Towards Machine Comprehension of Spoken Content

李宏毅

Hung-yi Lee

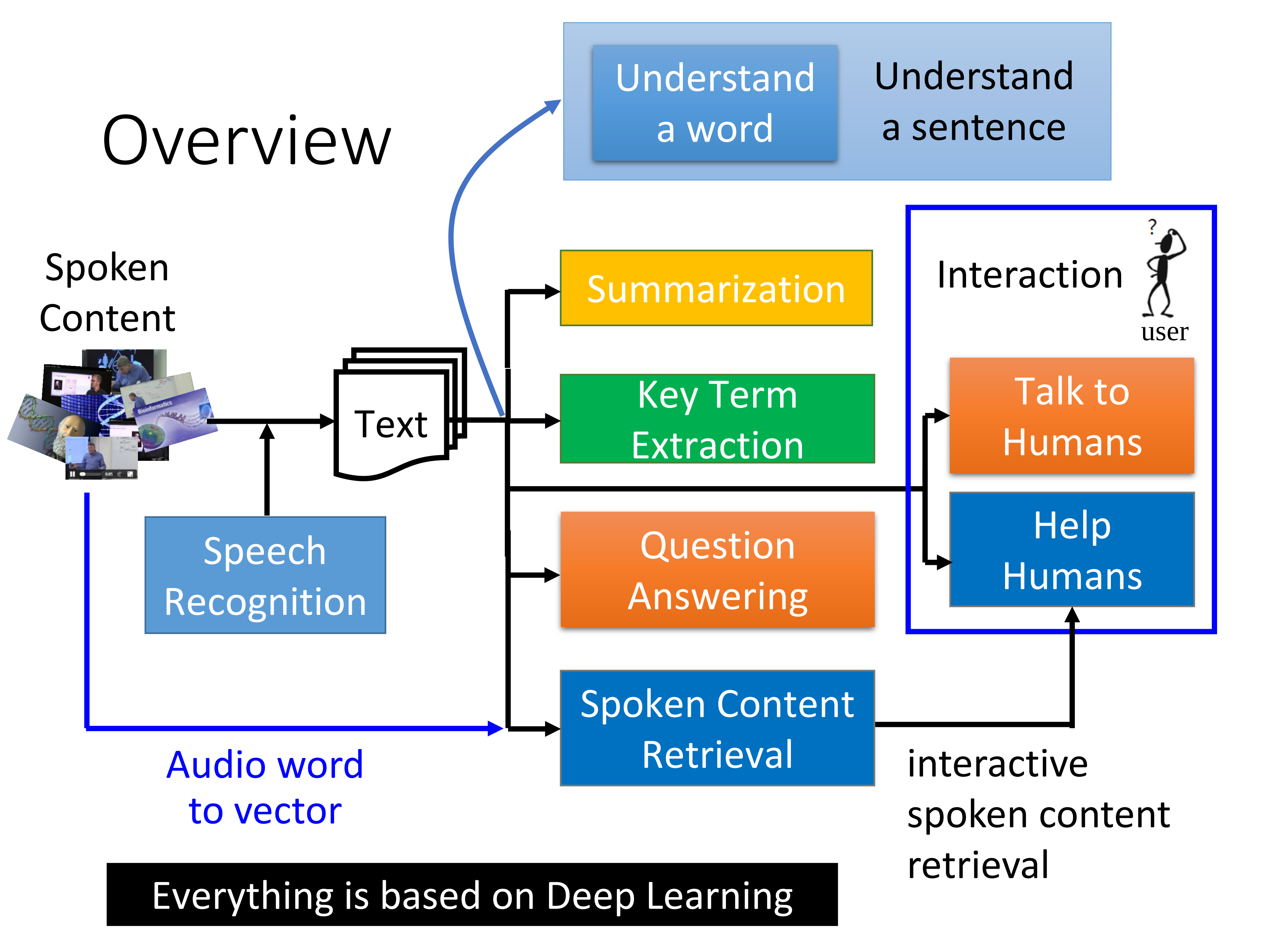
National Taiwan University

Machine Comprehension of Spoken Content

300 hrs multimedia is uploaded per minute. (2015.01)

More than 2000 courses on Coursera

- 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.

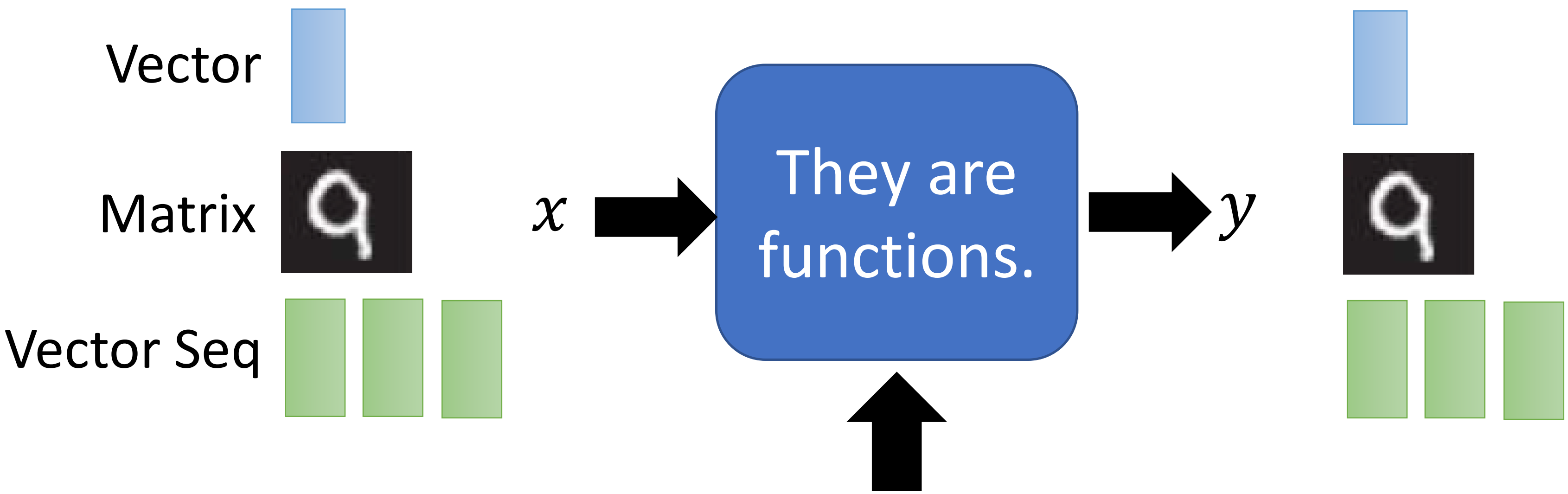


Overview
Understand a word
Understand a sentence
Spoken Content
Text
Speech Recognition
Summarization
Key Term Extraction
Question Answering
Spoken Content Retrieval
Interaction
user
Talk to Humans
Help Humans
Audio word to vector
interactive spoken content retrieval
Everything is based on Deep Learning

Deep Learning in One Slide

Many kinds of networks:

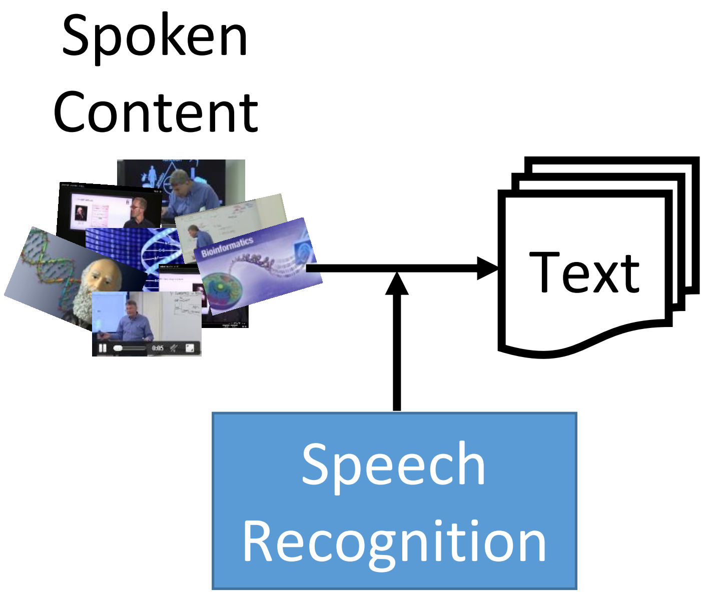
- Fully connected feedforward network
- Convolutional neural network (CNN)
- Recurrent neural network (RNN)

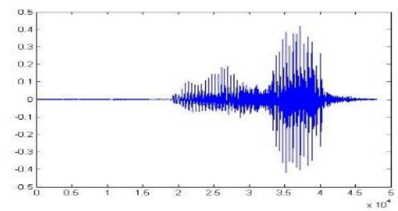


How to find the function?

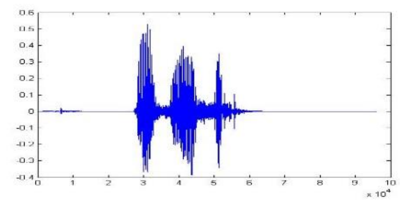
Given the examples of inputs/outputs as (**training data**): $\{(x_1,y_1),(x_2,y_2), \ldots\ldots, (x_{1000},y_{1000})\}$

Speech Recognition

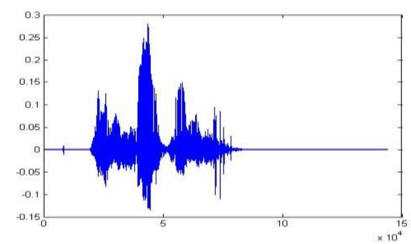




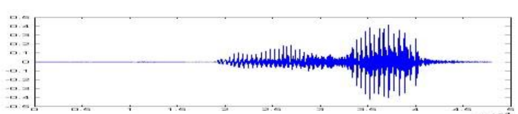
"Hi"



"I am fine"



"Good bye"

$f($  $) =$ "How are you"

Typical Deep Learning Approach

- The hierarchical structure of human languages

what do you think

Phoneme:

hh w aa t d uw y uw th ih ng k

Tri-phone:

…… t-d+uw d-uw+y uw-y+uw y-uw+th ……

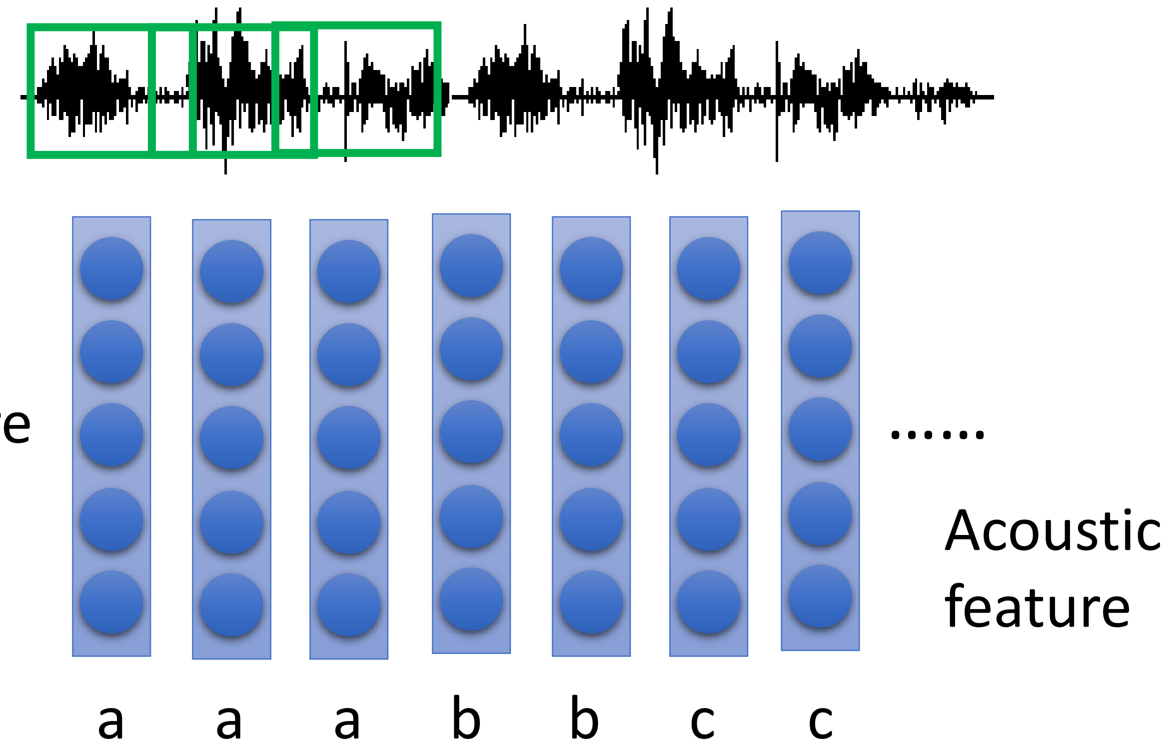
t-d+uw1 t-d+uw2 t-d+uw3 d-uw+y1 d-uw+y2 d-uw+y3

State:

Typical Deep Learning Approach

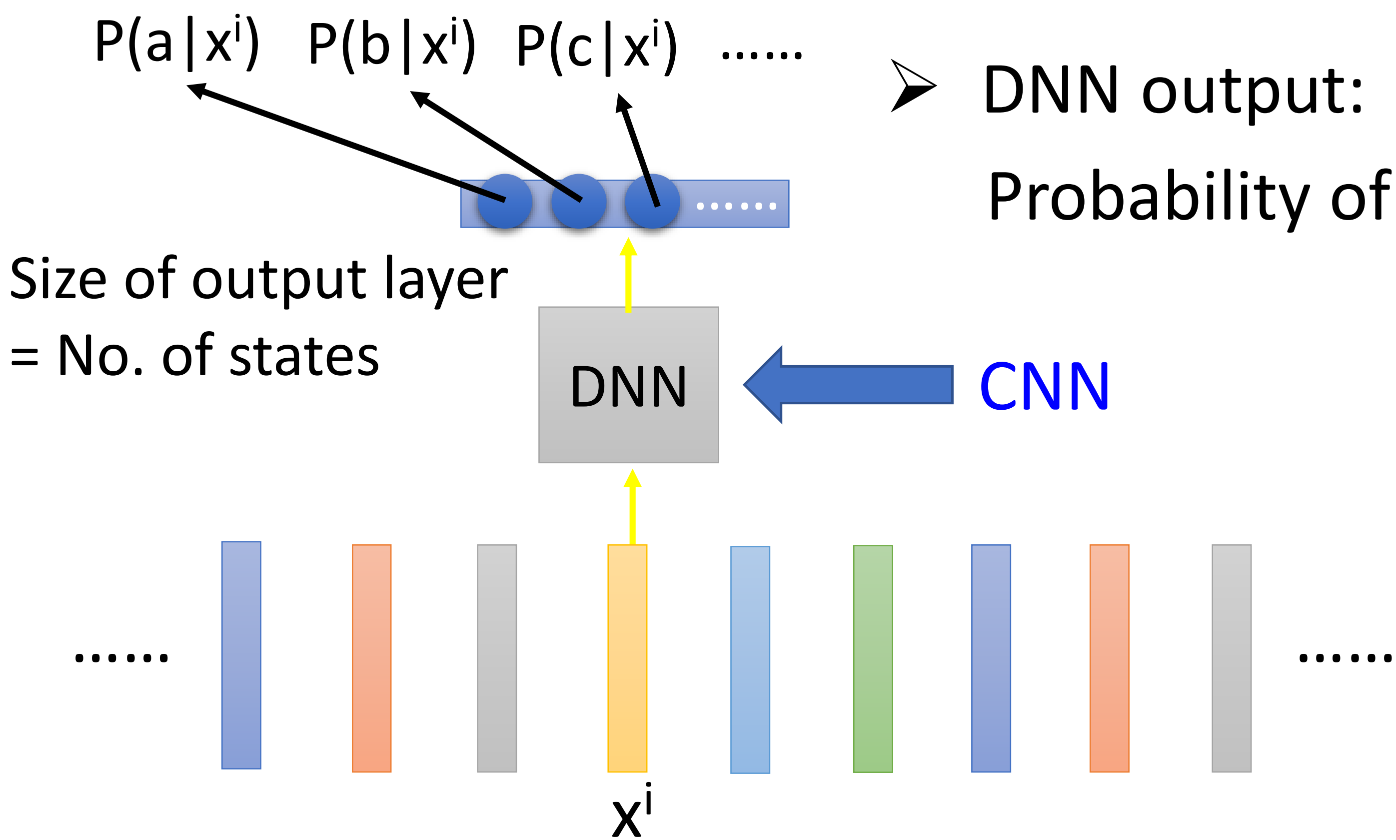
- The first stage of speech recognition
  - Classification: input → acoustic feature, output → state

Determine the state each acoustic feature belongs to



Typical Deep Learning Approach

- DNN input:
  One acoustic feature
- DNN output:
  Probability of each state



Very Deep

MSR

VGG Net (85M Parameters)	Residual-Net (38M Parameters)	LACE (65M Parameters)
14 weight layers	49 weight layers	22 weight layers
40x41 input	40x41 input	40x61 input
3 – conv 3x3, 96	3 – [conv 1x1, 64 conv 3x3, 64 conv 1x1, 256]	5 – conv 3x3, 128
Max pool	4 – [conv 1x1, 128 conv 3x3, 128 conv 1x1, 512]	5 – conv 3x3, 256
4 – conv 3x3, 192	6 – [conv 1x1, 256 conv 3x3, 256 conv 1x1, 1024]	5 – conv 3x3, 512
Max pool	3 – [conv 1x1, 512 conv 3x3, 512 conv 1x1, 2048]	5 – conv 3x3, 1024
4 – conv 3x3, 384	Average pool	1 – conv 3x4, 1
Max pool	Softmax (9000)	Softmax (9000)
2 – FC – 4096		
Softmax (9000)		

Human Parity!

- 微軟語音辨識技術突破重大里程碑：對話辨識能力達人類水準！(2016.10)
  - https://www.bnext.com.tw/article/41414/bn-2016-10-19-020437-216
  - Dong Yu, Wayne Xiong, Jasha Droppo, Andreas Stolcke , Guoli Ye, Jinyu Li , Geoffrey Zweig, “Deep Convolutional Neural Networks with Layer-wise Context Expansion and Attention”, Interspeech 2016

Machine 5.9% v.s. Human 5.9%

- IBM vs Microsoft: 'Human parity' speech recognition record changes hands again (2017.03)
  - http://www.zdnet.com/article/ibm-vs-microsoft-human-parity-speech-recognition-record-changes-hands-again/
  - George Saon, Gakuto Kurata, Tom Sercu, Kartik Audhkhasi, Samuel Thomas, Dimitrios Dimitriadis, Xiaodong Cui, Bhuvana Ramabhadran, Michael Picheny, Lynn-Li Lim, Bergul Roomi, Phil Hall, “English Conversational Telephone Speech Recognition by Humans and Machines”, arXiv preprint, 2017

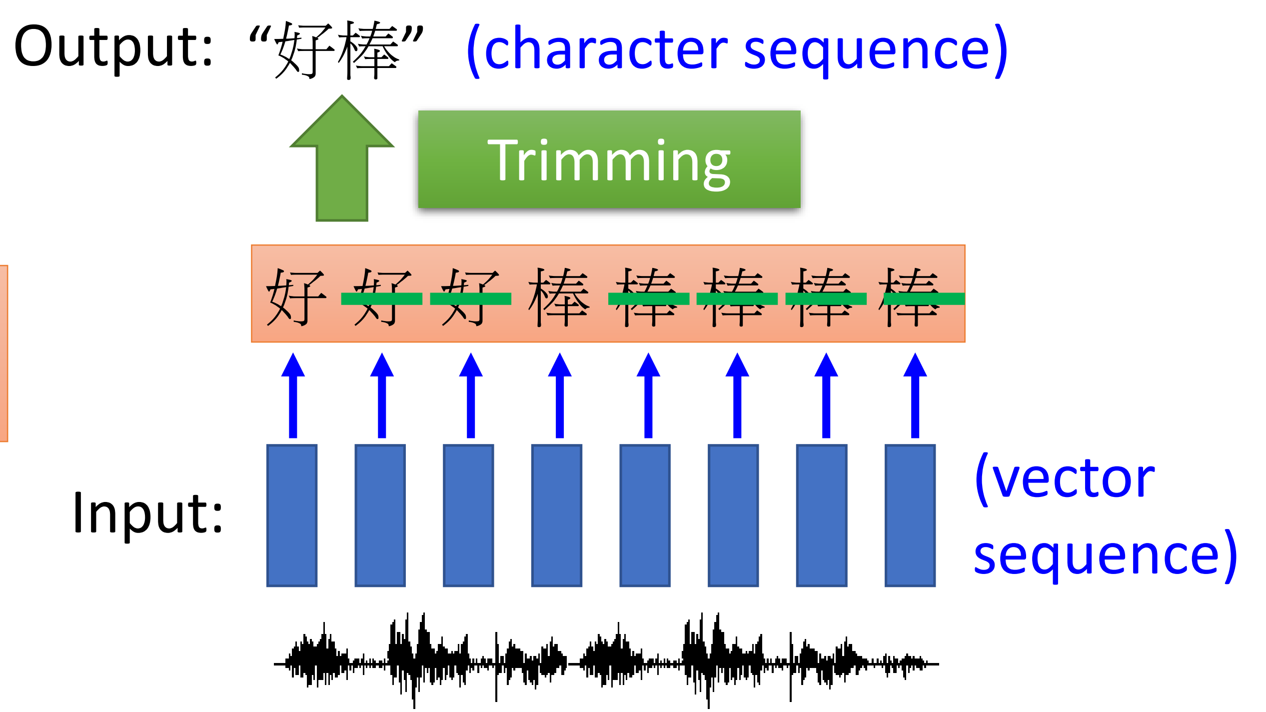
Machine 5.5% v.s. Human 5.1%

End-to-end Approach - Connectionist Temporal Classification (CTC)

- Connectionist Temporal Classification (CTC) [Alex Graves, ICML'06][Alex Graves, ICML'14][Haşim Sak, Interspeech'15][Jie Li, Interspeech'15][Andrew Senior, ASRU'15]

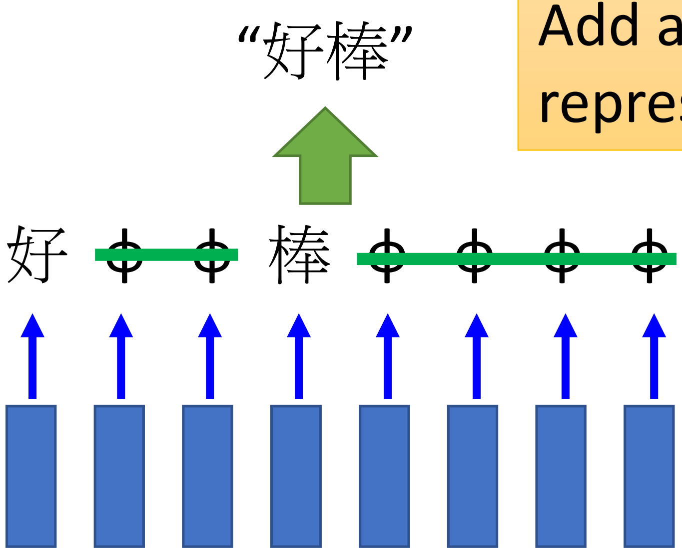
Problem?

