Towards Machine Comprehension of Spoken Content

Hung-yi Lee

李宏毅

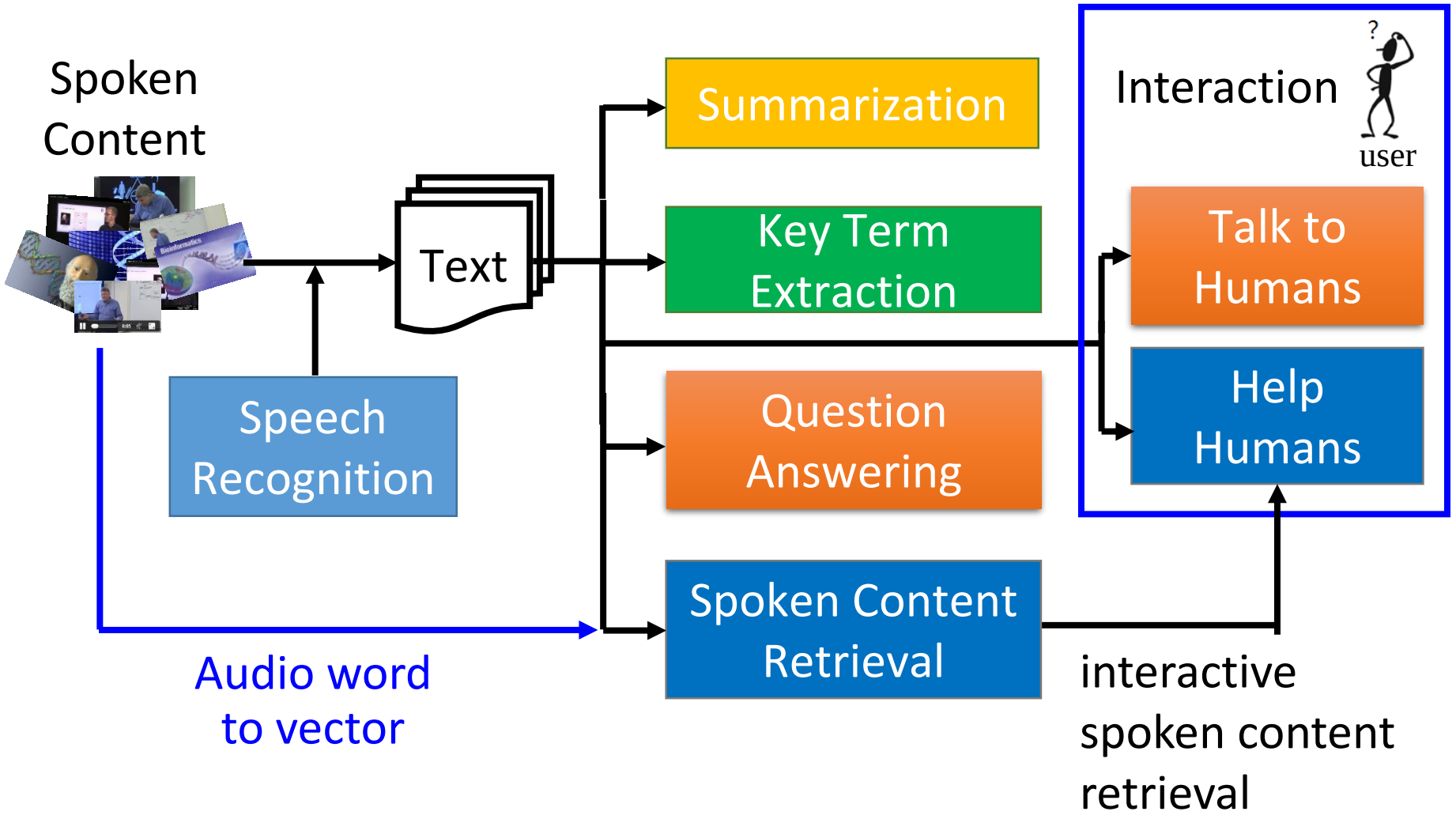
Machine Comprehension of Spoken Content

300 hrs multimedia is uploaded per minute. (2015.01)

2163 courses on Coursera (today)

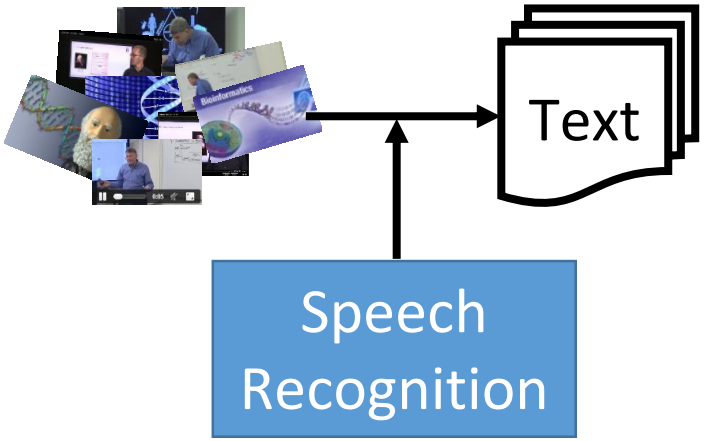
- Nobody is able to go through the data.
- In these multimedia, the spoken part carries very important information about the content.
- We need machine to listen to the audio data, understand it, and extract useful information for humans.

Overview

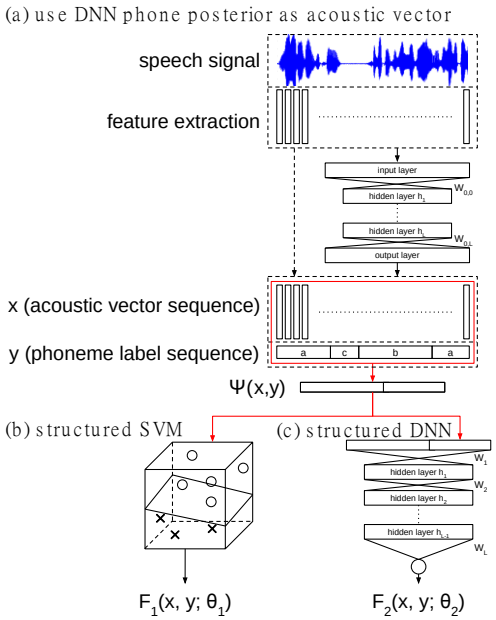

Spoken Content
Text
Speech Recognition
Summarization
Key Term Extraction
Question Answering
Spoken Content Retrieval
Audio word to vector
Interaction
user
Talk to Humans
Help Humans
interactive spoken content retrieval

Speech Recognition

Spoken Content

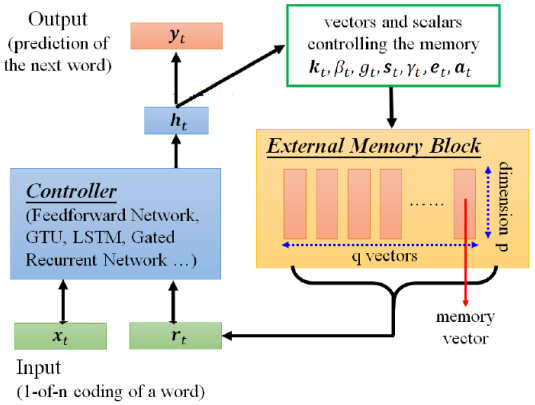


Acoustic Models



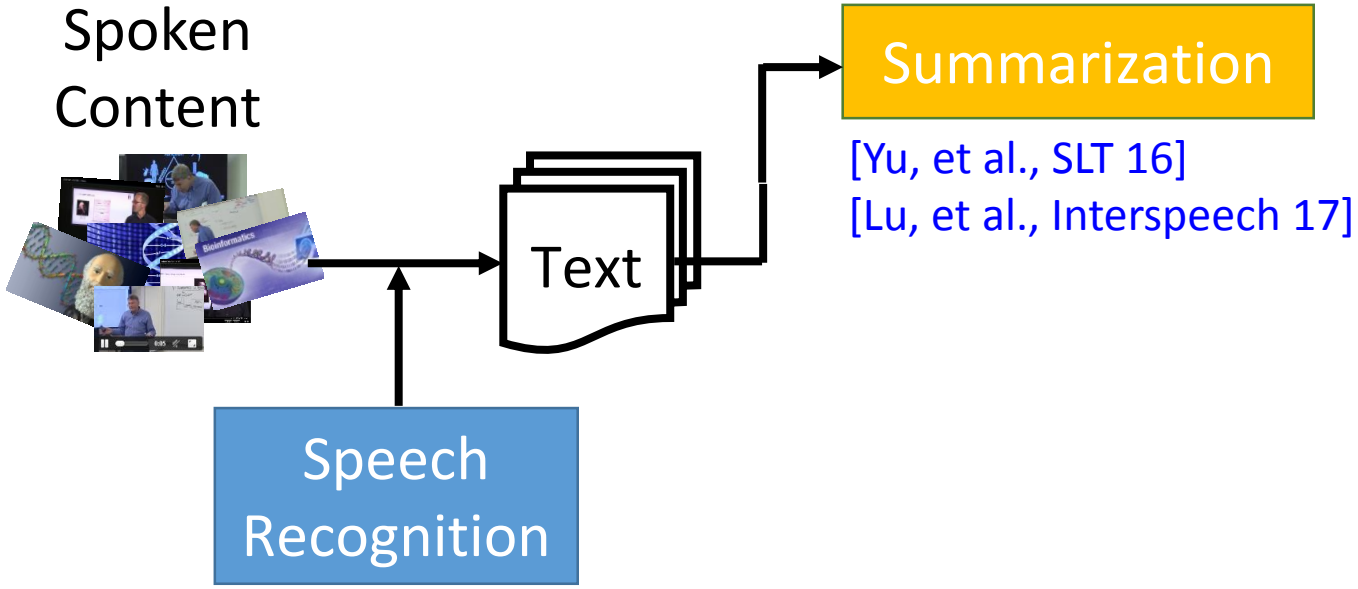
[Liao, et al., ASRU 15]

