Conversational Interfaces: Advances and Challenges

Victor W. Zue and James R. Glass

Proceedings of the IEEE'00

Introduction

- It is impractical for uses to use mouse or keyboard on handheld devices ➔ spoken language.

- Two speech-based interface:
  - Speech recognition
  - Speech synthesis

Introduction

- The degree with which the system maintains an active role in the conversation.
  - System-initiative
  - User-initiative
  - Mixed-initiative
    - Goal-oriented dialogue
    - It is necessary to examine human-human interactions.

C:	Yeah, [umm] I'm looking for the Buford Cinema.	*disfluency*
A:	OK, and you want to know what's showing there or ...	*interruption*
C:	Yes, please.	*confirmation*
A:	Are you looking for a particular movie?	
C:	[umm] What's showing.	*clarification*
A:	OK, one moment.	*back channel*
	...	
A:	They're showing A Troll In Central Park.	
C:	No.	*inference*
A:	Frankenstein.	*ellipsis*
C:	What time is that on?	*co-reference*
A:	Seven twenty and nine fifty.	
C:	OK, and the others?	*fragment*
A:	Little Giant.	
C:	No.	
A:	...	
C:	...	
A:	That's it.	
C:	Thank you.	
A:	Thanks for calling Movies Now.	

Fig. 1. Transcript of a conversation between an agent (A) and a client (C) over the phone. Typical conversational phenomena are annotated on the right.

Introduction

Table 1 Statistics of Human–Human Dialogues in a Movie Domain [29]. Annotated Dialogue Acts are Sorted by Customer Usage and Include Frequency of Occurence and Average World Length

Customer			Agent	
Act	Freq.	Words	Freq.	Words
Acknowledge	47.9	2.3	30.8	3.1
Request	29.5	9.0	15.0	12.3
Confirm	13.1	5.3	11.3	6.4
Inform	5.9	7.9	27.8	12.7
Statement	3.4	6.9	15.0	6.7

