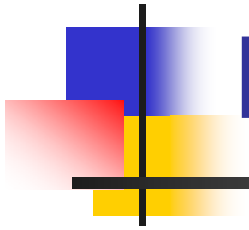
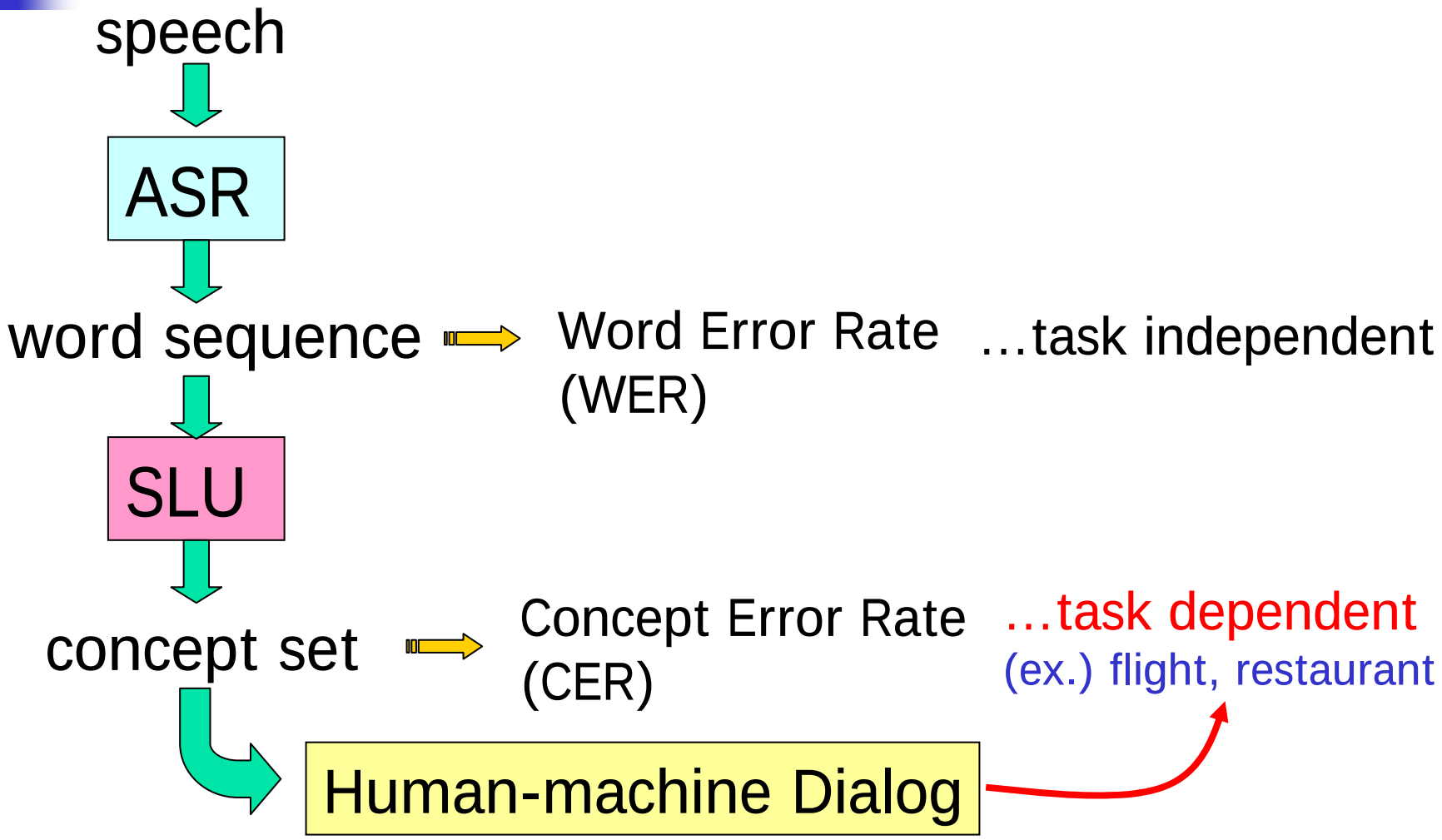


New Perspectives on Spoken Language Understanding (SLU)

Tatsuya Kawahara

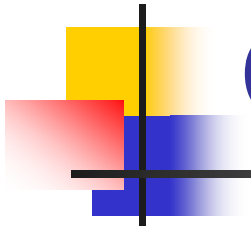
(Kyoto Univ., Japan)

Traditional View of SLU



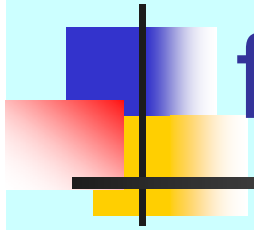
New Perspectives of SLU

- Target of SDS (Spoken Dialog System)
  - simple database query
    - Relational-structured data (ex.) flight, train
  - general information retrieval (IR/search)
    - Semantic slots cannot be well defined!!
- Target of ASR
  - Human-machine interface
  - Human-human communication
    - SLU is not limited to concept extraction