Language Models



[Ko, et al., ICASSP 17]

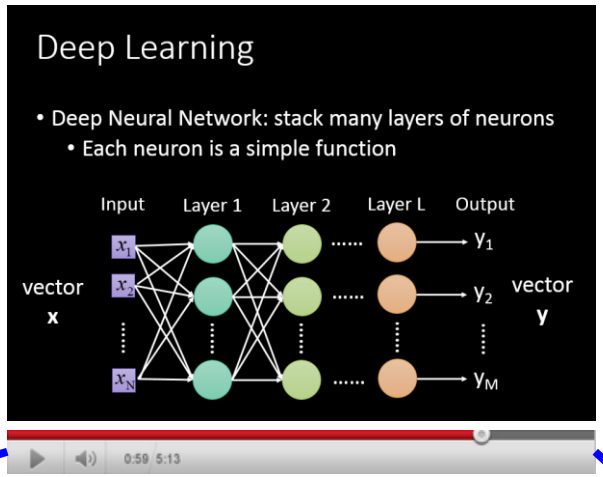
Summarization



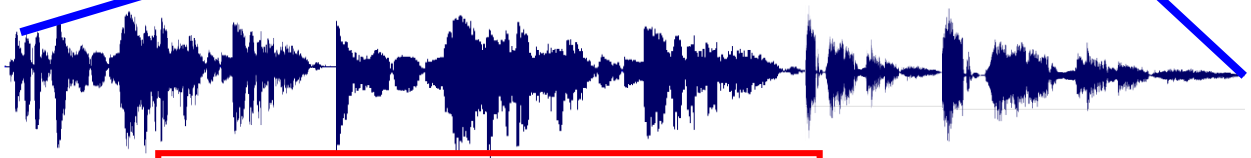
Summarization

Extractive Summaries

[Lee, et al., Interspeech 12]
[Lee, et al., ICASSP 13]
[Shiang, et al., Interspeech 13]



Audio File
to be summarized



…… deep learning is powerful …… …… ……

This is the summary.

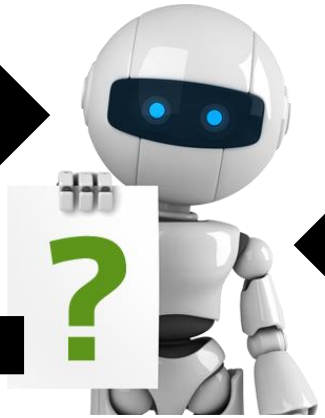
- Select the most informative segments to form a compact version
- Machine does not write summaries in its own words

Abstractive Summarization

- Now machine can do **abstractive summary** (write summaries in its own words)
  - *Title generation*: abstractive summary with one sentence

title generated by machine

(in its own words)



without hand-crafted rules

Training Data

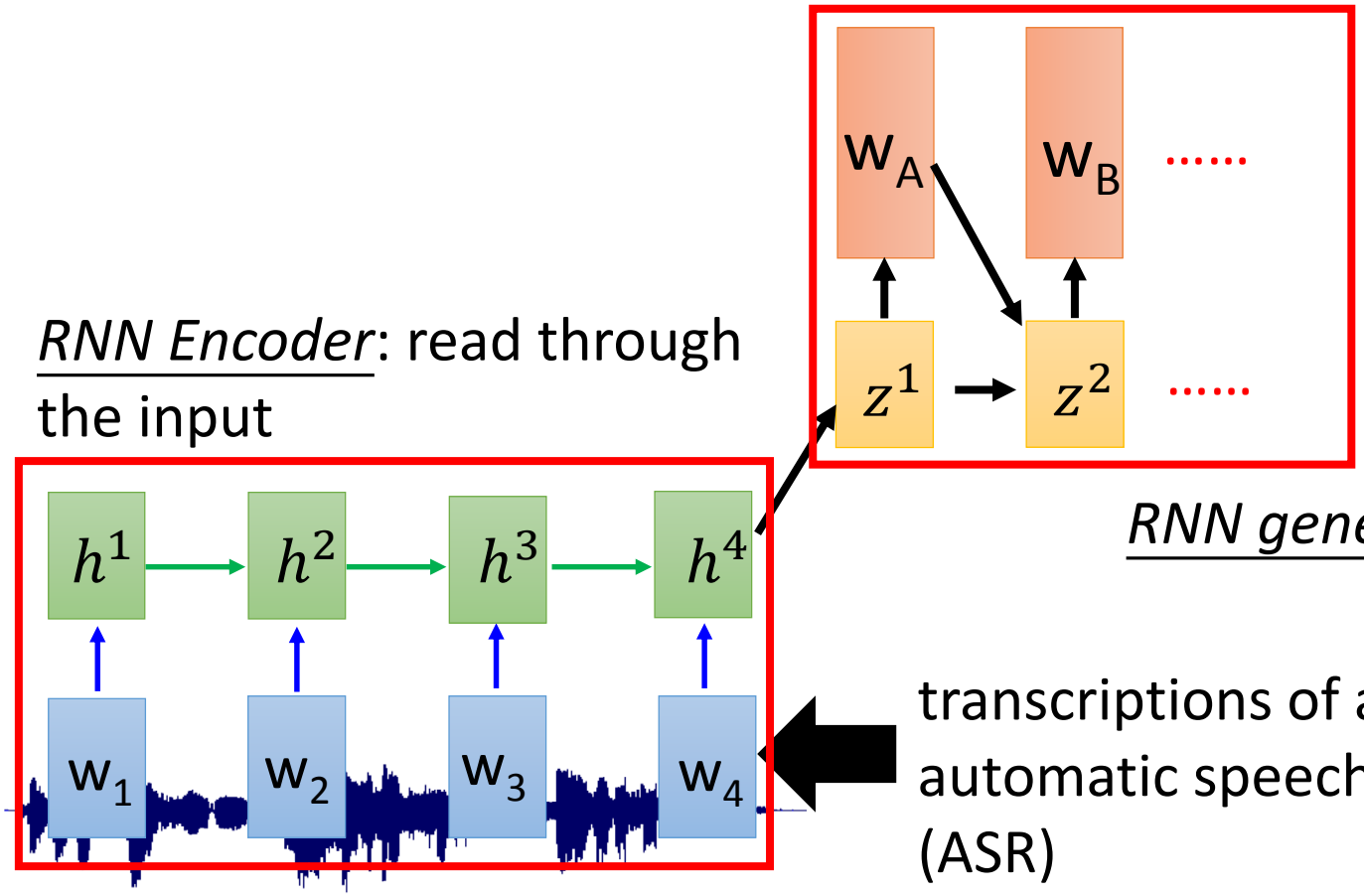
Title 1

Title 2

Title 3

Sequence-to-sequence

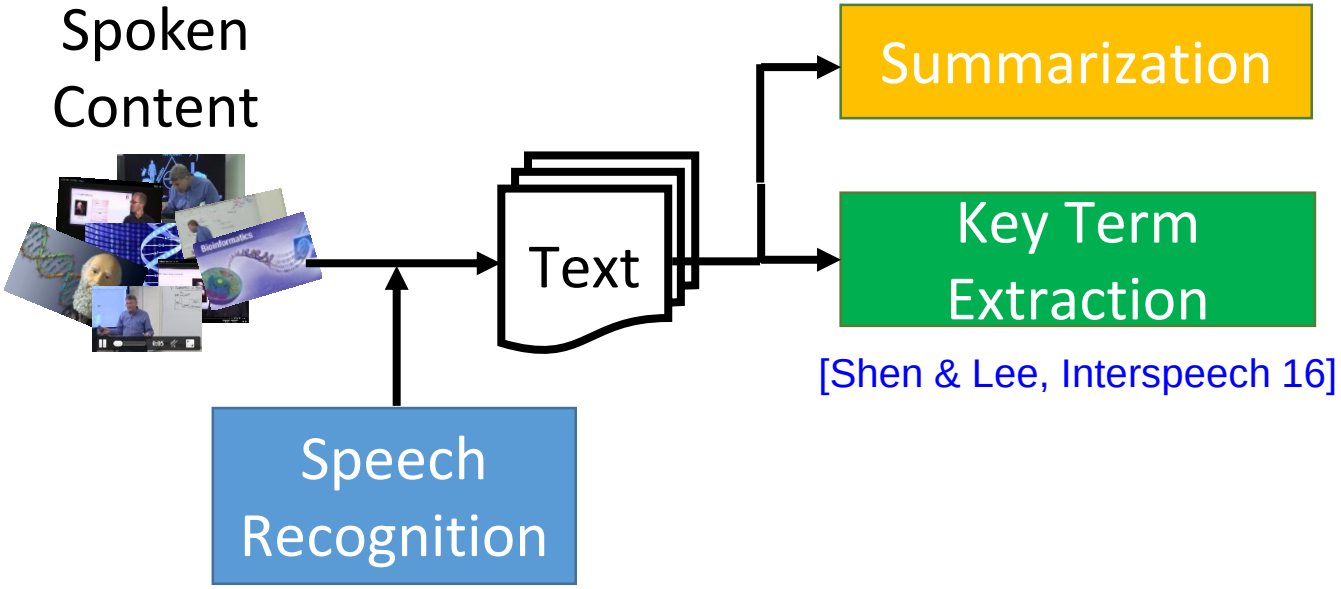
- Input: transcriptions of audio, output: title



