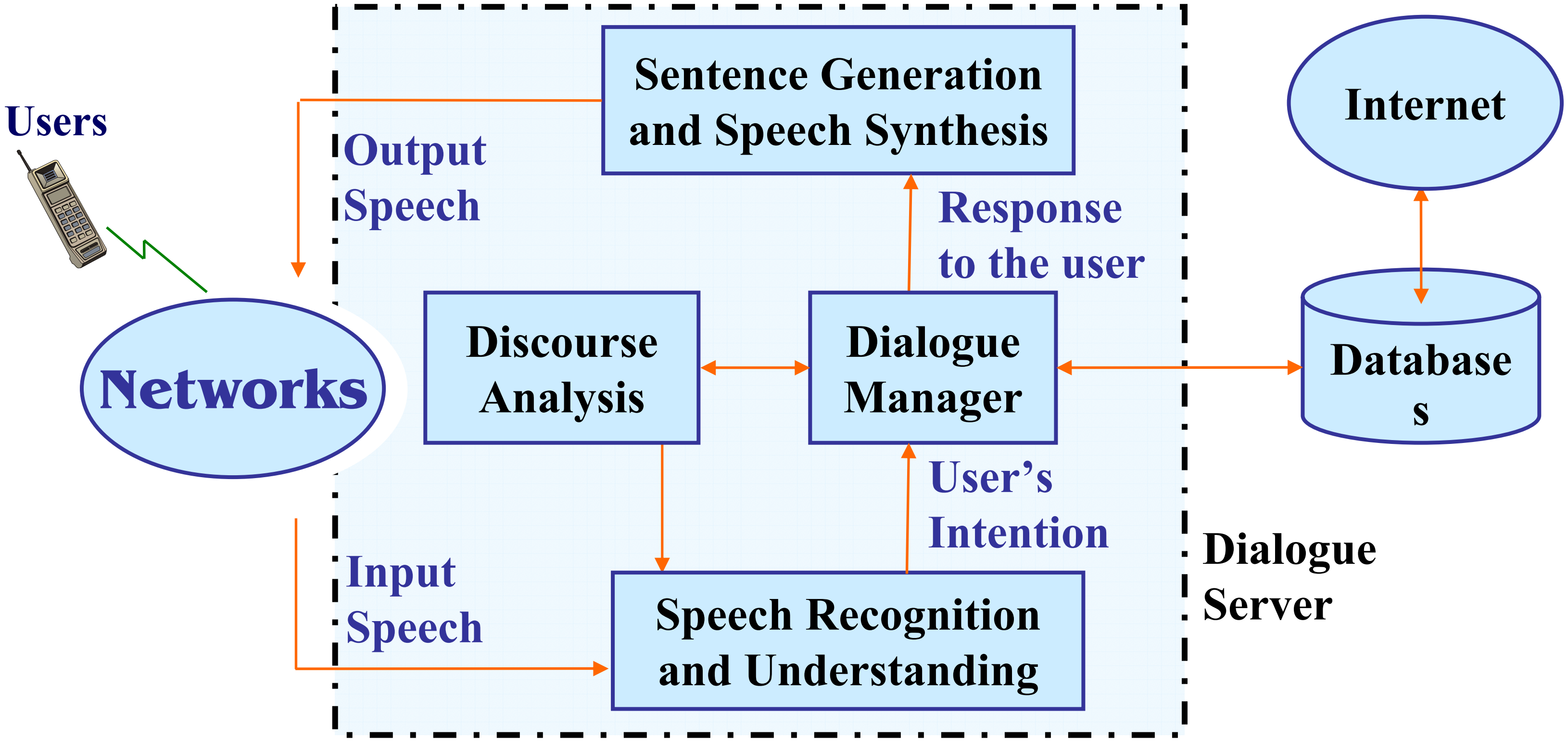
15.0 Spoken Dialogues

References: 1. 11.1 - 11.2.1, Chapter 17 of Huang
2. Sadek and De Mori, "Spoken Dialogues with Computers", Academic Press, 1998
3. "Special Issue on Language Modeling and Dialogue Systems", IEEE Trans. on Speech and Audio Processing, Jan 2000
4. "Conversational Interfaces: Advances and Challenges", Proceedings of the IEEE, Aug 2000

Spoken Dialogue Systems

- Almost all human-network interactions can be made by spoken dialogue
- Speech understanding, speech synthesis, dialogue management, discourse analysis
- System/user/mixed initiatives
- Reliability/efficiency, dialogue modeling/flow control
- Transaction success rate/average dialogue turns