Overview of the Talk

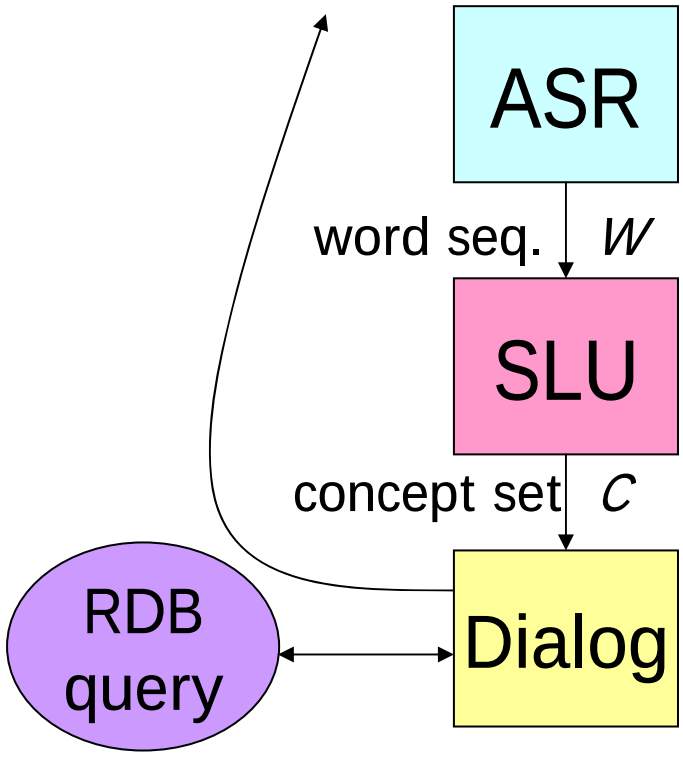
- SLU for new-generation SDS
  - IR (Information Retrieval)
  - QA (Question-Answering)
- SLU for human-human speech communication
  - Rich transcription
  - Hot-spot detection



SLU with IR for Human-Machine Dialog

- Conventional SLU
- From RDB to IR
- Interactive IR
- Interaction using QA

Conventional SLU



Acoustic Model
Language Model

Concept Model := semantic slots

(ex.) [destination, “Milan”]
[arrival_date “12_12_2009”]

Dialog Management

- Not filled → ask question
- Not confident → confirmation

Evaluation: Concept Error Rate (CER)

Typical Formulation of Conventional SLU

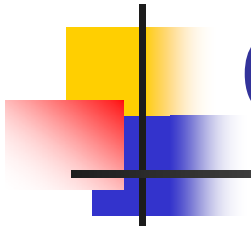
$$\arg\max\, p(C,W \mid X)$$

$$= \arg\max\, p(C) \cdot p(W \mid C) \cdot p(X \mid W)$$

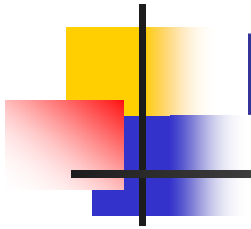
- P(C): statistical model of concepts
- P(W|C): language model dependent on concept
- P(X|W): acoustic model

See review by [DeMori:ASRU07] or [Wang:SPSmagazine05]

Assumptions of Conventional SLU



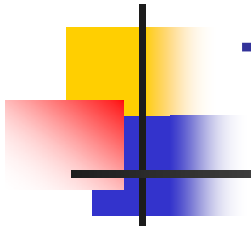
- Set of concepts definite, given task-domain
  - Otherwise, concept model & concept error rate cannot be defined
  - Consistent with back-end DB system
    - Semantic slots → SQL query



From RDB Query to IR

- Search target (in general)
  - structural RDB → general text
- Task of SDS
  - SQL (slot filling) → IR (search)
- Approach
  - symbolic → statistical (Vector Space Model)
  - Cannot assume definite semantic slots & dialog states based on the slots

From RDB query To Text Search (IR)



Backend	ASR	SLU	Dialog
Relational Database (RDB) (flight, bus..)	Finite State Grammar	Mapping to SQL	state-action pair (voice XML)
Text base (KB) (Wikipedia, Web…)	SLM (N-gram)	Statistical Matching (VSM)	dynamic clarification & recommendation

Vector Space Model

SLU for IR

- When IR assumes some structure

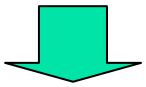
  (ex.) directory search

  → noisy-channel model [Wang:SPSmagazine08]

- When search space is too large

  (ex.) Web, newspaper

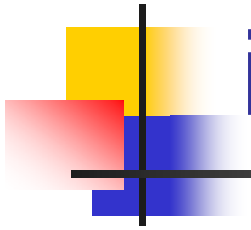
  → little room for SLU in addition to VSM



