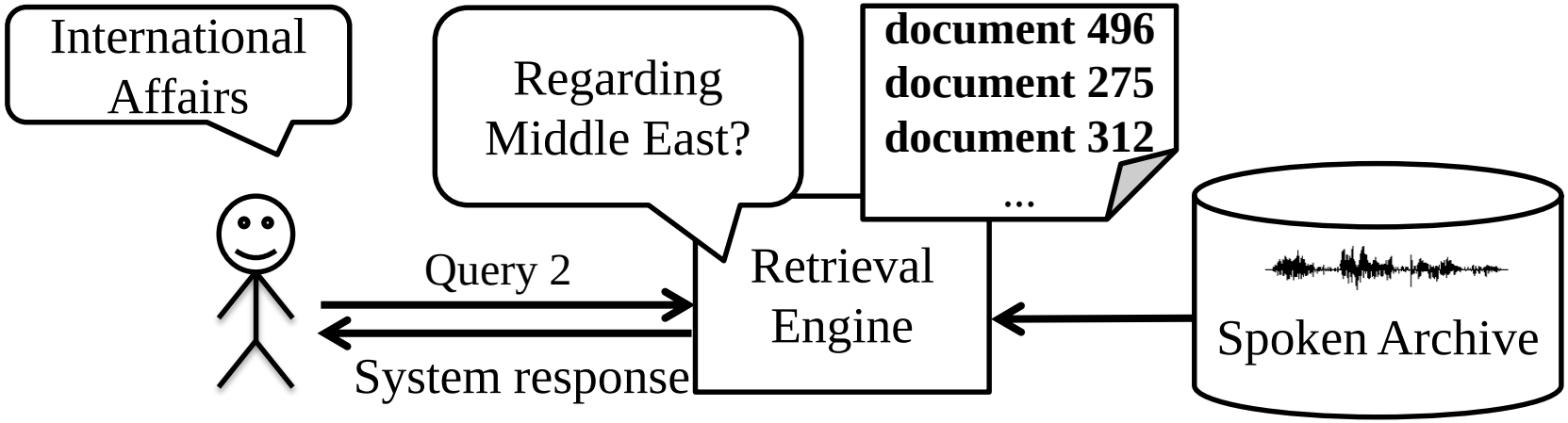
- Jointly Learn the Sound (Acoustic Models), Vocabulary (Lexicon) and Sentence Structure (Language Model)
  - rather than trained separately with different criteria
- One example
- A 70-year-old person has heard roughly no more than 0.6 million of hrs of voice in his life
  - machines can be trained with more than this quantity of data in very short time

Multi-modal Interactive Dialogue



- **Interactive dialogue: retrieval engine interacts with the user to find out more precisely his information need**
  - User entering the query
  - When the retrieved results are divergent, the system may ask for more information rather than offering the results