Key Processes in A Spoken Dialogue

- **A Basic Formulation**

$$A_n^* = \underset{A_n}{\arg\max}\ \text{Prob}\,(A_n | X_n, S_{n-1})$$

X_n: speech input from the user in the n-th dialogue turn
S_n: discourse semantics (dialogue state) at the n-th dialogue turn
A_n: action (response, actions, etc.) of the system (computer, hand-held device, network server, etc.) after the n-th dialogue turn

  – goal: the system takes the right actions after each dialogue turn and complete the task successfully finally

$$A_n^* \approx \underset{A_n, S_n}{\arg\max}\ \underbrace{P(A_n | S_n)}_{\text{by dialogue management}} \sum_{F_n} \underbrace{P(S_n | F_n, S_{n-1})}_{\text{by discourse analysis}}\ \underbrace{P(F_n | X_n, S_{n-1})}_{\text{by speech recognition and understanding}}$$

F_n: semantic interpretation of the input speech X_n

- **Three Key Elements**
  – speech recognition and understanding: converting X_n to some semantic interpretation F_n
  – discourse analysis: converting S_{n-1} to S_n, the new discourse semantics (dialogue state), given all possible F_n
  – dialogue management: select the most suitable action A_n given the discourse semantics (dialogue state) S_n

Dialogue Structure

- **Turns**
  - an uninterrupted stream of speech(one or several utterances/sentences) from one participant in a dialogue
  - speaking turn: conveys new information
    back-channel turn: acknowledgement and so on(e.g. O. K.)
- **Initiative-Response Pair**
  - a turn may include both a response and an initiative
  - system initiative: the system always leads the interaction flow
    user initiative: the user decides how to proceed
    mixed initiative: both acceptable to some degree
- **Speech Acts(Dialogue Acts)**
  - goal or intention carried by the speech regardless of the detailed linguistic form
  - forward looking acts
    - conversation opening(e.g. May I help you?), offer(e.g. There are three flights to Taipei…), assert(e.g. I'll leave on Tuesday), reassert(e.g. No, I said Tuesday), information request(e.g. When does it depart?), etc.
  - backward looking acts
    - accept(e.g. Yes), accept-part(e.g. O.K., but economy class), reject(e.g. No), signal not clear(e.g. What did you say?), etc.
  - speech acts ↔ linguistic forms : a many-to-many mapping
    - e.g. "O.K." — request for confirmation, confirmation
  - task dependent/independent
  - helpful in analysis, modeling, training, system design, etc.
- **Sub-dialogues**
  - e.g. "asking for destination", "asking for departure time", …..

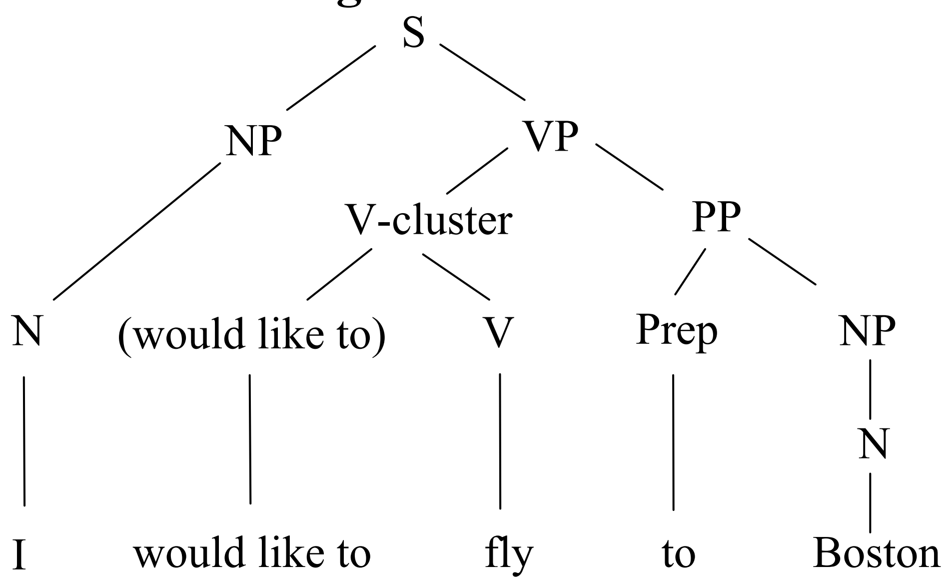
Language Understanding for Limited Domain