Introduction

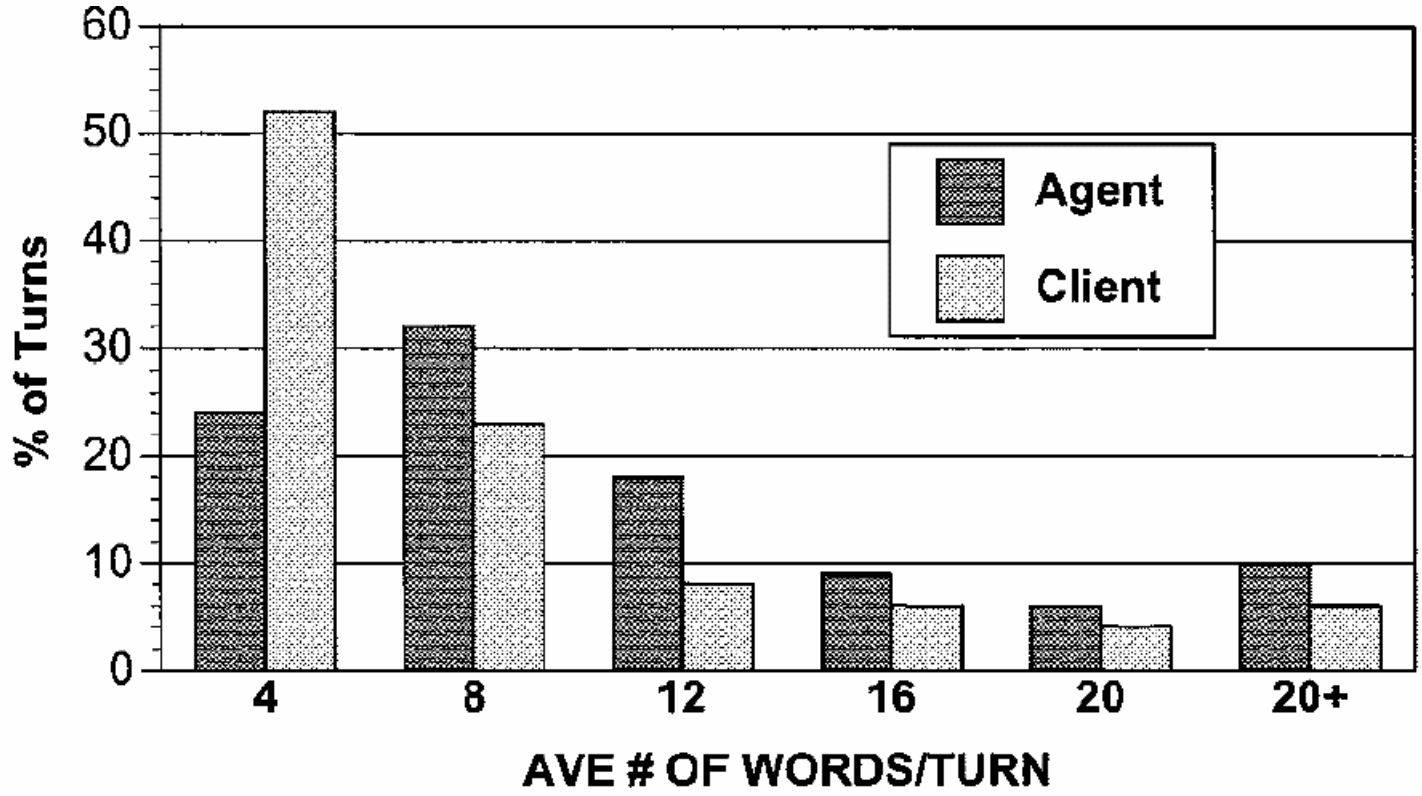

Fig. 2. Histograms of utterance length for agents and clients in tasks of information access over the phone.

Introduction

- It is important to note that some of the spontaneous speech phenomena serve useful roles in human-human communication.
  - Initial disfluent speech can serve an attention-getting function
  - Filled pauses and back-channel acknowledgement provide reassurances that the utterance is understood or one partner of the conversation is still working on the problem.

Underlying Technologies and Research Issues

- System Architecture
- Spoken Input: From Signal to Meaning
  - Automatic Speech Recognition
  - Natural Language Understanding
  - Discourse
- Output Processing: From Information to Signal
  - Natural Language Generation
  - Speech Synthesis Dialogue Management