- IR from documents in restricted domain

  (ex.) software manual, cooking recipe, tourist guide

IR from Knowledge Base (KB) in restricted domain

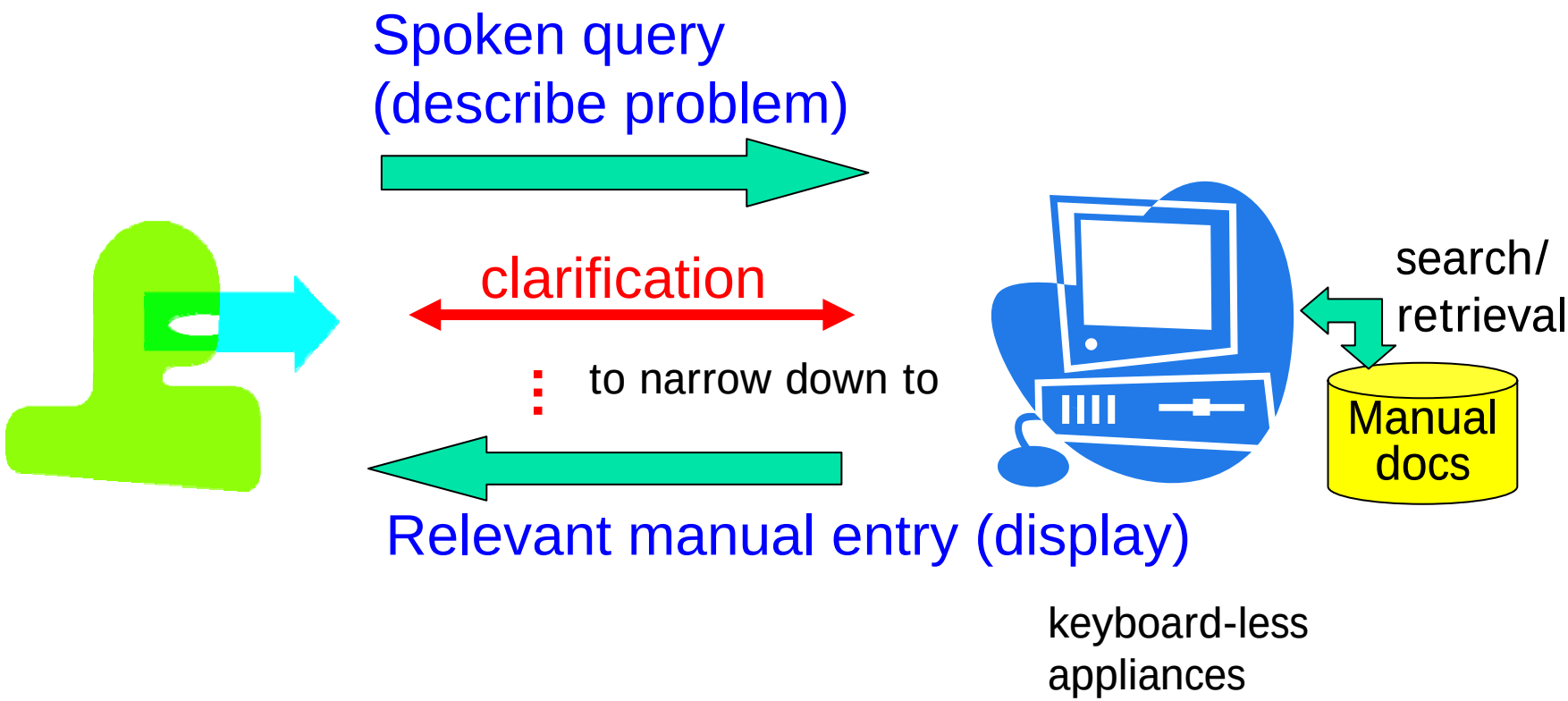


- Knowledge Base (KB) → Problem solving
  - software manuals → trouble shooting
  - cooking recipes → cooking assistant
  - tourist guide → tour planning
- Longer interaction than simple search

  → novel features in SLU and dialog
- Access to entire document set

  → exploit knowledge for SLU and dialog

Interactive IR from Software Support Manuals [Misu:SPECOM06]



Example Dialog by Interactive Software Manual Retrieval

query update = SLU state

S1: What is your problem?

U1: I cannot open my Excel file.

(query: “I cannot open an Excel file") → 50 docs.

S3: What is the version of your Excel?

U3: It is version 2002.

(query: “I cannot open an Excel 2002 file") → 30 docs.

S4: When did this trouble occur?

U4: When I tried to open it with Explore.

(query: “I cannot open an Excel 2002 file when I try to open with Explore.") → **3 docs.**

S5: Here are matched documents possibly useful for this trouble.

Dialog generated on-the-fly without pre-defined flow!

On-the-fly Clarification

- Select from a set of questions to maximize Information Gain (IG)
  - Expected to eliminate matched docs

$$IG(S) = -\sum_{i=0}^{n} P(i) \cdot \log P(i)$$

$$P(i) = \frac{|C_i|}{\sum_{i=0}^{n} |C_i|}$$

$$|C_i| = \sum_{D_k \in i} CM(D_k)$$

Ci: number of docs classified to category ***i*** by the question ***S***
CM(D): matching score of doc ***D***

Automatic Acquisition of a set of Clarification Questions

1. Dependency structure analysis for modifier-head pairs in all documents
2. Calculate entropy for head words

Install:

Application program	→ 20%
Service Pack	→ 10%
Device driver	→ 10%
External device	→ 8%
Client program	→ 6%
. .	

Large entropy → effective
"what did you install?"

