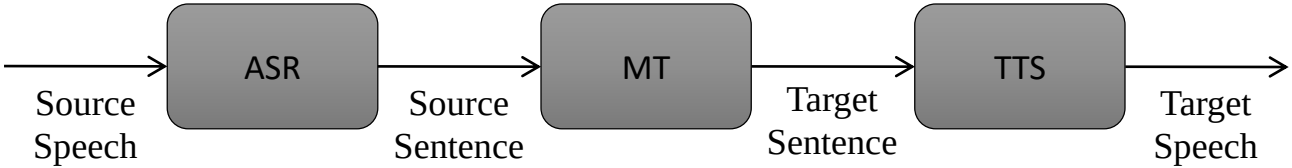
Language Understanding

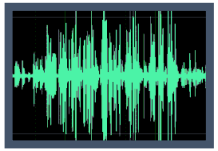
16 – Applications of Spoken Language Processing



Spoken Language Translation

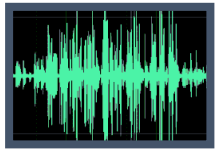
- Spoken language translation (SLT) is to directly translate spoken utterances into another language.
- Major components
  - Automatic Speech Recognition (ASR)
  - Machine Translation (MT)
  - Text-to-Speech (TTS)





ایستگاه اتوبوس کجاست؟

Where is the bus stop?



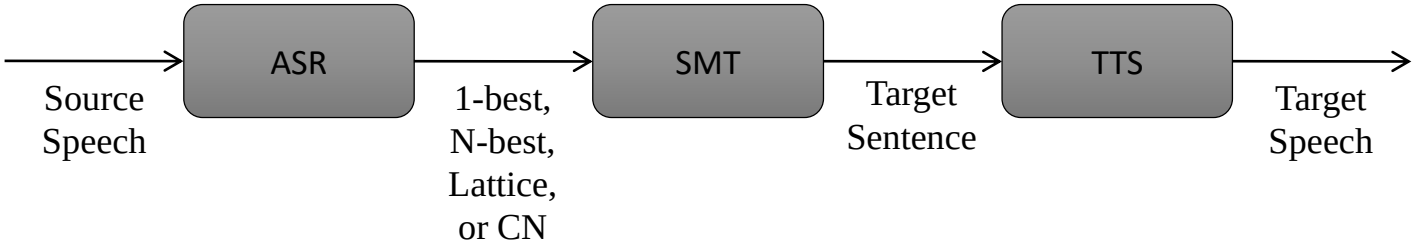
Coupling ASR to MT

- Motivation
  - ASR cannot secure an error-free system
    - One best of ASR could be wrong
    - SLT must be designed robust to speech recognition errors
    - MT could be benefited from wide range of supplementary information provided by ASR
  - MT quality may depend on WER of ASR
    - Strong correlation between recognition and translation quality
    - WER of ASR decreases in a set of hypotheses
    - Idea : Exploitation of more transcriptions

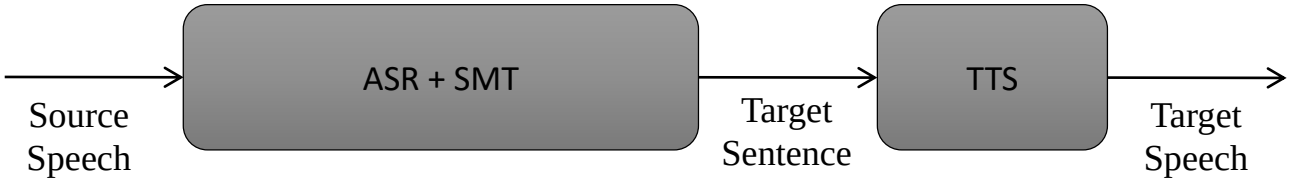
- SLT systems vary in the degree to which SMT and ASR are integrated within the overall translation process.