Why can't it be "好棒棒"

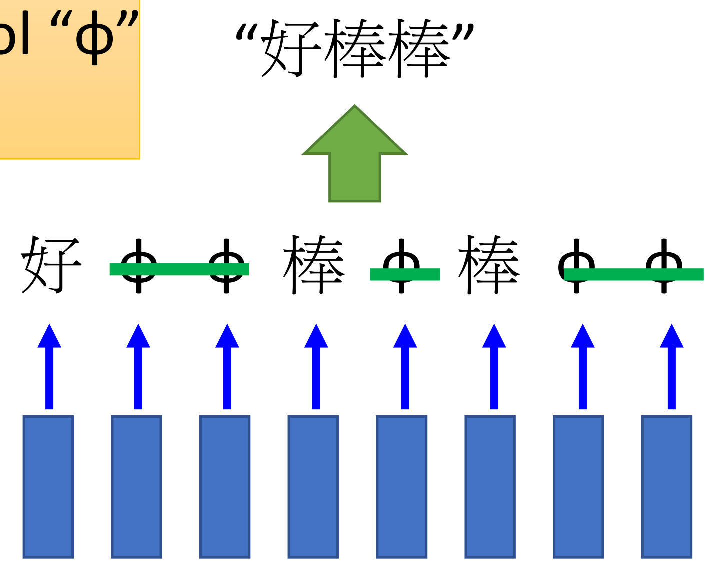


End-to-end Approach - Connectionist Temporal Classification (CTC)

- Connectionist Temporal Classification (CTC) [Alex Graves, ICML'06][Alex Graves, ICML'14][Haşim Sak, Interspeech'15][Jie Li, Interspeech'15][Andrew Senior, ASRU'15]

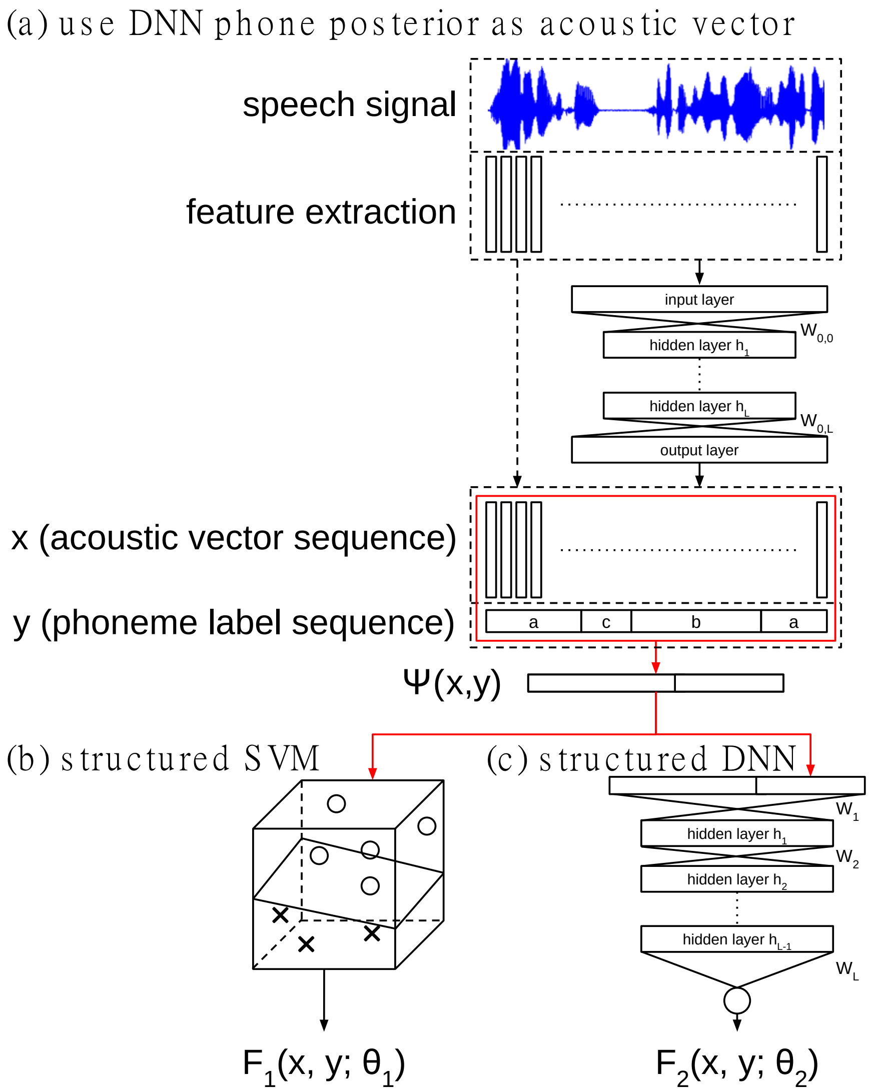


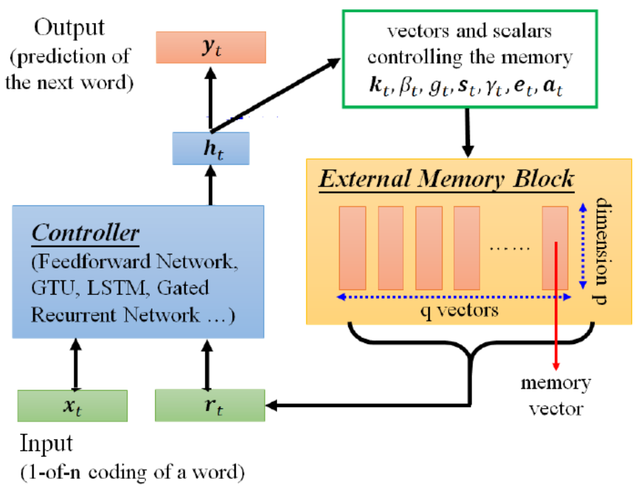
Add an extra symbol “ϕ” representing “null”



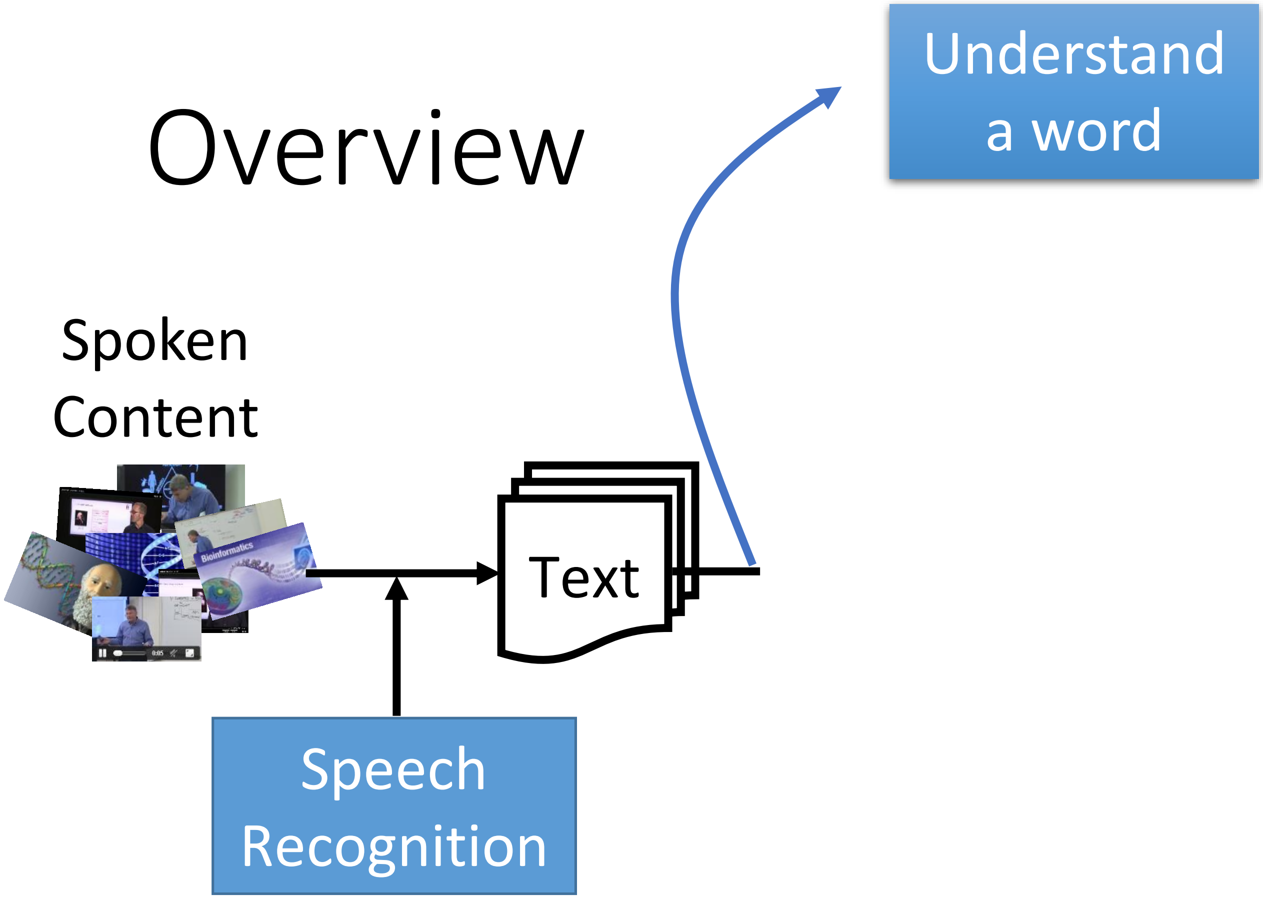
Proposed Approaches

- DNN + structured SVM
  - [Meng & Lee, ICASSP 10]
- DNN + structured DNN
  - [Liao & Lee, ASRU 15]
- Neural Turing Machine
  - [Ko & Lee, ICASSP 17]





Overview



Understand a word
Spoken Content
Bioinformatics
Text
Speech Recognition

1-of-N encoding

How to represent each word as a vector?

1-of-N Encoding lexicon = {apple, bag, cat, dog, elephant}

The vector is lexicon size.

Each dimension corresponds to a word in the lexicon

The dimension for the word is 1, and others are 0

apple = [1 0 0 0 0]

bag = [0 1 0 0 0]

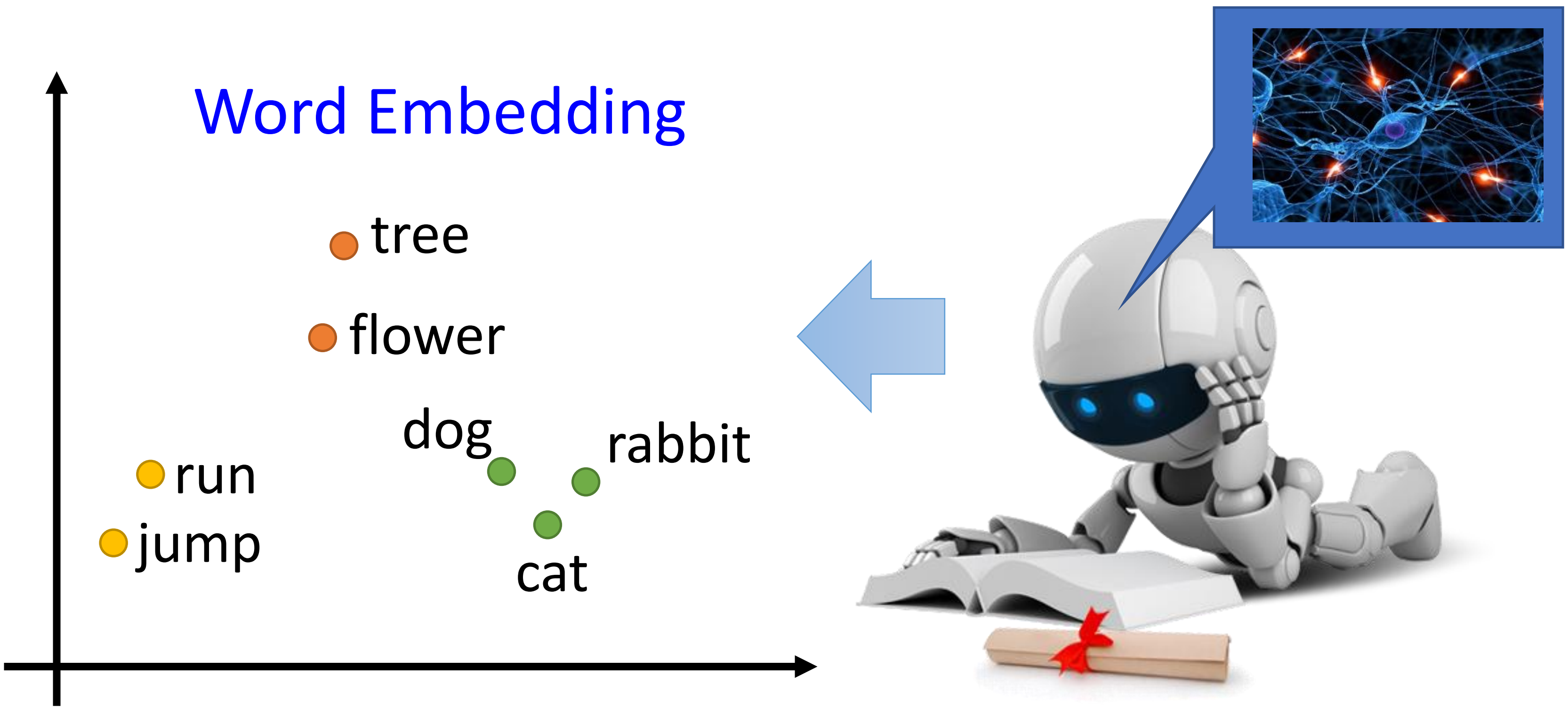
cat = [0 0 1 0 0]

dog = [0 0 0 1 0]

elephant = [0 0 0 0 1]

Word Embedding

- Machine learns the meaning of words from reading a lot of documents without supervision



Word Embedding

- Machine learns the meaning of words from reading a lot of documents without supervision
- A word can be understood by its context

蔡英文、馬英九 are something very similar

You shall know a word by the company it keeps

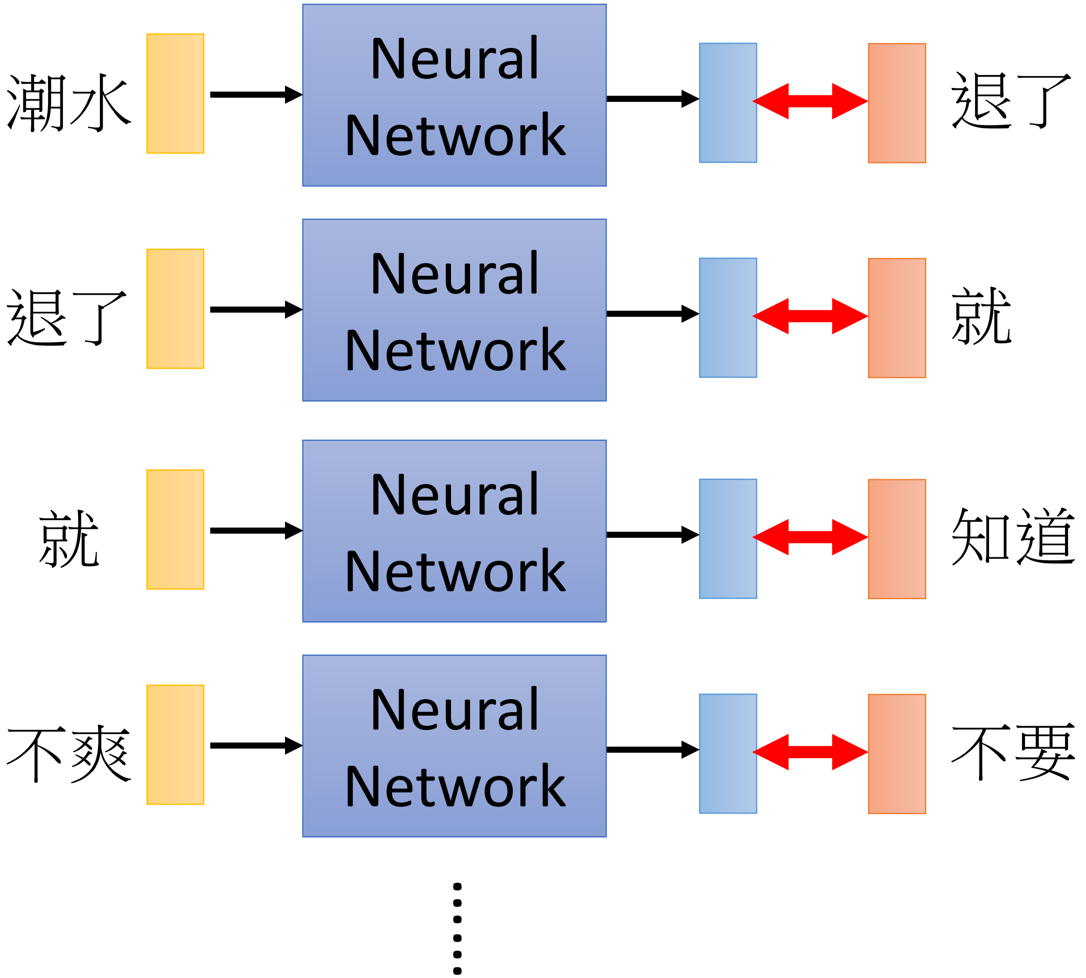
馬英九 520宣誓就職

蔡英文 520宣誓就職

Prediction-based

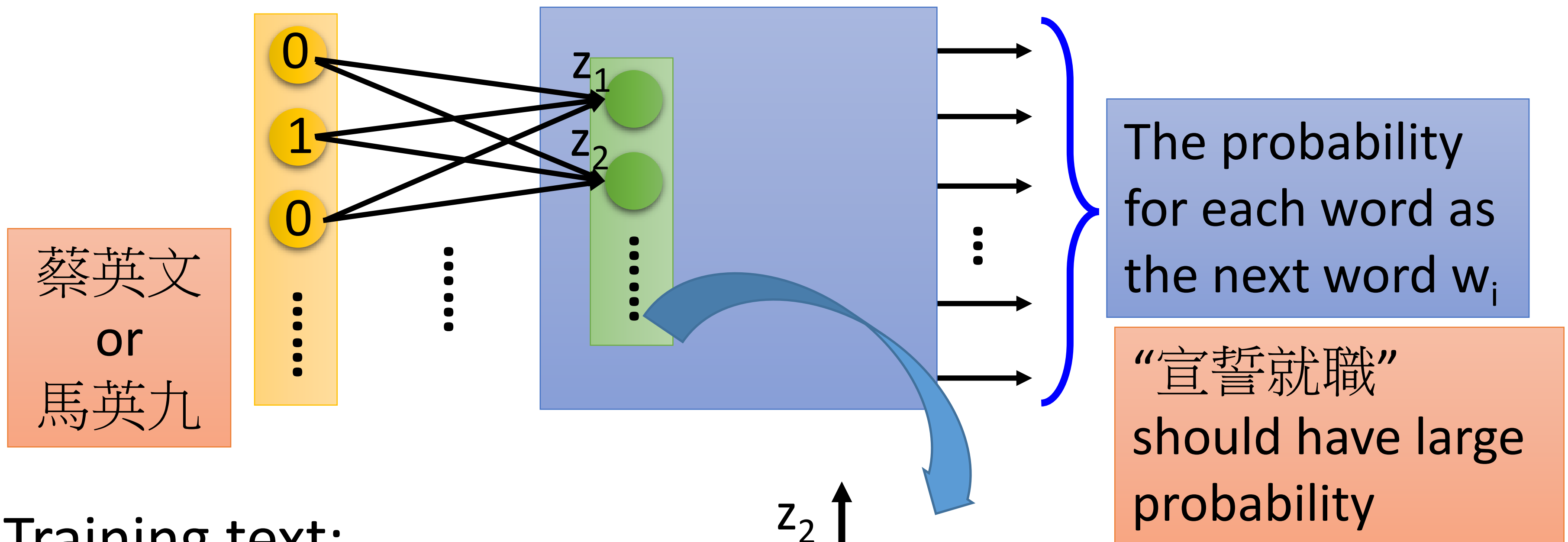
Collect data:

潮水 退了 就 知道 …
不爽 不要 買 …
公道價 八萬 一 …
………



Prediction-based

You shall know a word by the company it keeps



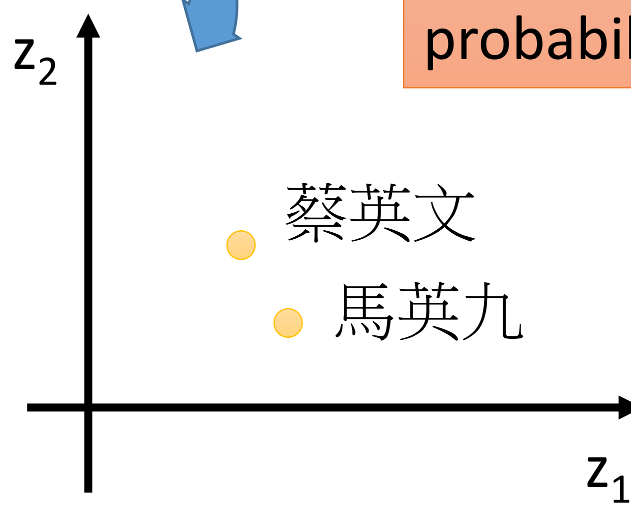
Training text:

…… 蔡英文 宣誓就職 ……

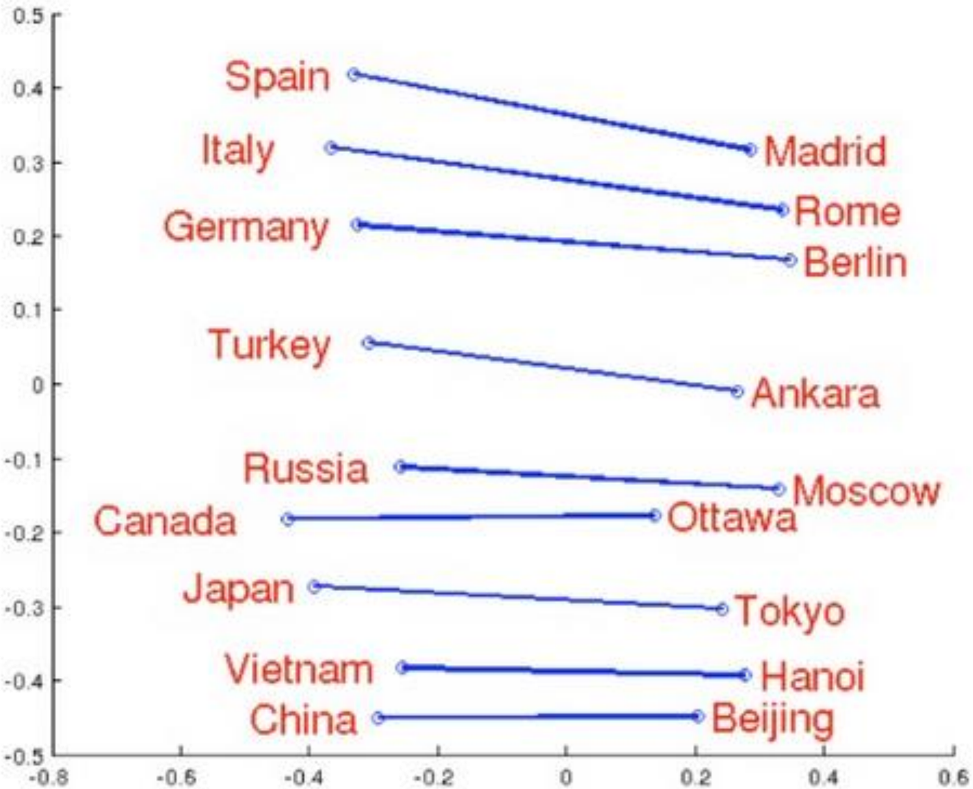
w_{i-1} w_i

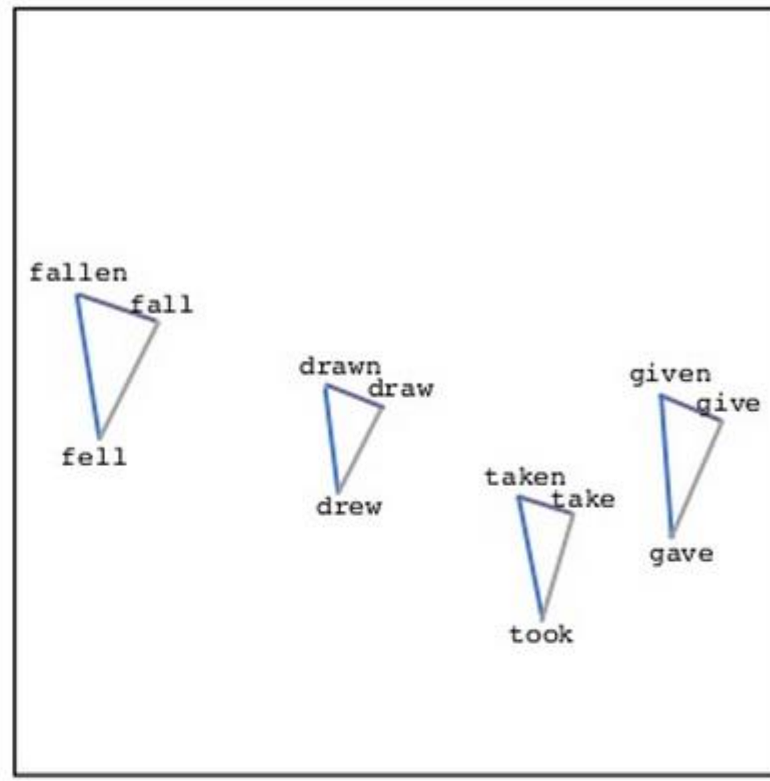
…… 馬英九 宣誓就職 ……

w_{i-1} w_i



Word Embedding





Source: http://www.slideshare.net/hustwj/cikm-keynotenov2014

Word Embedding

- Characteristics

$$V(Germany) \approx V(Berlin) - V(Rome) + V(Italy)$$

$$V(hotter) - V(hot) \approx V(bigger) - V(big)$$

$$V(Rome) - V(Italy) \approx V(Berlin) - V(Germany)$$

$$V(king) - V(queen) \approx V(uncle) - V(aunt)$$

- Solving analogies

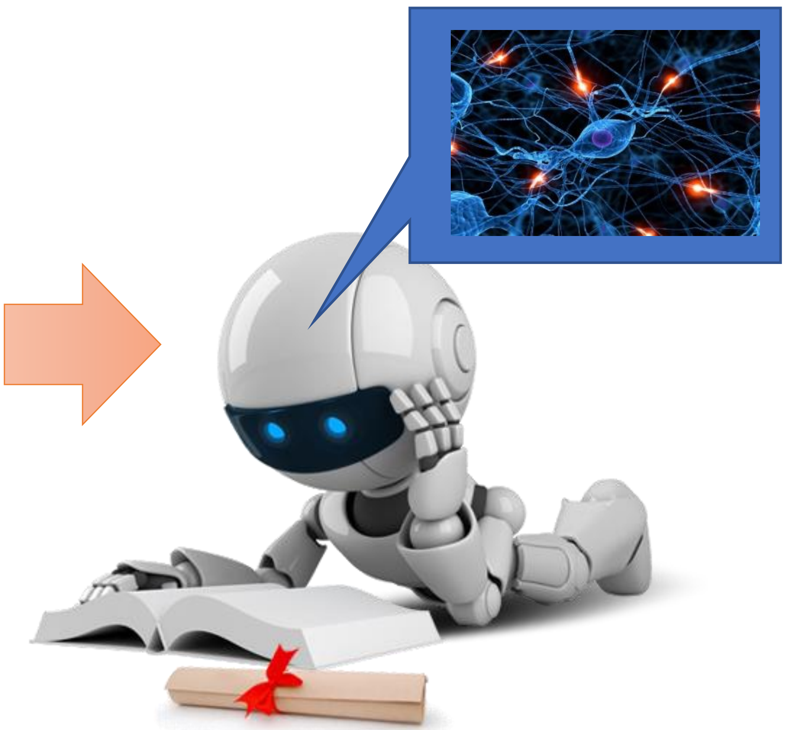
Rome : Italy = Berlin : ?