System Architecture

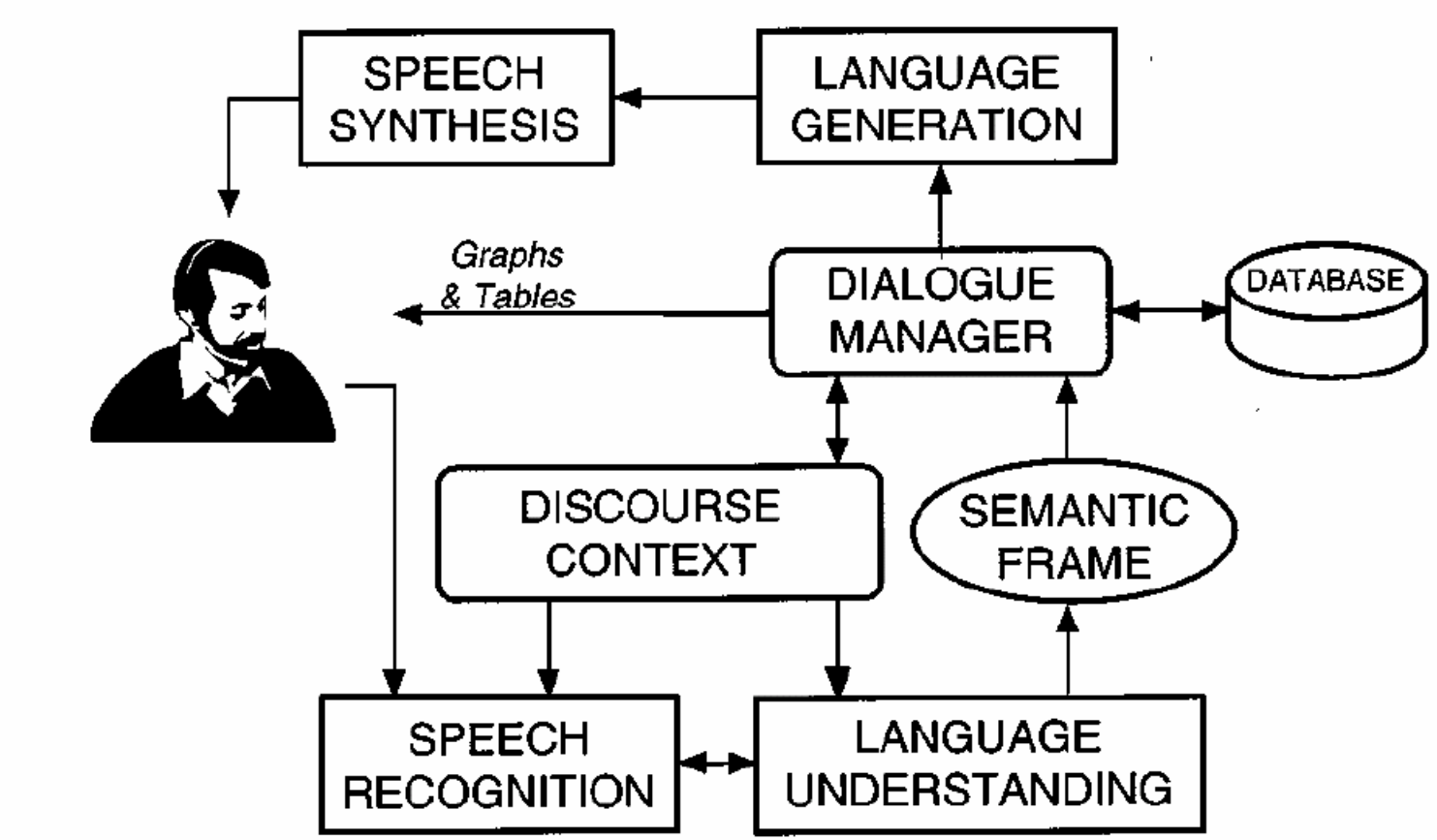

Fig. 3. Generic block diagram for a typical conversational interface.

Spoken Input: From Signal to Meaning

- Two Steps:
  - The conversion of the signal to a set of words.
  - The derivation of the meaning from the word hypotheses
- A discourse component is often used to properly interpret the meaning of an utterance in the larger context of the interaction.

Automatic Speech Recognition

- Deal with the problem of disfluency ➔ filler model.

- An issue is arising : The recognition of telephone-quality speech.

Natural Language Understanding

- How to handle unknown words, novel linguistic constructs, recognition errors and spontaneous speech events such as false starts.

- Syntax-driven approach ➔ Semantic-driven approach

Natural Language Understanding

- How should the speech-recognition component interact with the natural language component in order to obtain the correct meaning representation?
  - N-best interface
  - Tightly coupled integration strategy

Discourse

- A discourse ability allows a conversational system to understand an utterance in the context of the previous interaction.

- Example:
  - “Show me only United flights” + “I want to go from Boston to Denver”

Output Processing: From Information to Signal

- Two Steps:
  - The information is converted into well-formed sentences.
  - Then the sentences are fed through a text-to-speech (TTS) system to generate the verbal responses.
- Concept-to-speech generation:
  - It produce higher quality output speech tan decoupled system, since it permits finer control of prosody.

Natural Language Generation

- Spoken language generation serves two important roles:
  - It provides a verbal response to the user's queries, essential in applications where visual displays are unavailable.
  - It can provide feedback to the user in the form of a paraphrase, confirming the system's proper understanding of the input query.

Speech Synthesis

- Rule-driven approach:
  - Intelligible but suffer in naturalness.

- Corpus-based approach:
  - Units excised from recorded speech are concatenated to form an utterance.
  - The selection of the units is based on a search procedure subject to a predefined distortion measure.

Dialogue Management

- In the early stages of the conversation:
  - Gather information from users ➔ complete query
  - Retrieve database.
- In the later stages of the conversation:
  - Might be involved in some negotiation with the user.
- Inform and guide the user by suggesting subsequent subgoals.