Coupling ASR to MT

- Loose coupling
  - SMT uses ASR output (1-best, N-best, lattice, or confusion network) as input for 1-way module communication

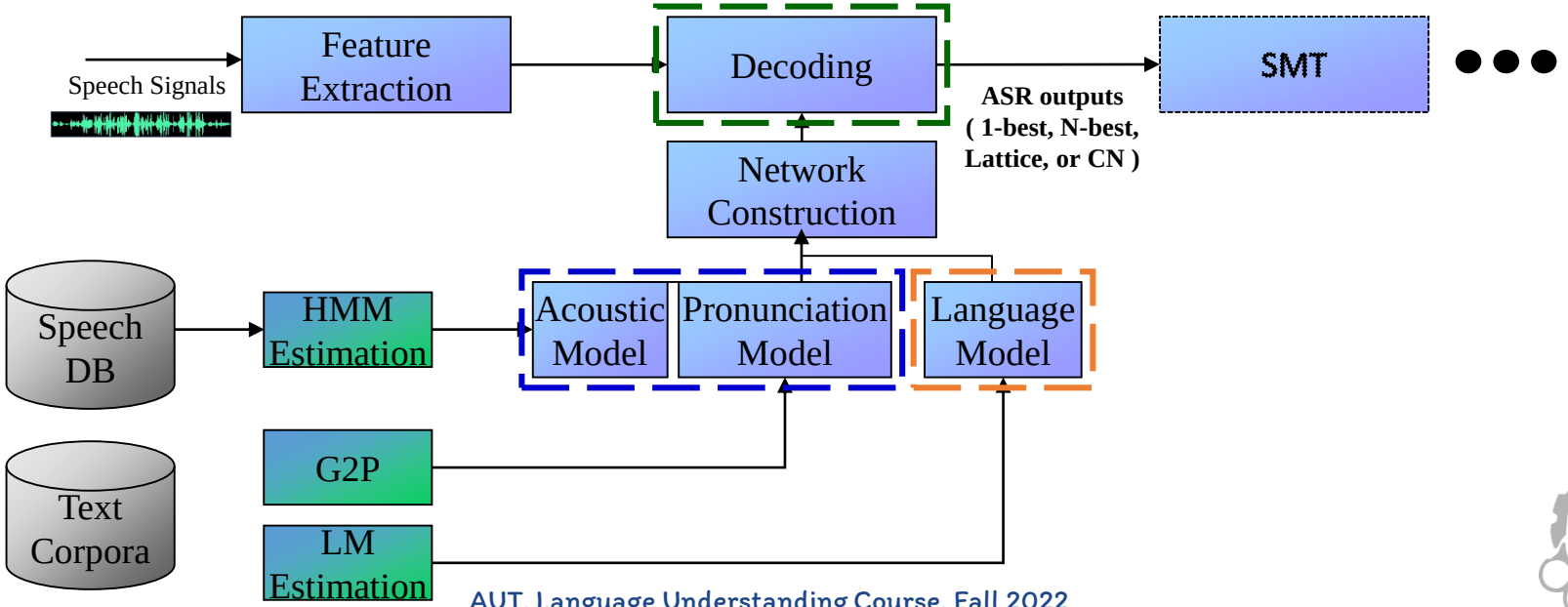


- Tight coupling
  - The whole search space of ASR and MT is integrated



ASR Outputs

- Automatic speech recognition (ASR) is a process by which an acoustic speech signal is converted into a set of words.