Demo

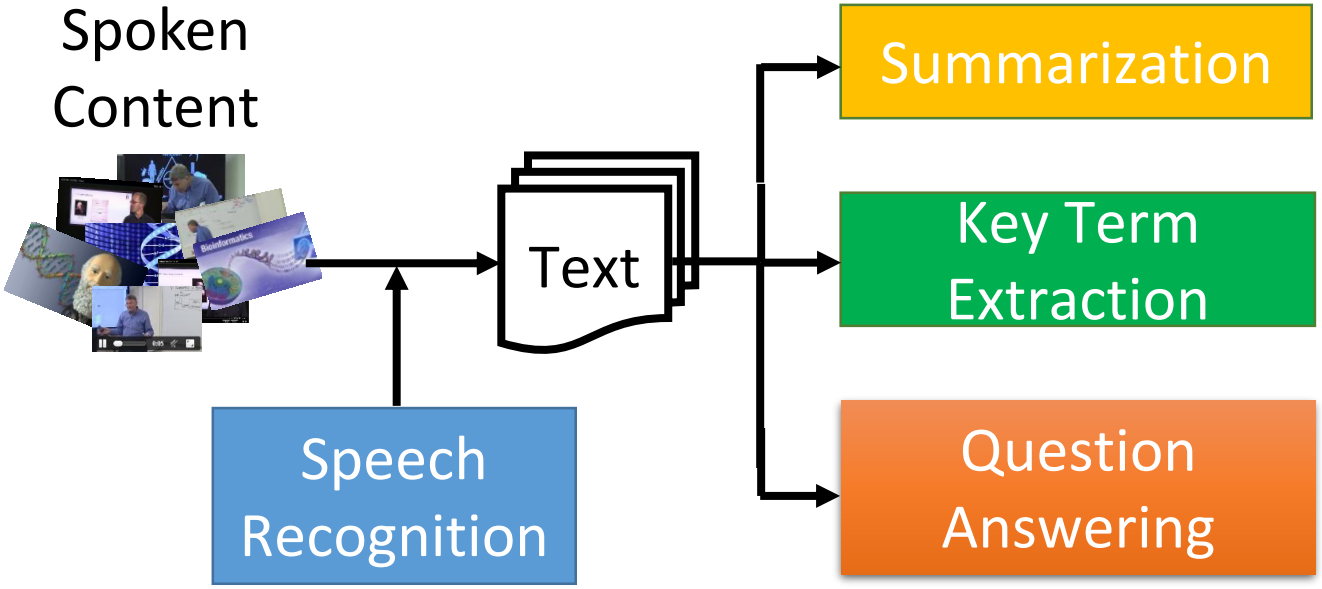
- 作者:葉政杰、周儒杰
- http://140.112.30.37:2401/
- https://www.youtube.com/watch?v=X3BapMl7Wv8
- From SONG TUYEN NEWS: https://www.youtube.com/channel/UC-P4mEcWZVrFfdZIuiODiTg

Key Term Extraction



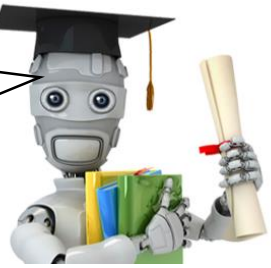
[Shen & Lee, Interspeech 16]

Speech Question Answering



Speech Question Answering

What is a possible origin of Venus' clouds?



Gases released as a result of volcanic activity

Speech Question Answering: Machine answers questions based on the information in spoken content

New task for Machine Comprehension of Spoken Content

- **TOEFL Listening Comprehension Test by Machine**

Audio Story:  (The original story is 5 min long.)

Question: “ What is a possible origin of Venus’ clouds? ”

Choices:

(A) gases released as a result of volcanic activity

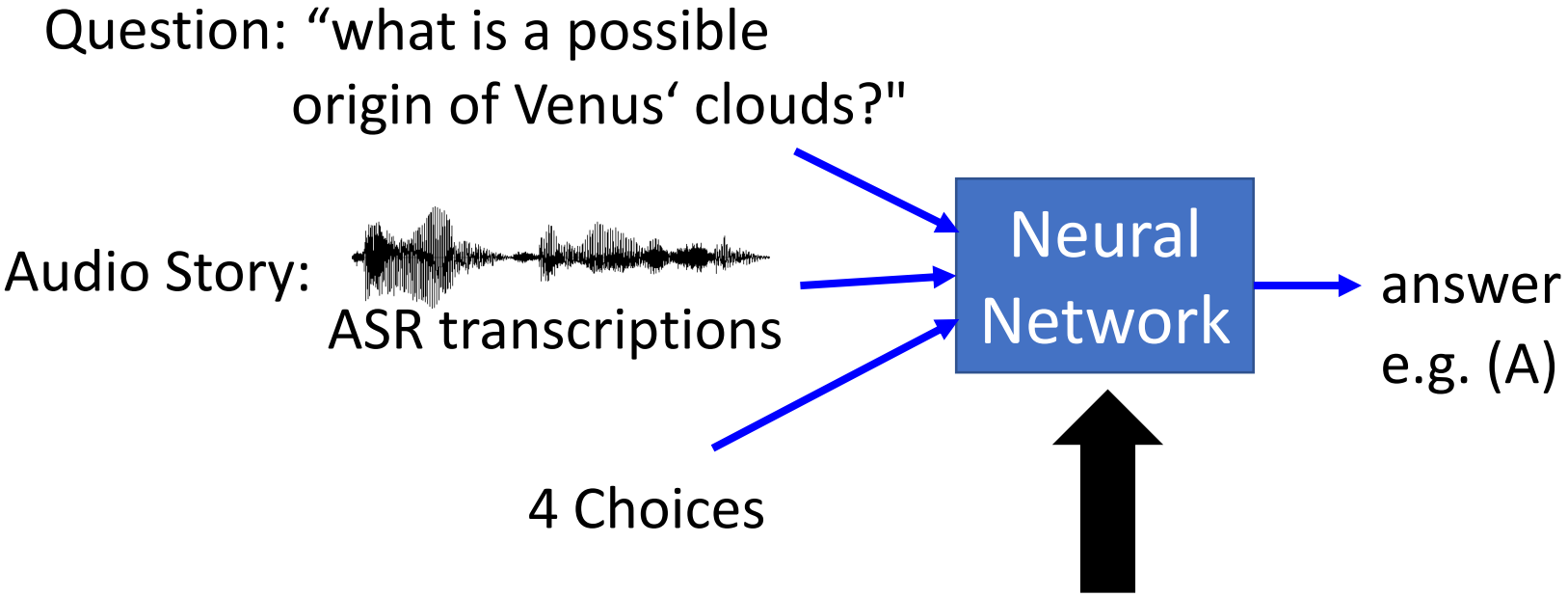
(B) chemical reactions caused by high surface temperatures

(C) bursts of radio energy from the plane's surface

(D) strong winds that blow dust into the atmosphere

New task for Machine Comprehension of Spoken Content

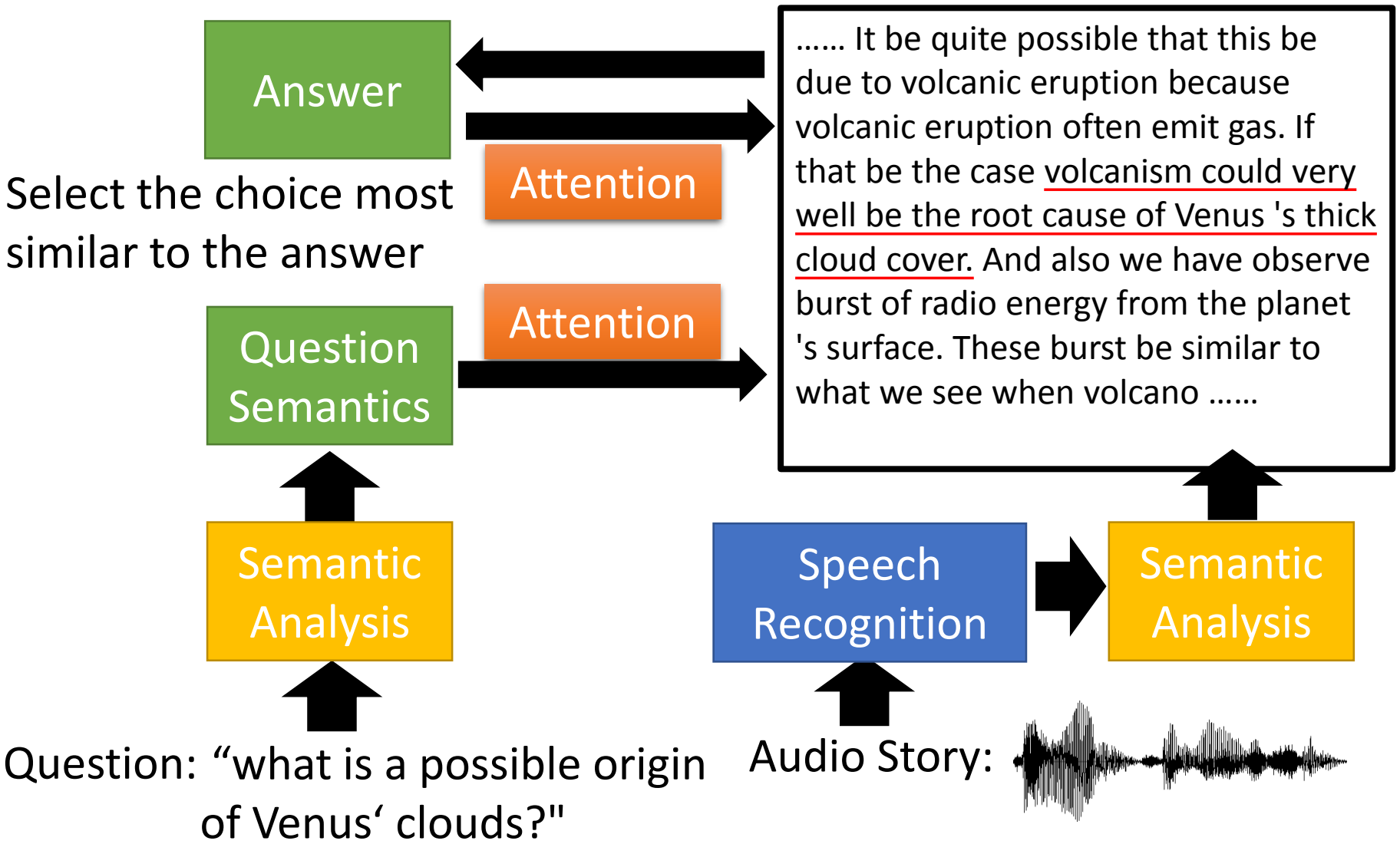
- **TOEFL Listening Comprehension Test by Machine**



