CONVERSATIONAL INTERFACES: ADVANCES AND CHALLENGES[1]

Victor Zue

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139 USA
zue@mit.edu

ABSTRACT

The last decade has witnessed the emergence of a new breed of human computer interfaces that combines several human language technologies to enable information access and transactional processing using spoken dialogue. In this paper, I discuss my view on the research issues involved in the development of such interfaces, describe the recent work done in this area at the MIT Laboratory for Computer Science, and outline some of the unmet research challenges, including the need to work in *real* domains, spoken language generation, and portability across domains and languages.

1. INTRODUCTION

Computers are fast becoming a ubiquitous part of our lives, brought on by their rapid increase in performance and decrease in cost. With their increased availability comes the corresponding increase in our appetite for information. Today, there are more than 600K web servers hosting in excess of 30M publicly accessible homepages, and the growth is continuing at an astronomical rate. One can now obtain a plethora of online data, ranging from New York Times stories to Dilbert trivia, and services, such as purchasing airline tickets and scheduling package pickups. Vast amounts of useful information are being made widely available, and people are utilizing it routinely for education, decision-making, finance, and entertainment.

The advent of the information age places increasing demands on technologists to provide "universal access." For information to be truly accessible to all – especially the technologically naive – anytime, anywhere, one must seriously address the problem of user interfaces. A promising solution to this problem is to impart human-like capabilities onto machines, so that they can speak and hear, just like the users with whom they need to interact. Spoken language is attractive because it is the most natural, efficient, flexible, and inexpensive means of communication among humans.

When one thinks about a speech-based interface, two technologies immediately come to mind: speech recognition and speech synthesis. There is no doubt that these are important and as yet unsolved problems in their own right, with a clear set of applications that include document preparation and audio indexing. However, many applications that lend themselves to spoken input/output – inquiring about weather or making travel arrangements – are in fact exercises in information access and/or interactive problem solving. The solution is often built up incrementally, with both the user and the computer playing active roles in the "conversation." Therefore, several language-based input and output technologies must be developed and integrated to reach this goal. The resulting *conversational interface* is the subject of this paper.

Many speech-based interfaces can be considered conversational, and they differ primarily in the degree with which the system maintains an active role in the conversation. In one extreme, the computer can take complete control of the interaction by requiring that the user answer a set of prescribed questions, much like the touch-tone implementation of interactive voice responses (IVR) systems. In the case of air travel planning, for example, the system could ask the user to "Please say just the departure city." Since the user's options are severely restricted, successful completion of such system-initiated transactions is easier to attain, and indeed some successful demonstration has been made [40]. But this may be accomplished at the cost of user annoyance due to its inflexibility. At the other extreme, the user can take total control of the interaction (e.g., "I want to visit my grandmother") while the system remains passive. In this case, the user may feel uncertain as to what capabilities exist, and may, as a consequence, stray quite far from the domain of competence of the system, leading to great frustration because nothing is understood. Instead, this paper is concerned with a *mixed-initiative* goal-oriented dialogue, in which both the user and the computer participate actively to solve a problem interactively using a conversational paradigm.

What is the nature of such mixed initiative interaction? One way to answer the question is to examine human-*human* interactions during joint problem solving. Figure 1 shows the transcript of a conversation between an agent (A) and a client (C) over the phone. As illustrated by this example, sponta-

[1] This research was supported by DARPA under contract N66001-94-C-6040, monitored through Naval Command, Control and Ocean Surveillance Center.

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.	

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

neous dialogue is replete with disfluencies, interruption, confirmation, clarification, ellipsis, co-reference, and sentence fragments. Some of the utterances cannot be understood properly without knowing the context in which they appear. As we shall see, while present systems cannot handle *all* these phenomena satisfactorily, some of them are being dealt with in a limited fashion.

Should one build conversational interfaces by mimicking human-human interactions? Opinion in this regard is somewhat divided. Some researchers argue that human-human dialogues can be quite variable, containing frequent interruptions, speech overlaps, incomplete or unclear sentences, incoherent segments, and topic switches. Some of these variabilities may not contribute directly to goal-directed problem solving. However, one may argue that users could feel more comfortable with an interface that posseses some of the characteristics of a human agent. In our case and to the extent possible, we have taken the approach of developing a human-machine interface based on analyses of human-human interactions when solving the same tasks. Regardless of the approach, we believe, as do others, that studying human-human dialogue and comparing it to human-machine dialogue can provide valuable insights [2]. As an example, consider the histograms of the lengths of the utterances per turn for agents and clients shown in Figure 2. The statistics were gathered from the transcripts of over 100 hours of conversation, in more than 1000 interactions, between agents and clients over the phone on a variety of information access tasks. Over 80% of the clients' utterances are 12 words or less, with a preponderance of very short utterances. Closer examination of the data reveals that these short utterances are mostly back channel communications.