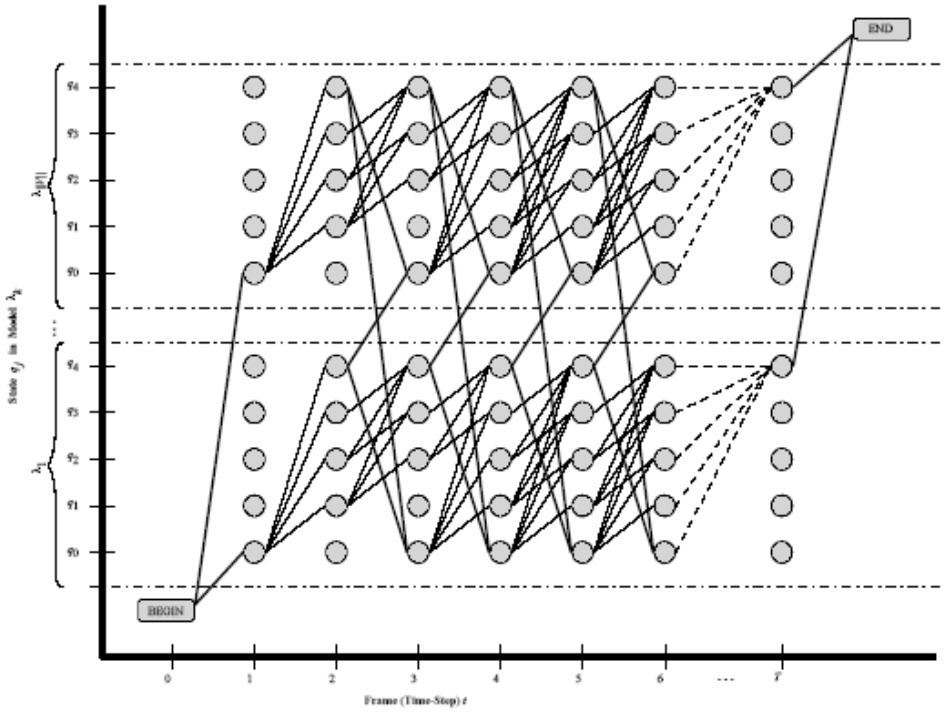


ASR Outputs

- 1-best
  - The best path could find from back tracking
  - Why a 1-best “word” sequence?
    - Storing the backtracking pointer table for state sequence takes a lot of memory
    - Usually a backtrack pointer storing : The previous words before the current word

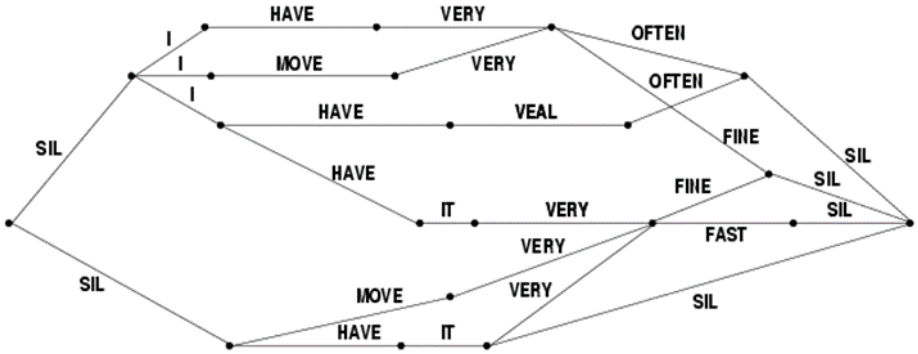
- N-best
  - Traceback not only from the 1st-best, also from the 2nd best and 3rd best, etc.
  - Methods
    - Directly from search backtrack pointer table – Exact N-best algorithm, Word pair N-best algorithm, A* search using Viterbi score as heuristic
    - Generate lattice first, then generate N-best from lattice

State Lattice



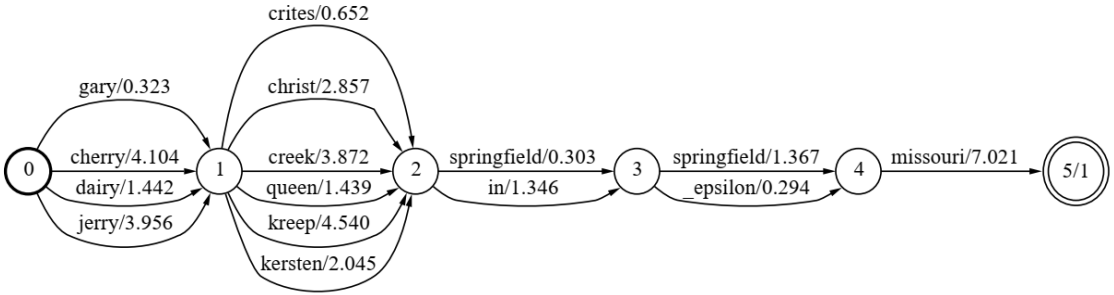
ASR Outputs

- Lattice
  - A word-based lattice
    - A compact representation of state-lattice
    - Only word node are involved
  - From the decoding backtracking pointer table
    - Only record all the links between word nodes
  - From N-best list
    - Become a compact representation of N-best