Compute $\underline{V(Berlin) - V(Rome) + V(Italy)}$

Find the word w with the closest V(w)

Demo

- Machine learn the meaning of words from reading a lot of documents without supervision



Overview

Understand a word

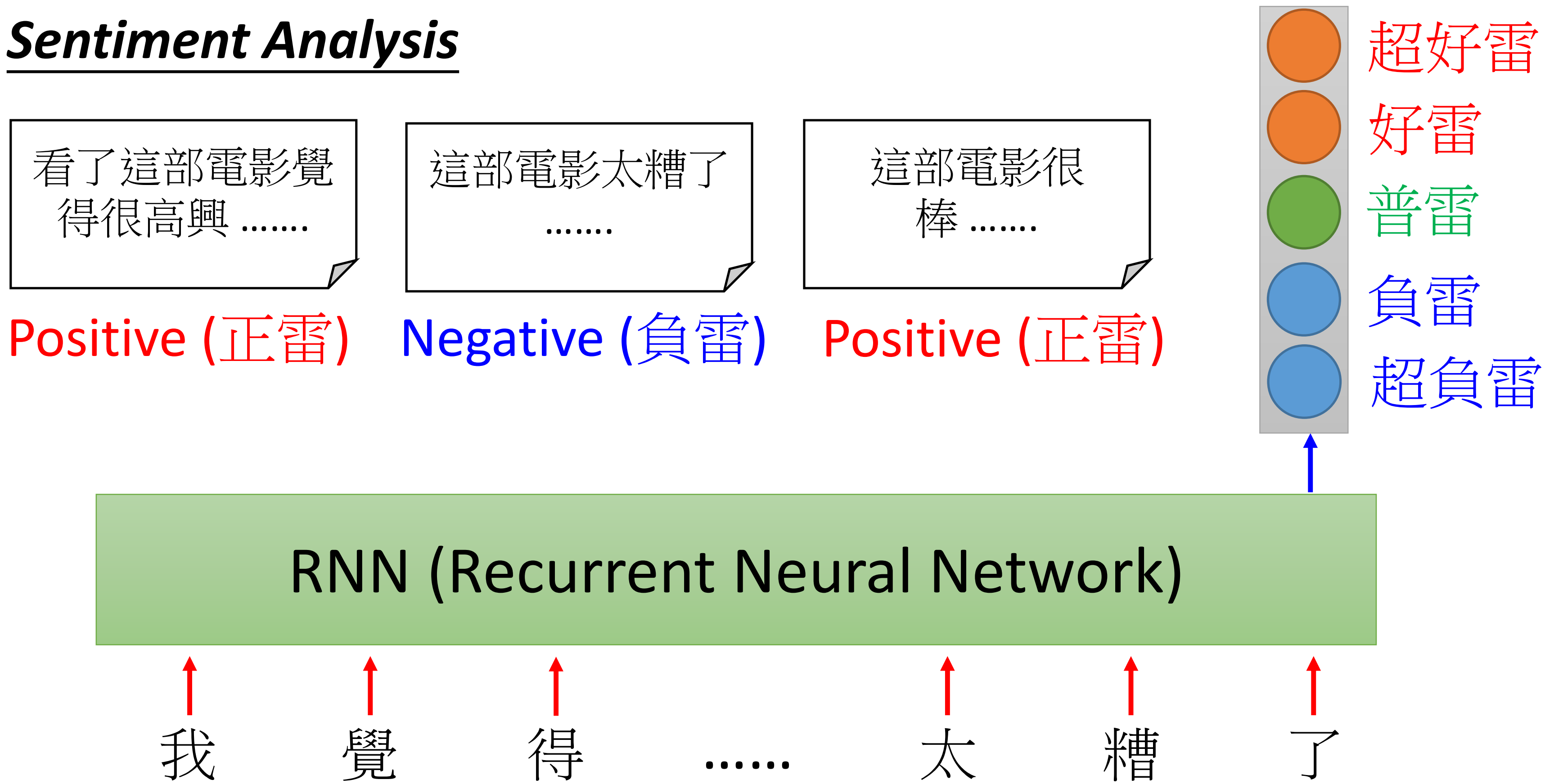
Understand a sentence

Spoken Content

Text

Speech Recognition

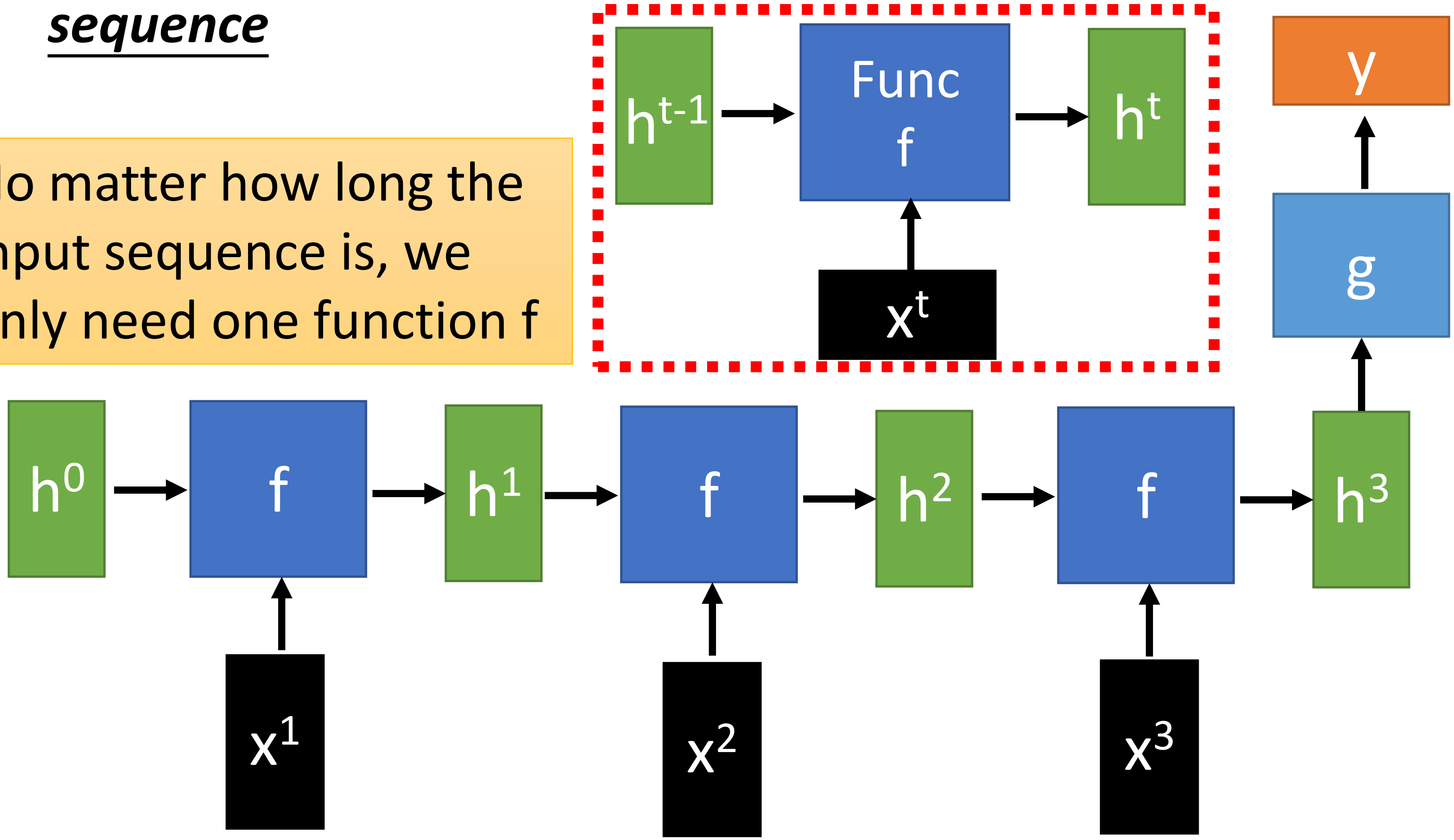
Sentiment Analysis

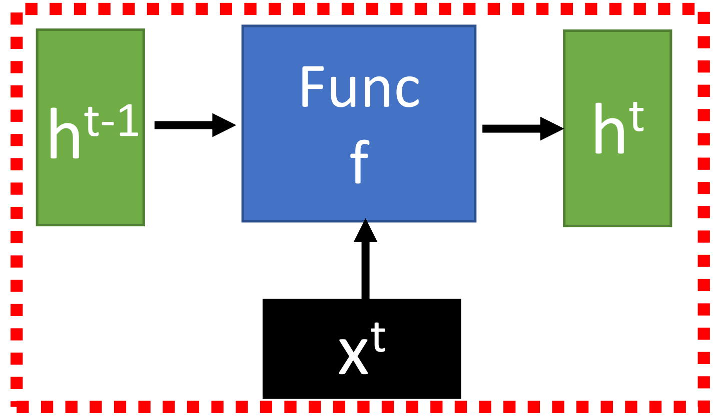


Recurrent Neural Network

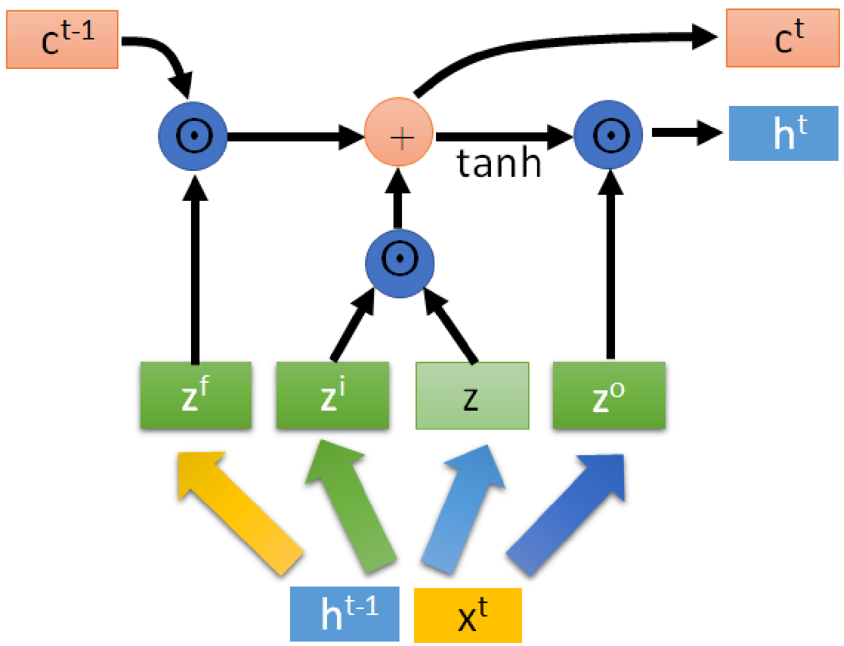
- Recurrent Structure: usually used when the ***input is a sequence***

No matter how long the input sequence is, we only need one function f



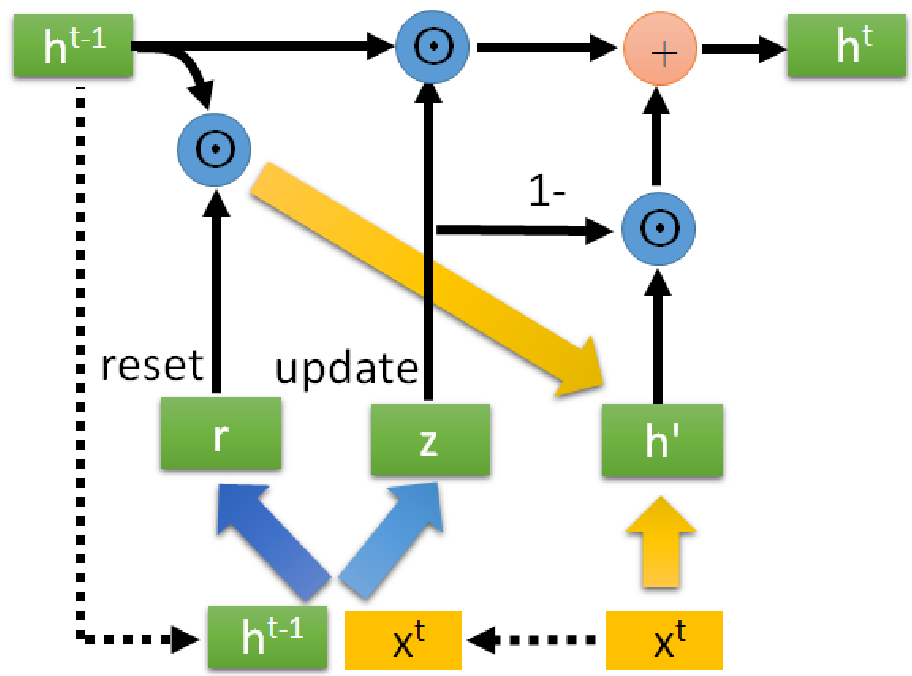


h^t-1
Func
f
h^t
x^t



c^t-1
c^t
tanh
h^t
z^f
z^i
z
z^o
h^t-1
x^t

LSTM



h^t-1
h^t
1-
reset
update
r
z
h'
h^t-1
x^t
x^t

GRU

Demo

Summarization

Spoken Content

Text

Speech Recognition

Summarization

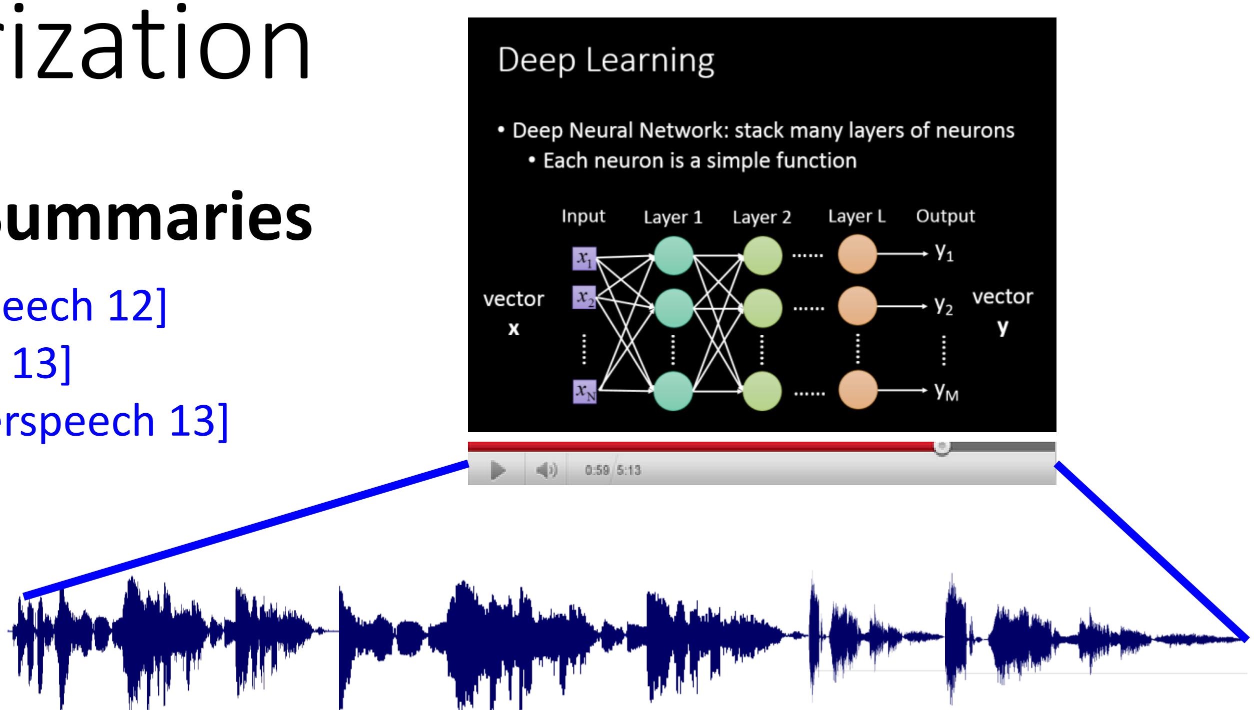
[Yu, et al., SLT 16]

[Lu, et al., Interspeech 17]

Summarization

Extractive Summaries

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



Audio File
to be summarized

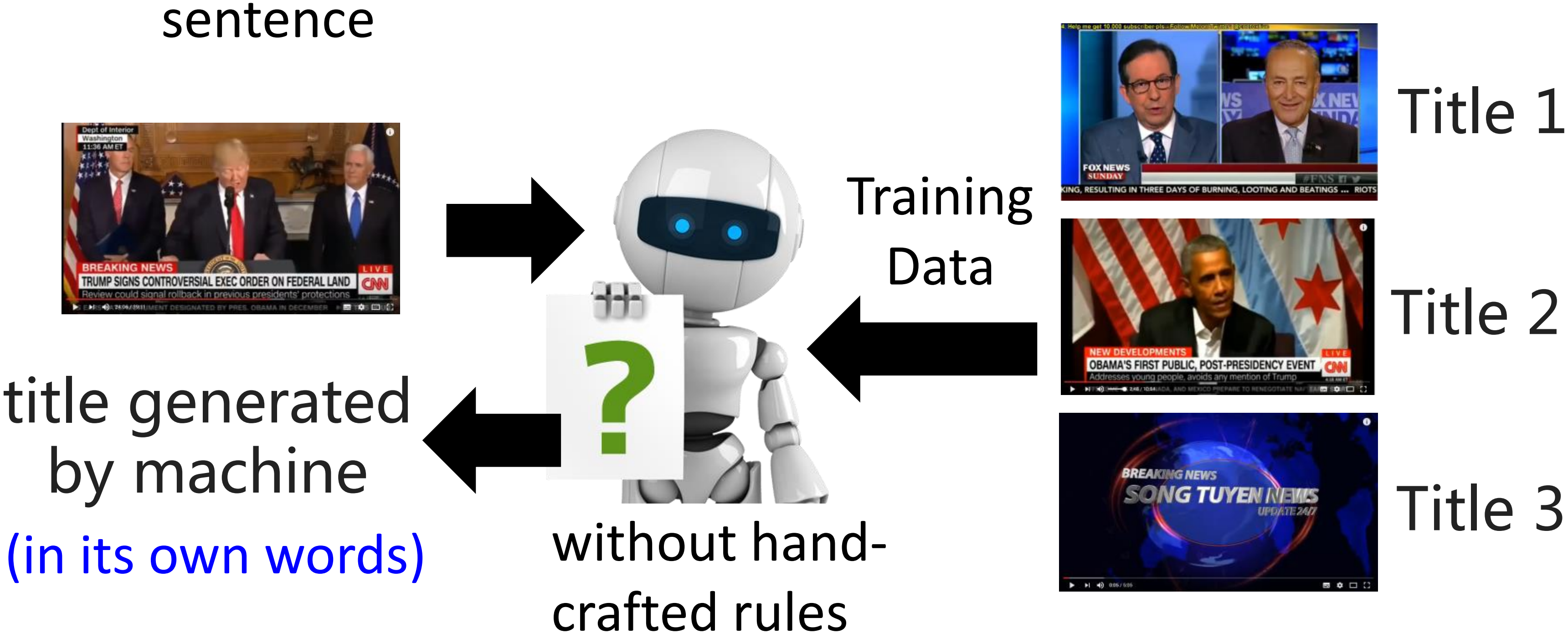
…… deep learning is powerful …… …… ……

This is the summary.

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

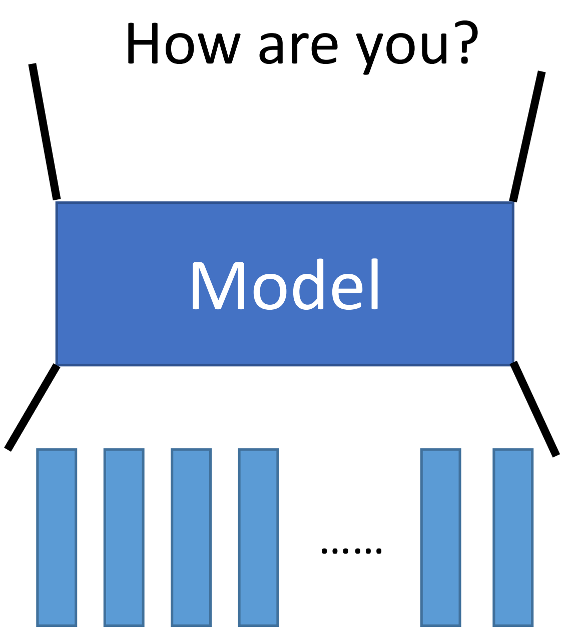
Abstractive Summarization

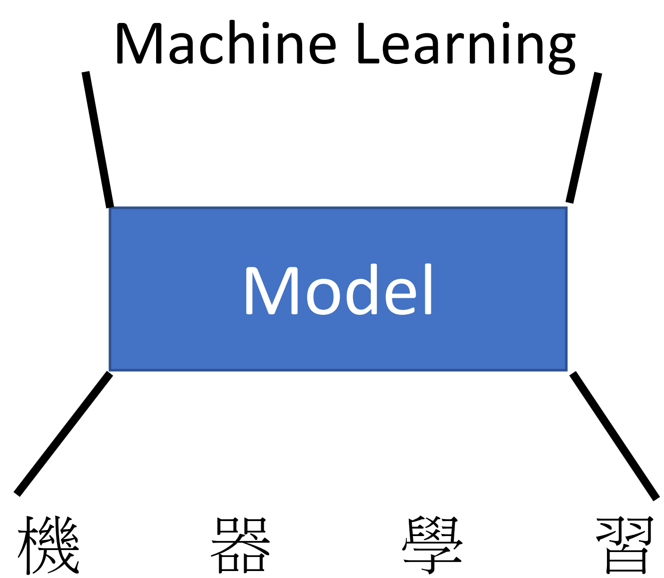
- Now machine can do **abstractive summary** (write summaries in its own words)
  - *Title generation*: abstractive summary with one sentence



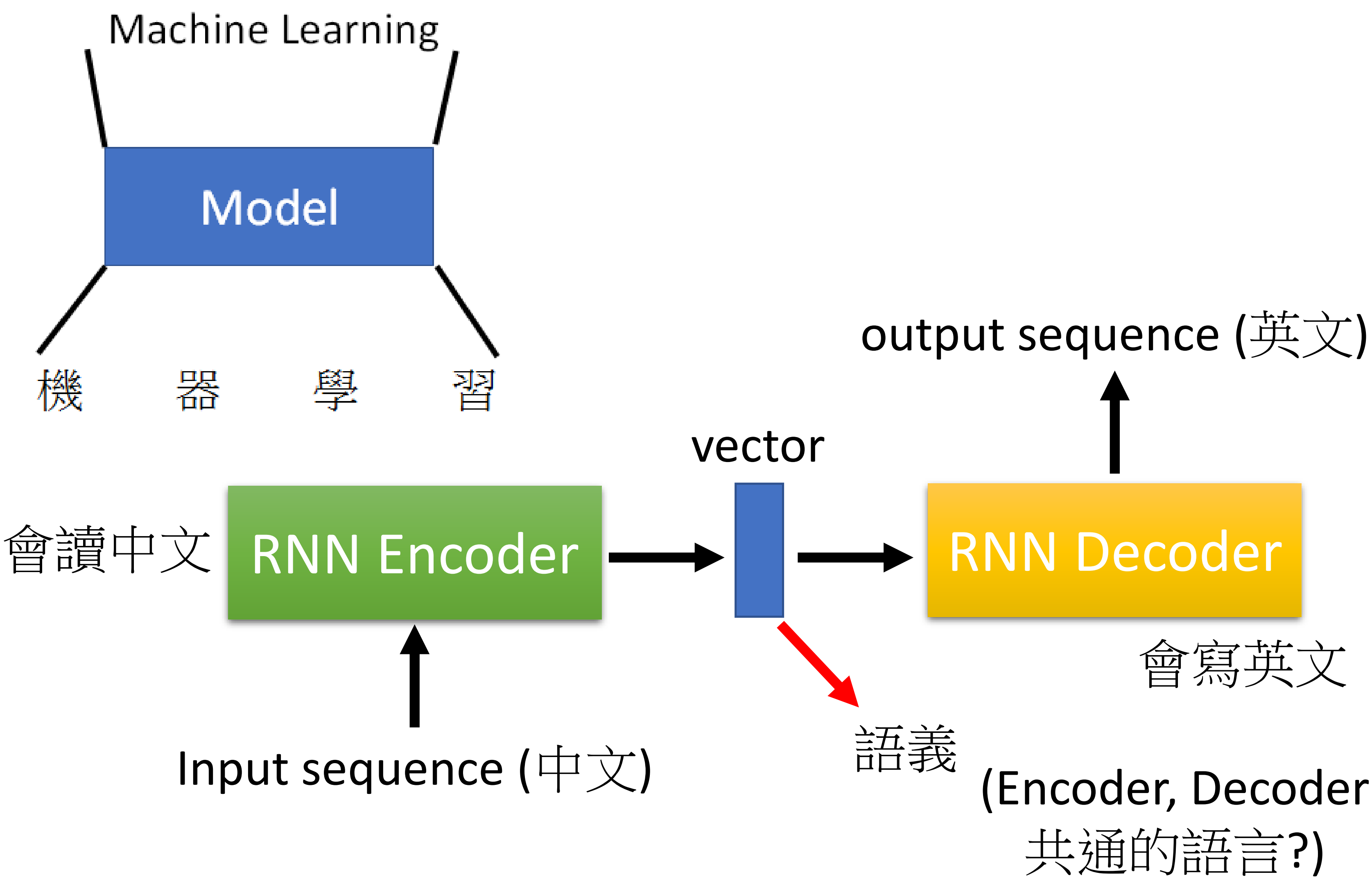
Sequence-to-sequence Learning

- Sequence to sequence learning: Both input and output are both sequences ***<u>with different lengths</u>***.



