Evaluation

May 23rd, 2017

# Intelligent Conversational Bot

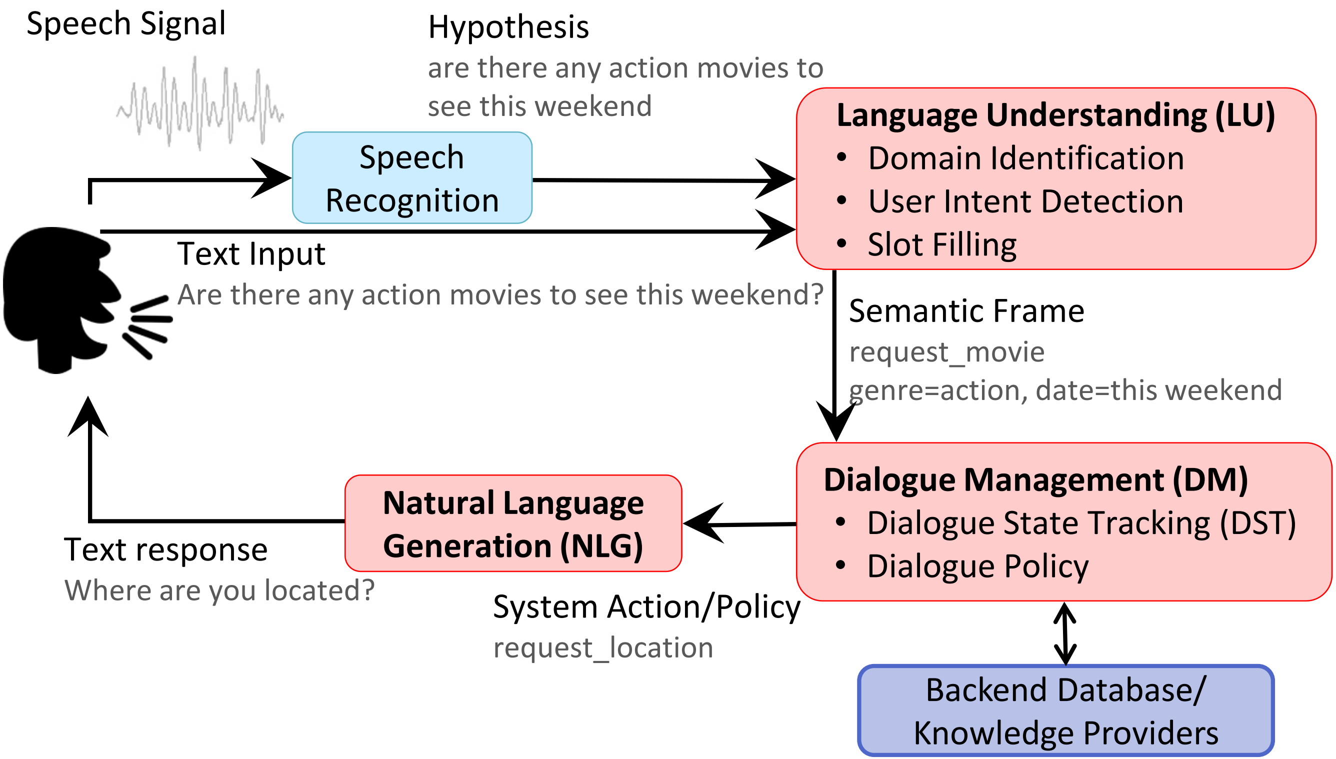
YUN-NUNG (VIVIAN) CHEN

WWW.CSIE.NTU.EDU.TW/~YVCHEN/S105-ICB

國立臺灣大學
National Taiwan University

# Task-Oriented Dialogue System (Young, 2000)

http://rsta.royalsocietypublishing.org/content/358/1769/1389.short



Speech Signal

Hypothesis

are there any action movies to see this weekend

Speech Recognition

Text Input

Are there any action movies to see this weekend?

**Language Understanding (LU)**

- Domain Identification
- User Intent Detection
- Slot Filling

Semantic Frame

request_movie

genre=action, date=this weekend

**Dialogue Management (DM)**

- Dialogue State Tracking (DST)
- Dialogue Policy

Backend Database/ Knowledge Providers

System Action/Policy

request_location

**Natural Language Generation (NLG)**

Text response

Where are you located?