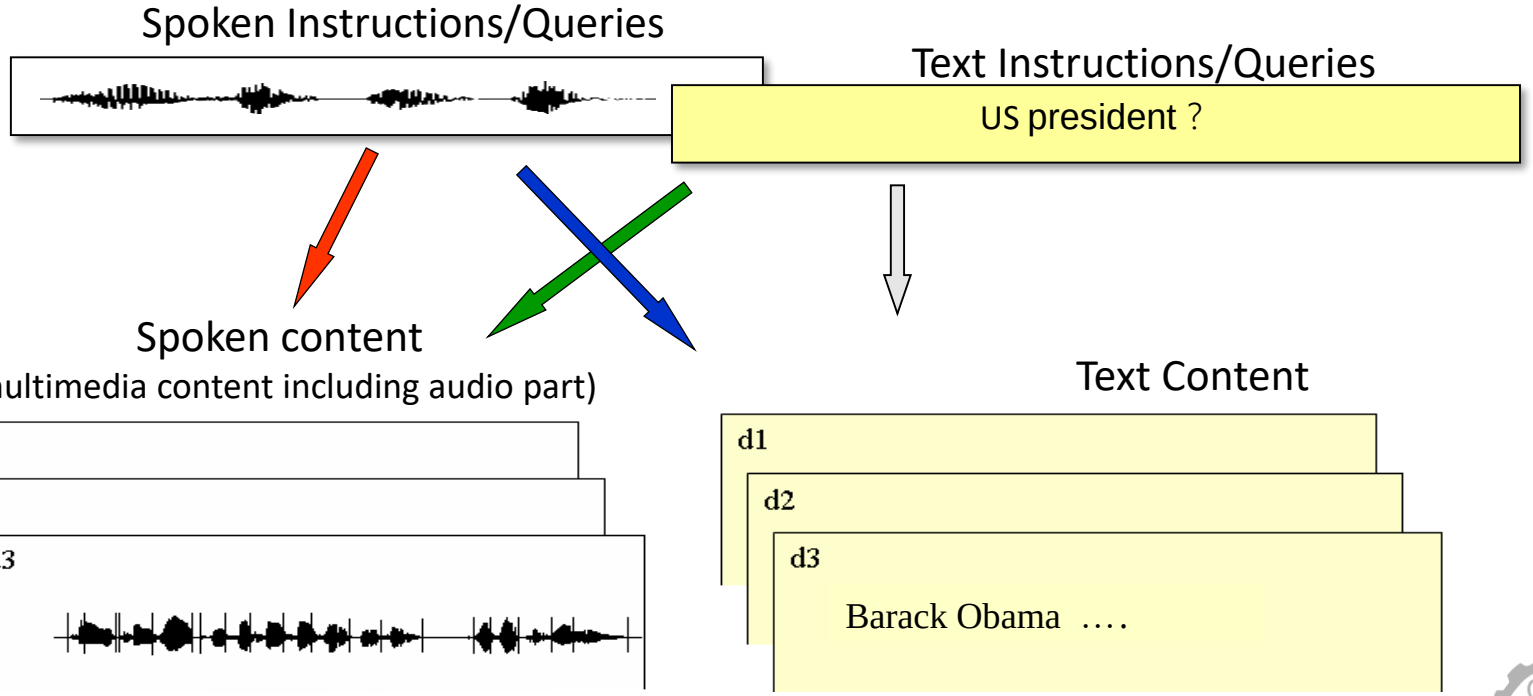


ASR Outputs

- Confusion Network
  - Or “Consensus Network”
  - A weighted directed graph with a start node, an end node, and word labels over its edges
  - Each path from the start node to the end node goes through all the other nodes
  - From lattice
    - Multiple alignment