Multi-modal Interactive Dialogue



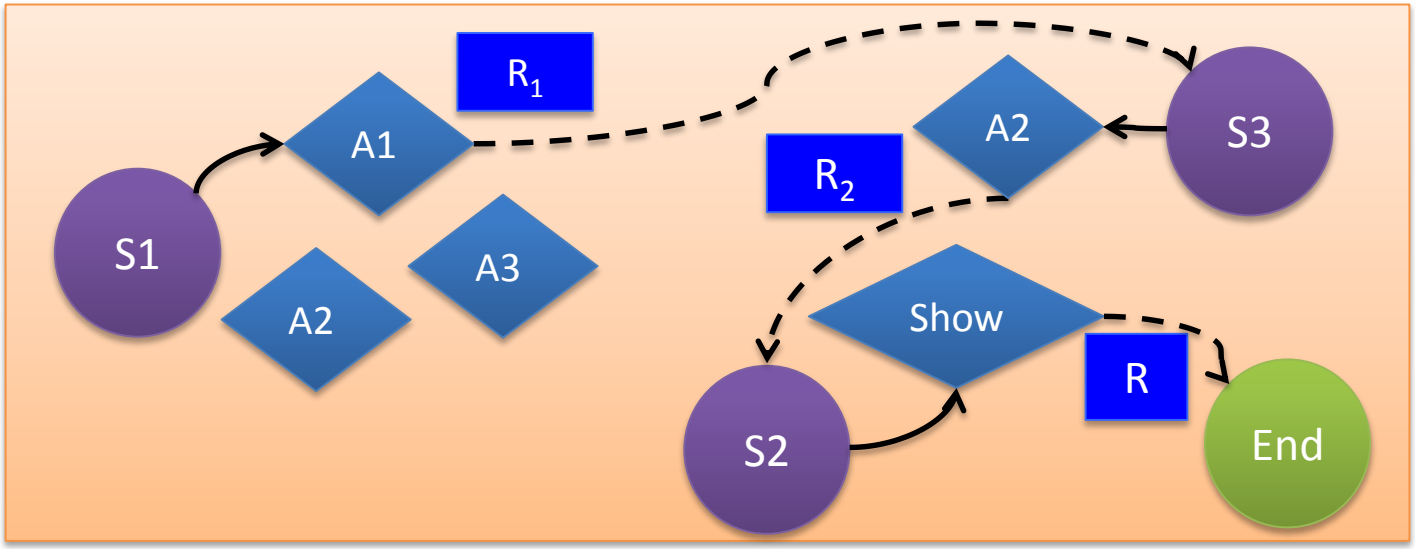
- **Interactive dialogue: retrieval engine interacts with the user to find out more precisely his information need**
  - User entering the second query
  - when the retrieved results are still divergent, but seem to have a major trend, the system may use a key word representing the major trend asking for confirmation
  - User may reply： “Yes” or “No, Asia”

Markov Decision Process (MDP)

- **A mathematical framework for decision making, defined by (S,A,T,R,π)**
  - S: Set of states, current system status
    $\{s_1, s_2, s_3,$
  - A: Set of actions the system can take at each state
    $\{A_1, A_2, A_3,$
  - T: transition probabilities between states when a certain action is taken
  - R: reward received when taking an action
    $\{R_1, R_2, R_3,$
  - π: policy, choice of action given the state
    $\{\pi: s_i \rightarrow A_j\}$
- **Objective : Find a policy that maximizes the expected total reward**

Multi-modal Interactive Dialogue

Model as ***Markov Decision Process (MDP)***



- After a query entered, the system starts at a certain state
- States: retrieval result quality estimated as a continuous variable (e.g. MAP) plus the present dialogue turn
- Action: at each state, there is a set of actions which can be taken: asking for more information, returning a keyword or a document, or a list of keywords or documents asking for selecting one, or showing results….
- User response corresponds to a certain negative reward (extra work for user)
- when the system decides to show to the user the retrieved results, it earns some positive reward (e.g. MAP improvement)
- Learn a policy maximizing rewards from historical user interactions(π: $S_i \rightarrow A_j$)

Reinforcement Learning

- **Example approach: Value Iteration**
  - Define value function: $Q^{\pi} : S \times A \rightarrow \mathbb{R}$

    $Q^{\pi}(s,a) = E\left[\Sigma_{k=0}^{\infty} \gamma^{k} r_{k} \middle| s_0 = s, a_0 = a\right]$

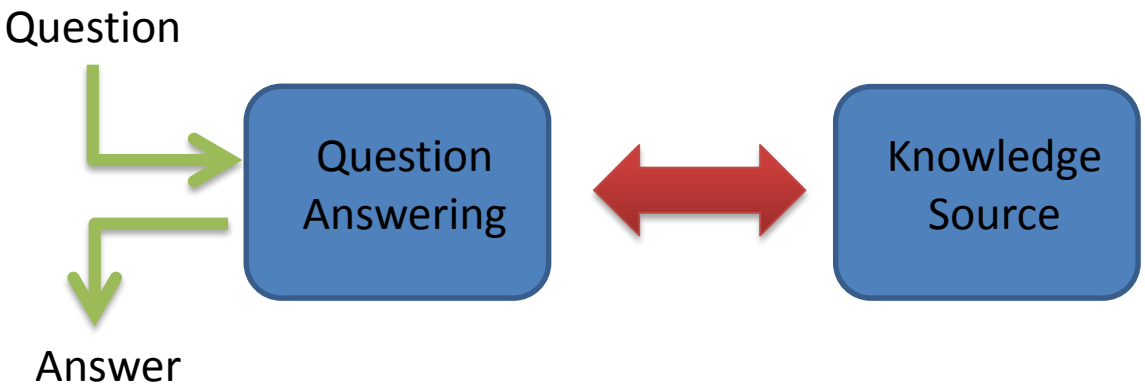
    the expected discounted sum of rewards given π