Shutdown:

System	→ 40%
Computer	→ 50%
Server	→ 5%
. .	
. .	

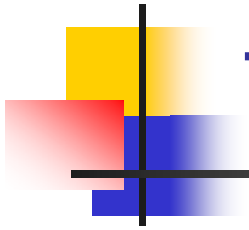
Small entropy → too obvious
"what did you shutdown?"

Evaluation

- Microsoft software manuals of 40K entries

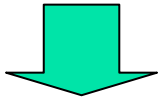
	success rate	#extra turns
baseline	70.7	
Modifier-head pair	74.5	0.38
Heuristic	74.5	0.97
Meta-data (application, version)	76.1	0.89
Combined	83.3	2.24

From Simple IR to Information Navigation



- IR: search unique answer documents

  (ex.) Manual entry for specific problem



- Speech interface without GUI (Agent/Robot)
  - Cannot read out documents

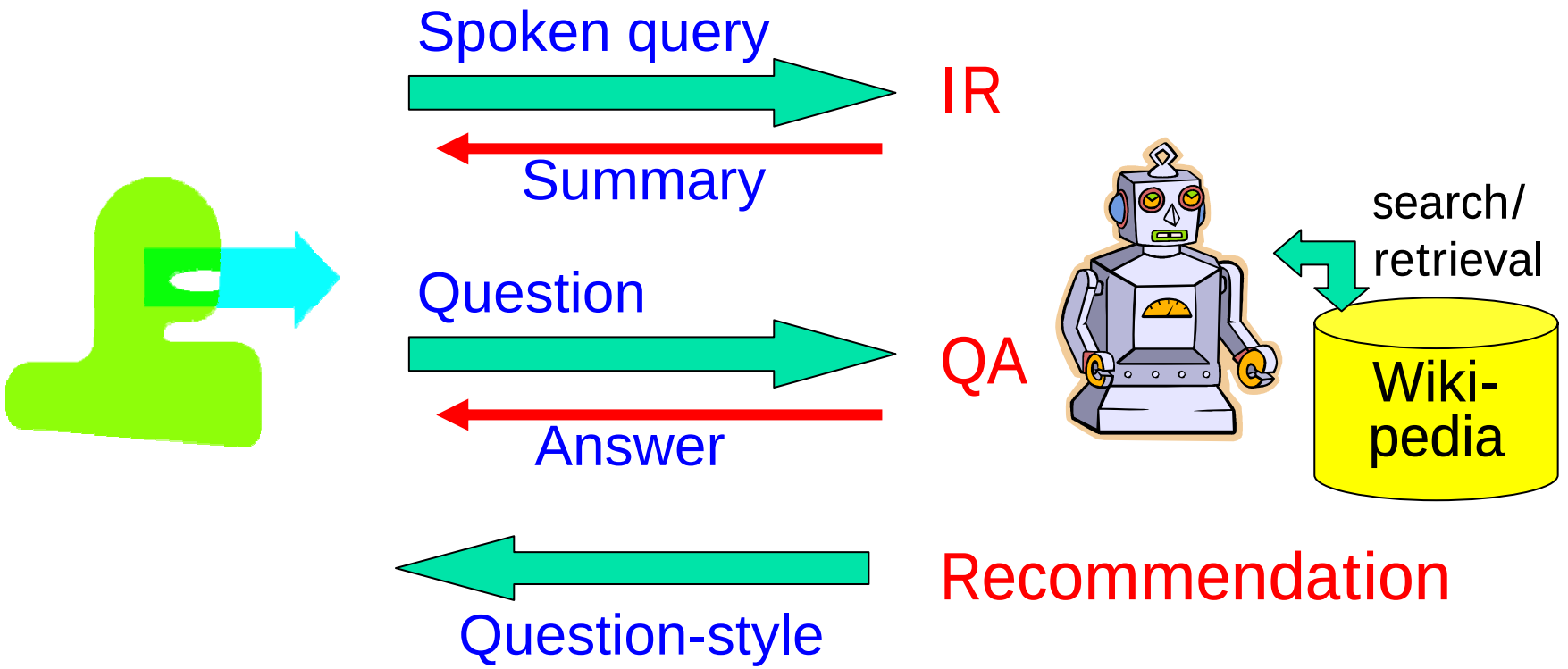
    → Summarize
  - Need to handle pin-point queries

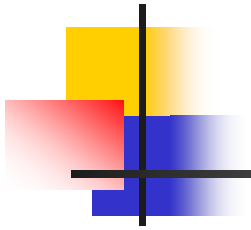
    → Question-Answer (QA)
  - Need to talk with users without definite goals

    (ex.) “something interesting”, “something tasty”

    → Proactive Recommendation

Tour Guide Agent based on IR+QA Techniques [Misu:SPECOM10]





IR and QA

- IR: query → documents

  (ex.) tell me about “Kiyomizu temple”.

- QA: wh-question → named entity (NE)

  (ex.) when was it built?

  (ex.) who built it?