Recent Progress

Table 2 A Comparison of Several Conversational Systems that have been Deployed and Used by Real Users

Domain	Language	Vocabulary Size	Average Words/Utt	Average Utts/Dialogue
CSELT Train Timetable Info	Italian	760	1.6	6.6
SpeechWorks Air Travel Reservation	English	1000	1.9	10.6
Philips Train Timetable Info	German	1850	2.7	7.0
CMU Movie Information	English	757	3.5	9.2
CMU Air Travel Reservation	English	2851	3.6	12.0
LIMSI Train Timetable Info	French	1800	4.4	14.6
MIT Weather Information	English	1963	5.2	5.6
MIT Air Travel Reservation	English	1100	5.3	14.1
AT&T Operator Assistance	English	4000	7.0	3.0
Air Travel Reservations (human)	English	?	8.0	27.5

Development Issues

- Working in Real Domains
- Data Collection
- Evaluation

Working in Real Domains

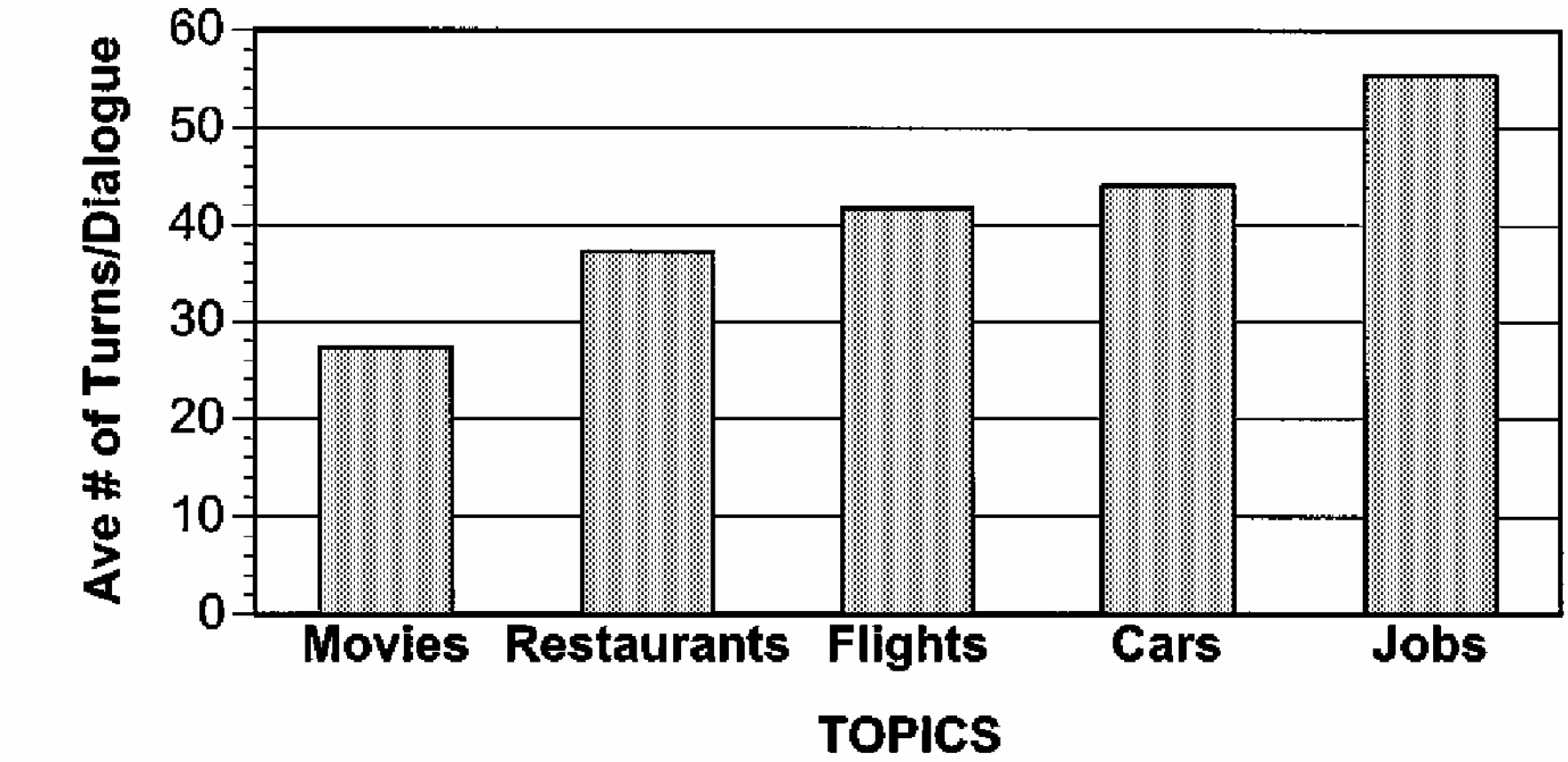

Fig. 4. Averaged number of dialogue turns for several application domains.