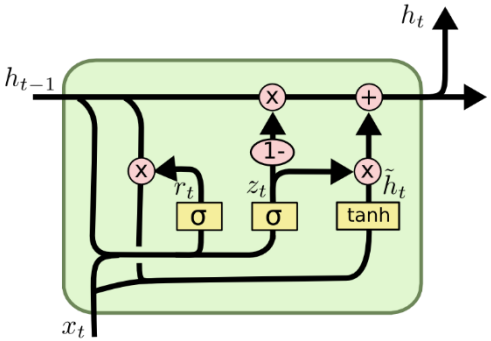
Using **previous exams** to train the network

Model Architecture

The whole model learned end-to-end.



More Details

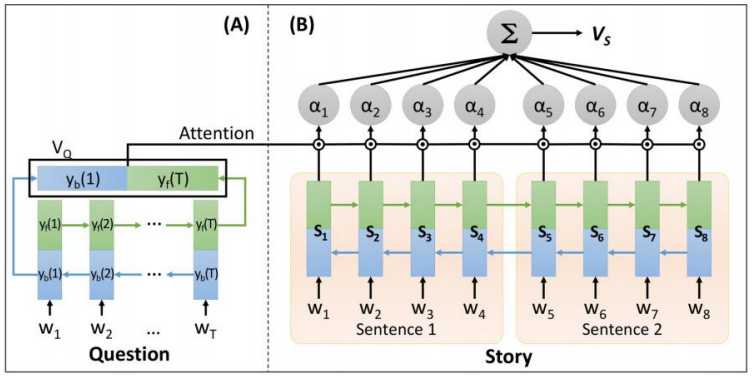


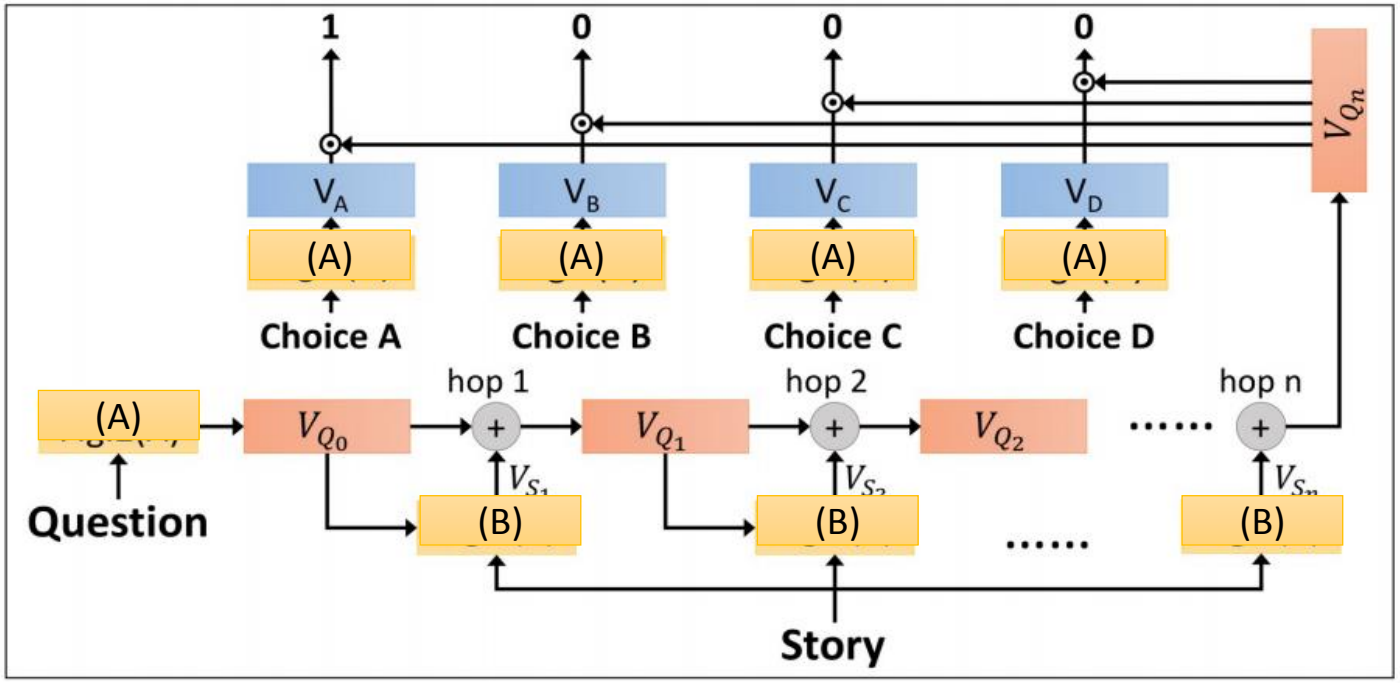
$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

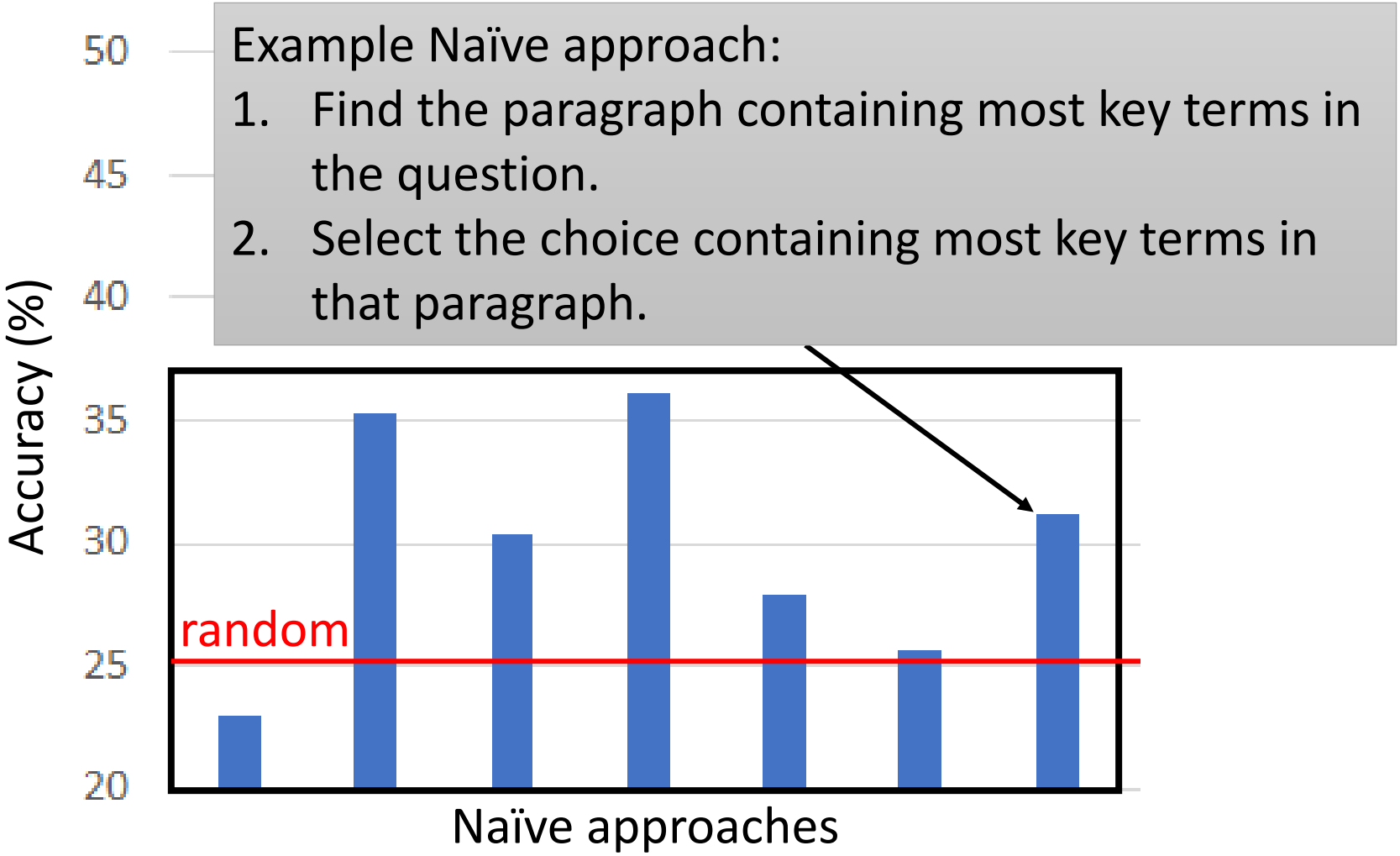
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