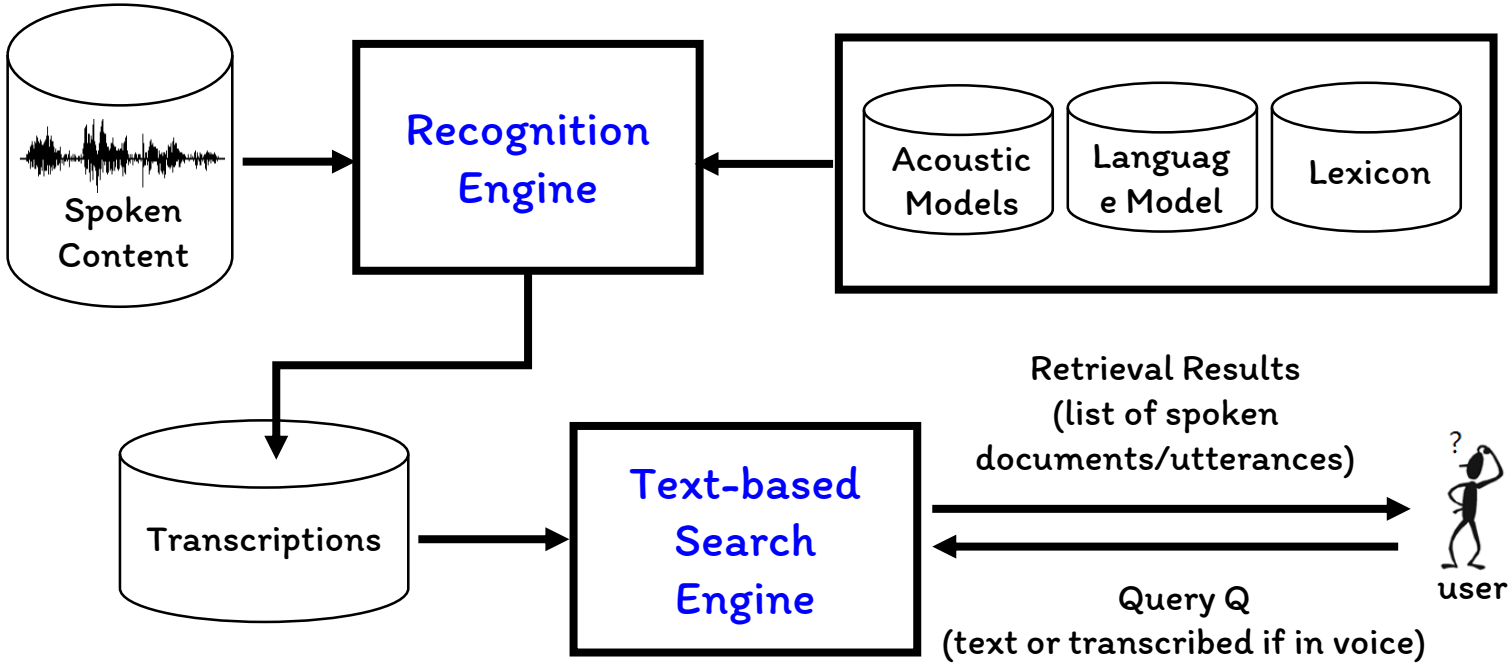
Speech-based Information Retrieval



Speech-based Information Retrieval

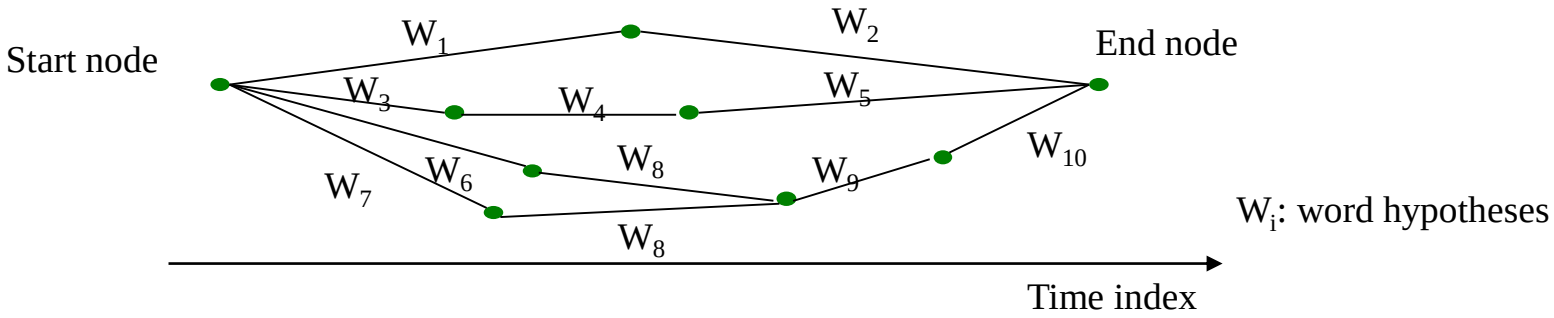
- **User instructions and/or network content can be in form of voice**
  - text queries/spoken content : spoken document retrieval, spoken term detection
  - spoken queries/text content : voice search
  - spoken queries/spoken content : query by example