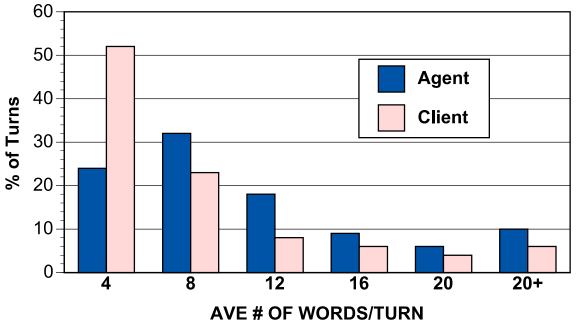


Figure 2: Histograms of utterance length for agents and clients in tasks of information access over the phone.

The last decade has witnessed the emergence of some conversational systems with limited capabilities. Despite our moderate success, the ultimate deployment of such interfaces will require continuing improvement of the core human language technologies and the exploration into many uncharted research territories. The purpose of this paper is to outline some of these new research challenges. To set the stage, I will first briefly introduce the components of a spoken language system, and discuss some of the research issues. I will then provide a thumb-nail sketch of the recent landscape, drawing heavily from my own experience in developing such systems at MIT over the past eight years. Interested readers are referred to the recent proceedings of the Eurospeech Conference, the International Conference of Spoken Language Processing, the International Conference of Acoustics, Speech, and Signal Processing, the International Symposium on Spoken Dialogue, and other relevant publications (e.g., [8]).

2. RESEARCH ISSUES

2.1. System Architecture

Figure 3 shows the major components of a typical conversational interface. The spoken input is first processed through the speech recognition component. The natural language component, working in concert with the recognizer, produces a meaning representation. For information retrieval applications illustrated in this figure, the meaning representation can be used to retrieve the appropriate information in the form of text, tables and graphics. If the information in the utterance is insufficient or ambiguous, the system may choose to query the user for clarification. Natural language generation and text-to-speech synthesis are then used to produce spoken responses that may serve to clarify the tabular information. Throughout the process, discourse information is maintained and fed back to the speech recognition and language understanding components, so that sentences can be properly understood in context.

Figure 3 does not convey the notion that a conversational interface may include input and output modal-

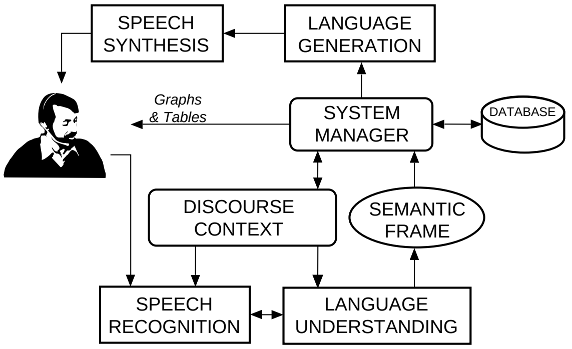


Figure 3: A generic block diagram for a typical conversational interface.

ities other than speech. While speech may be the interface of choice, as is the case with phone-based transactions and hands-busy/eyes-busy settings, human communication is inherently multimodal, employing facial, gestural, and other cues to convey the underlying linguistic message. In this view, speech interfaces should be complemented by visual and sensory motor channels. The user should be able to choose among many modalities, including gesturing, pointing, writing, and typing on the input side [26, 38], and graphics and a talking head on the output side [22], to achieve the task in hand in the most natural and efficient manner.

The development of conversational interfaces offers a set of significant challenges to speech and natural language researchers, and raises several important research issues, three of which will be discussed in the remainder of this section.

2.2. Spoken Language Understanding

Spoken language understanding involves the transformation of the speech signal into a meaning representation that can be used to interact with the specific application back-end. This is typically accomplished in two steps, the conversion of the signal to a set of words (i.e., speech recognition), and the derivation of the meaning from the word hypotheses (i.e., language understanding).

2.2.1. Speech Recognition

Historically, speech recognition systems have long been developed with the assumption that the speech material is read from prepared text. Input to conversational interfaces, however, is typically generated extemporaneously, containing disfluencies (i.e., unfilled and filled pauses such as "umm" and "aah," as well as word fragments) and words outside the system's working vocabulary. Thus far, some attempts have been made to deal with these problems. For example, researchers have improved their system's recognition performance by introducing explicit acoustic models for the filled pauses [42, 5]. Similarly, "trash" models have been introduced to detect the presence of unknown words, and procedures have been devised to learn the new words once they have been detected [1].