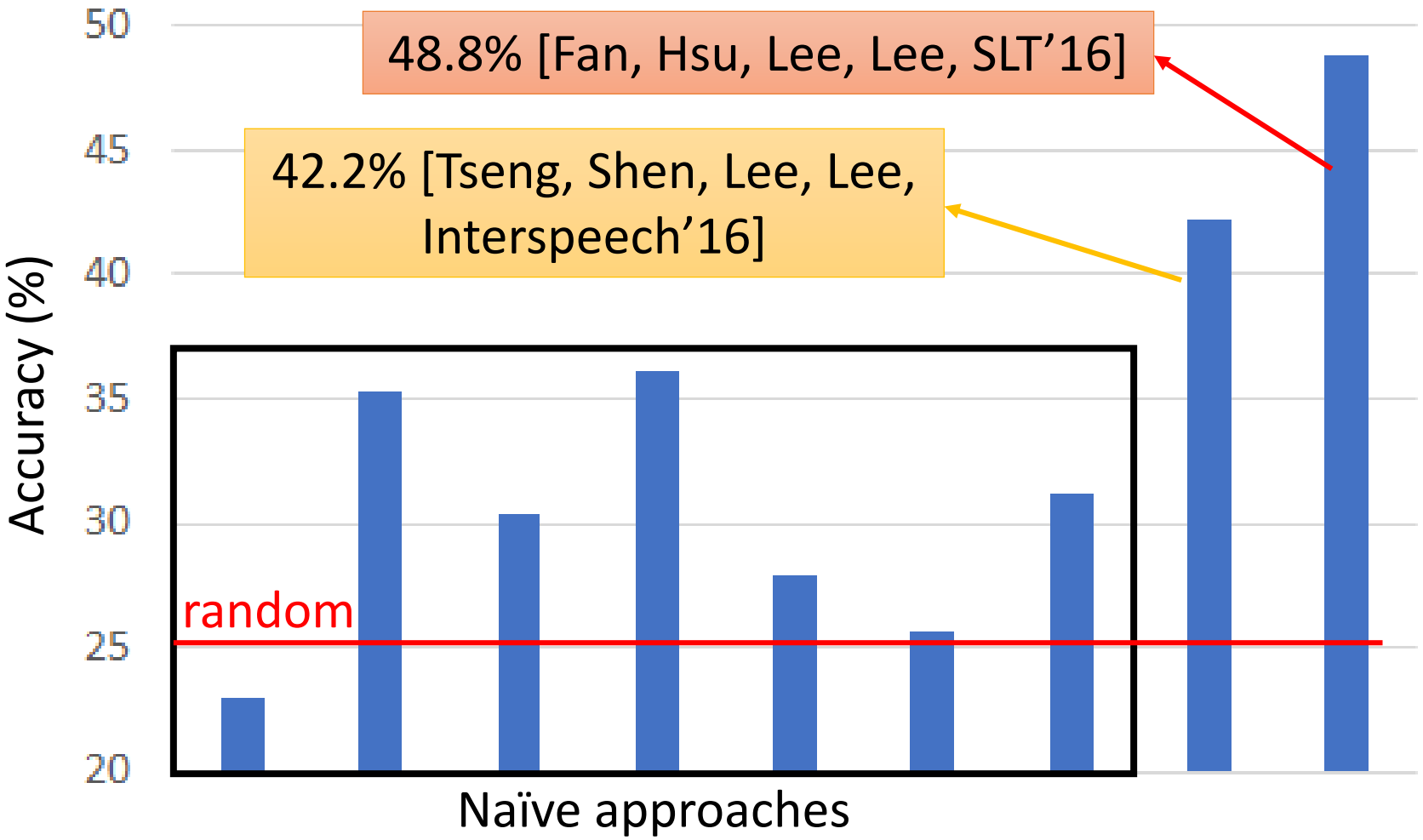




Experimental Results



Experimental Results

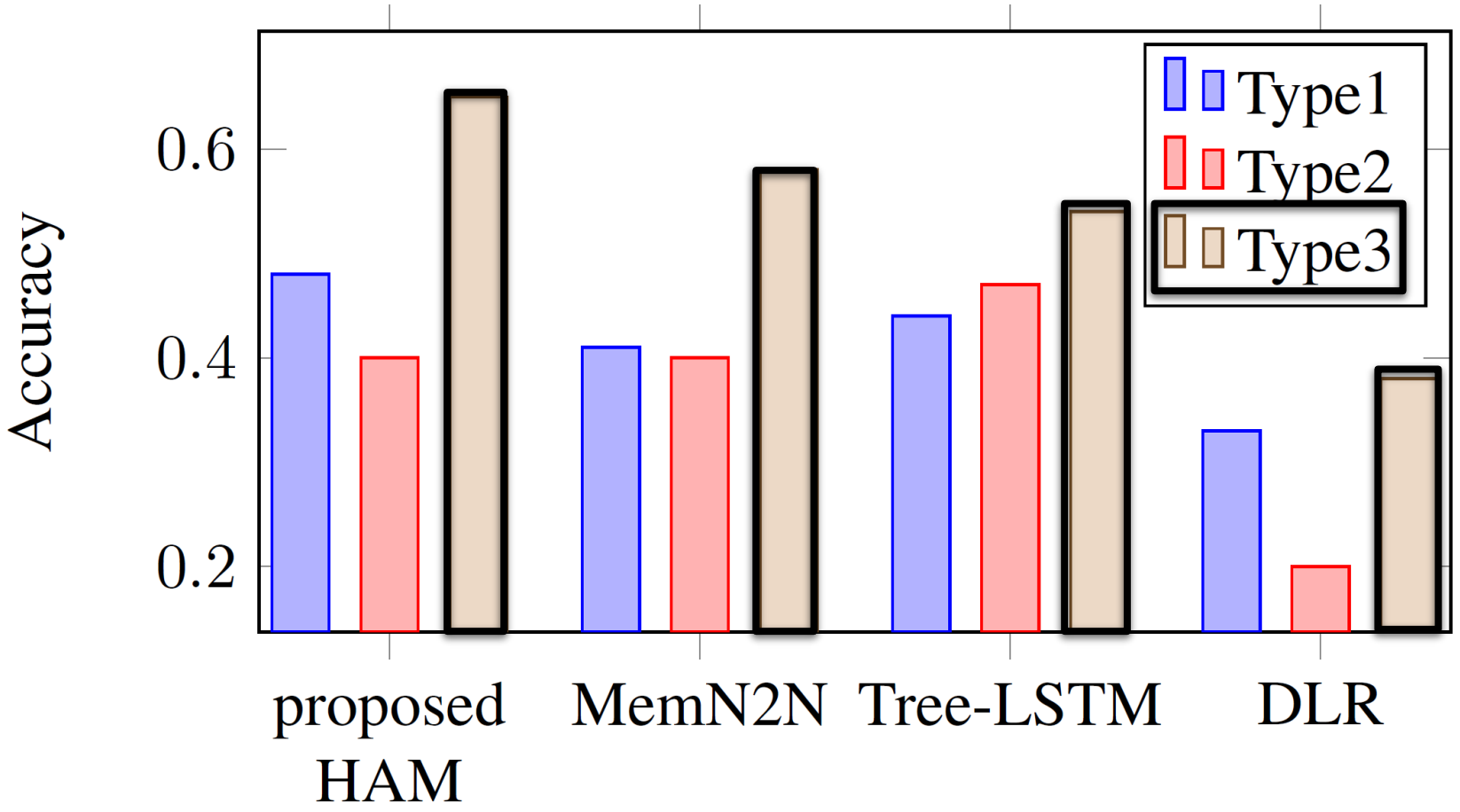


Analysis

Type 3: **Connecting Information**

- Understanding Organization
- Connecting Content
- Making Inferences

- There are three types of questions

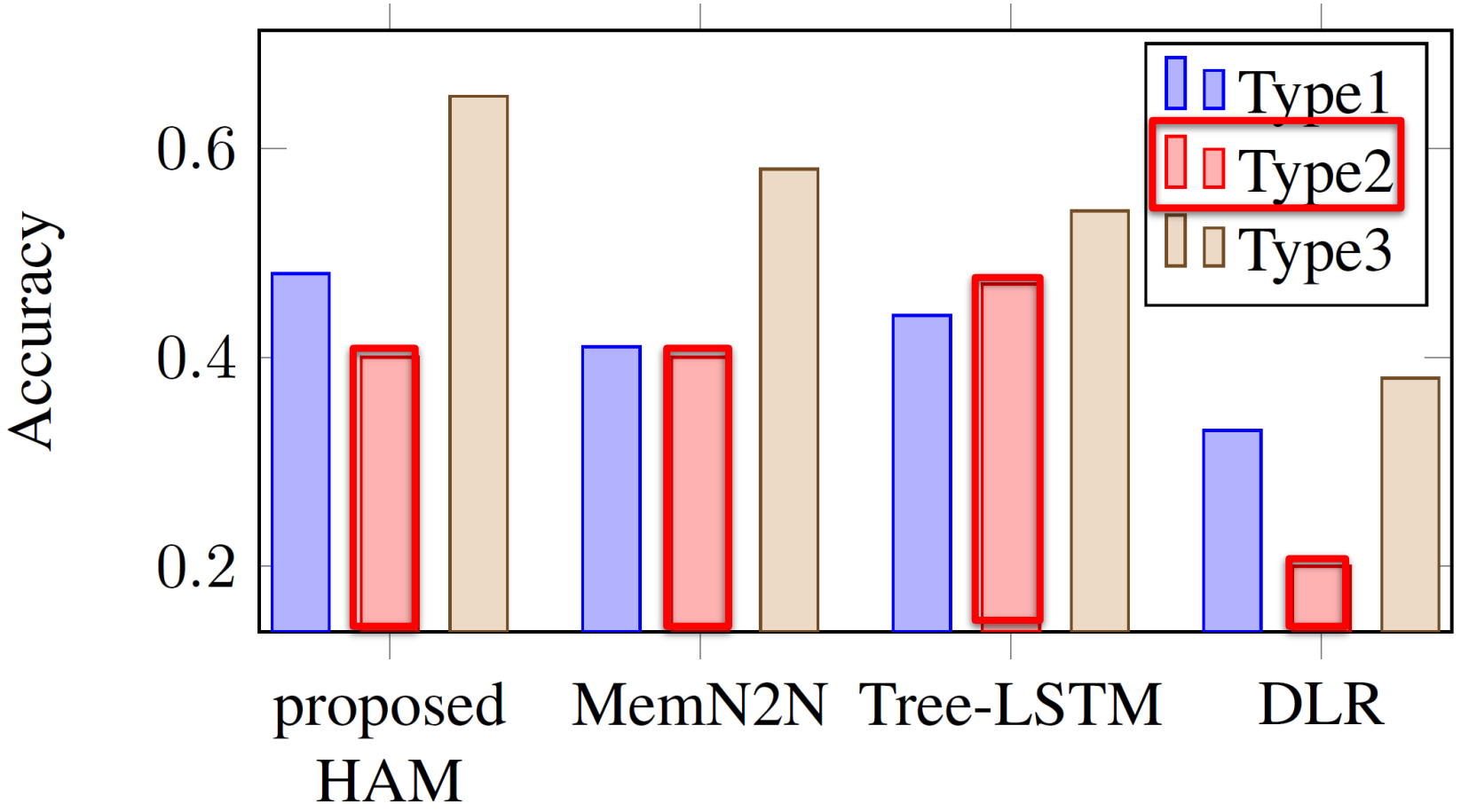


Analysis