    started from (s,a)
  - The real value of Q can be estimated iteratively from a training set:

    $Q^{*}(s,a) = E_{s'|s,a}[R(s,a,s') + \gamma_{b\in A}^{max} Q^{*}(s',b)]$

    $Q^{*}(s,a)$:estimated value function based on the training set
  - Optimal policy is learned by choosing the best action given each state such that the value function is maximized

Question-Answering (QA) in Speech



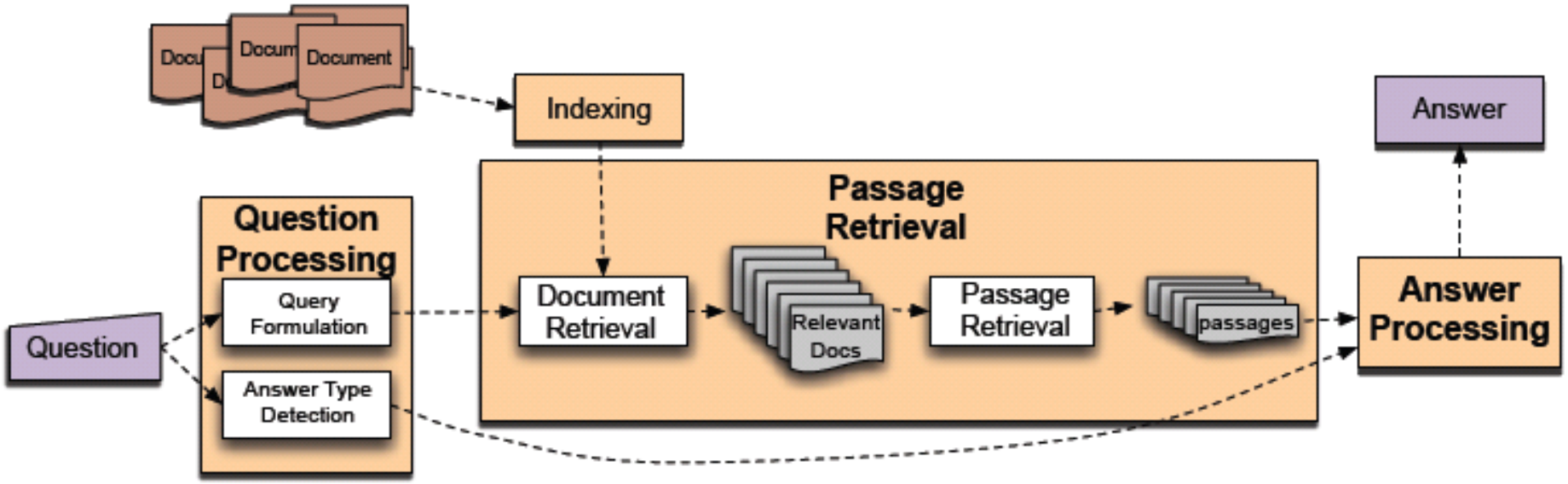
- **Question, Answer, Knowledge Source can all be in text form or in Speech**
- **Spoken Question Answering becomes important**
  - spoken questions and answers are attractive
  - the availability of large number of on-line courses and shared videos today makes spoken answers by distinguished instructors or speakers more feasible, etc.
- **Text Knowledge Source is always important**

Three Types of QA

- **Factoid QA:**
  - What is the name of the largest city of Taiwan? Ans: Taipei.
- **Definitional QA :**
  - What is QA?
- **Complex Question:**
  - How to construct a QA system?