An issue that is receiving increasing attention by the research community is the recognition of telephone quality speech. It is highly likely that the first several conversational interfaces to become available to the general public will be accessible via telephone, in many cases replacing presently existing IVR systems. Telephone quality speech is significantly more difficult to recognize than high quality recordings, both because of the limited bandwidth and the noise and distortions introduced in the channel.

2.2.2. Language Understanding

Speech recognition systems typically implement linguistic constraints as a statistical grammar (i.e., n-gram) that specifies the probability of a word given its predecessors. While these language models have been effective in reducing the search space and improving performance, they do not begin to address the issue of speech understanding. On the other hand, most natural language systems are developed with text input in mind; it is usually assumed that the entire word string is known with certainty. This assumption is clearly false for speech input, where many words are competing for the same time span (e.g., "euthanasia" and "youth in Asia,") and some words may be more reliable than others because of varying signal robustness. Furthermore, spoken language is often agrammatical, containing fragments, disfluencies and partial words. Language understanding systems designed for text input may have to be modified in fundamental ways to accommodate spoken input.

Natural language analysis has traditionally been predominantly syntax-driven – a complete syntactic analysis is performed which attempts to account for *all* words in an utterance. However, when working with spoken material, researchers quickly came to realize that such an approach [4, 34] can break down dramatically in the presence of unknown words, novel linguistic constructs, recognition errors, and spontaneous speech events such as false starts.

Due to these problems, many researchers have tended to favor more semantic-driven approaches, at least for spoken language tasks in limited domains. In such approaches, a meaning representation is derived by "spotting" key words and phrases in the utterance [43]. While this approach loses the constraint provided by syntax, and may not be able to adequately interpret complex linguistic constructs, the need to accommodate spontaneous speech input has outweighed these potential shortcomings. At the present time, almost all viable systems have abandoned the notion of achieving a complete syntactic analysis of every input sentence, favoring a more robust strategy that can still answer when a full parse is not achieved [19, 35, 39]. This can be accomplished by identifying parsable phrases and clauses, and providing a separate mechanism for gluing them together to form a complete meaning analysis [35]. Ideally, the parser includes a probabilistic framework with a smooth transition to parsing fragments when full linguistic analysis is not achievable. Examples of systems that incorporate such *stochastic* modelling techniques can be found in [30, 24].

2.2.3. SR/NL Integration

How should the speech recognition component interact with the natural language component in order to obtain the correct meaning representation? At present, the most popular strategy is the so-called N-best interface [7], in which the recognizer proposes its best N complete sentence hypotheses one by one, stopping with the first sentence that is successfully analyzed by the natural language component. In this case, the natural language component acts as a filter on *whole sentence* hypotheses.

In the N-best interface, many of the candidate sentences may differ minimally in regions where the acoustic information is not very robust. While confusions such as “an” and “and” are acoustically reasonable, one of them can often be eliminated on linguistic grounds. In fact, many of the top N sentence hypotheses could have been eliminated before reaching the end if syntactic and semantic analyses had taken place early on in the search. One possible solution, therefore, is for the speech recognition and natural language components to be tightly coupled, so that only the acoustically promising hypotheses that are linguistically meaningful are advanced. For example, partial theories can be arranged on a stack, prioritized by score. The most promising partial theories are extended using the natural language component as a predictor of all possible next-word candidates; none of the other word hypotheses are allowed to proceed. Therefore, any theory that completes is guaranteed to parse. Researchers are beginning to find that such a tightly coupled integration strategy can achieve higher performance than an N-best interface, often with a considerably smaller stack size [15, 13, **?**, 25]. The future is likely to see increasing instances of systems making use of linguistic analysis at early stages in the recognition process.

2.3. Spoken Language Generation

On the output side, a conversational interface must be able to convey the information to the user in natural sounding sentences. This is typically accomplished in two steps: the information is converted into well formed sentences, which are then fed through a text-to-speech (TTS) system to generate the verbal responses.

Spoken language generation serves two important roles: it provides a verbal response to the user’s queries and it can also provide a paraphrase of the user’s input, which can serve as a confirmation of the system’s proper understanding of the input query. Research in language generation for conversational systems has not received nearly as much attention as has language understanding, perhaps due to the funding priorities set forth by the major government sponsors. The language generation component of a conversational system typically produces the response one sentence at a time, without paragraph level planning. One effective approach for sentence generation is to concatenate templates after filling slots by applying recursive rules along with appropriate constraints (person, gender, number, etc.) [10].

Currently, the language generation and text-to-speech components on the output side of conversational systems are not closely coupled; the same text is generated whether it is to be read or spoken. Furthermore, current systems typically expect the language generation component to produce a textual surface form of a sentence (throwing away valuable linguistic and prosodic knowledge) and then require the text-to-speech component to produce linguistic analysis anew. Clearly, these two components would benefit from a shared knowledge base.