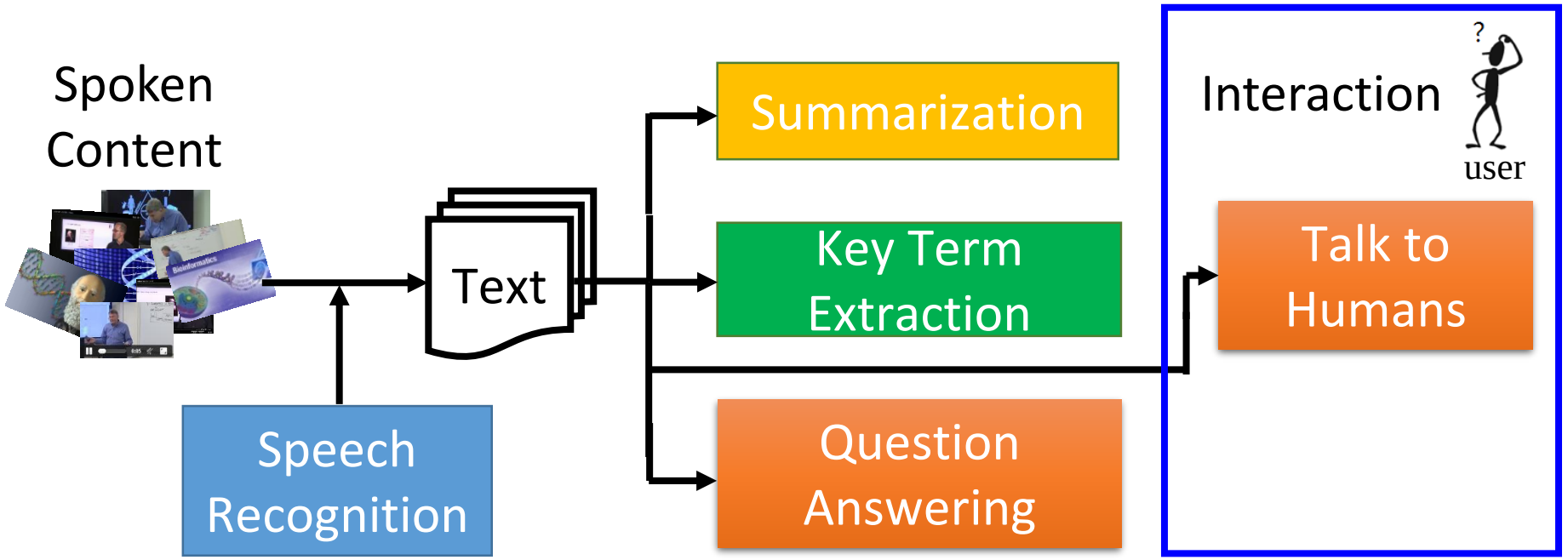
Type 3: Pragmatic Understanding

- Understanding the *Function of What Is Said*
- Understanding the *Speaker's Attitude*

- There are three types of questions



Talk to Humans

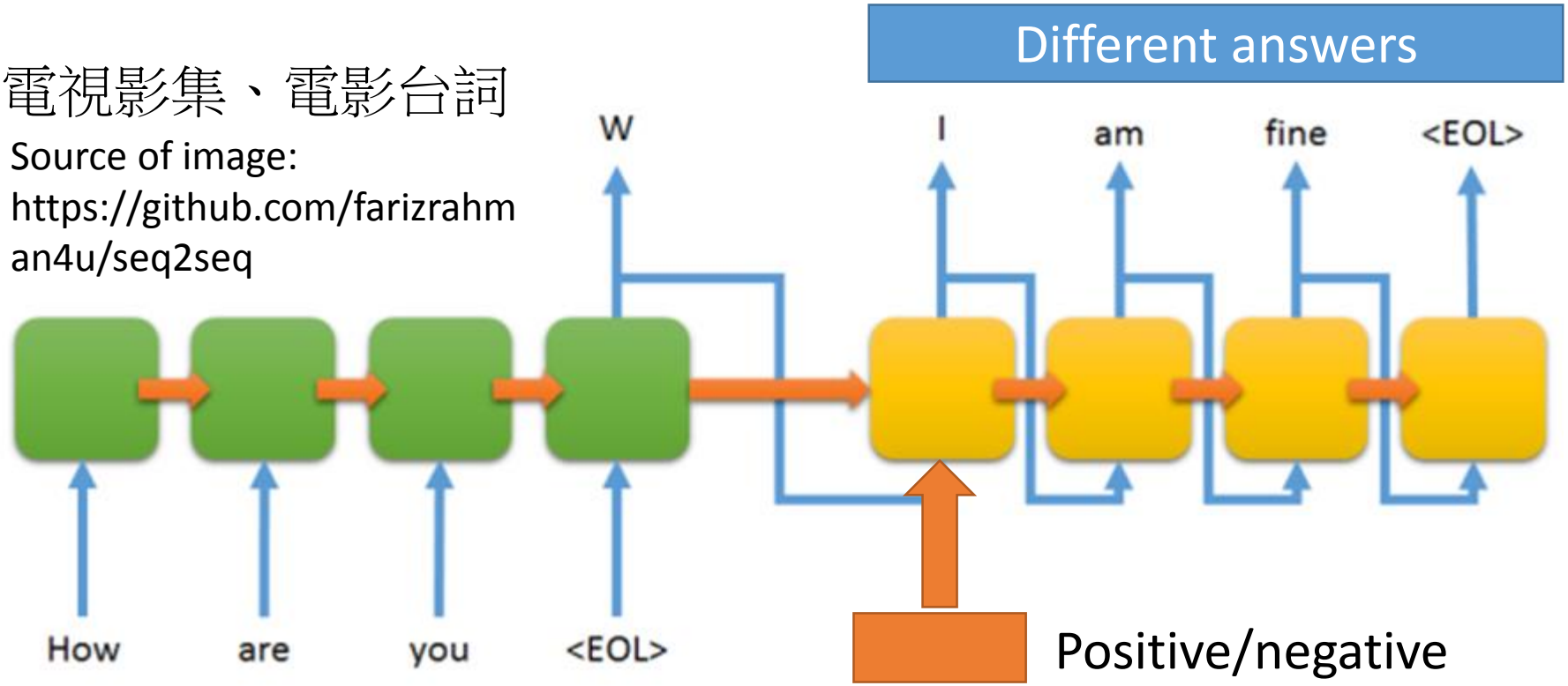


Chat-bot

Sequence-to-sequence learning from human conversation without hand-crafted rules.