- **Semantic Frames — An Example for Semantic Representation**
  – a semantic class defined by an entity and a number of attributes(or slots)

  e.g. [Flight]:

  [Airline] → (United)
  [Origin] → (San Francisco)
  [Destination] → (Boston)
  [Date] → (May 18)
  [Flight No] → (2306)

  – "slot-and-filler" structure
- **Sentence Parsing with Context-free Grammar (CFG) for Language Understanding**



Grammar(Rewrite Rules)

S → NP VP
NP → N
VP → V-cluster PP
V-cluster → (would like to) V
V → fly| go
PP → Prep NP
N → Boston | I
Prep → to

  – extension to Probabilistic CFG, integration with N-gram(local relation without semantics), etc.

Robust Parsing for Speech Understanding

- **Problems for Sentence Parsing with CFG**
  - ungrammatical utterances
  - speech recognition errors (substitutions, deletions, insertions)
  - spontaneous speech problems: um–, cough, hesitation, repetition, repair, etc.
  - unnecessary details, irrelevant words, greetings, unlimited number of linguistic forms for a given act

    e.g. to Boston
    I'm going to Boston, I need be to at Boston Tomorrow
    um– just a minute– I wish to – I wish to – go to Boston

- **Robust Parsing as an Example Approach**
  - small grammars for particular items in a very limited domain, others handled as fillers

    e.g. Destination→ Prep CityName
    Prep → to |for| at
    CityName → Boston |Los Angeles|...

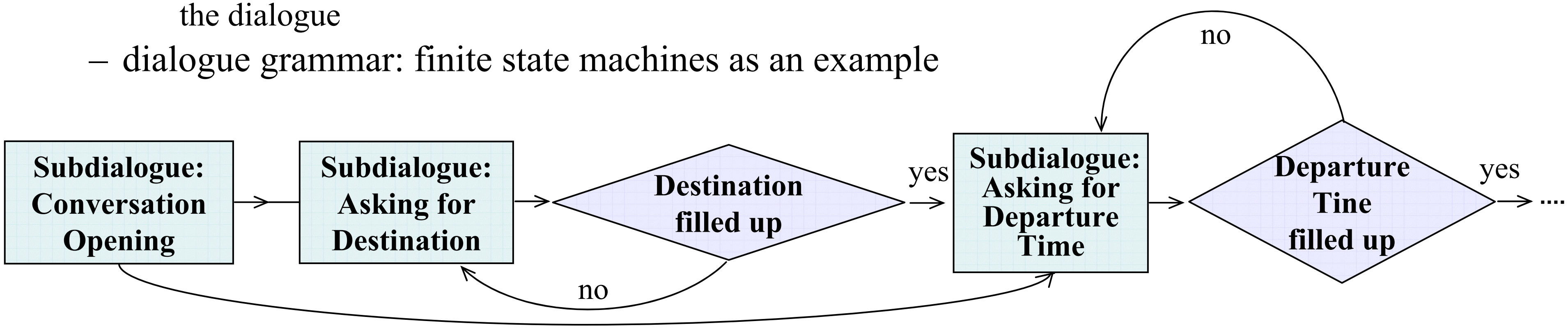
  - different small grammars may operate simultaneously
  - keyword spotting helpful
  - concept N-gram may be helpful

    $Prob(c_i|c_{i-1})$, c_i: concept
    CityName (Boston,...) direction (to, for...)
    similar to class-based N-gram

- **Speech Understanding**
  - two-stage: speech recognition (or keyword spotting) followed by semantic parsing (e.g. robust parsing)
  - single-stage: integrated into a single stage