  - IR + NE extraction = SLU

Recommendation in Question form Generated from Key Sentences

[Original text]

By the way, **Queen Elizabeth** praised this stone garden very much, when she once …

↓

By the way, **who** praised this stone garden a lot, when she …

↓

Who praised this stone garden a lot

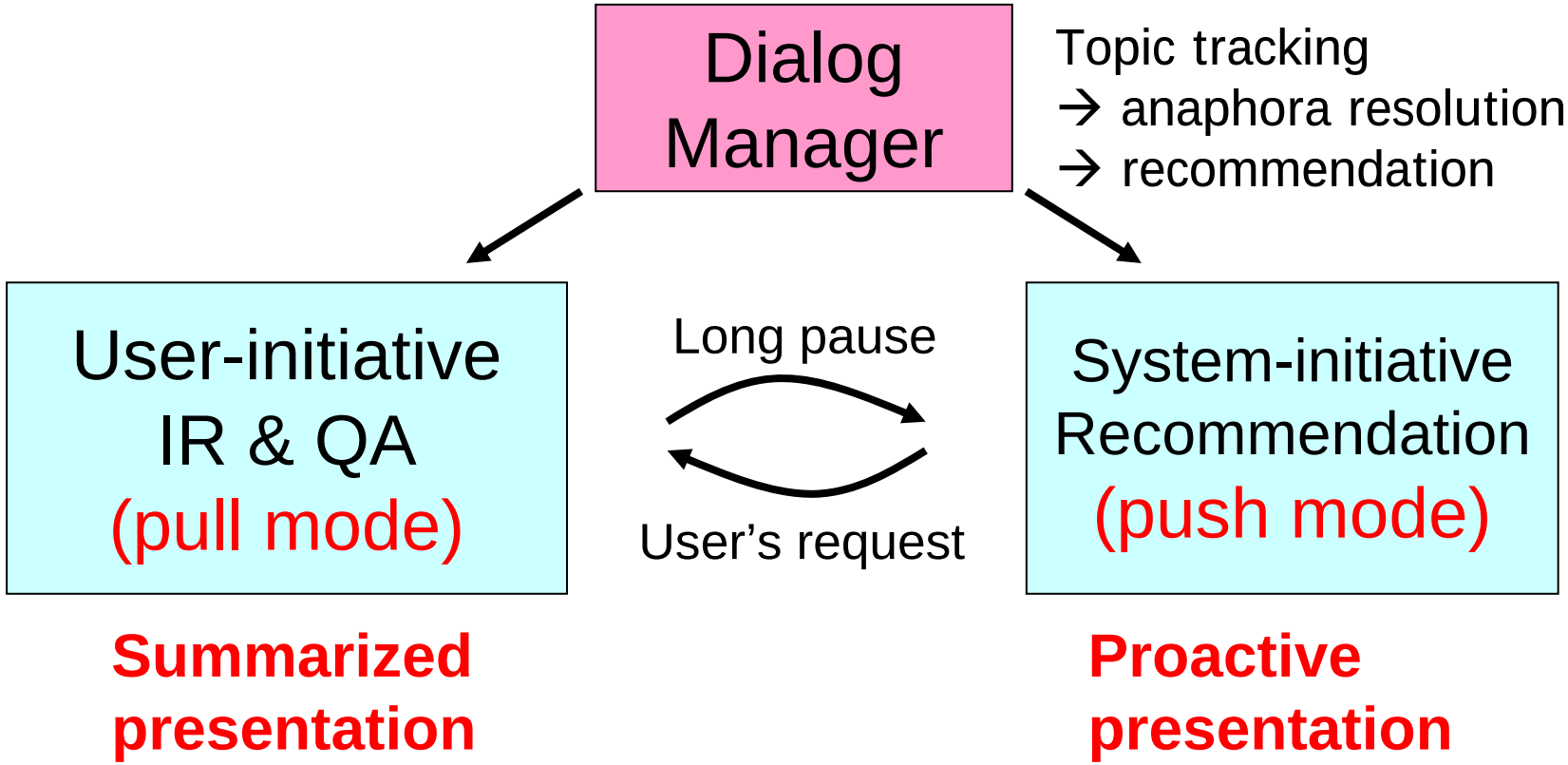
↓

Inverse process of QA

[Question]

Do you know **who** praised this stone garden very much?

Dialog Strategy



Example Dialog in Kyoto Navigator

U1: Please explain Golden Pavilion.

S1: Golden Pavilion is one of the buildings in the Rokuon-ji in Kyoto, and is the main attraction of the temple sites. The entire pavilion except the basement floor is covered with a pure gold leaf.

U2: When was it built?

S2: Golden Pavilion was originally built in 1397 to serve as a retirement villa for Shogun Ashikaga Yoshimitsu.

..(silence)..

S3: Well then, do you know what was awarded to this temple in 1994?

U3: No, please tell me.

S4: It was awarded as listing on the UNESCO World Heritage in 1994.

U4: How can I get there?

Dialog generated on-the-fly from Wikipedia documents!

Kyoto Navigator using Wikipedia as backend KB

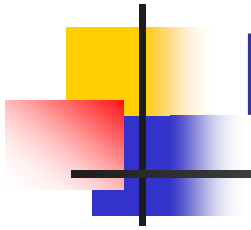
Question about date

U2: "When was it built?"