電視影集、電影台詞

Source of image:
https://github.com/farizrahman4u/seq2seq



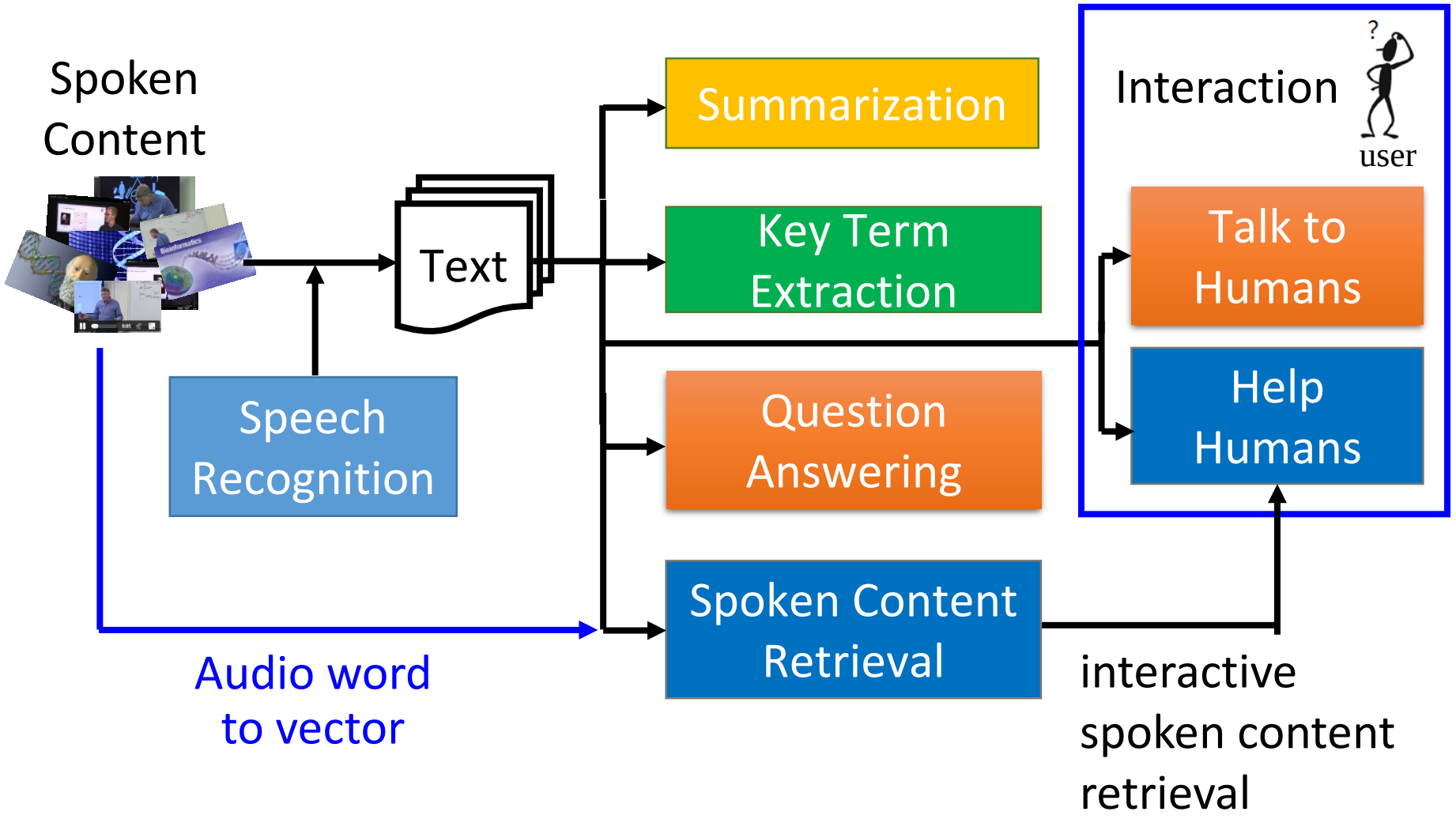
On-going project:

- Training by reinforcement learning
- Training by generative adversarial network (GAN)

Demo - Towards Characterization

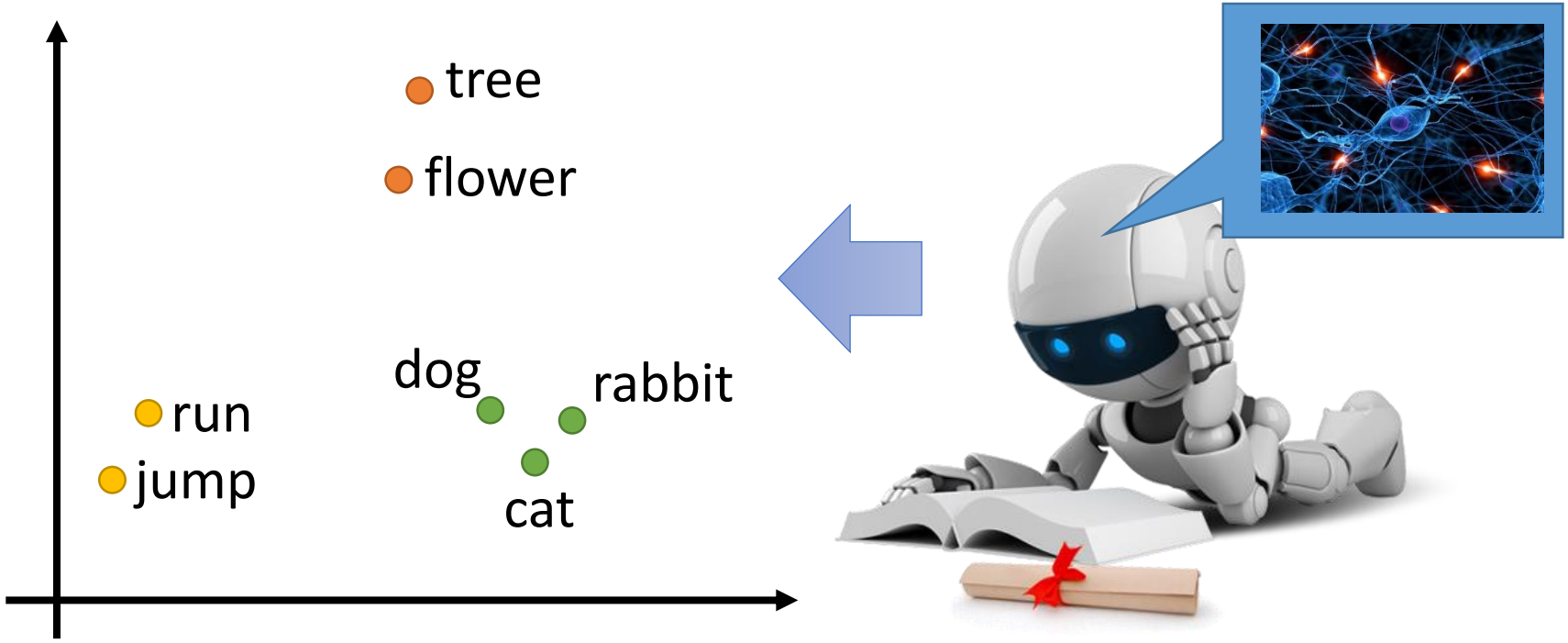
- 作者：王耀賢
- https://github.com/yaushian/simple_sentiment_dialogue
- https://github.com/yaushian/personal-dialogue

Audio Word to Vector



Typical Word to Vector

- Machine represents each word by a vector representing its meaning
- Learning from lots of text without supervision



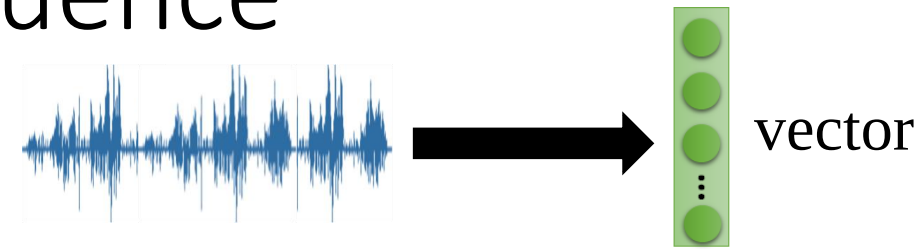
Audio Word to Vector

- Machine represents each audio segment also by a vector

Learn from lots of audio without supervision

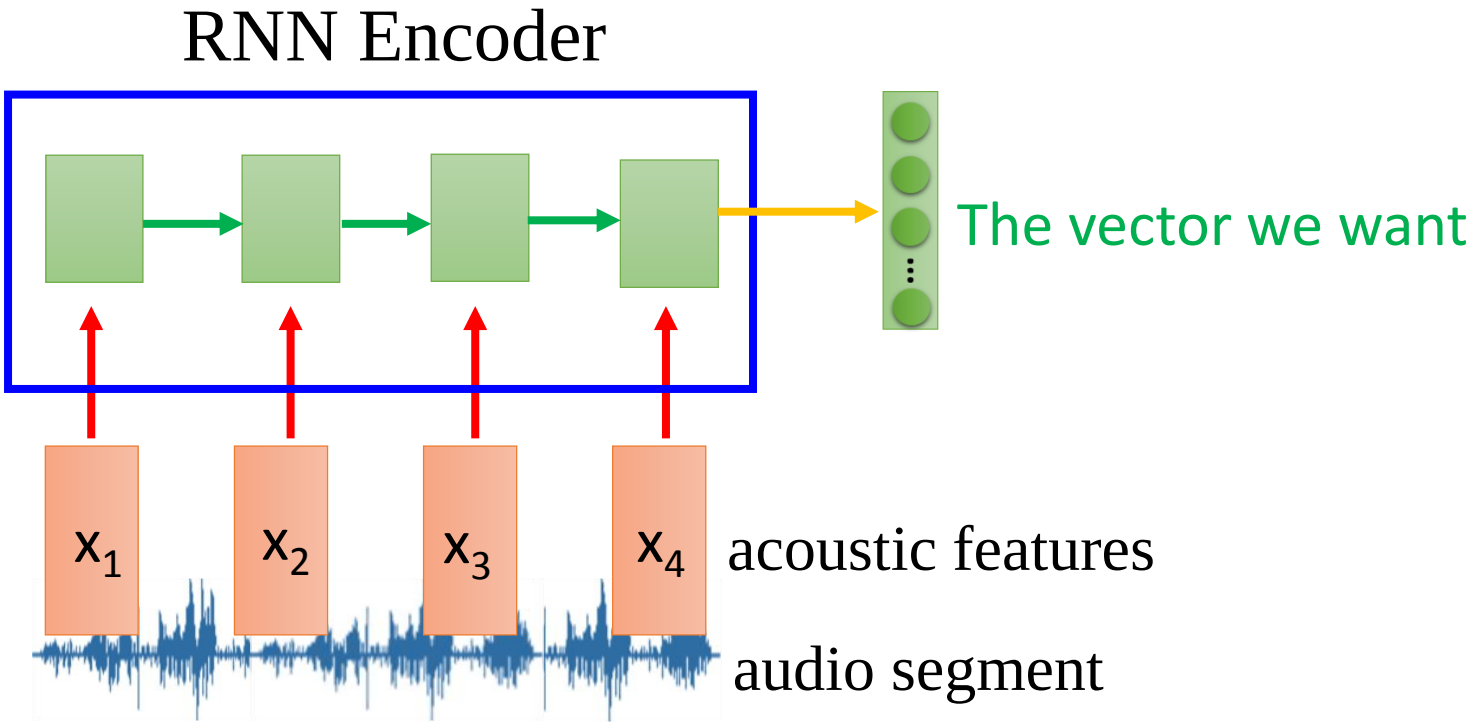
[Chung, et al., Interspeech 16)