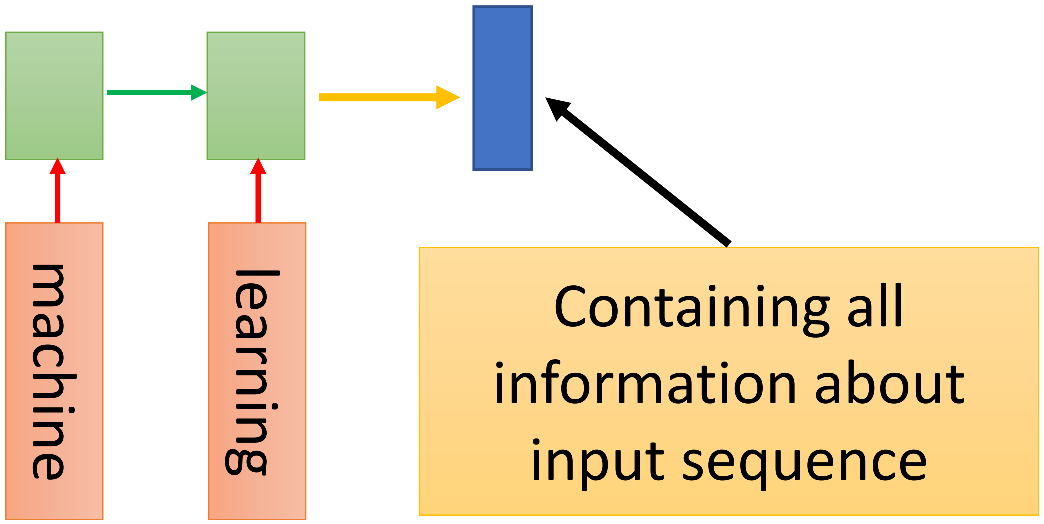


Sequence-to-sequence Learning



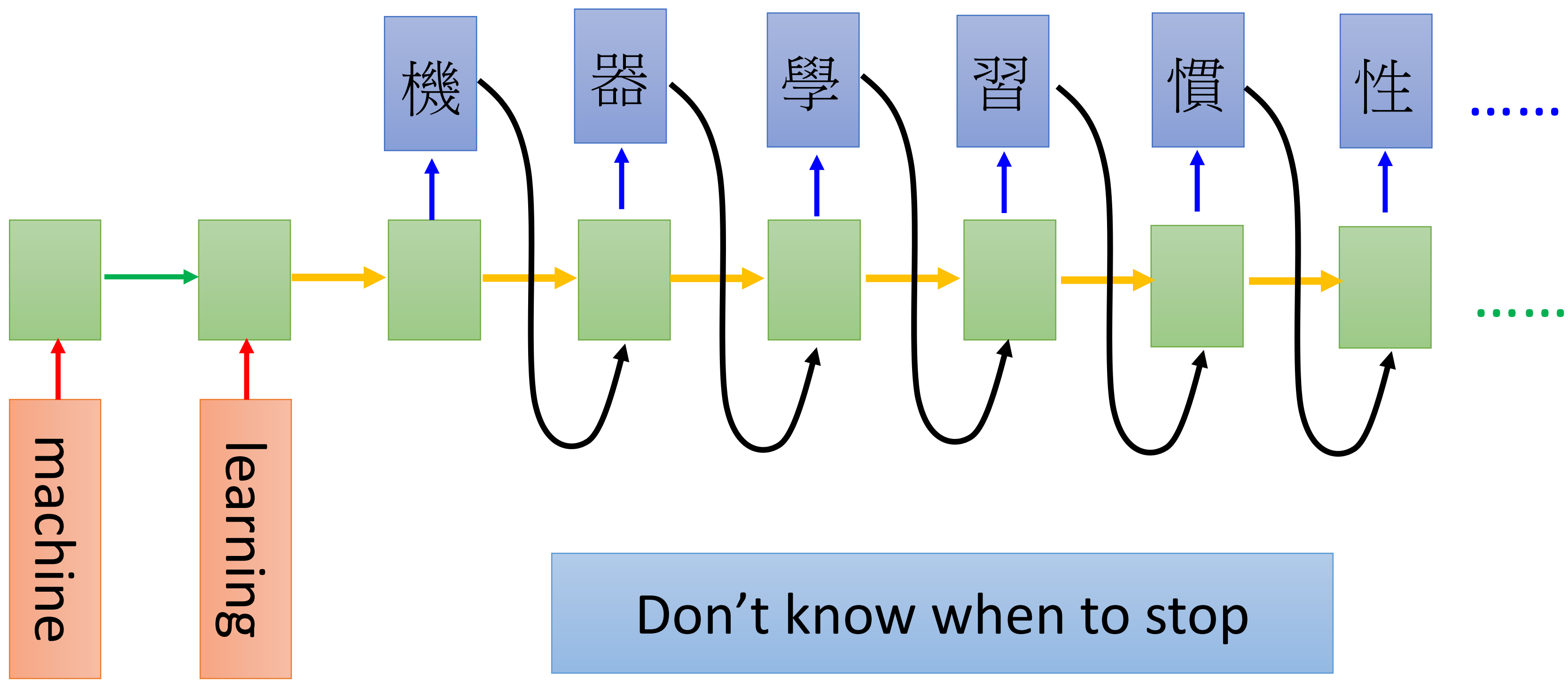
Sequence-to-sequence Learning

- Both input and output are both sequences ***with different lengths***. → ***Sequence to sequence learning***
  - E.g. ***Machine Translation*** (machine learning→機器學習)

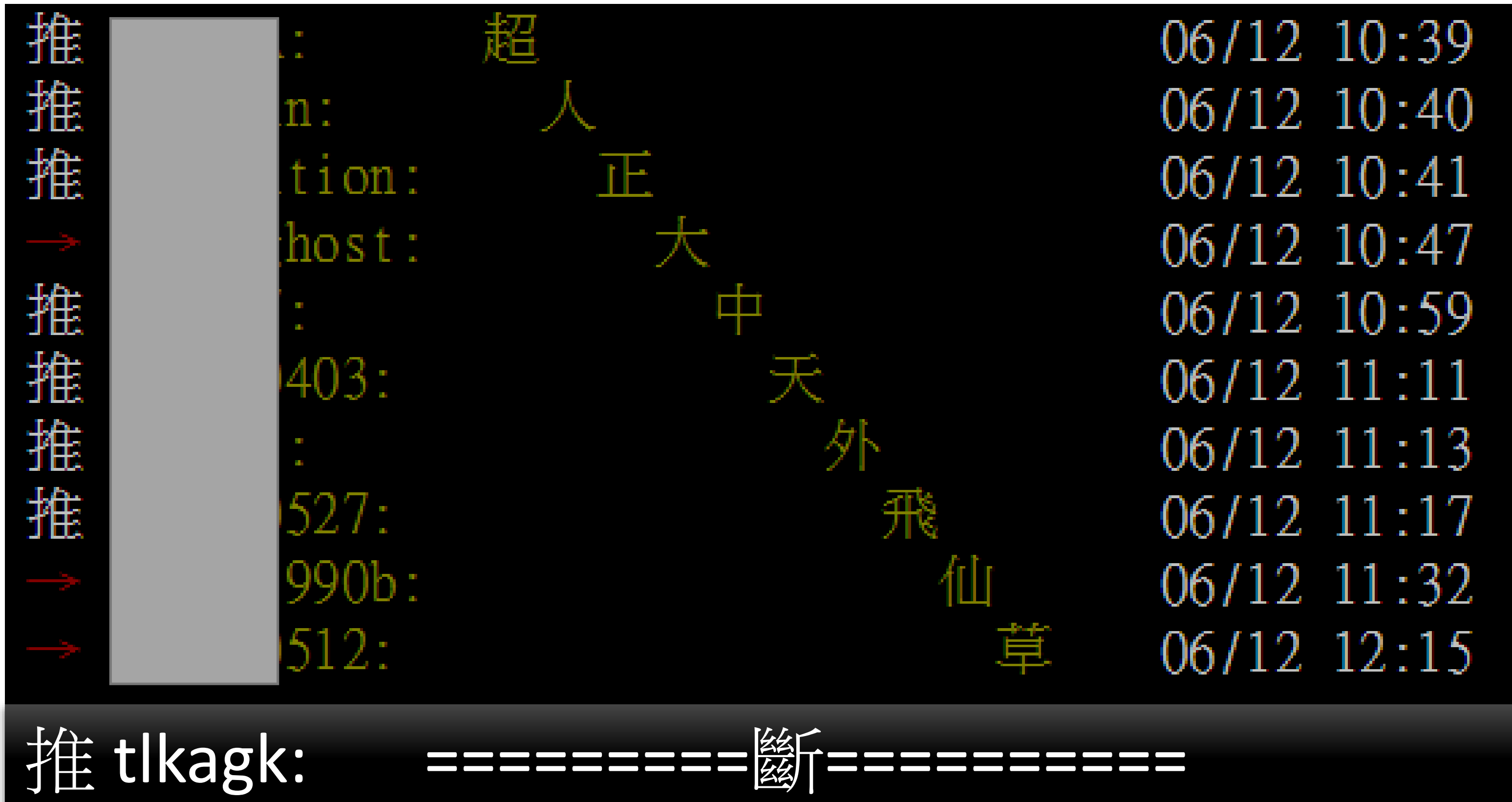


Sequence-to-sequence Learning

- Both input and output are both sequences ***with different lengths***. → ***Sequence to sequence learning***
  - E.g. ***Machine Translation*** (machine learning→機器學習)



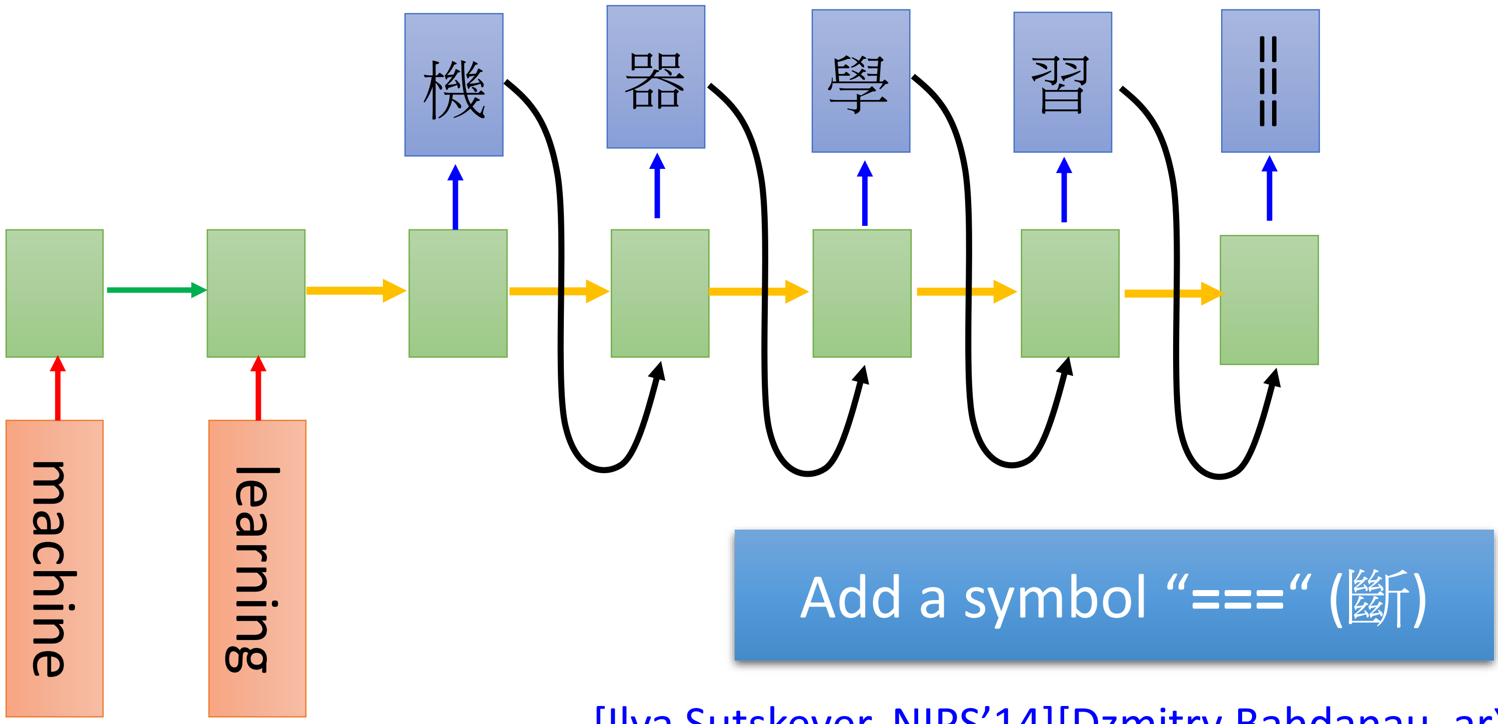
Sequence-to-sequence Learning



接龍推文是ptt在推文中的一種趣味玩法，與推齊有些類似但又有所不同，是指在推文中接續上一樓的字句，而推出連續的意思。該類玩法確切起源已不可知(鄉民百科)

Sequence-to-sequence Learning

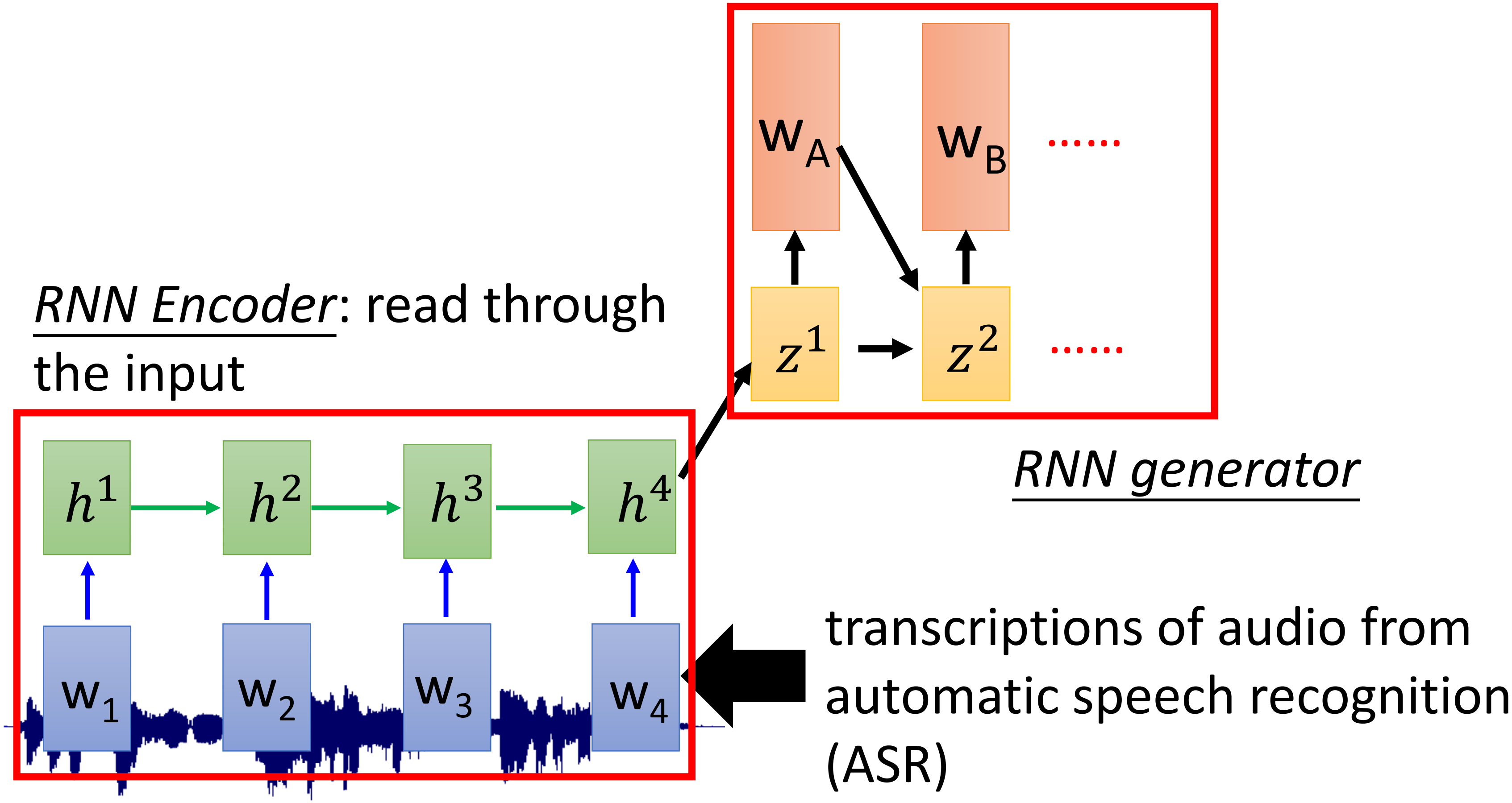
- Both input and output are both sequences ***with different lengths***. → ***Sequence to sequence learning***
  - E.g. ***Machine Translation*** (machine learning→機器學習)



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

Summarization

- Input: transcriptions of audio, output: title



感謝 盧柏儒 提供實驗結果

Summarization

Document: 據印度報業托拉斯報道印度北方邦22日發生一起小公共汽車炸彈爆炸事件造成15 人死亡 3 人受傷 ……

Human: 印度汽車炸彈爆炸造成15人死亡

Machine: 印度發生汽車爆炸事件

Document: 刑事局偵四隊今天破獲一個中日跨國竊車集團，根據調查國內今年七月開放重型機車上路後 ……

Human:跨國竊車銷贓情形猖獗直得國內警方注意

Machine:刑事局破獲中國車集

CTC

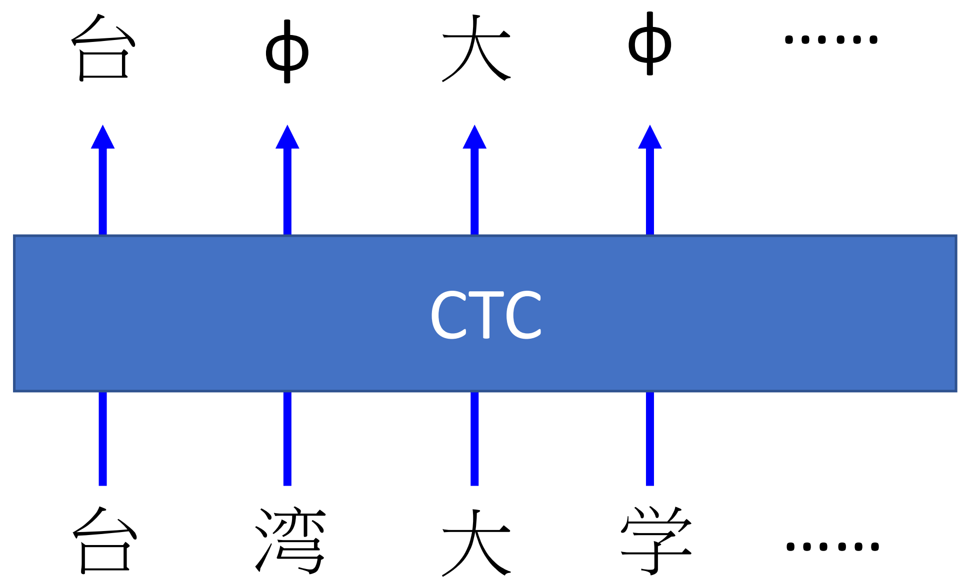


Table 1: *Performance comparison of various models and input sequence elements over Chinese Gigaword (no ASR errors).*

Model			Input	Output	k	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-L
Baseline	Seq2Seq	(a)	word	character	1	34.47	18.30	8.82	31.26
		(b)	character	character	1	36.33	18.58	8.78	32.39
	Attentive Seq2Seq	(c)	word	character	1	36.37	20.23	10.23	32.98
		(d)	character	character	1	37.97	20.47	10.27	33.88
Proposed: CTC		(e)	word	character	1	25.36	9.20	3.43	24.49
		(f)	word	character	2	33.58	15.70	7.34	32.20
		(g)	character	character	1	**42.71**	**24.62**	**14.24**	**40.56**

Experiments

- Training data: Chinese Gigaword
  - Text documents
  - 2M story-headline pairs
- Testing data: TV News
  - Spoken documents
  - 50 hours (1k spoken documents)
  - Character Error Rate = 28.7% (our system), 36.5% (wit.ai)
- Input and output of the model are both Chinese characters

	ROUGE-1	ROUGE-2	ROUGE-L
Manual (Oracle)	26.8	6.5	23.9
ASR	21.3	4.8	20.0

Pseudo ASR error

- Adding pseudo ASR error into training data
  - Analyze the error patterns of ASR system
  - Changing some characters in training text documents by probability

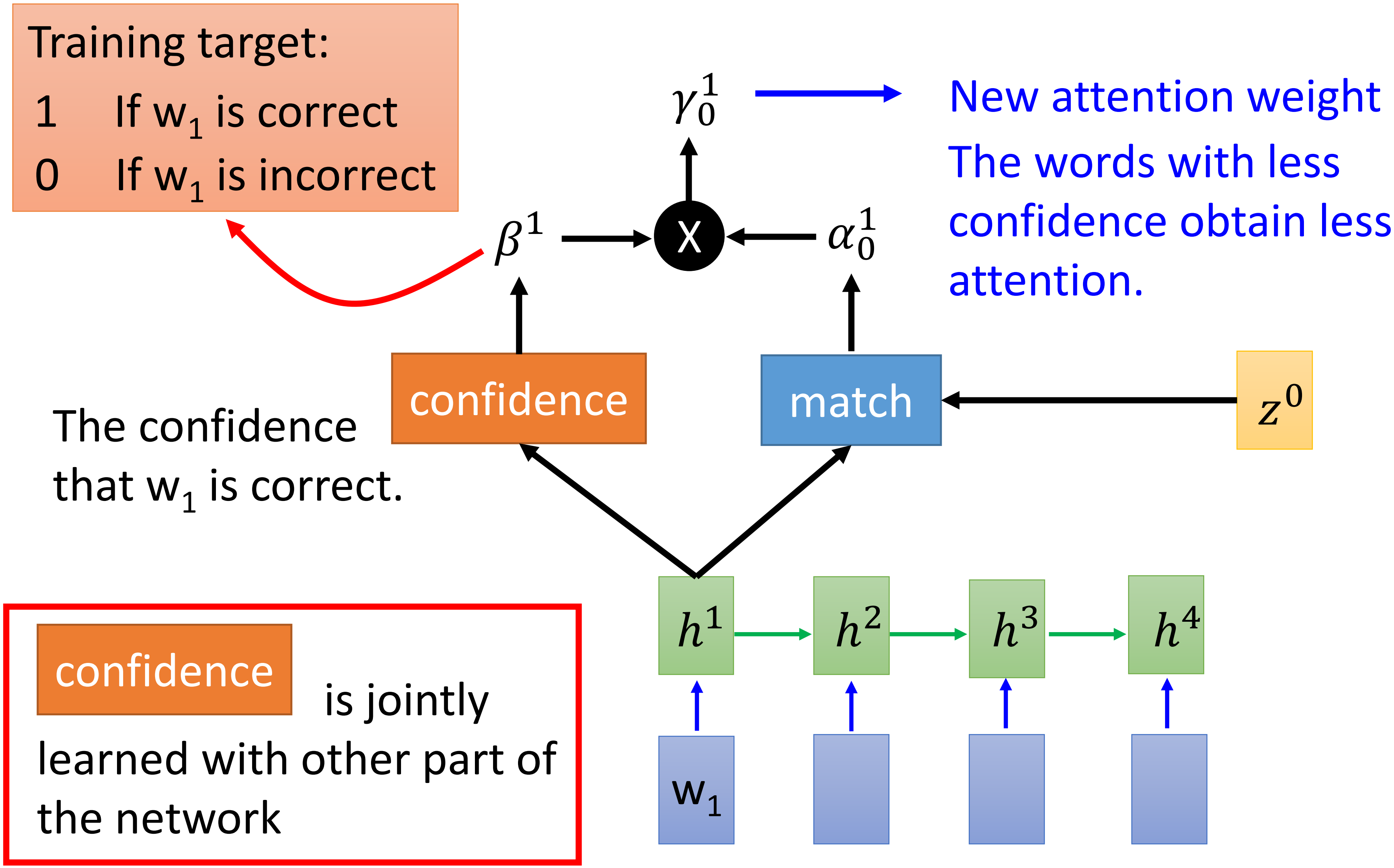
Training Data	ROUGE-1	ROUGE-2	ROUGE-L
Text	21.3	4.8	20.0
Text + pseudo error	20.9	3.4	19.1

(Testing spoken documents have ASR errors)

➢ Even worse after adding pseudo error

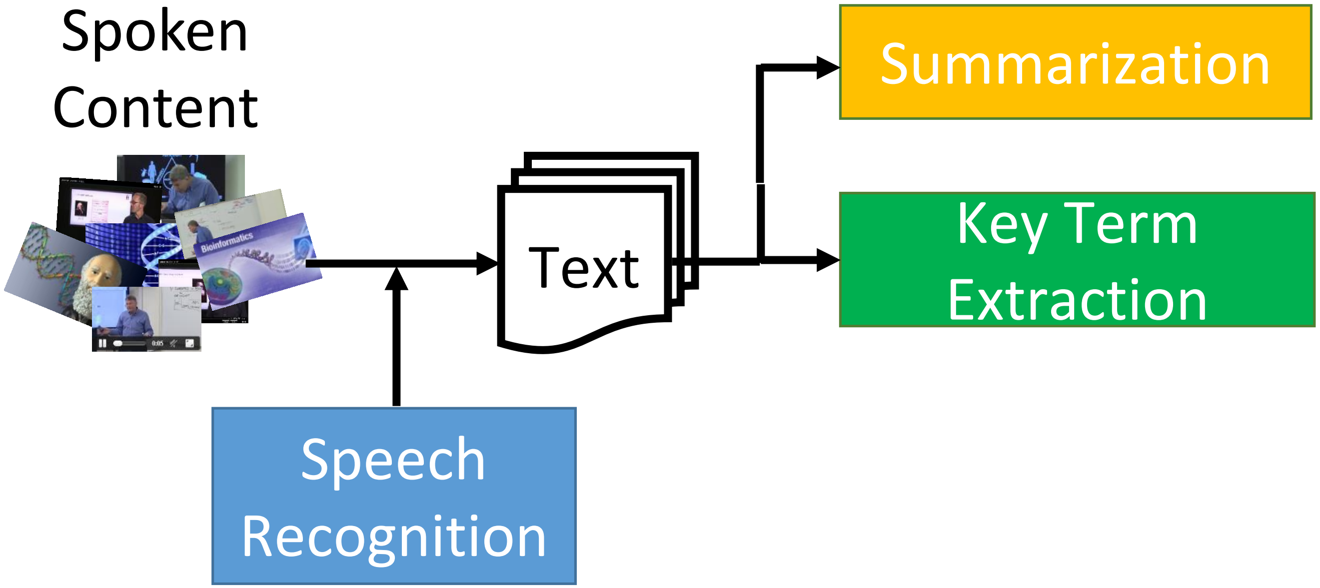
➢ The model learns to correct the ASR error in input document, which is difficult and causes over-fitting

Learn to Ignore ASR Errors



Ch				ROUGE-1	ROUGE-2	ROUGE-L
Text	BSL		Seq-2-seq	21.87	4.93	20.52
			w/ att.	21.32	4.84	20.05
Consider-ing ASR Error	Character	naï	Seq-2-seq	19.50	3.57	18.50
			w/ att.	20.86	3.40	19.09
		Proposed		22.89	5.01	20.86
	Phoneme	naï	Seq-2-seq	19.32	3.13	17.79
			w/ att.	19.46	3.25	18.06
		Proposed		24.01	5.16	22.13
	Initial/Final	Naï	Seq-2-seq	19.87	3.36	17.42
			w/ att.	20.41	3.24	18.60
		Proposed		24.56	5.73	22.41
	Syllable	Naï	Seq-2-seq	19.37	2.72	17.34
			w/ att.	19.64	2.71	17.49
		Proposed		22.62	4.46	20.60
Text	Oracle		Seq-2-seq	26.60	5.68	23.70
			w/ att.	26.75	6.54	23.91