System-initiative question

S2: "It was built in 1397 as the villa for Shogun Yoshimitsu. The entire pavilion except the basement floor is covered with pure gold leaf."

*System scripts are taken from English version of Wikipedia



Evaluation

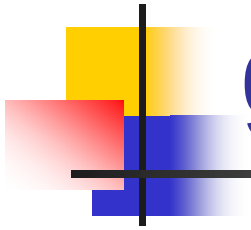
- QA
  - Contextual information important

  ← anaphora resolution
- Recommendation
  - Question-style preferred

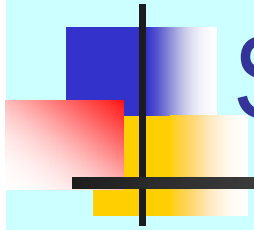
  ← more likely to be accepted
- Dialog
  - Longer interaction suggests satisfaction with the system

  ← different criteria from task-oriented SDS

[Summary] SLU for New-generation Dialog System

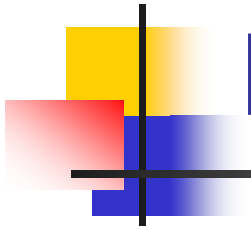


- SLU in interactive IR
  - Vector Space Model (VSM)
  - Query update
- SLU in interactive QA
  - IR + NE extraction
  - Topic/focus detection
- Understanding?
  - Maybe NO in conventional sense
  - Still, important to extract structures such as dependency and discourse



SLU for Human-Human Speech Communication

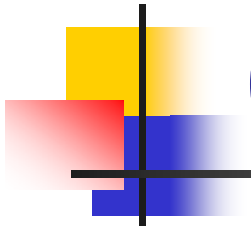
- Rich Transcription
- Hot-spot Detection



Rich Transcription (RT)

- Enhance transcript of spontaneous speech, which is not readable
  - Disfluency detection
  - Punctuation insertion
- Machine learning approach
  - A set of features: lexical, prosodic…
  - Classifiers: SVM, CRF..

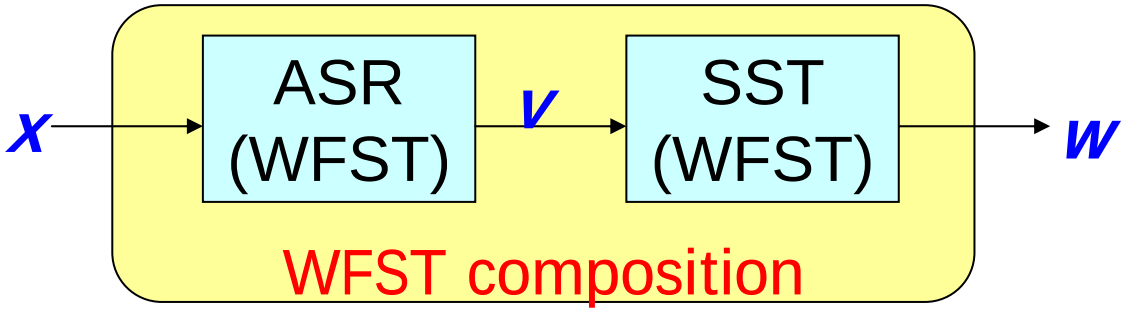
Speaking-style Transformation (SST)



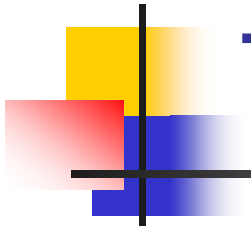
- Convert faithful transcript into formal document-style
  - Deletion of redundant words
  - Correction (substitution) of colloquial expressions
  - Recovery (insertion) of omitted words
- SMT (Statistical Machine Translation) approach [Neubig:IS2009]
  - Log-linear model
  - WFST decoder

Automatic Transcription System for Japanese Congress [Akita:IS2009]

- **Deployed in 2010**
- Evaluation measure: WER
  - NOT against faithful transcript ***V*** (as-is)
  - BUT for final proceeding text ***W*** (should-be)
  - Consistent with system's goal
  - Faithful transcript costly → cannot make everyday



Does Understanding help Transcription?



- Apparently, YES for human
- But NO for machine (ASR)

- Stenographers:

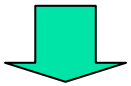
  are NOT sure if they “understand” the speech during shorthand transcription, but do NOT “hear” disfluencies.

High-level Annotation

- Dialog Act tagging [Shriberg:SIGDIAL04]
  - Identify intent type of utterances

    (ex.) request-info, greeting…
- Information Extraction [Ramshaw:ICASSP05]
  - Identify named entities (NE) and their relationships

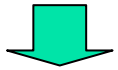
    (ex.) [A sell B] [A acquire C]



New Direction of SLU

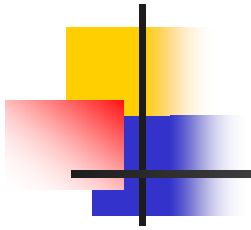
Speech Summarization
[Furui:SLT2006]

- Extract important portions and generate compact output
- MMR (Maximum Marginal Relevance) [Carbonell:1998]
  - Similarity defined with VSM
  - Extract sentences which best match the entire document and differ each other