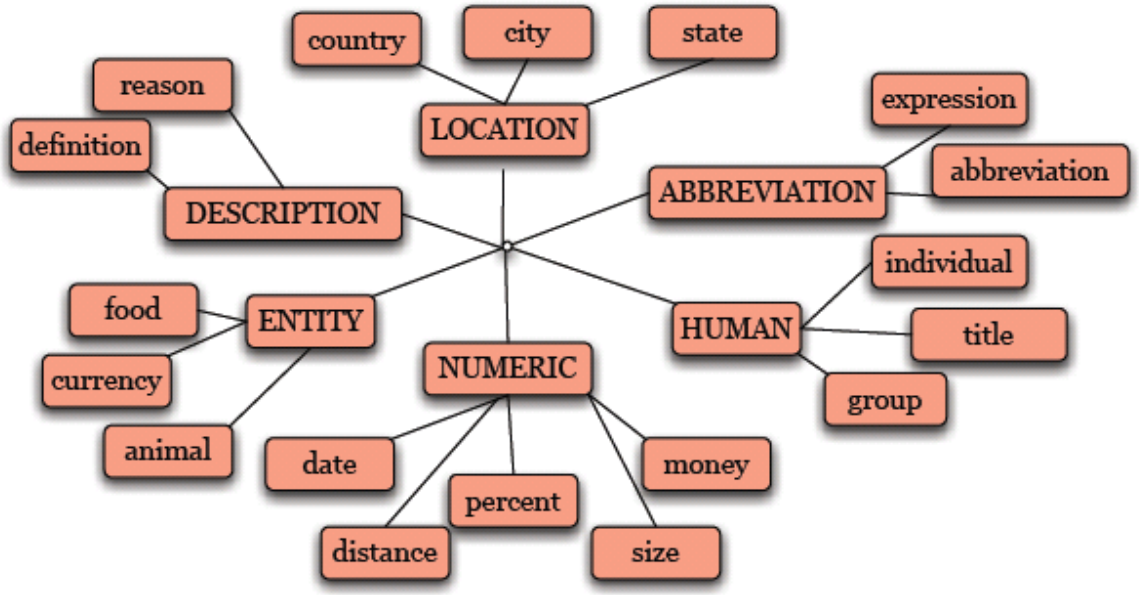
Factoid QA

- **Question Processing**
  - Query Formulation: transform the question into a query for retrieval
  - Answer Type Detection (city name, number, time, etc.)
- **Passage Retrieval**
  - Document Retrieval, Passage Retrieval
- **Answer Processing**
  - Find and rank candidate answers



Factoid QA – Question Processing

- **Query Formulation: Choose key terms from the question**
  - Ex: What is the name of the largest city of Taiwan?
  - "Taiwan", "largest city " are key terms and used as query
- **Answer Type Detection**
  - "city name" for example
  - Large number of hierarchical classes hand-crafted or automatically learned



An Example Factoid QA

- **Watson: a QA system develop by IBM (text-based, no speech), who won "Jeopardy!"**

Definitional QA

- **Definitional QA ≈ Query-focused summarization**
- **Use similar framework as Factoid QA**
  - Question Processing
  - Passage Retrieval
  - Answer Processing is replaced by Summarization

References

- **Key terms**
  - "Automatic Key Term Extraction From Spoken Course Lectures Using Branching Entropy and Prosodic/Semantic Features", IEEE Workshop on Spoken Language Technology, Berkeley, California, U.S.A., Dec 2010, pp. 253-258.
  - "Unsupervised Two-Stage Keyword Extraction from Spoken Documents by Topic Coherence and Support Vector Machine", International Conference on Acoustics, Speech and Signal Processing, Kyoto, Japan, Mar 2012, pp. 5041-5044.
- **Title Generation**
  - "Automatic Title Generation for Spoken Documents with a Delicate Scored Viterbi Algorithm", 2nd IEEE Workshop on Spoken Language Technology, Goa, India, Dec 2008, pp. 165-168.
  - "Abstractive Headline Generation for Spoken Content by Attentive Recurrent Neural Networks with ASR Error Modeling" IEEE Workshop on Spoken Language Technology (SLT), San Diego, California, USA, Dec 2016, pp. 151-157.