Sequence-to-sequence Auto-encoder



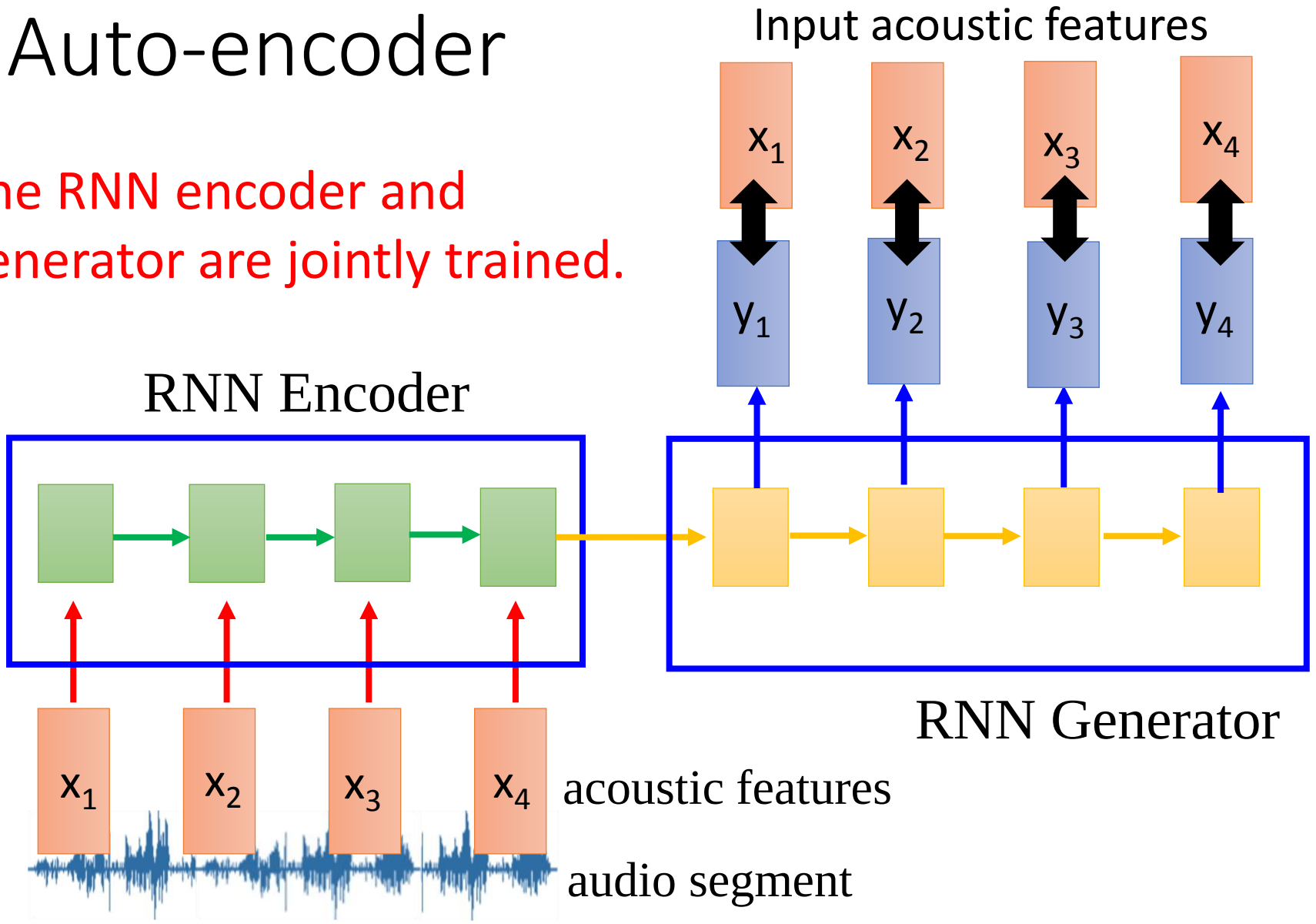
We use ***sequence-to-sequence auto-encoder*** here

The training is unsupervised.



Sequence-to-sequence Auto-encoder

The RNN encoder and generator are jointly trained.



What does machine learn?

- Typical word to vector:

$$V(Rome) - V(Italy) + V(Germany) \approx V(Berlin)$$
$$V(king) - V(queen) + V(aunt) \approx V(uncle)$$

- Audio word to vector (phonetic information)

V() - V() + V() = V()

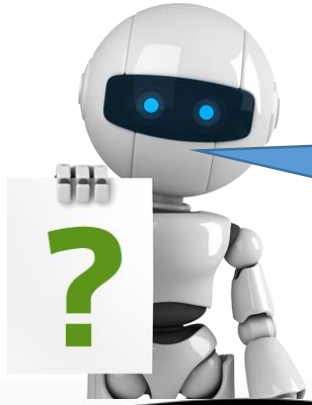
GIRL PEARL PEARLS GIRLS

V() - V() + V() = V()

Demo

Application:
Video Caption Generation

Visual + Audio



A girl is running.

A group of people is knocked by a tree.

A group of people is walking in the forest.

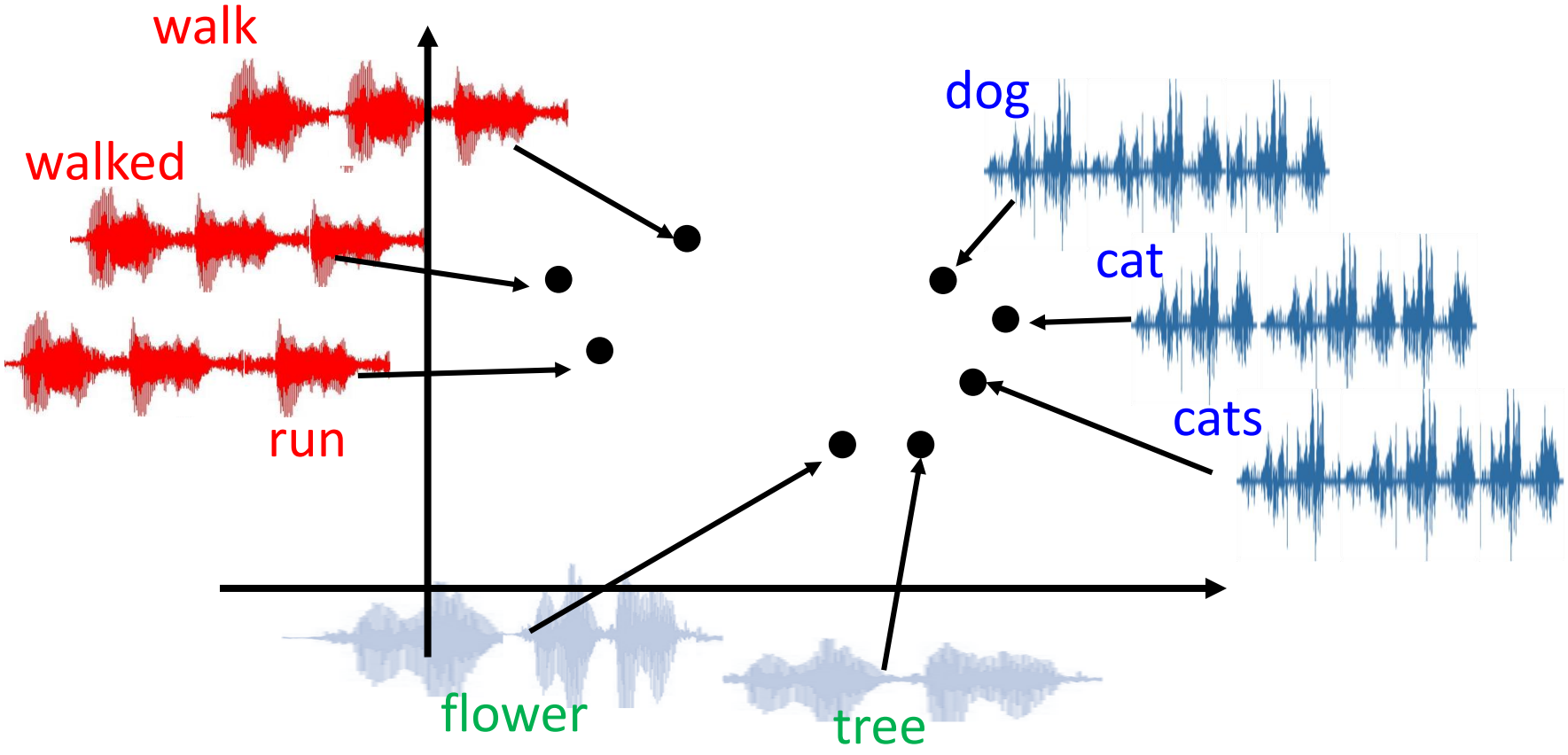
Demo

- Can machine describe what it see from video?
- 作者: 莊舜博、楊棋宇、黃邦齊、萬家宏

Next Step ……

One day we can build all spoken language understanding applications directly from ***audio word to vector***.

- Audio word to vector with semantics



Concluding Remarks