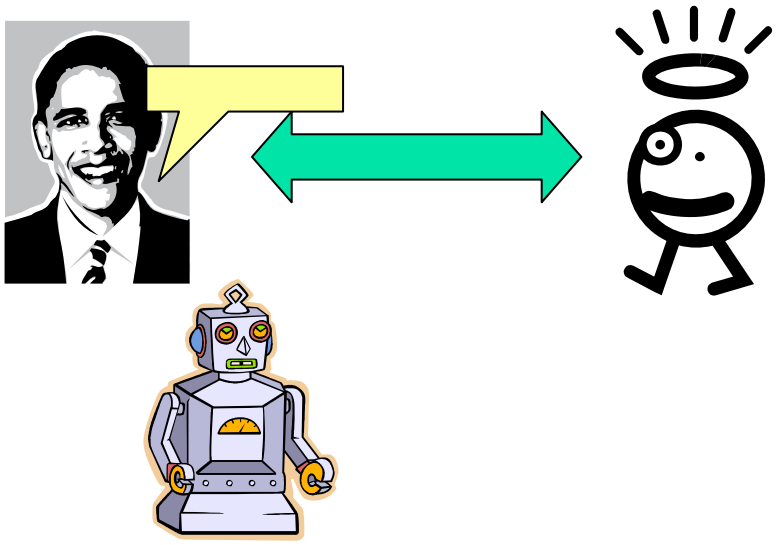
  - SLU with VSM

From Content-based Approach to Interaction-based Approach



- Content-based approach
  - try to understand & annotate content of speech
  - Actually hardly “understand”
- Interaction-based approach
  - give up “understanding” of speaker’s utterances
  - look into reaction of listeners/audience, who understand the content
  - More oriented for human cognitive process

From Content-based Approach to Interaction-based Approach

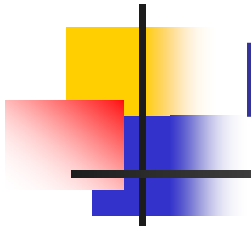


- Even if we do not understand the talk, we can see funny/important parts by observing audience's laughing/nodding
- Page rank is determined by the number of links rather than by the content

From Content-based Approach to Interaction-based Approach

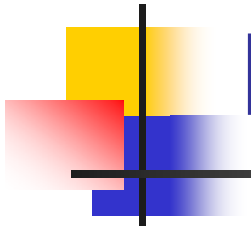
	Focus	Features	Annotation
Content -based	Main speaker's utterances	lexical, prosodic …	Important segment
Interaction -based	Listener's reaction	non-verbal, multi-modal	Impressive segment (interesting)

Multi-modal Corpus of Poster Sessions [Kawahara:IS2008]



- Norm in conferences & open-houses
- Mixture characteristics of lectures [CSJ] and meetings [AMI]
  - One main speaker, with small audience
  - audience can take initiative
- Interactive: real-time feedback by audiences
  - Nodding & backchannels
  - Comments and questions
- Multi-modal (truly)
  - Standing & moving

Multi-modal Sensing Environment



- Wire-less head-worn microphone
- Distant microphone
- Microphone array mounted on poster stand