Basic Approach for Spoken Content Retrieval



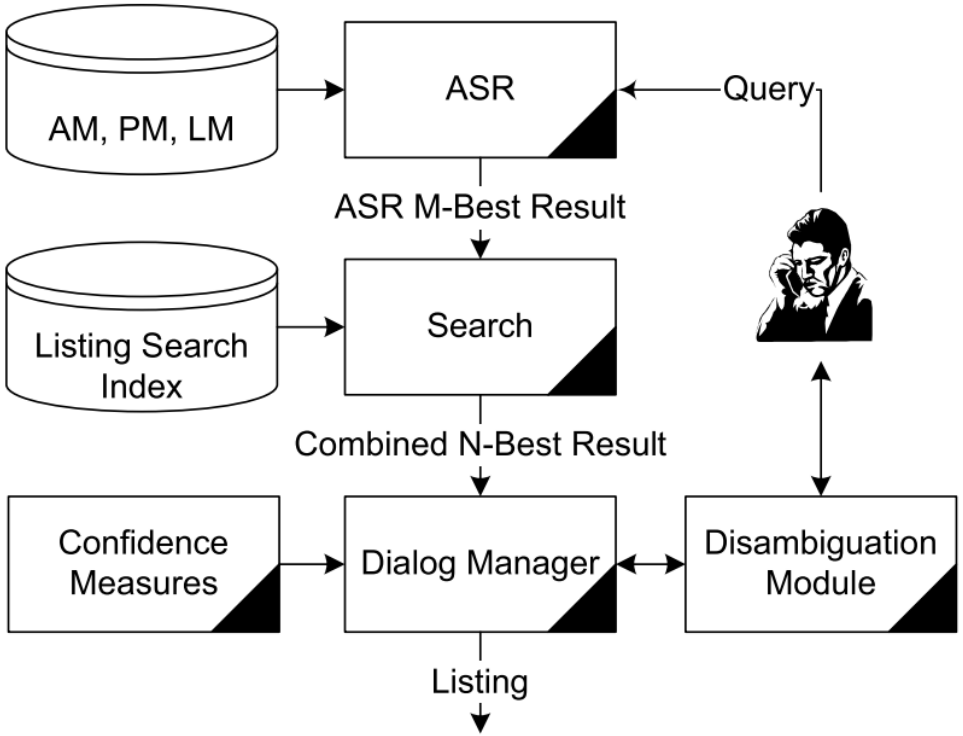
Lattices for Spoken Content Retrieval

- Low recognition accuracies for spontaneous speech including Out-of-Vocabulary (OOV) words under adverse environment
  - considering lattices with multiple alternatives rather than 1-best output



  - higher probability of including correct words, but also including more noisy words
  - correct words may still be excluded (OOV and others)
  - huge memory and computation requirements

Voice Search System



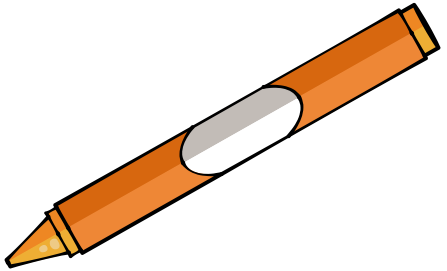
Voice Search System

- A user's utterance is first recognized with an automatic speech recognizer (ASR)
  - It utilizes an acoustic model (AM), a pronunciation model (PM) and a language model (LM)
- The *m*-best results from the ASR are passed to a search component to obtain the *n*-best semantic interpretations.
  - I.e. a list of up to *n* entries in the database
- The interpretations are passed to a dialogue manager (DM) subsequently.
  - It utilizes confidence measures, which indicate the certainty of the interpretations, to decide how to present the *n*-best results.
- If the system has high confidence on a few entries, it directly presents them to the user.
  - Otherwise, a disambiguation module is exploited to interact with the user to understand his true intent.

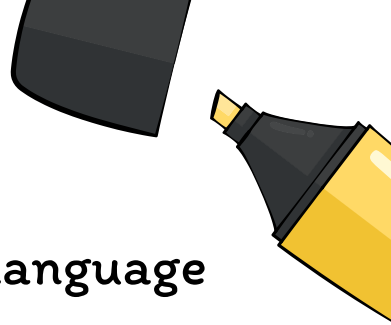
Voice Search System

SLU task	User input utterances		Target semantic representation	
	Naturalness	Input space	Resolution	Semantic space
Frame-based/directed dialogue	Low	Small	Low	Small
Frame-based/mixed-initiative	Low-medium	Small	High	Small
Call routing	High	Large	Low	Small
Voice search	Medium-high	Large	Low	Large

Thanks for your attention



References and IP Notice



- Some slides from Gary Geunbae Lee's slides on Spoken language translation.
- Some slides from Lin-shan Lee's slides on Speech-based Information Retrieval.
- Tur, Gokhan, and Renato De Mori. "Spoken language understanding", 2011
- Some graphics from Slidesgo online template.