# Speech Recognition / Multimodality

- Speech recognition
  - Word error rate $$WER = \frac{S + D + I}{N}$$ #words in the reference
  - Word accuracy $$WACC = 1 - WER$$

Hyp: A A B D C K

Ref: A C D A C

$$WER = \frac{1+1+2}{5} = 80\%$$

$$WACC = 1 - 80\% = 20\%$$

- Emotion recognition
  - Accuracy

# Language Understanding Evaluation

- Data
  - Training and testing should be *split*
    - Testing data should be real data collected from human to make evaluation results convincing
- Metrics
  - Sub-sentence-level: intent accuracy, slot F1
  - Sentence-level: whole frame accuracy

# Dialogue State Tracking Evaluation

- Metric
  - Tracked state accuracy with respect to user goal
  - Recall/Precision/F-measure individual slots

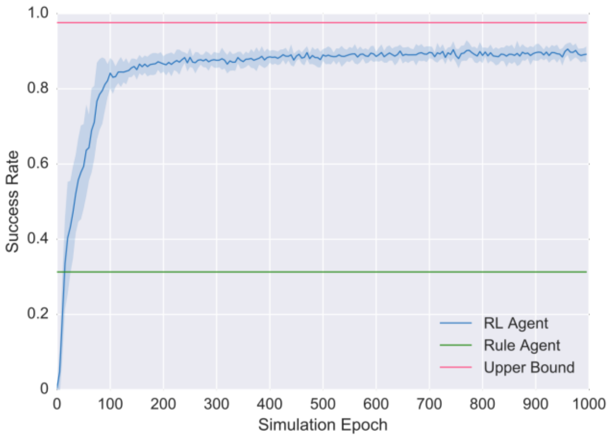
# Dialogue Policy Evaluation

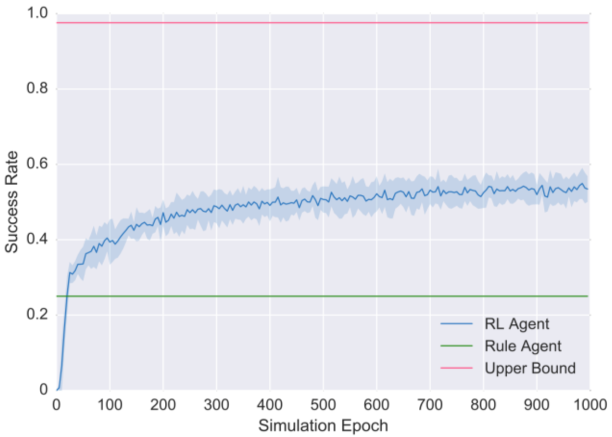
- Metrics
  - Turn-level evaluation: system action accuracy
  - Dialogue-level evaluation: task success rate, reward, #dialogue turn

# Reinforcement Learning Policy

- □ Frame-level semantics
- □ Natural language

Note: check whether the interactions can be satisfied by the system’s functionality





If your RL agent cannot outperform the rule-based agent, please consider to increase the complexity of system functionality and the simulated user.

X. Li, Y.-N. Chen, L. Li, and J. Gao, “End-to-End Task-Completion Neural Dialogue Systems,” preprint arXiv: 1703.01008, 2017.

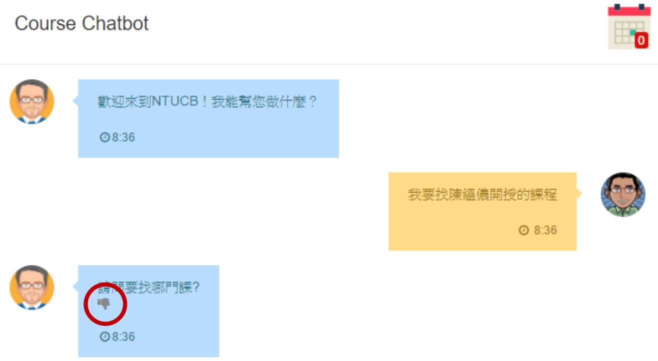
# Natural Language Generation Evaluation

- Metrics
  - Subjective: human judgement (Stent et al., 2005)
    - Adequacy: correct meaning
    - Fluency: linguistic fluency
    - Readability: fluency in the dialogue context
    - Variation: multiple realizations for the same concept
  - Objective: automatic metrics
    - Word overlap: BLEU (Papineni et al, 2002), METEOR, ROUGE
    - Word embedding based: vector extrema, greedy matching, embedding average

There is a gap between human perception and automatic metrics

# User Study

- System performance from real users
  1) Allow others to interact with the system
  2) Record the dialogues and compute the success rate, satisfaction degree
  3) Analyze where the errors come from



Course Chatbot

歡迎來到NTUCB！我能幫您做什麼？

8:36

我要找陳縕儂開授的課程

8:36

請問要找哪門課?

8:36

# Concluding Remarks

- Evaluate all components of the system in detail
  - Speech recognition: word accuracy
  - Language understanding: frame accuracy
  - Dialogue state tracking: frame accuracy
  - Dialogue policy: success rate
  - Natural language generation: BLEU
- User study
  - Subjective: satisfaction
  - Objective: success rate