Data Collection

- Developing conversational interfaces is a classic chicken and egg problem.
  - In order to develop the system capabilities, one needs to have a large corpus of data for system development, training, and evaluation.
  - In order to collect data that reflect actual usage, one needs to have a system that users can speak to.

Data Collection

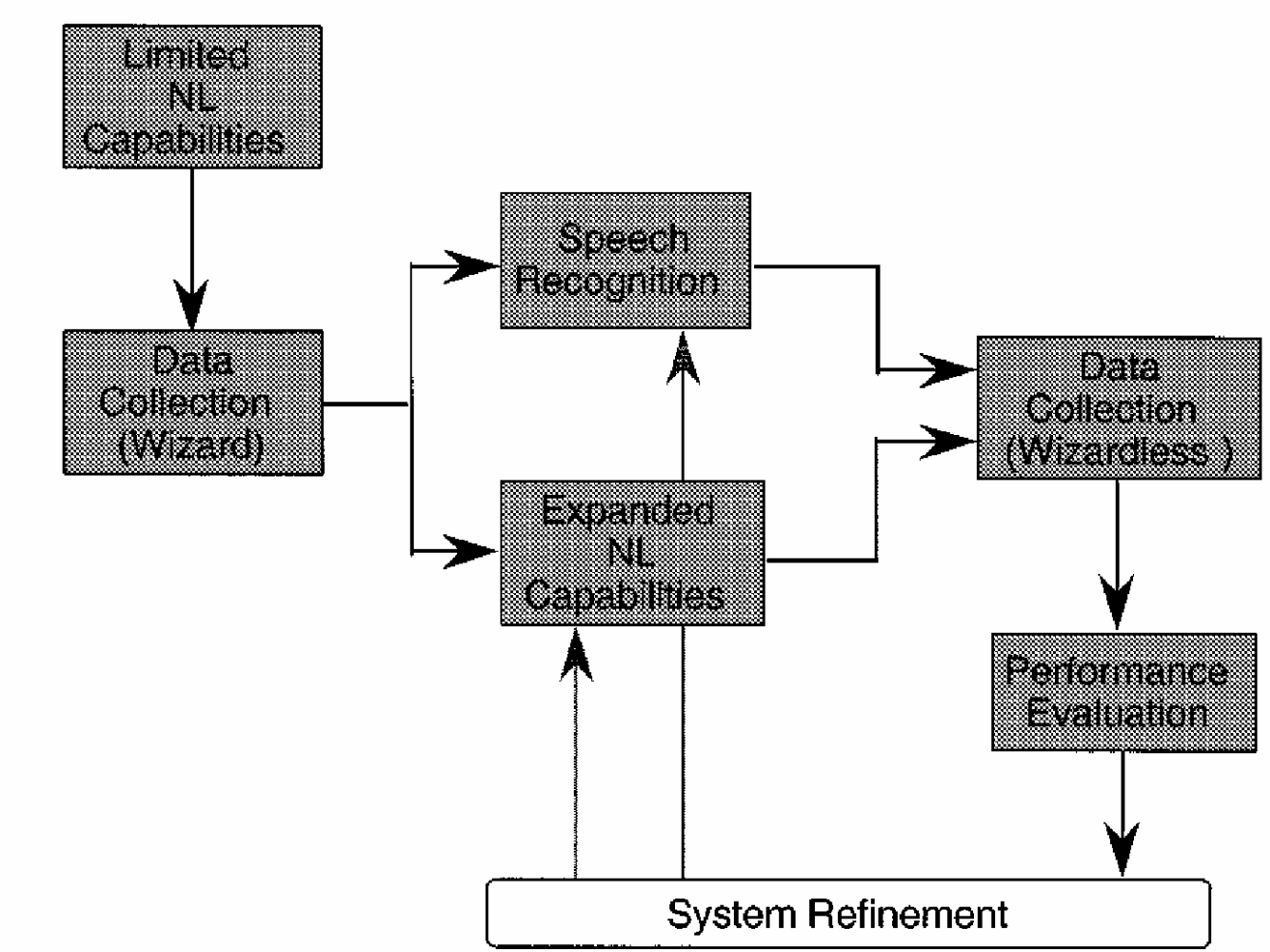

Fig. 5. Illustration of data collection procedures.