Audio

- 8 cameras installed in the room

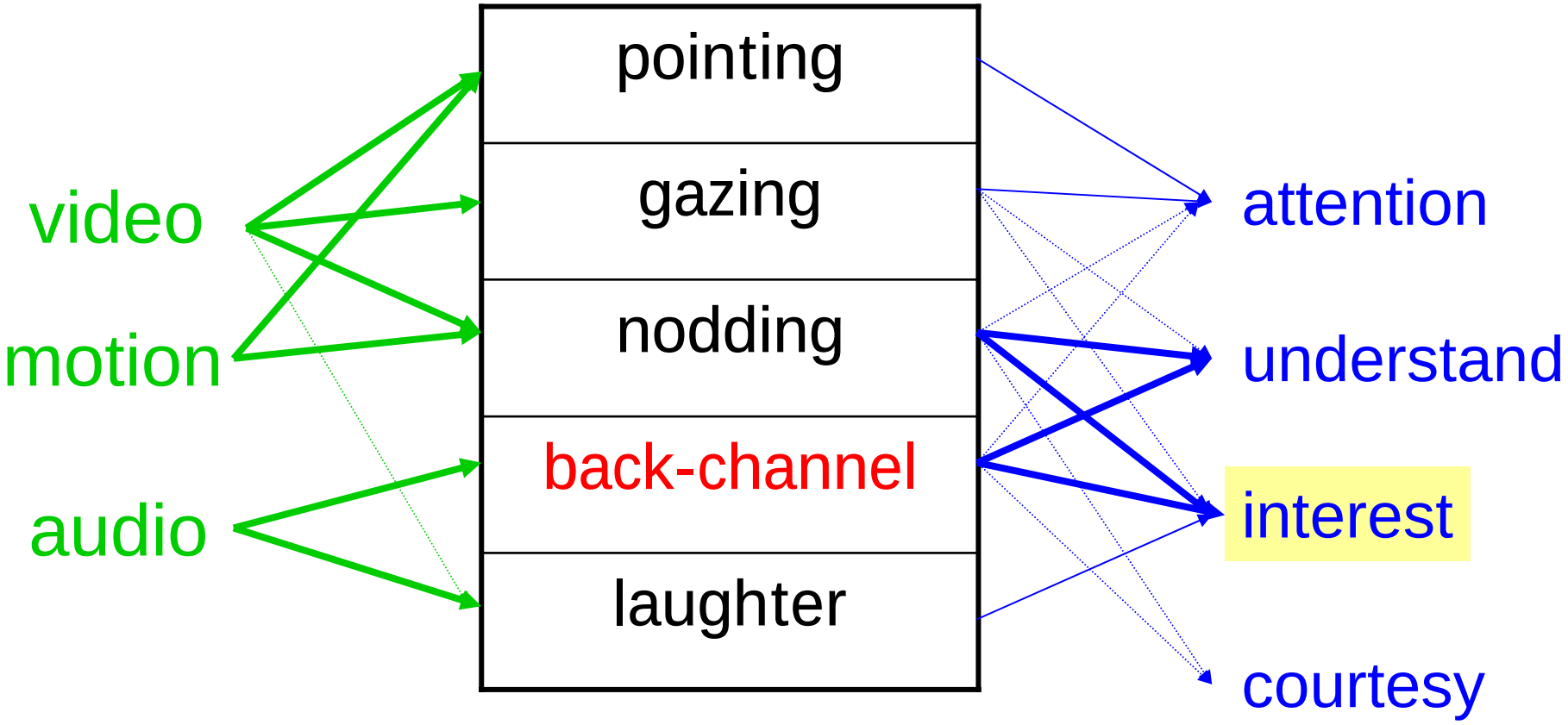
Video

- Motion-capturing system
- Accelerometer

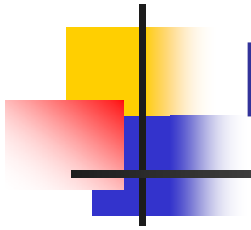
Motion

- Eye-tracking recorders → Gazing

Multi-modal Annotation



Hot-spot Detection based on Listener's Backchannel Response

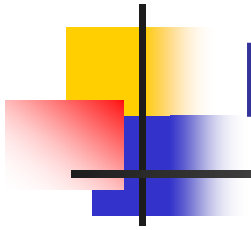


- Hot-spot: where audience was impressed
- Backchannel *(aizuchi)*
  - Short verbal responses made in real-time

    (cf.) Twitter
  - often non-lexical

    (ex.) "yeah", "uh-huh"…
  - indicate "I hear you, understand you…"
  - change syllabic & prosodic patterns, according to state of mind

Identification of Backchannel Patterns related with Interest-level



- Occurrence frequency patterns
  - "*hai* (yes)"
    - frequent when listening to reply to his own question
    - Acknowledgment & courtesy
- Prosodic patterns (F0, power, duration)
  - "*he:*", "*hu:N*", "*a:*"
    - Large variation
    - Large correlation (in some prosodic patterns) with interest-level by subjective evaluation

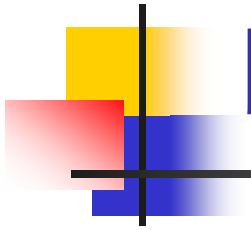
Identification of Backchannel Patterns related with Interest-level

Reactive token	prosody	interest	surprise
へー *he:*	duration	○	○
	F0max	○	○
	F0range	○	○
	Pmax	○	○
あー *a:*	duration		
	F0max	○	○
	F0range		
	Pmax	○	
ふーん *fu:N*	duration	○	○
	F0max		
	F0range		
	Pmax		

Acoustic Event Detection
[Sumi:IS2009]

- Target of Detection
  - Reactive tokens used in backchannel
  - Laughter
- Method
  - BIC-segmentation & GMM classification
  - Dedicated verifier: prosodic information
- Performance
  - F-measure: 70%
  - Precision of reactive tokens: 85%

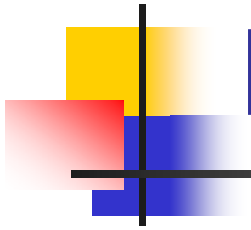
Audio Indexing of Hot-spots based on Listener's Reactions



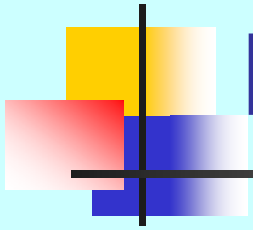
- Detection of reactive tokens & laughter
- Classification of interest-level

- Browser interface

[Summary] SLU for Human-human Communication



- Hard to “understand”
- Content-based approach
  - Feature vectors…lexical, prosodic
  - Information extraction, dialog act tagging
- Interaction-based approach
  - NOT understand main speaker’s utterance
  - BUT watch reactions of audience
  - to be combined with content-based approach



Conclusions: New Perspectives of SLU

- Paradigm shift in human-machine dialog
  - Simple DB query → general IR
  - Semantic slot extraction → VSM (vector space model)
  - Robust extraction of shallow structures useful
- Exploration to human-human communication
  - Hard to “understand”
  - New approach focusing on human understanding process