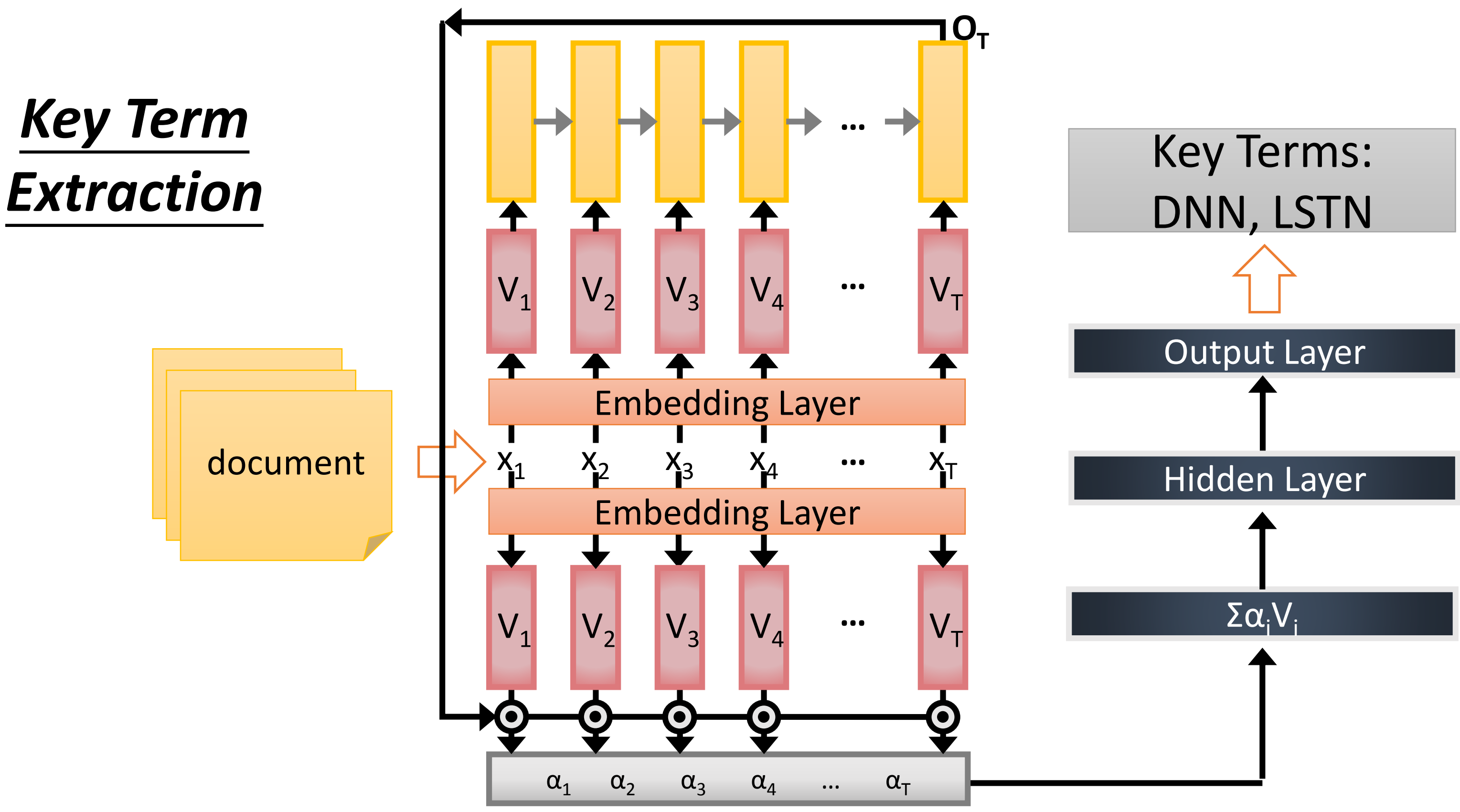
Key Term Extraction



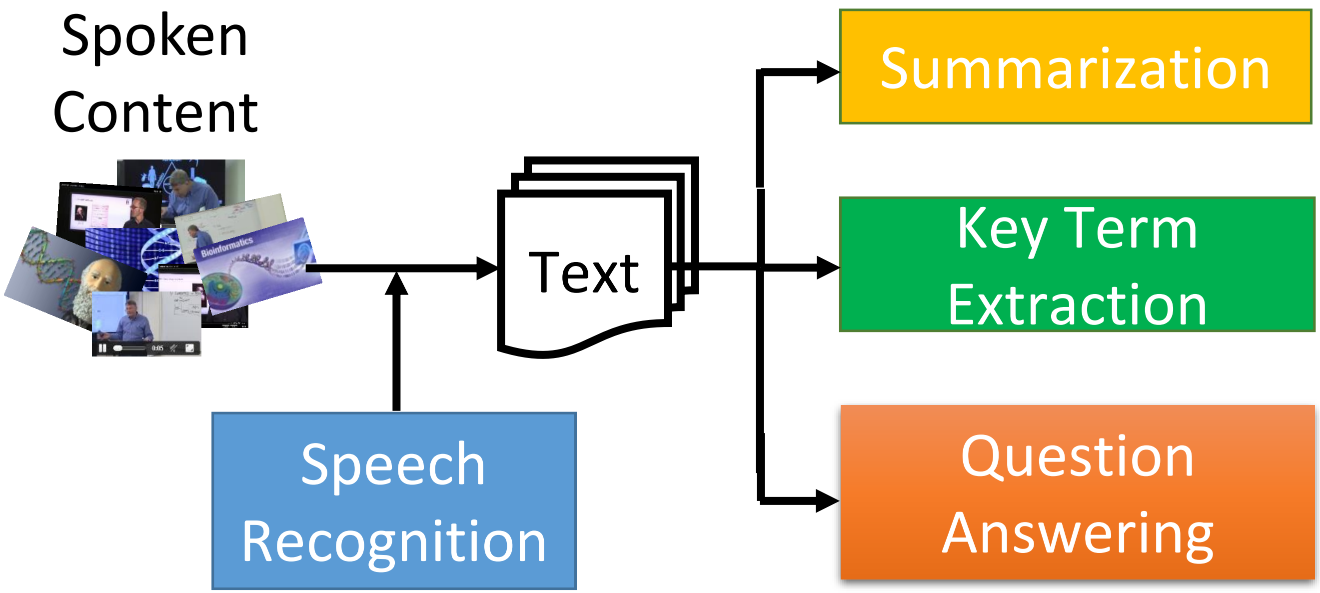
Key Term Extraction

[Shen & Lee, Interspeech 16]

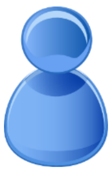
- Input is a vector sequence, but output is only one vector



Speech Question Answering



Speech Question Answering



What is a possible origin of Venus' clouds?

Gases released as a result of volcanic activity

<u>Speech Question Answering</u>: 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

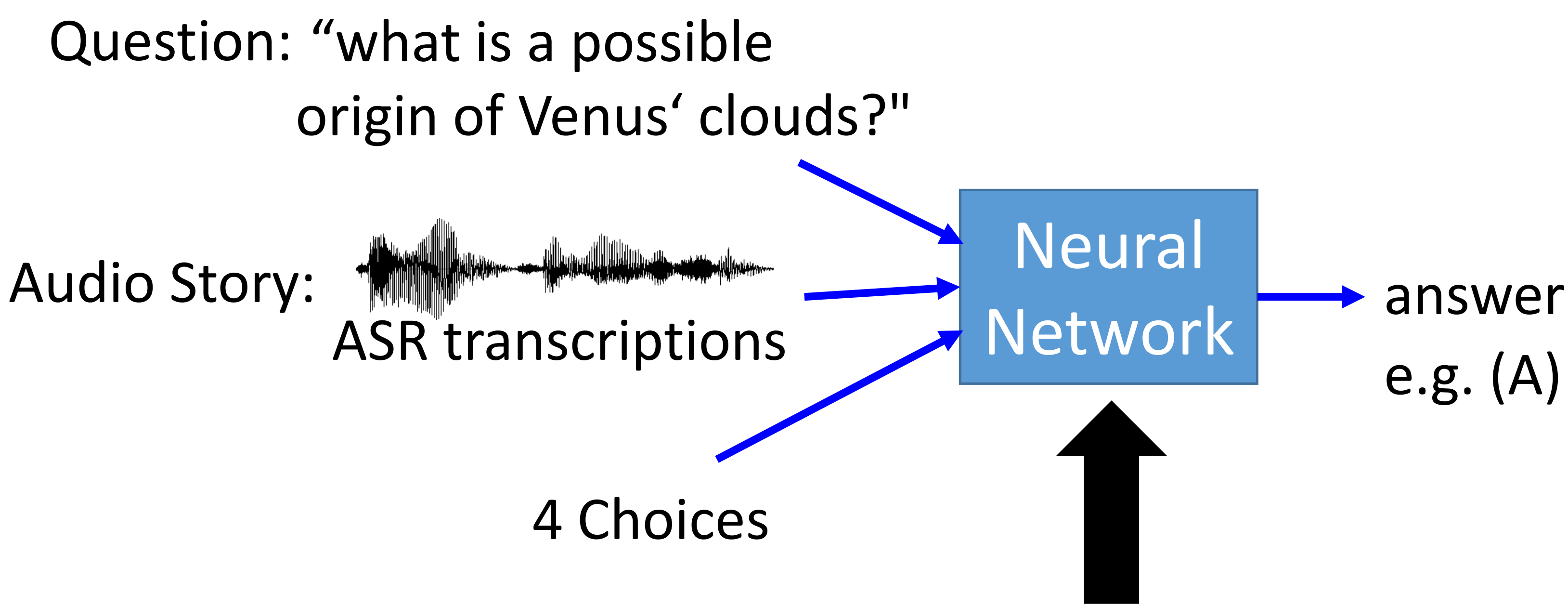
(B) chemical reactions caused by high surface temperatures

(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

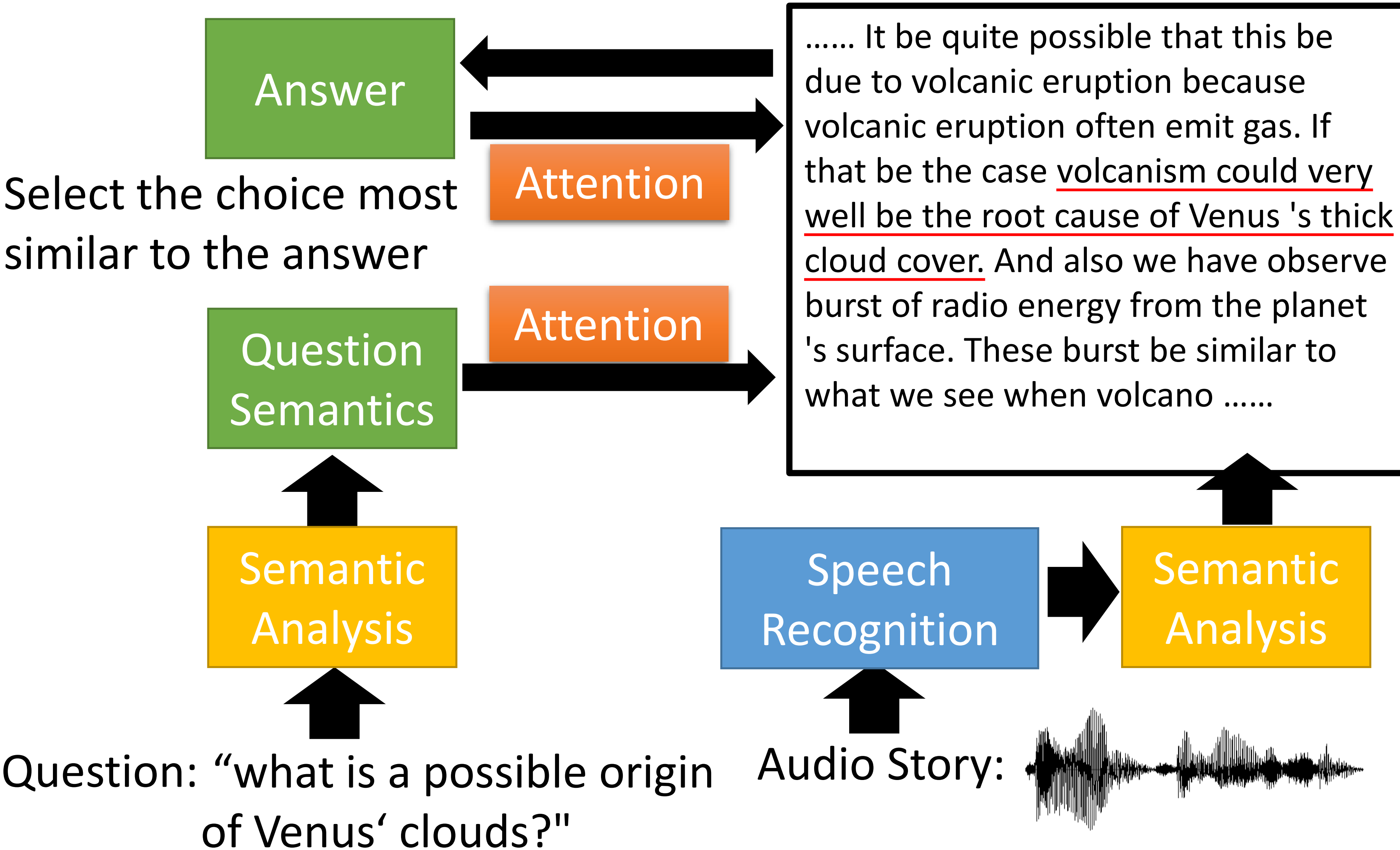
- **TOEFL Listening Comprehension Test by Machine**



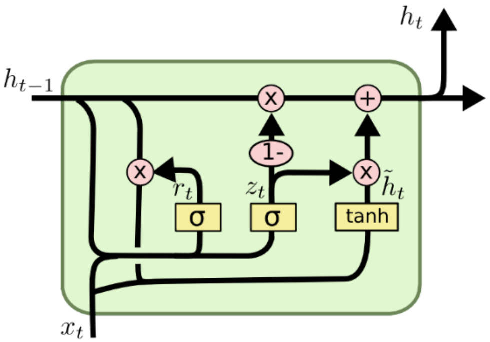
Using **previous exams** to train the network

Model Architecture

The whole model learned end-to-end.



Model Details

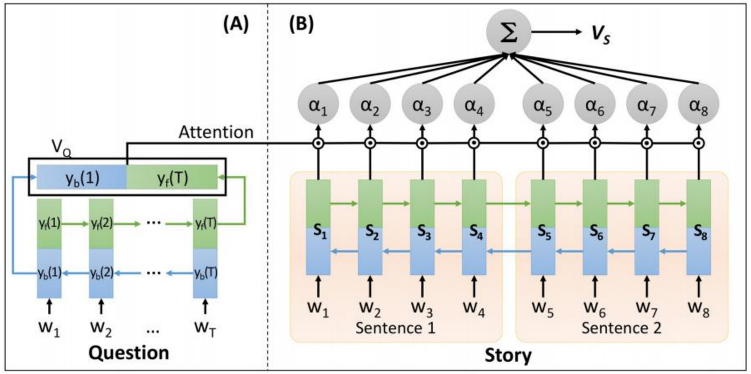


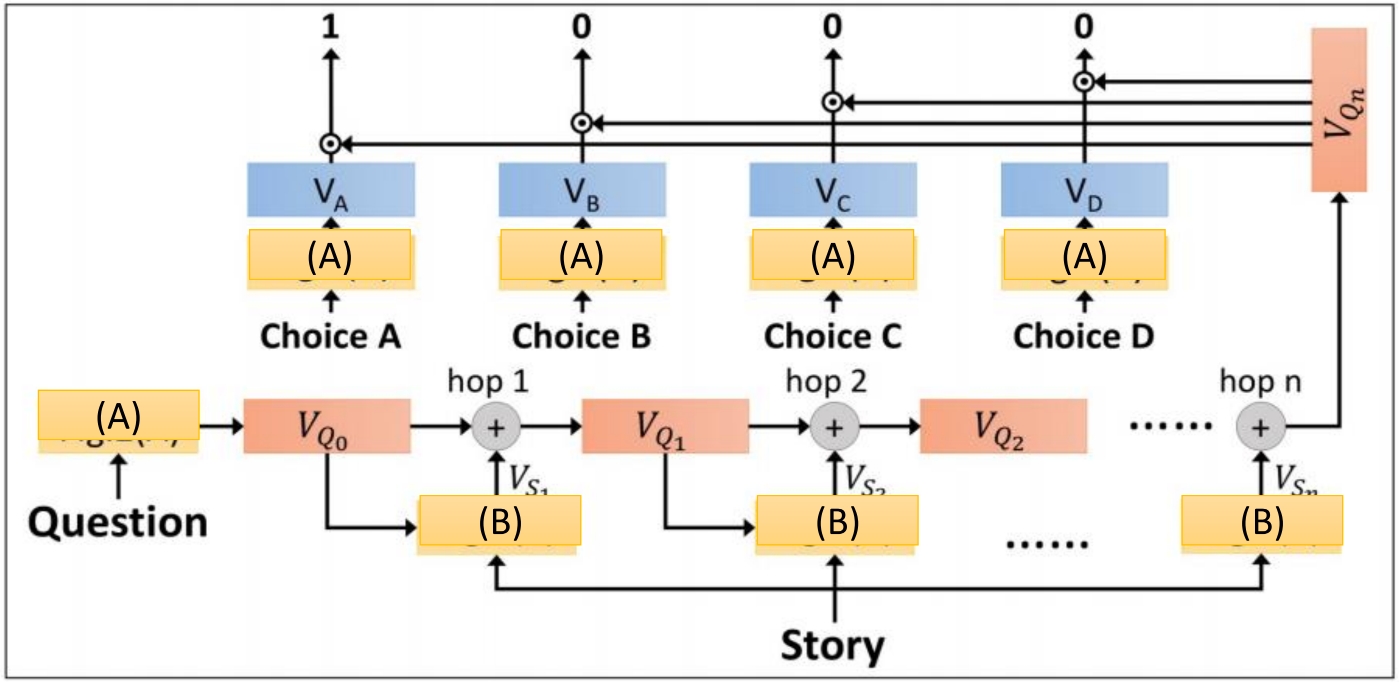
$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

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

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

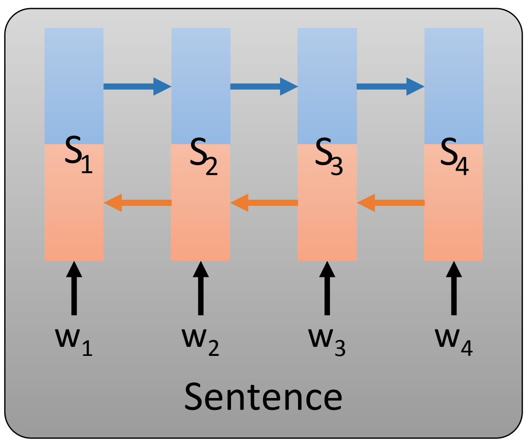
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$



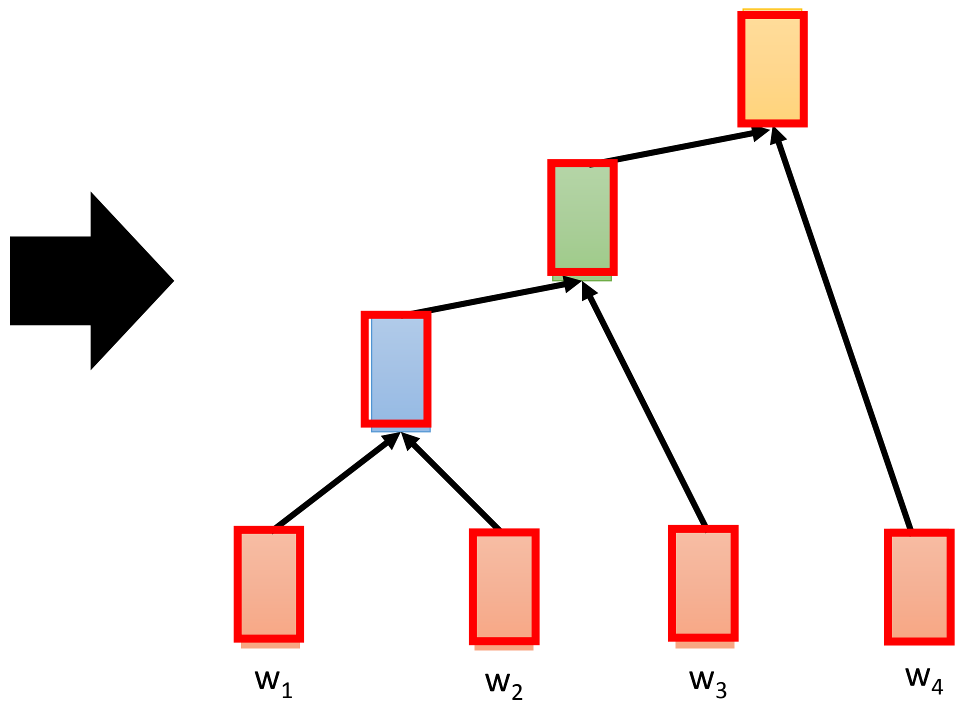


Sentence Representation

Bi-directional RNN



Tree-structured Neural Network

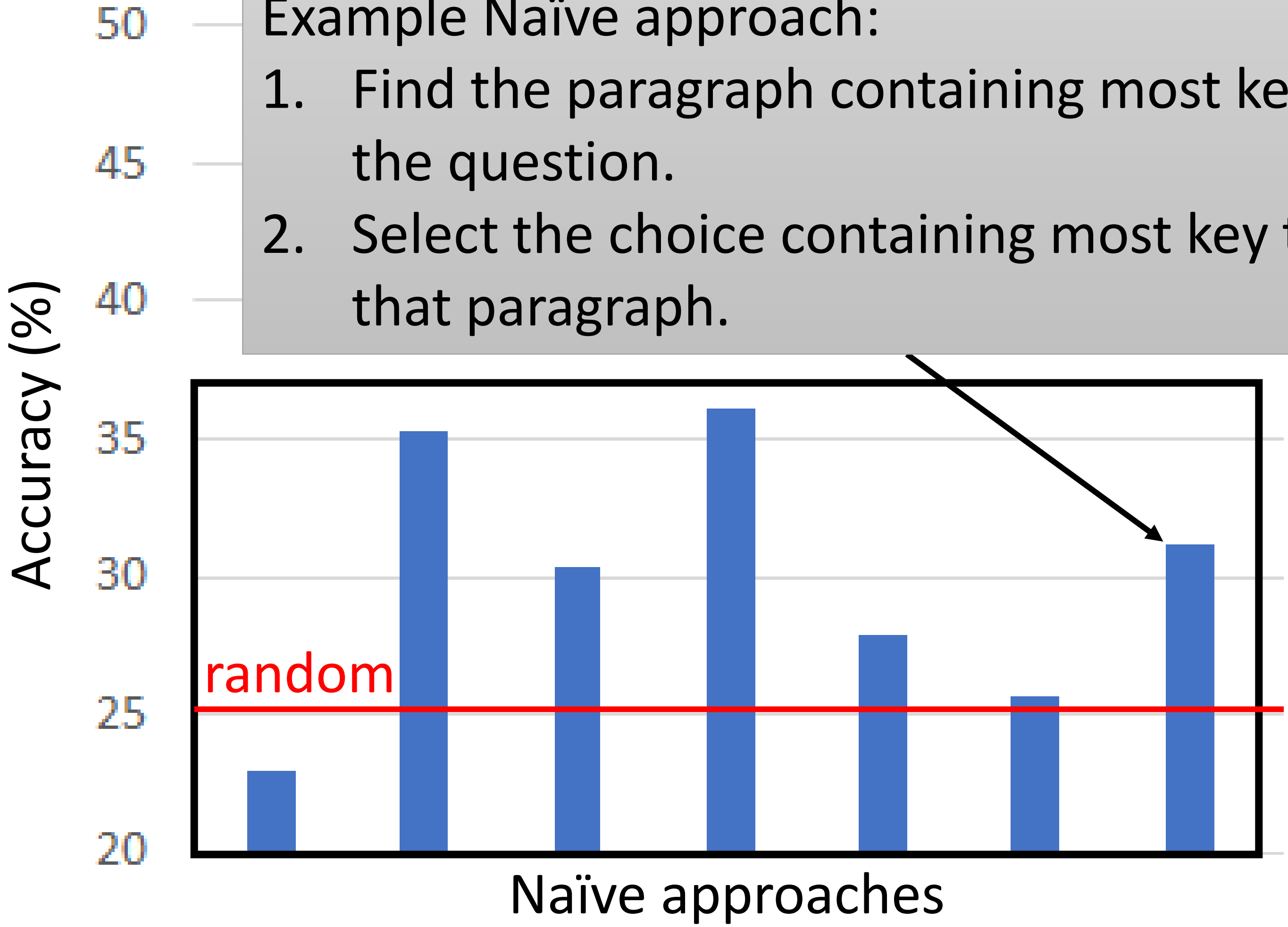


Attention on all phrases

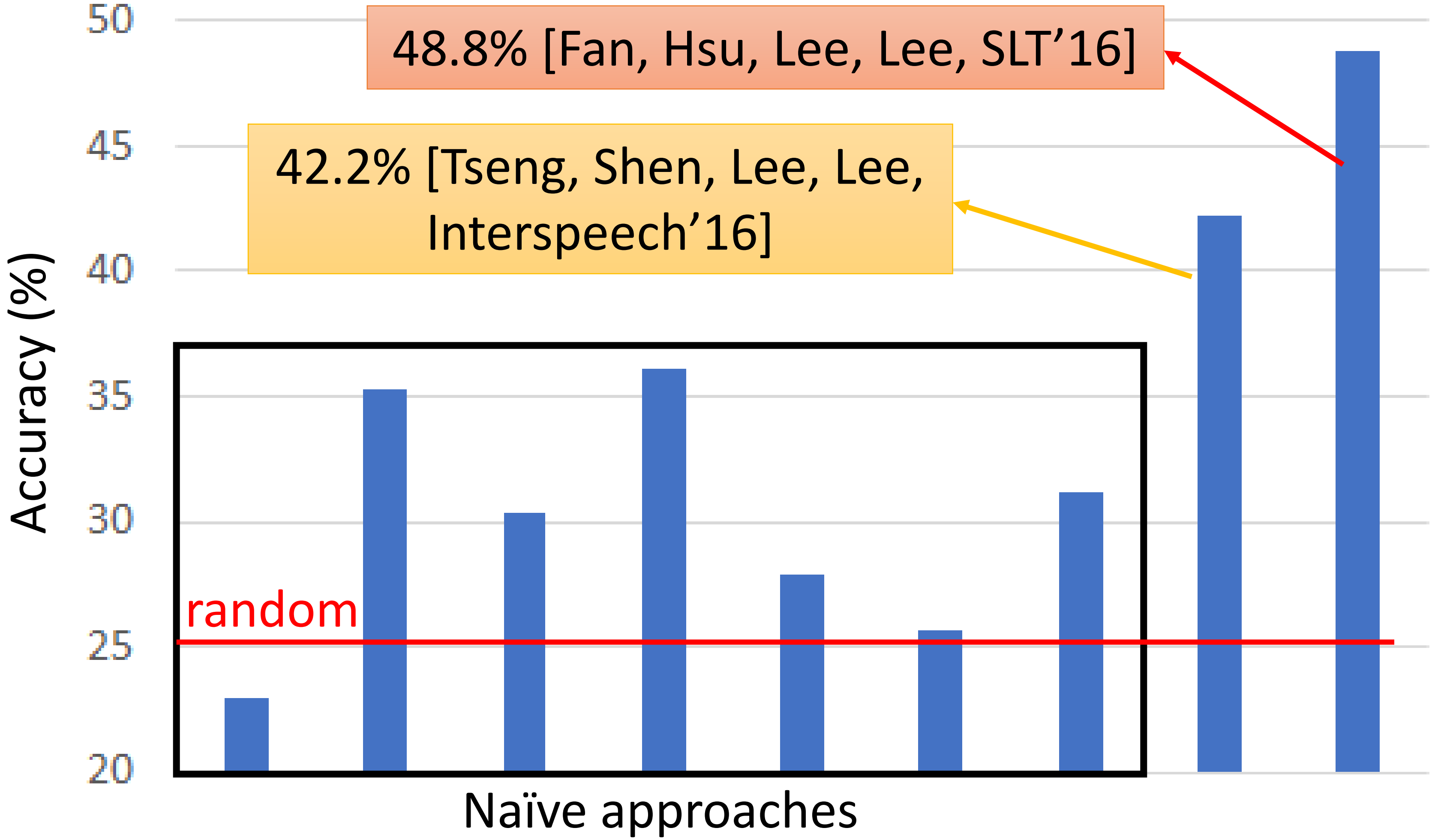
Experimental Results

Example Naïve approach:

1. Find the paragraph containing most key terms in the question.
2. Select the choice containing most key terms in that paragraph.