References

- **Summarization**
  - “Supervised Spoken Document Summarization Jointly Considering Utterance Importance and Redundancy by Structured Support Vector Machine”, Interspeech, Portland, U.S.A., Sep 2012.
  - “Unsupervised Domain Adaptation for Spoken Document Summarization with Structured Support Vector Machine”, International Conference on Acoustics, Speech and Signal Processing, Vancouver, Canada, May 2013.
  - “Supervised Spoken Document Summarization Based on Structured Support Vector Machine with Utterance Clusters as Hidden Variables”, Interspeech, Lyon, France, Aug 2013, pp. 2728-2732.
  - “Semantic Analysis and Organization of Spoken Documents Based on Parameters Derived from Latent Topics”, IEEE Transactions on Audio, Speech and Language Processing, Vol. 19, No. 7, Sep 2011, pp. 1875-1889.
  - "Spoken Lecture Summarization by Random Walk over a Graph Constructed with Automatically Extracted Key Terms," InterSpeech 2011

References

- **Summarization**
  - "Speech-to-text and Speech-to-speech Summarization of Spontaneous Speech", IEEE Transactions on Speech and Audio Processing, Dec. 2004
  - "The Use of MMR, diversity-based reranking for reordering document and producing summaries" SIGIR, 1998
  - "Using Corpus and Knowledge-based Similarity Measure in Maximum Marginal Relevance for Meeting Summarization" ICASSP, 2008
  - "Opinosis: A Graph-Based Approach to Abstractive Summarization of Highly Redundant Opinions", International Conference on Computational Linguistics , 2010

References

- **Interactive Retrieval**
  - "Interactive Spoken Content Retrieval by Extended Query Model and Continuous State Space Markov Decision Process", International Conference on Acoustics, Speech and Signal Processing, Vancouver, Canada, May 2013.
  - "Interactive Spoken Content Retrieval by Deep Reinforcement Learning", Interspeech, San Francisco, USA, Sept 2016.
  - Reinforcement Learning: An Introduction, Richard S. Sutton and Andrew G. Barto, The MIT Press, 1999.
  - Partially observable Markov decision processes for spoken dialog systems, Jason D. Williams and Steve Young, Computer Speech and Language, 2007.

Reference

- **Question Answering**
  - Rosset, S., Galibert, O. and Lamel, L. (2011) Spoken Question Answering, in Spoken Language Understanding: Systems for Extracting Semantic Information from Speech
  - Pere R. Comas, Jordi Turmo, and Lluís Màrquez. 2012. “Sibyl, a factoid question-answering system for spoken documents.” ACM Trans. Inf. Syst. 30, 3, Article 19 (September 2012), 40
  - “Towards Machine Comprehension of Spoken Content: Initial TOEFL Listening Comprehension Test by Machine”, Interspeech, San Francisco, USA, Sept 2016, pp. 2731-2735.
  - “Hierarchical Attention Model for Improved Comprehension of Spoken Content”, IEEE Workshop on Spoken Language Technology (SLT), San Diego, California, USA, Dec 2016, pp. 234-238.

Reference

- **Sequence-to-sequence Learning and End-to-end Speech Recognition**
  - “Sequence to Sequence Learning with Neural Networks”, NIPS, 2014
  - “Listen, Attend and Spell: A Neural Network for Large Vocabulary Conversational Speech Recognition”, ICASSP 2016
  - Alex Graves, Santiago Fernandez, Faustino Gomez, and Jürgen Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in Proceedings of the 23rd international conference on Machine learning. ACM, 2006, pp. 369-376
  - Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio, “End-to-end attention-based large vocabulary speech recognition,” in Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016, pp. 4945-4949