Discourse Analysis and Dialogue Management

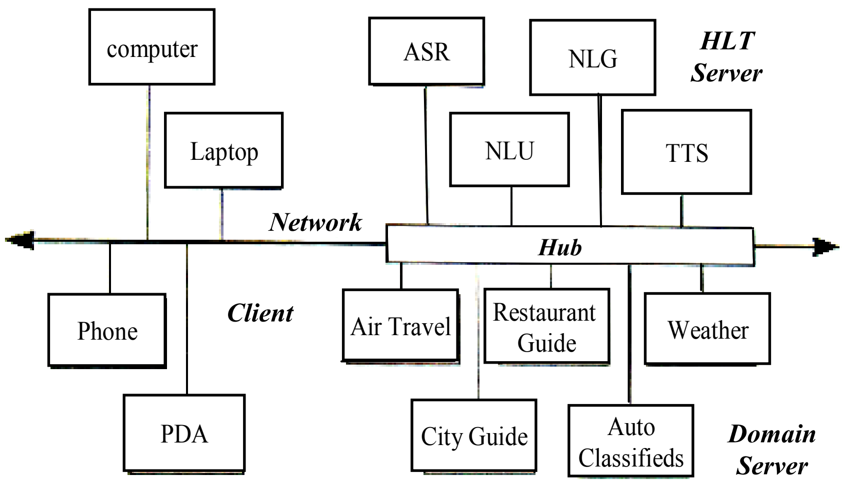
- **Discourse Analysis**
  - conversion from relative expressions(e.g. tomorrow, next week, he, it…) to real objects
  - automatic inference: deciding on missing information based on available knowledge(e.g. “how many flights in the morning? ” implies the destination/origin previously mentioned)
  - inconsistency/ambiguity detection (e.g. need clarification by confirmation)
  - example approach: maintaining/updating the dialogue states(or semantic slots)
- **Dialogue Management**
  - controlling the dialogue flow, interacting with the user, generating the next action
    - e.g. asking for incomplete information, confirmation, clarify inconsistency, filling up the empty slots one-by-one towards the completion of the task, optimizing the accuracy/efficiency/user friendliness of the dialogue
  - dialogue grammar: finite state machines as an example



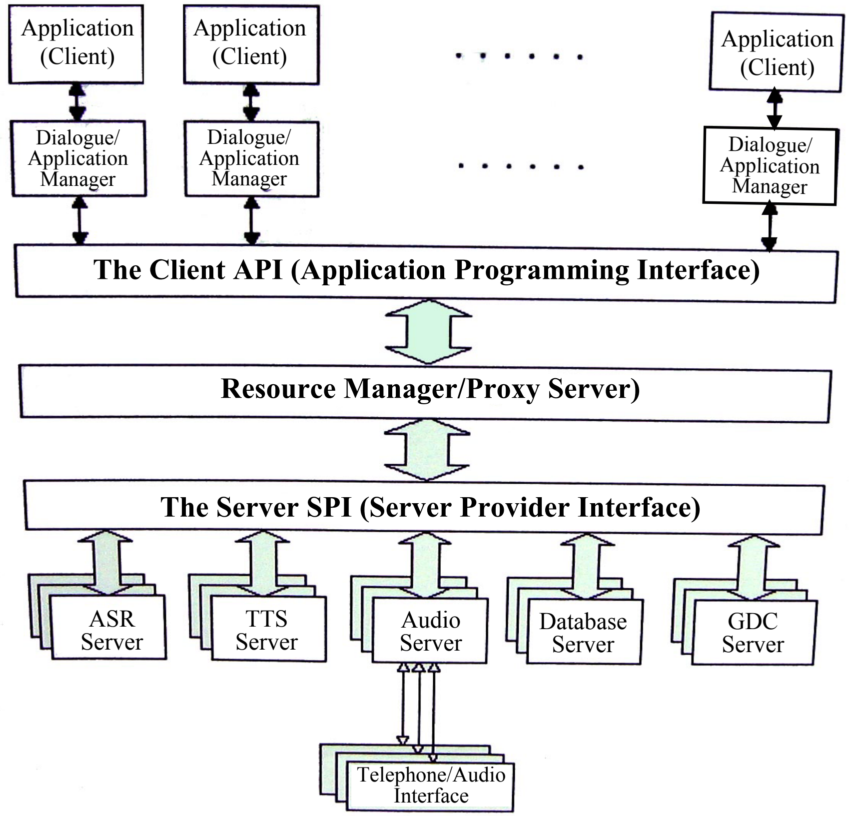
  - plan-based dialogue management as another example
  - challenging for mixed-initiative dialogues
- **Performance Measure**
  - internal: word error rate, slot accuracy (for understanding), etc.
  - overall: average success rate (for accuracy), average number of turns (for efficiency), etc.

Client-Server Architecture

- **Galaxy, MIT**



- **Integration Platform, AT& T**



- **Domain Dependent/Independent Servers Shared by Different Applications/Clients**
  - reducing computation requirements at user (client) by allocating most load at server
  - higher portability to different tasks