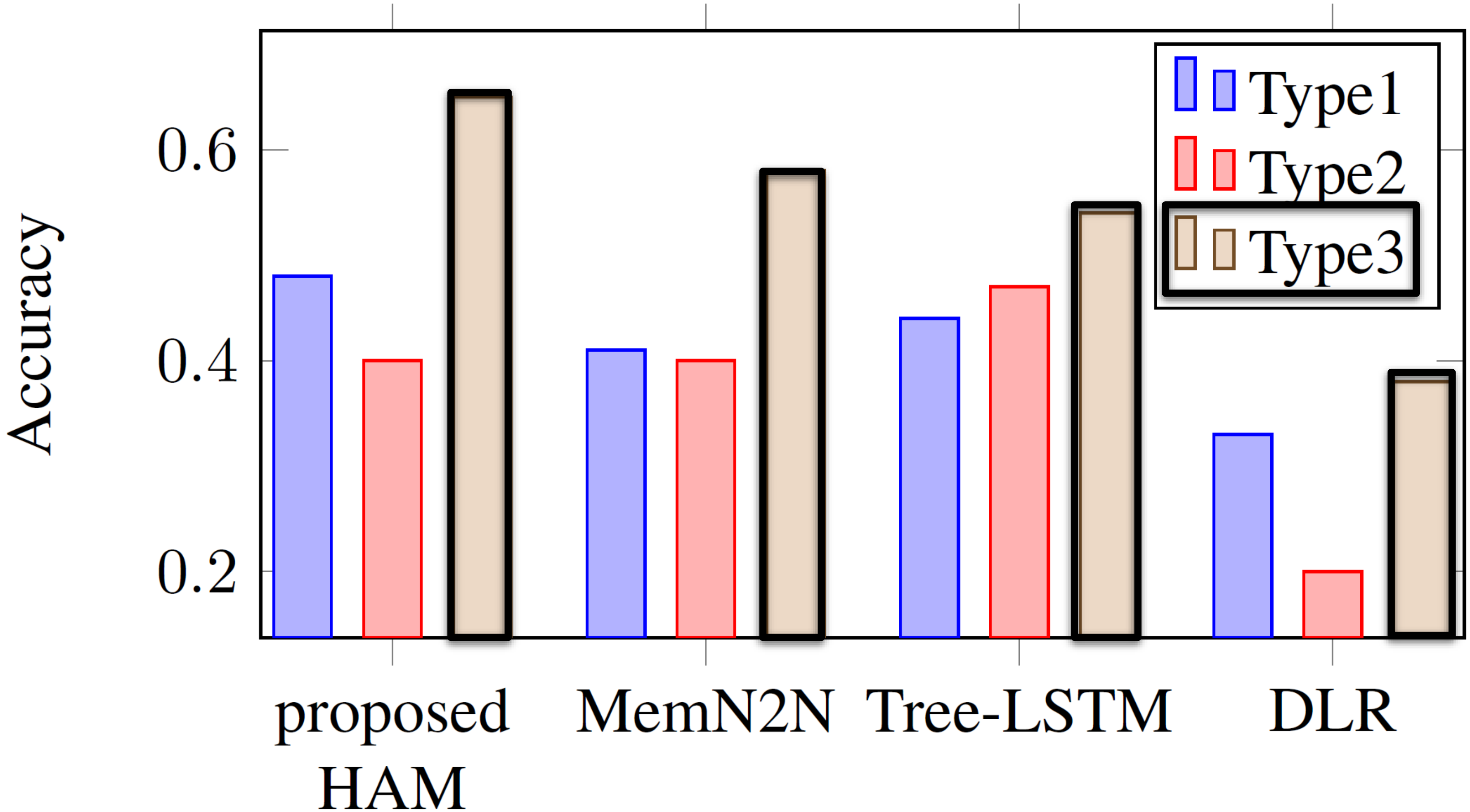
Experimental Results



Analysis

Type 3: Connecting Information
- ➢ Understanding Organization
- ➢ Connecting Content
- ➢ Making Inferences

- There are three types of questions

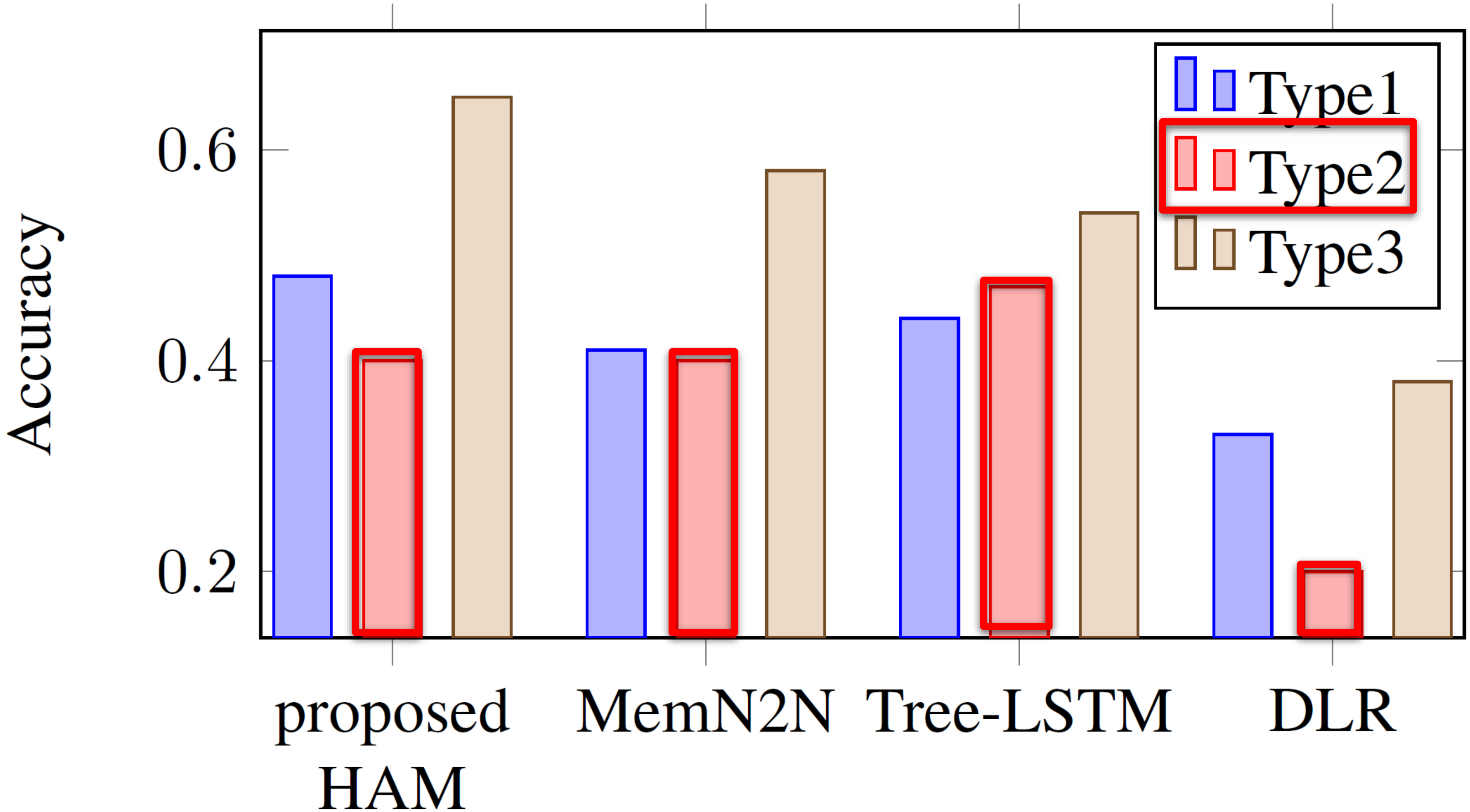


Analysis

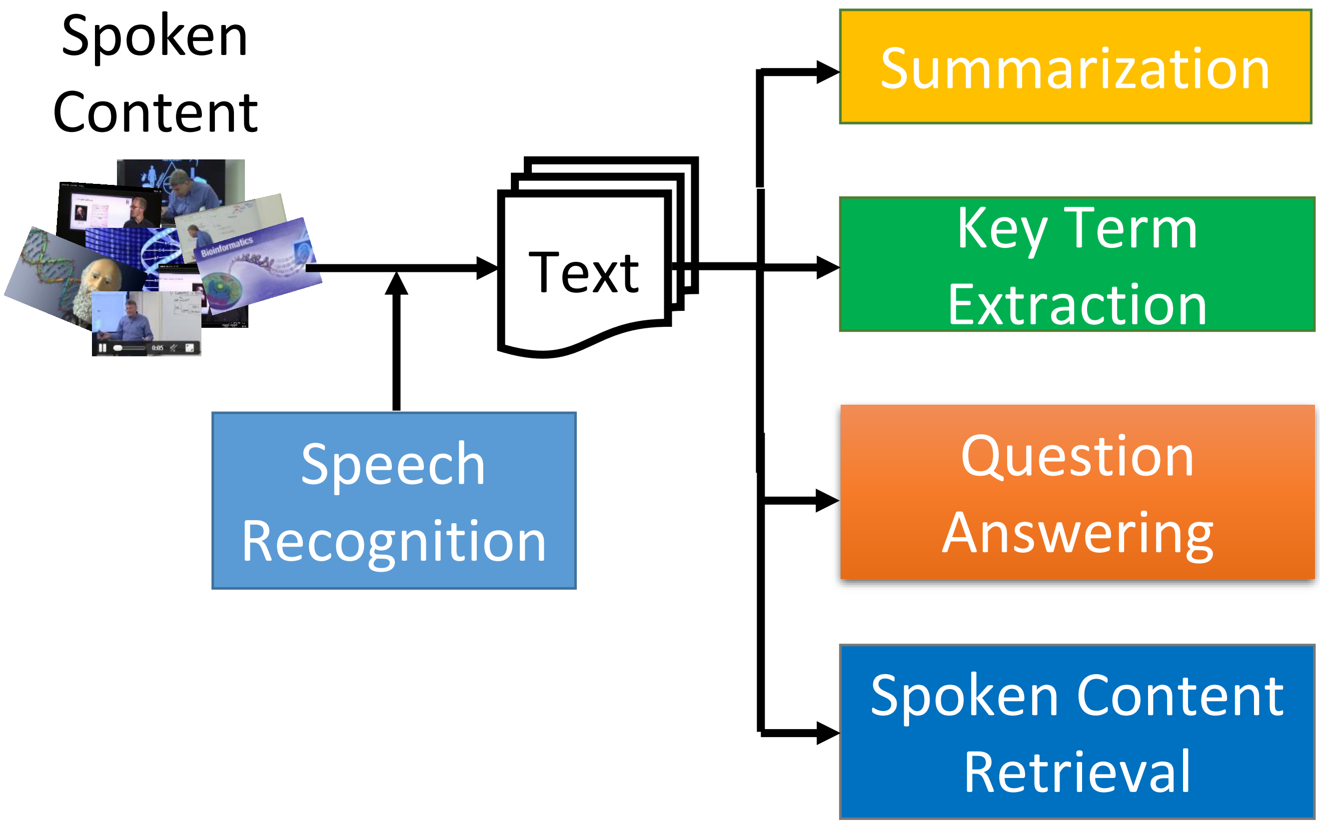
Type 3: Pragmatic Understanding

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

- There are three types of questions



Spoken Content Retrieval



Spoken Content Retrieval

- 3 hours tutorial at INTERSPEECH 2016 (with Prof. Lin-shan Lee)
  - Slide: http://speech.ee.ntu.edu.tw/~tlkagk/slide/spoken_content_retrieval_IS16.pdf
- Overview paper
  - Lin-shan Lee, James Glass, Hung-yi Lee, Chun-an Chan, "Spoken Content Retrieval —Beyond Cascading Speech Recognition with Text Retrieval," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.23, no.9, pp.1389-1420, Sept. 2015
  - http://speech.ee.ntu.edu.tw/~tlkagk/paper/Overview.pdf

One Slide Summarization

Spoken Content Retrieval

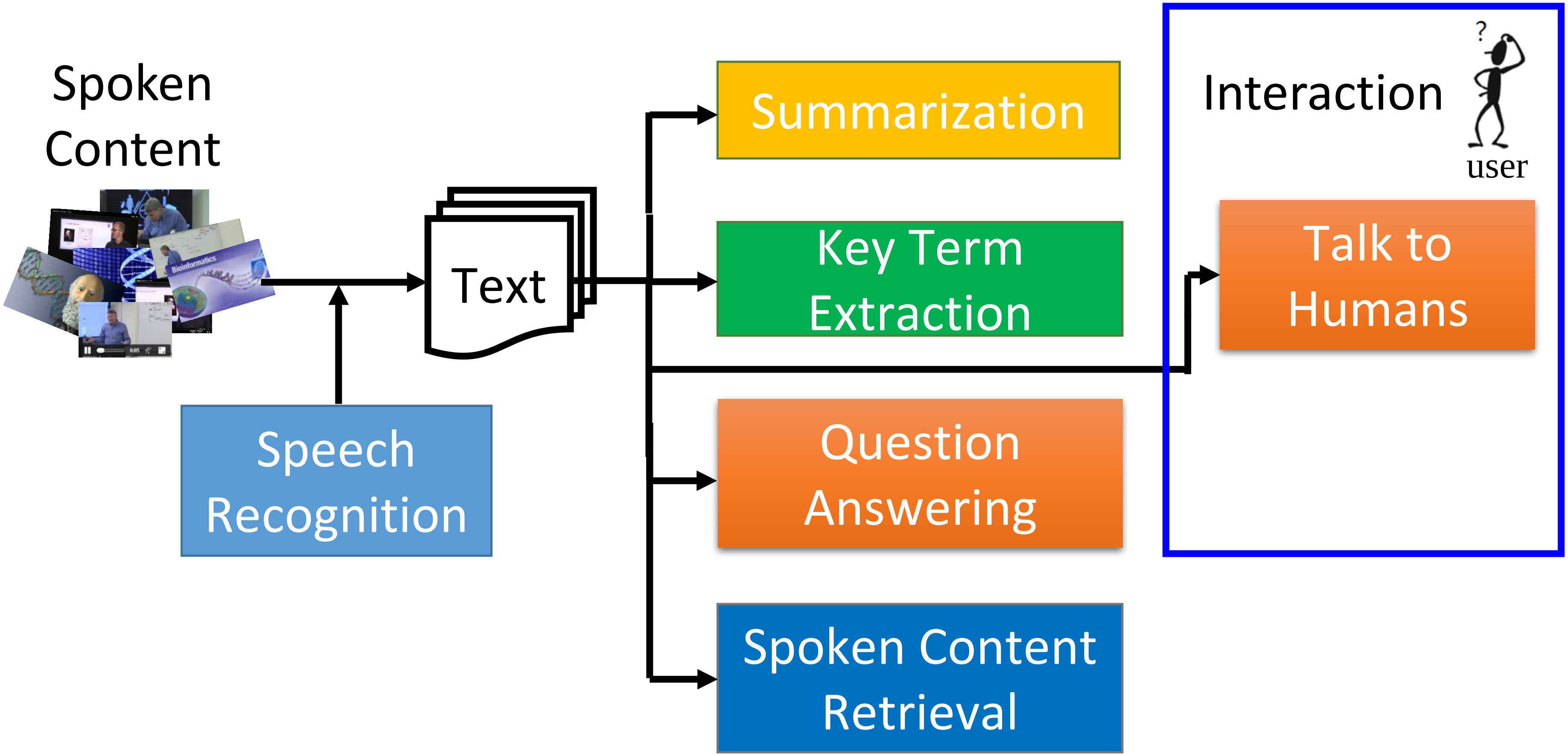
~~=~~

Speech Recognition

+

Text Retrieval

Talk to Humans

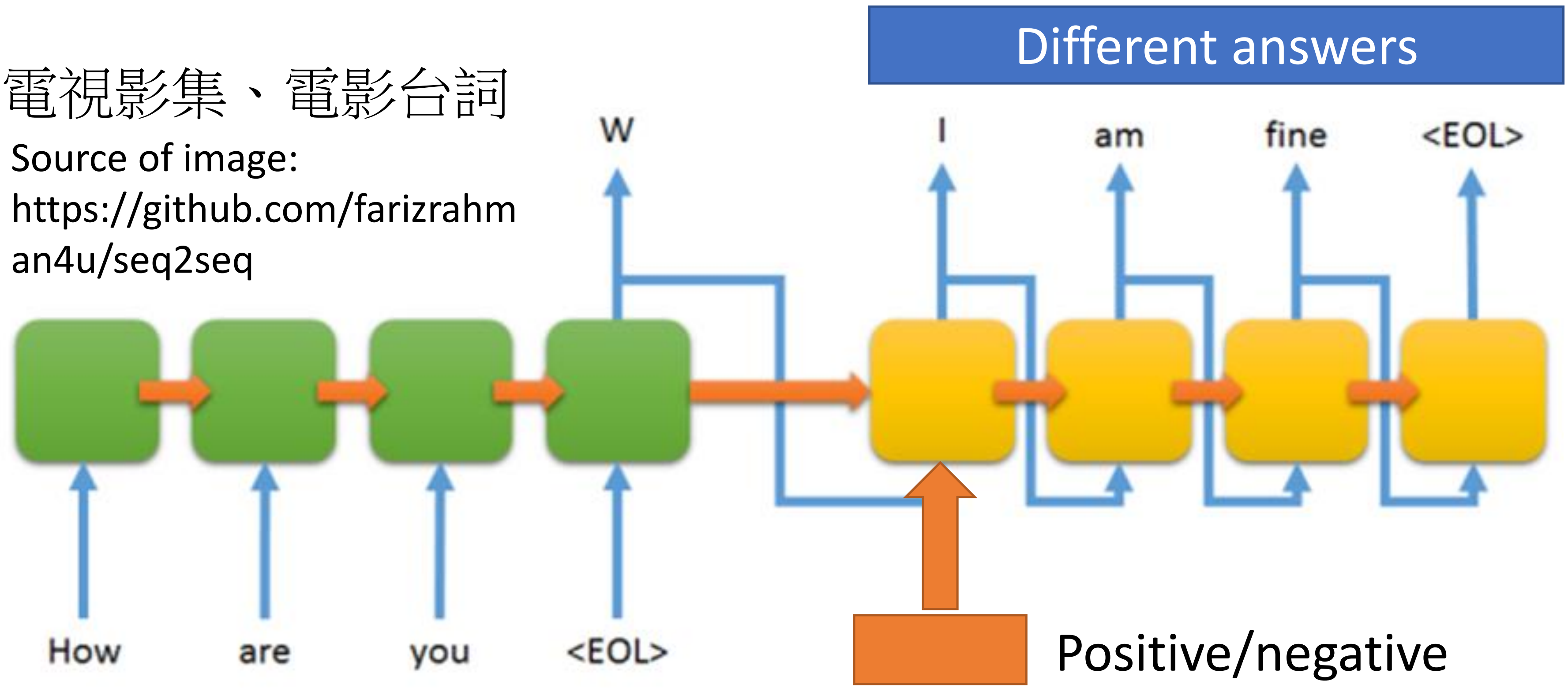


Chat-bot

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

Chat-bot with GAN

Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, Dan Jurafsky, "Adversarial Learning for Neural Dialogue Generation", arXiv preprint, 2017

Input sentence/history h → En De Chatbot → response sentence x

Input sentence/history h
response sentence x
→ Discriminator → Real or fake

human dialogues

Conditional GAN

Ref:一日搞懂 GAN https://www.slideshare.net/tw_dsconf/ss-78795326

感謝 段逸林 同學提供實驗結果

Example Results

input	I love you.

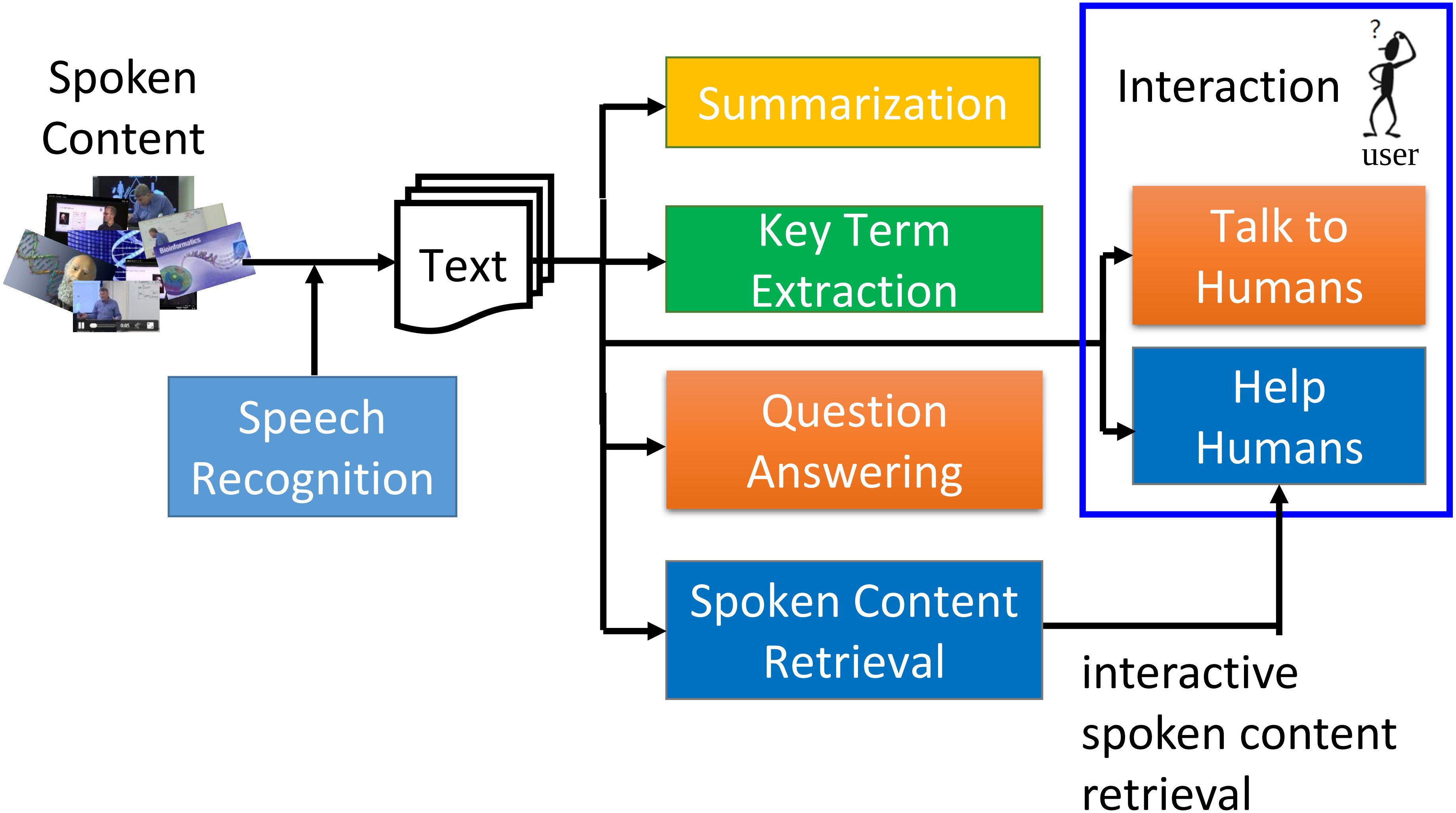
input	Do you like machine learning?

input	I thought I have met you before.