Data Collection

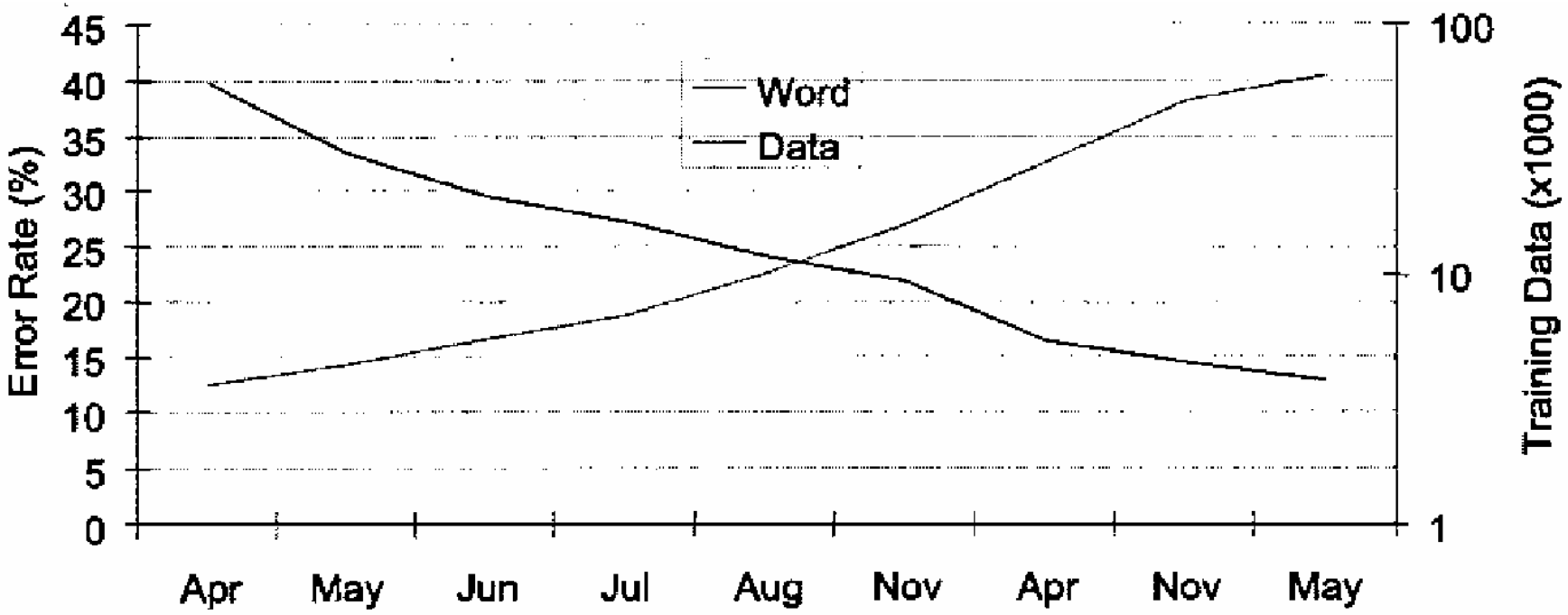

Fig. 6. Comparison of recognition performance and the number of utterances collected from real users over time in the MIT weather domain. Note that the x-axis has a nonlinear time scale, reflecting the time when new versions of the recognizer were released.

Evaluation

- Accompanying statistics
  - The length of time to complete the task
  - The number of turns
- Whether users liked the system.
- ASR and NLU evaluation
  - [ASR] Straightforward
  - [NLU] Comparing some form of meaning representation with a reference.
    - But there is no common meaning representation among different research siter.

Challenges

- Spoken Language Understanding
  - Speech Recognition : Robustness, Adaptation, …
  - Detection and Learning of new words.
  - Partial Parsing.
  - Keyword to meaning representation.
- Spoken Language Generation
  - Speech Synthesis : Prosody
- Dialogue Management
  - Outside the system's capabilities.
  - Misunderstandings → Robustness.
- Portability
  - On the commercial side, there has been a significant effort to develop the Voice eXtensible Markup Language(VoiceXML) as a standard to enable internet content and information access via voice and phone.

Conclusion

- Human language technology will play a central role in providing an interface that will dramatically change the human-human communication paradigm.