Evaluation

Intelligent Conversational Bot

YUN-NUNG (VIVIAN) CHEN

www.csie.ntu.edu.tw/~yvchen/S105-ICB

National Taiwan University
Task-Oriented Dialogue System (Young, 2000)

Speech Signal

Hypothesis
are there any action movies to see this weekend

Text Input
Are there any action movies to see this weekend?

Speech Recognition

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

System Action/Policy
request_location

Text response
Where are you located?

Natural Language Generation (NLG)

Backend Database/
Knowledge Providers

http://rsta.royalsocietypublishing.org/content/358/1769/1389.short
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
If your RL agent cannot outperform the rule-based agent, please consider to increase the complexity of **system functionality** and the **simulated user**.
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
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