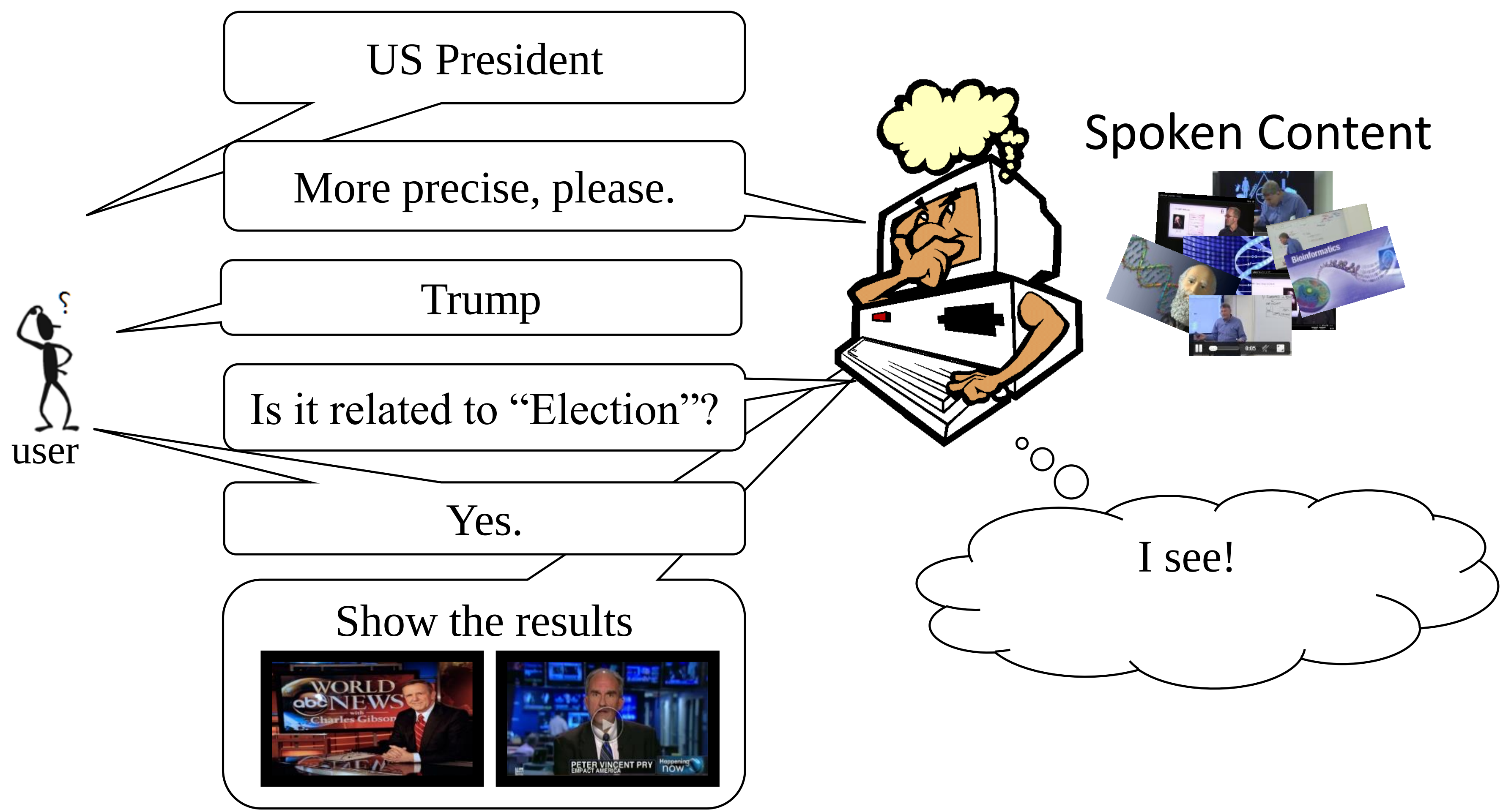
input	Let's go to the party.

input	How do you feel about the president?

Talk to Humans

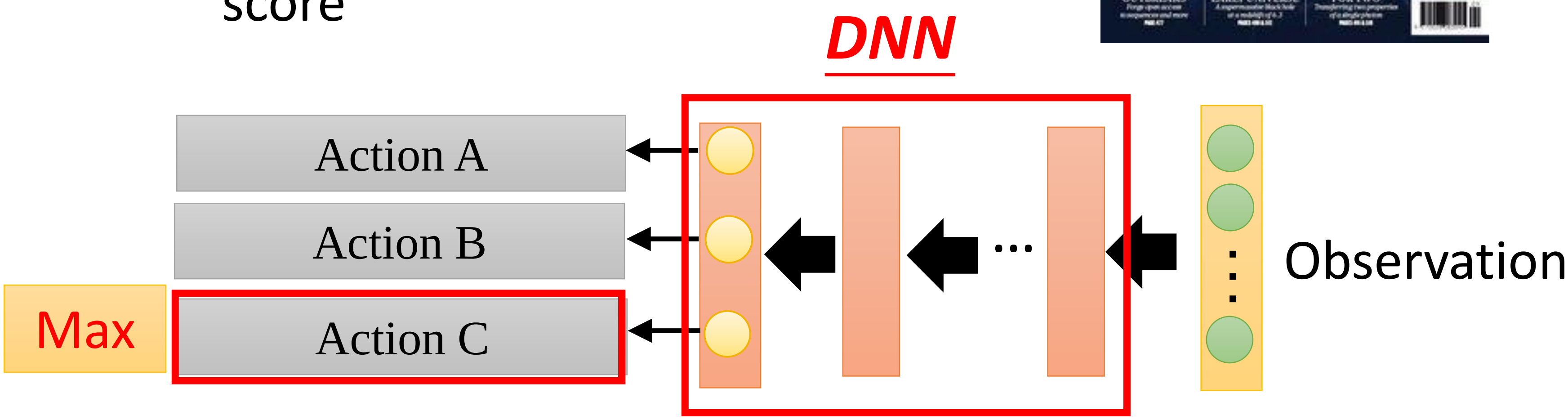


Scenario of Interactive Retrieval



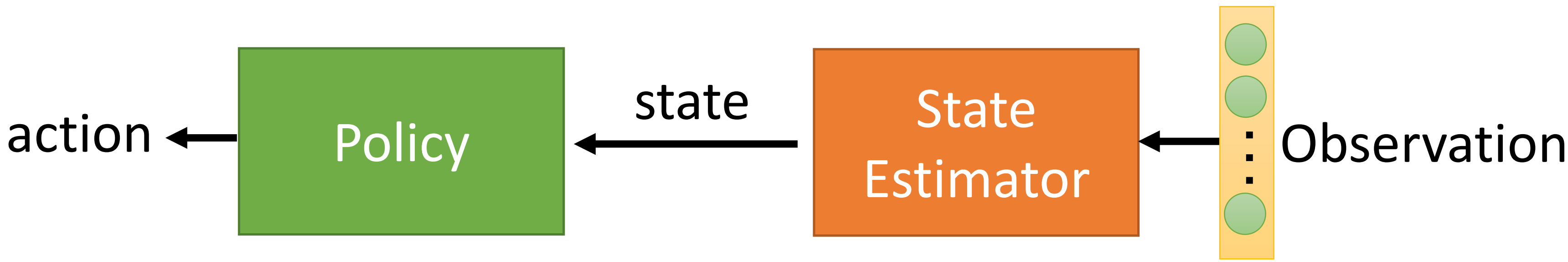
Deep Reinforcement Learning

- The actions are determined by a neural network
  - Input: information to help to make the decision
  - Output: which action should be taken
    - Taking the action with the highest score



Deep Reinforcement Learning v.s. Previous Work

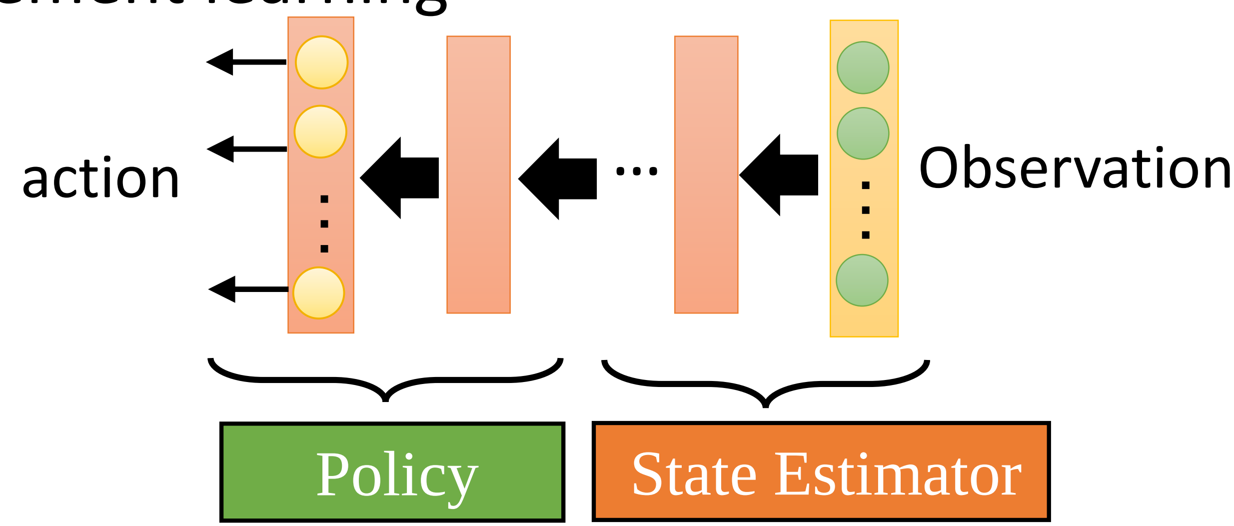
- Previous work [Wen & Lee, Interspeech 12][Wen & Lee, ICASSP 13]



- Deep reinforcement learning

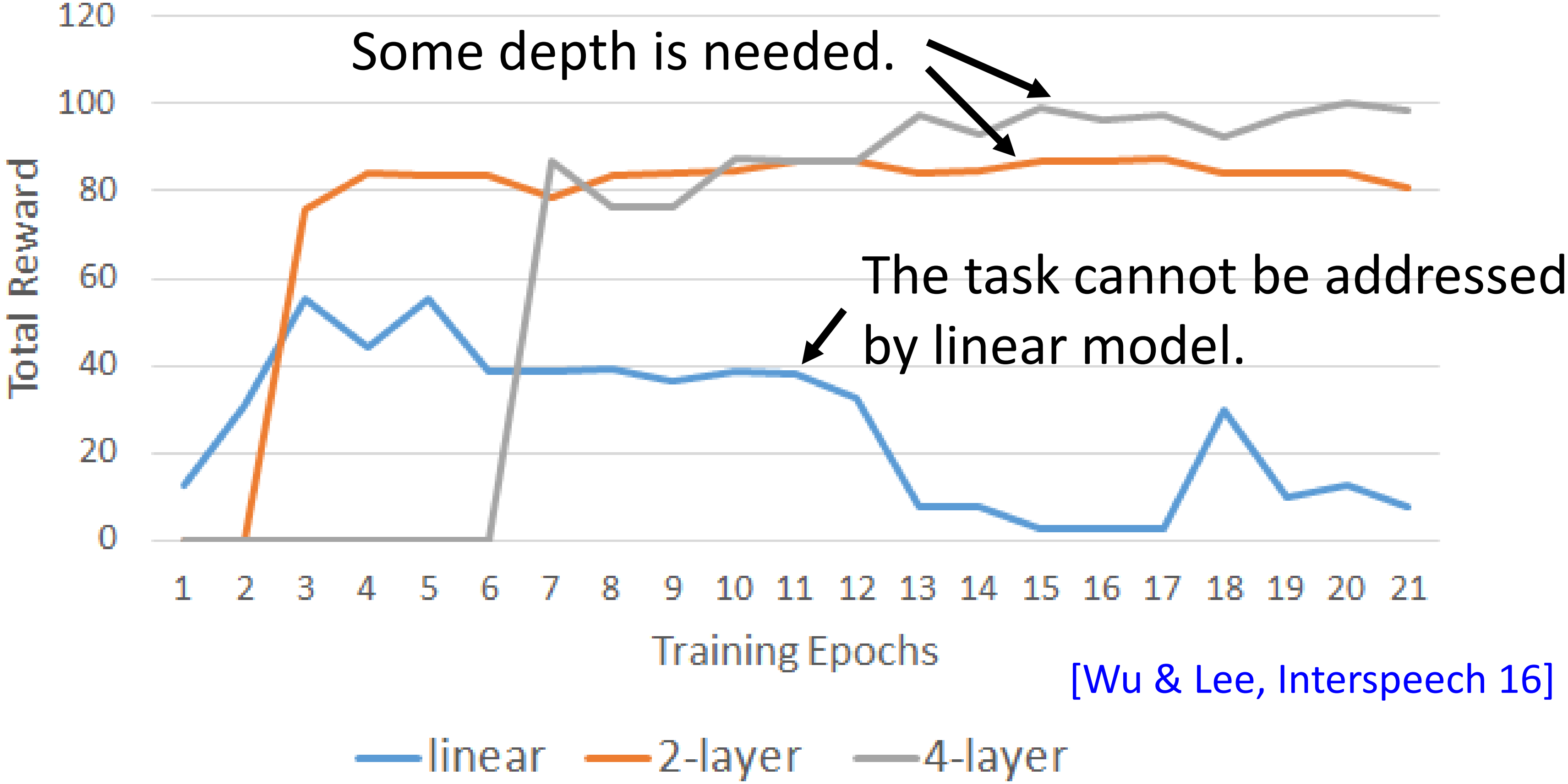
End-to-end learning

No hand-crafted states



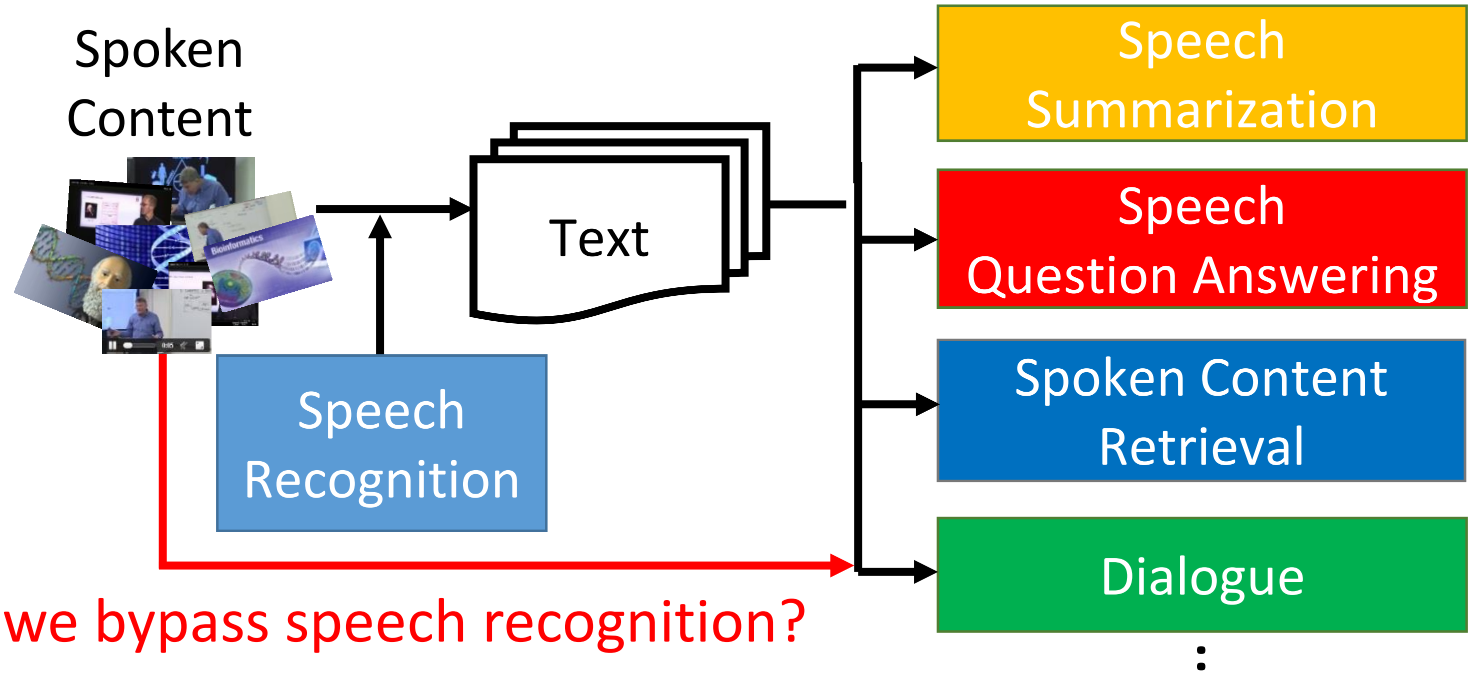
Experimental Results

- Different network depth, raw features



[Wu & Lee, Interspeech 16]

Audio Word to Vector



Can we bypass speech recognition?

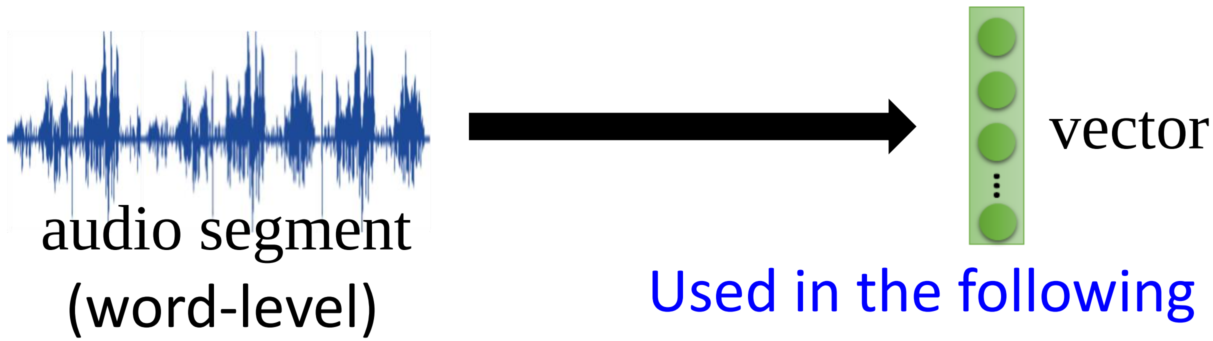
Why? Need the manual transcriptions of lots of audio to learn.

Most languages have little transcribed data.

New Research Direction: **Audio Word to Vector**

Audio Word to Vector

- Machine represents each audio segment also by a vector



Used in the following spoken language understanding applications

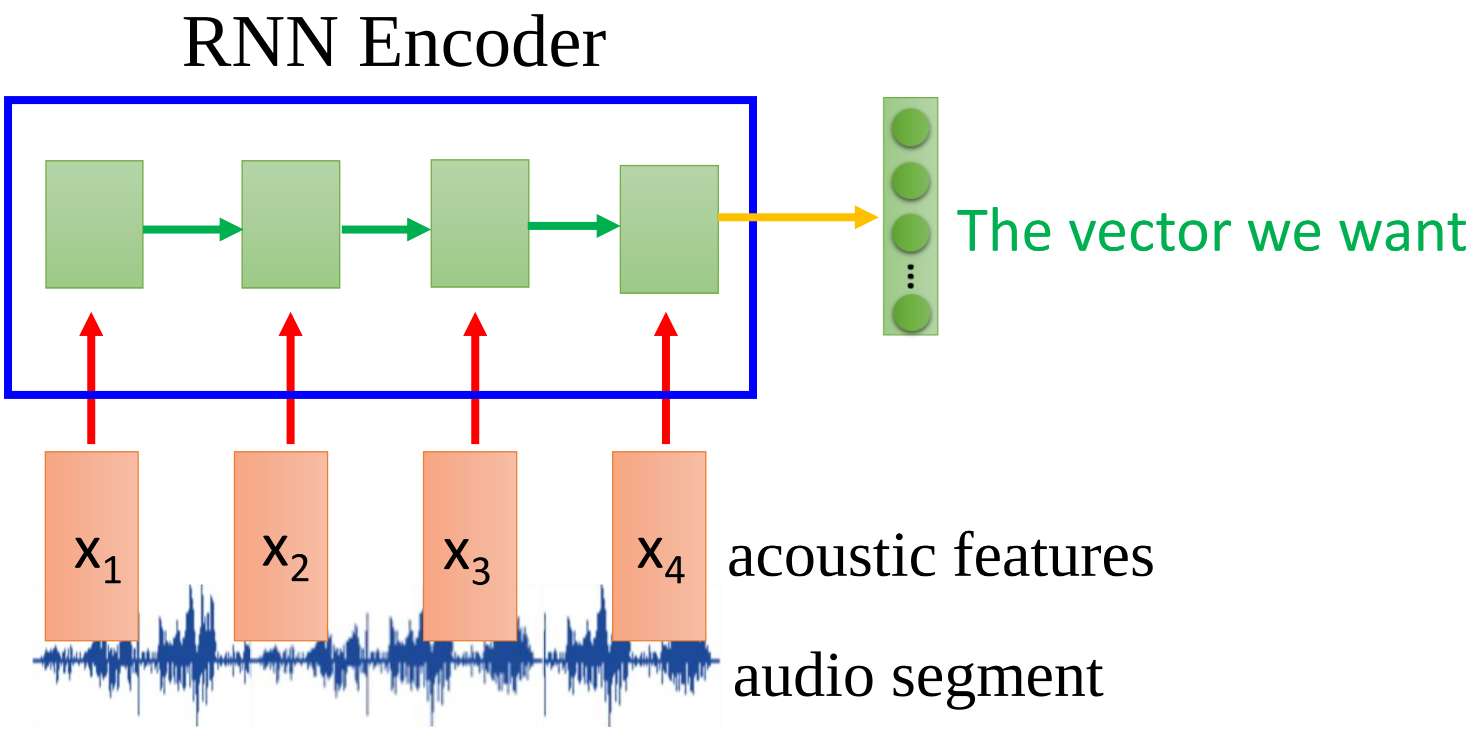
Learn from lots of audio without supervision

[Chung, Wu, Lee, Lee, Interspeech 16)

Sequence-to-sequence Auto-encoder

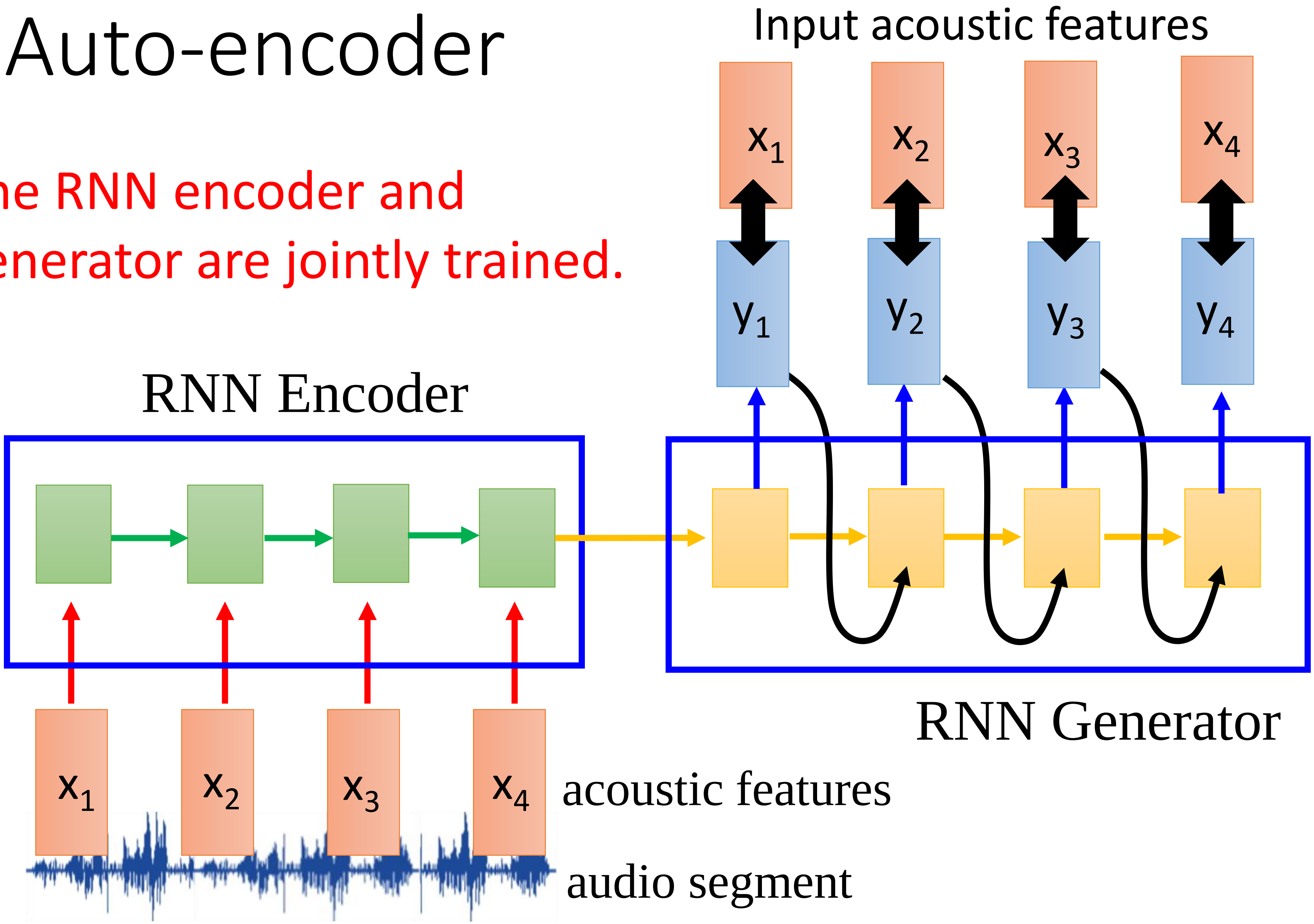
We use ***<u>sequence-to-sequence auto-encoder</u>*** here

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:

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

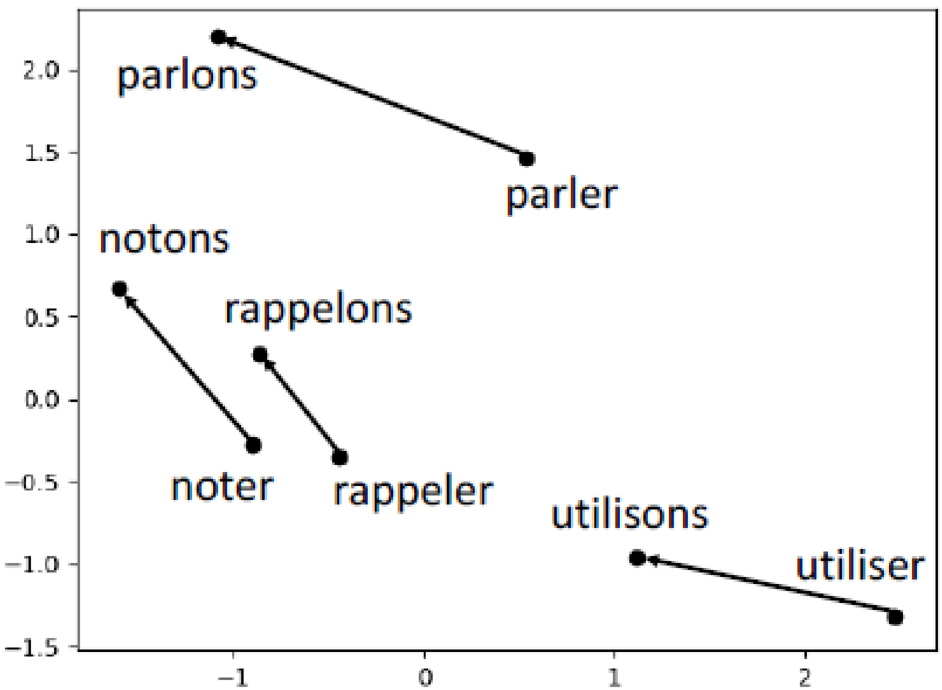
- Audio word to vector (phonetic information)

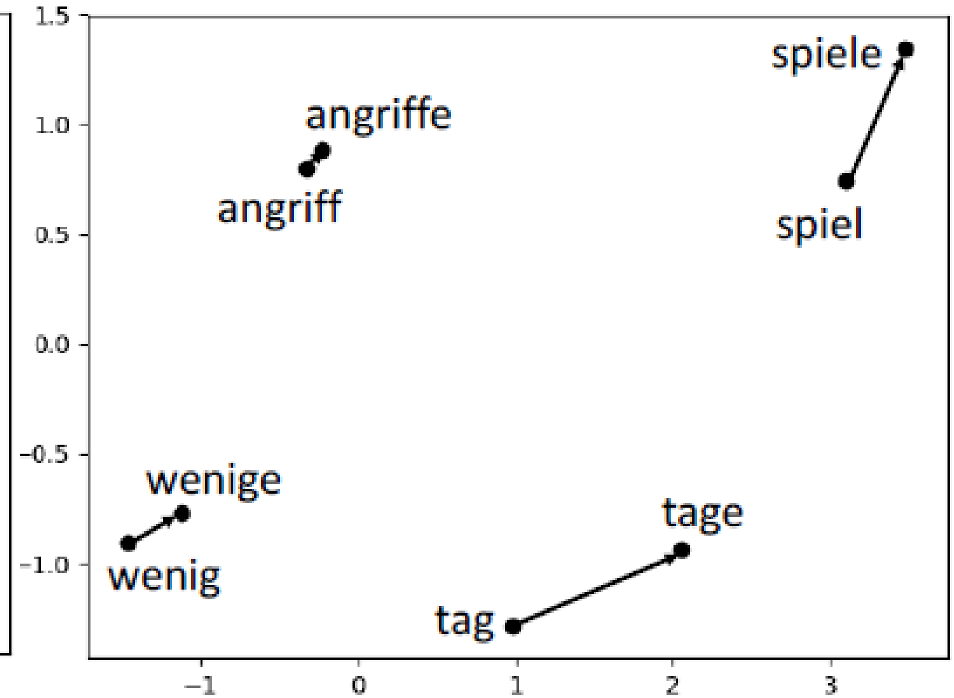
V( ) - V( ) + V( ) = V( )

GIRL PEARL PEARLS GIRLS

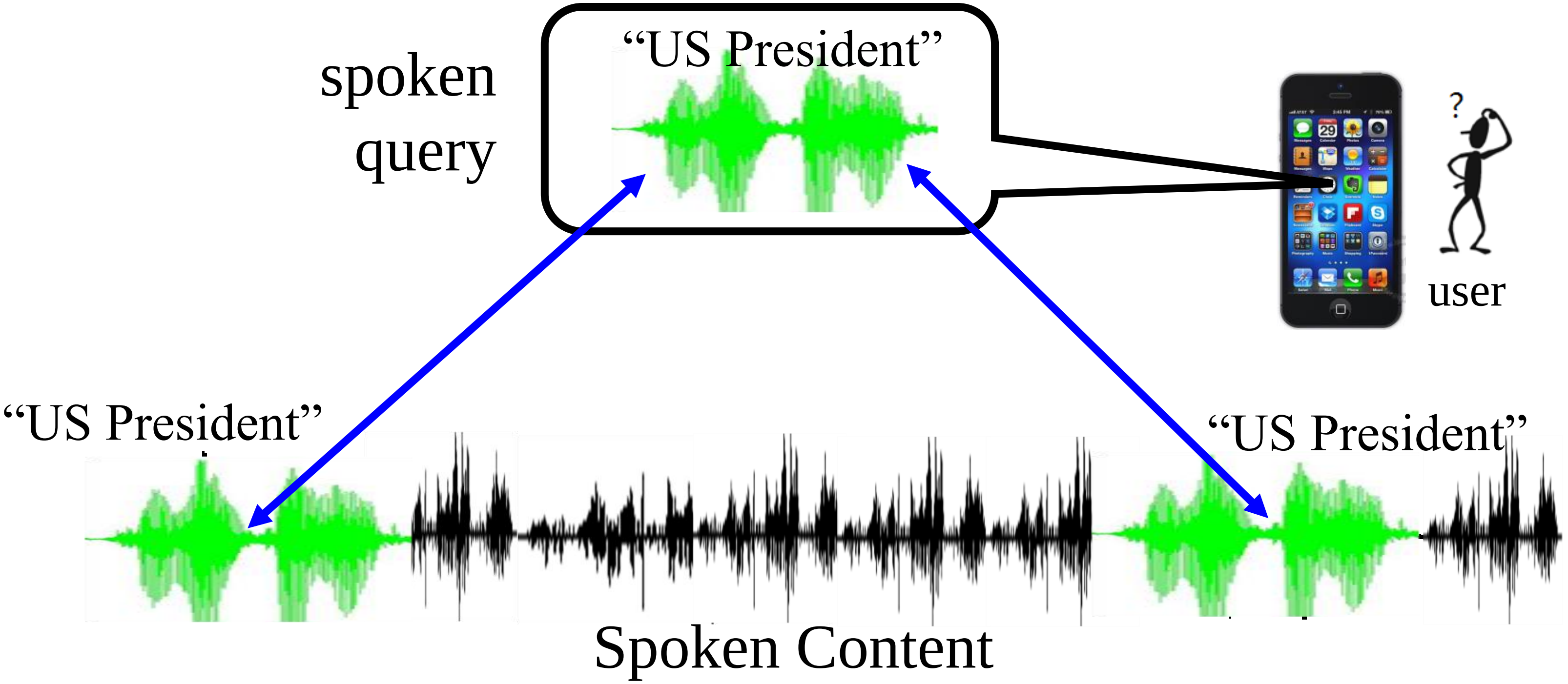
V( ) - V( ) + V( ) = V( )

New Languages





Audio Word to Vector – Application

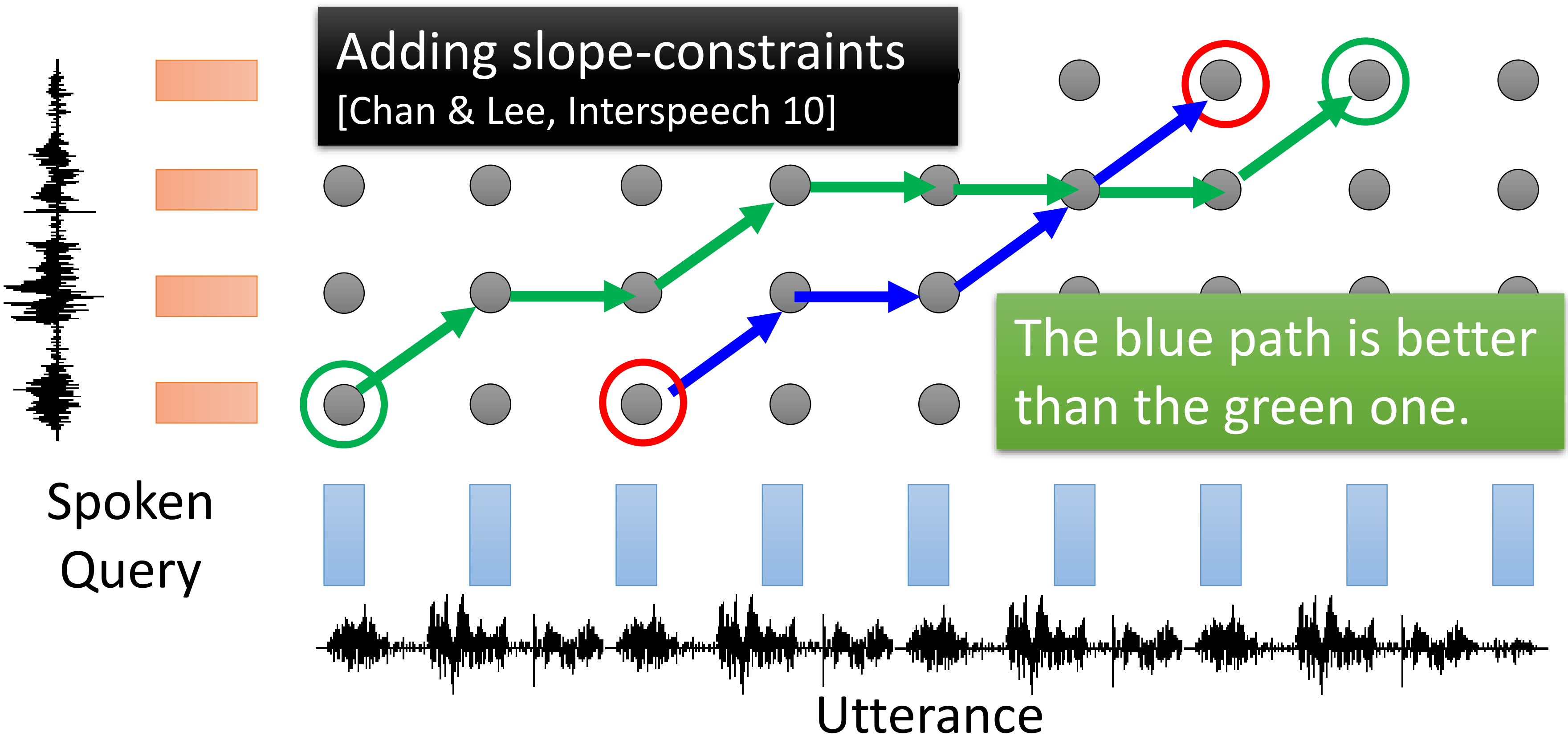


Compute similarity between spoken queries and audio files on acoustic level, and find the query term

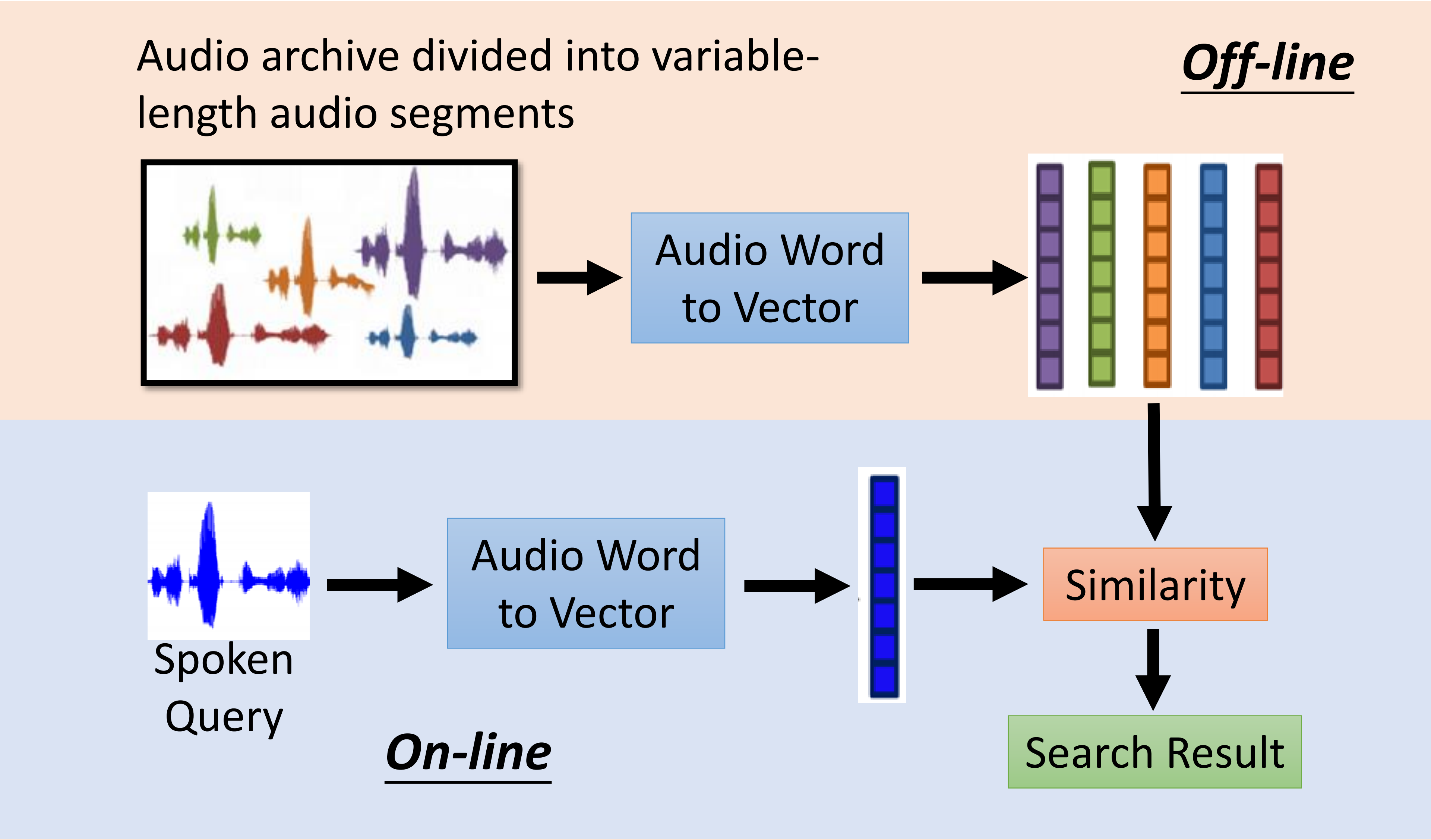
Audio Word to Vector – Application

- DTW for query-by-example

Segmental DTW [Zhang, ICASSP 10], Subsequence DTW [Anguera, ICME 13][Calvo, MediaEval 14]

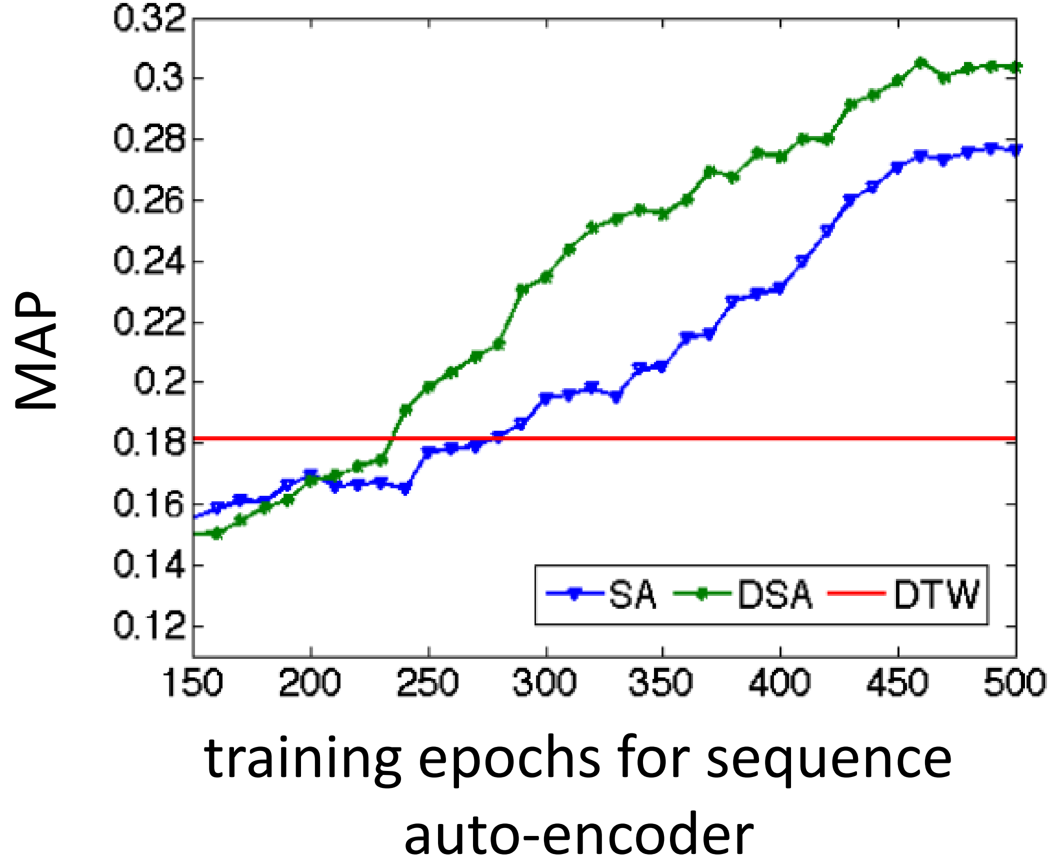


Audio Word to Vector – Application



Audio Word to Vector –Application

- Query-by-Example Spoken Term Detection



SA: sequence auto-encoder

DSA: de-noising sequence auto-encoder

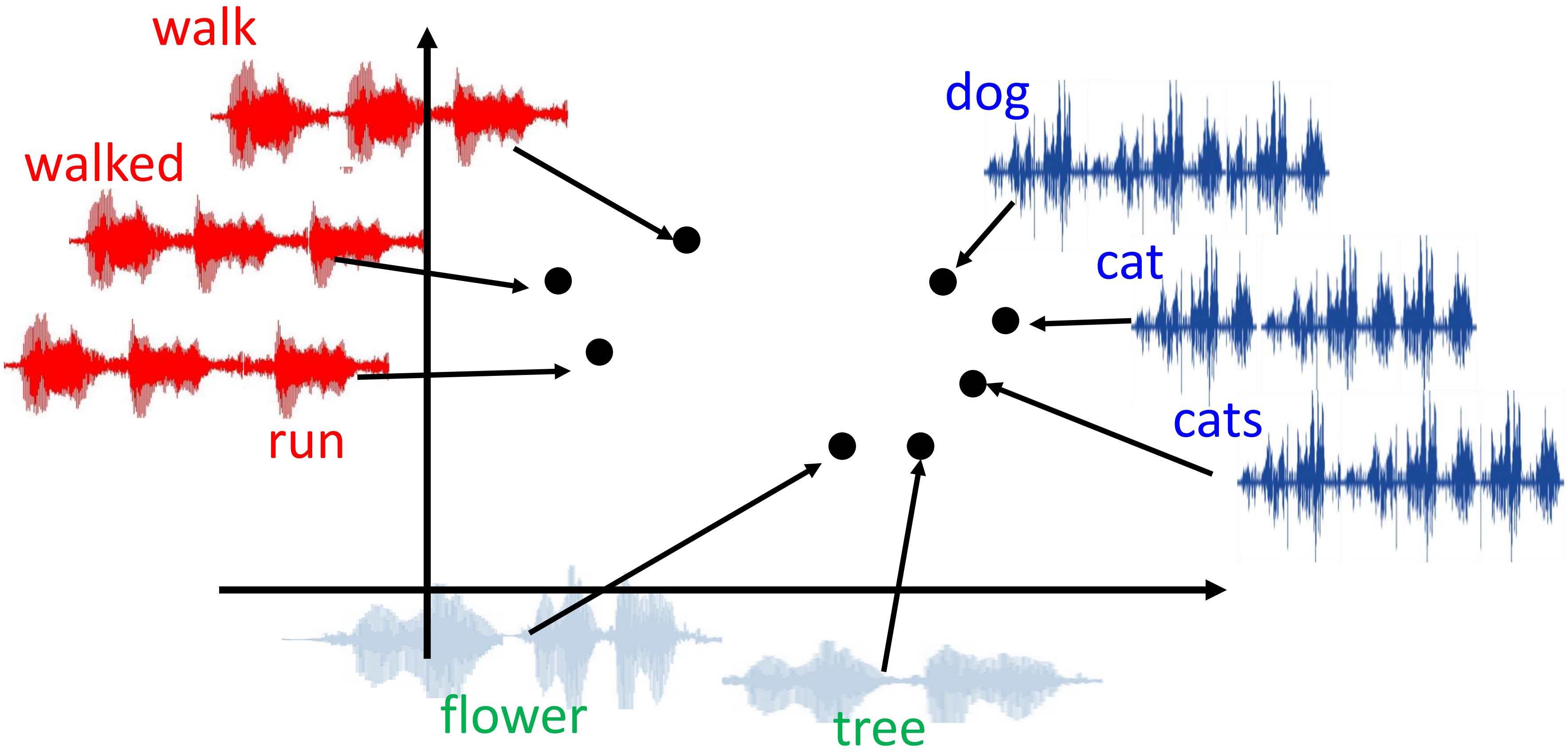
Input: clean speech + noise

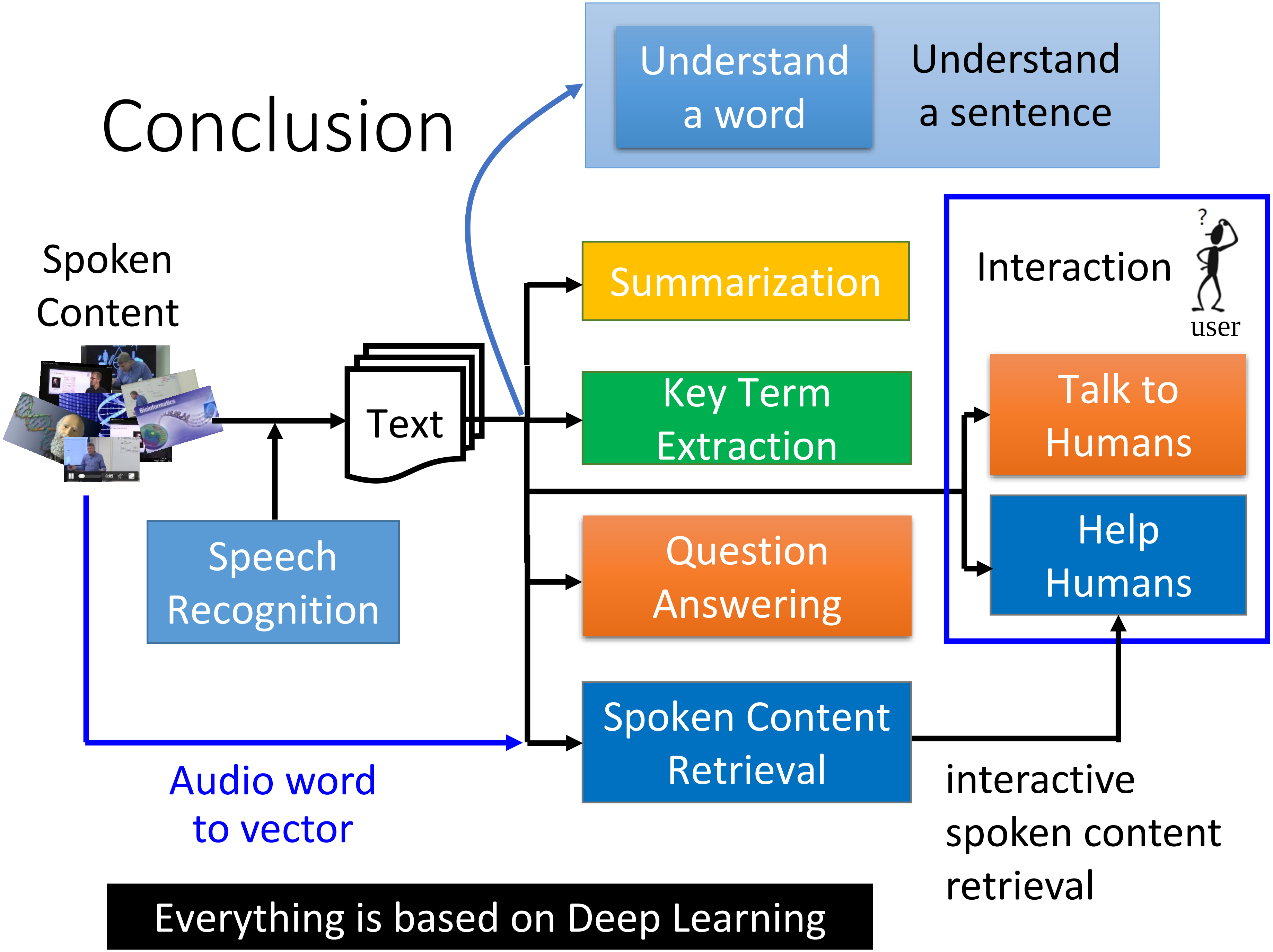
output: clean speech

Next Step ……

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

• Audio word to vector with semantics




Conclusion
Understand a word
Understand a sentence
Spoken Content
Bioinformatics
Text
Speech Recognition
Summarization
Key Term Extraction
Question Answering
Spoken Content Retrieval
Interaction
?
user
Talk to Humans
Help Humans
Audio word to vector
interactive spoken content retrieval
Everything is based on Deep Learning