2.4. Discourse and Dialogue

Human verbal communication is a two-way process involving multiple, active participants. Mutual understanding is achieved through direct and indirect speech acts, turn taking, clarification, and pragmatic considerations. An effective conversational interface for information retrieval and interactive transactions must incorporate extensive and complex dialogue modelling – initiating appropriate clarification sub dialogues based on partial understanding, and taking an active role in directing the conversation towards a valid conclusion. Although there has been some theoretical work on the structure of human-human dialogue [16], this has not yet led to effective insights for building human-machine interactive systems.

It is essential that a system be able to interpret a user’s queries in context. For instance, if the user says, “I want to go from Boston to Denver,” followed with, “show me only United flights,” they clearly don’t want to see *all* United flights, but rather just the ones that fly from Boston to Denver. The ability to inherit information from preceding sentences is particularly helpful in the face of recognition errors. The user may have asked a complex question involving several restrictions, and the recognizer may have misunderstood a single word, such as a flight number or an arrival time. If a good context model exists, the user can now utter a very short correction phrase, and the system will be able to replace just the misunderstood word, preventing the user from having to re-utter the entire sentence, running the risk of further recognition errors.

3. RECENT LANDSCAPE

3.1. Overview

Conversational systems are a relatively new technology, having first come into existence in the late 1980’s as a result of two major government-funded efforts on both sides of the Atlantic: the DARPA Spoken Language Systems (SLS) program in the United States and the Esprit SUNDIAL (Speech UNderstanding and DIALog) program in Europe [29]. These two programs were remarkably parallel in that both involved database access for travel planning, with the European one including both flight and train schedules, and the American one being restricted to air travel. The European program was a multilingual effort involving four languages (English, French, German, and Italian), whereas the American effort was, understandably, restricted to English. All of the systems focused within a narrowly defined area of expertise, and vocabulary sizes are generally limited to several thousand words. Nowadays, these types of systems can typically run in real-time on standard workstations and PCs with no additional hardware.

The DARPA-SLS program cannot be considered conversational in that its attention focused entirely on the input side. The Program adopted the approach of developing the underlying input technologies within a common domain called Air Travel Information Service, or ATIS [31]. ATIS permits users to verbally query for air travel information, such as flight schedules from one city to another, obtained from a small relational database excised from the Official Airline Guide. By requiring that all system developers use the same database, it has been possible to compare the performance of various spoken language systems based on their ability to extract the correct information from the database, using a set of prescribed training and test data, and a set of interpretation guidelines. Indeed, common evaluations have occurred at regular intervals, and steady performance improvements have been observed for all systems. At the end of the Program in 1995, the best system achieved a word error rate of 2.3% and a sentence error rate of 15.2% [27]. Additionally, the best system achieved an understanding error rate of 5.9% and 8.9% for text and speech input, respectively.[2]

Whereas the DARPA-SLS program emphasized competition through periodic common evaluations, the European SUNDIAL program promoted cooperation and plug compatibility by requiring different sites to contribute distinct components to a single multisite system. More significantly, the Program designated dialogue modelling and spoken language generation as integral parts of the research program. As a result, the emphasis on dialogue in Europe led to some interesting advances in dialogue control mechanisms. While the program terminated in 1993, some of the systems it spawned have continued to flourish. A notable example is the Philips Automatic Train Timetable Information System [9], which is capable of communicating with the user solely by voice over the telephone. The system has a vocabulary of about 1800 words, 1200 of which are distinct railway station names. The dialogue relies heavily on confirmation requests to permit correction of recognition errors, but the overall success rate for usage is very high.

There are several conversational systems that fall outside the two government sponsored programs mentioned above. The Office Manager system developed at the Carnegie Mellon University is designed to provide users with voice access to a set of application programs for the office of the future [32]. The Berkeley Restaurant Project (BeRP) [20], developed at the University of California acts as a restaurant guide in the Berkeley area. Another novel system is WAXHOLM, developed by researchers at Royal Institute of Technology in Sweden [3]. WAXHOLM provides timetables for ferries in the Stockholm archipelago, as well as port locations, hotels, camping sites, and restaurants that can be found on the islands. The WAXHOLM developers designed a flexible, easily-controlled dialogue module, based on a scripting language that describes dialogue flow [6].

3.2. The MIT Experience

3.2.1. Early Systems

Since 1989, our group has been conducting research leading to the development of prototypical conversational systems. The first such system, VOYAGER, can engage in verbal dialogues with users about a restricted geographical region within Cambridge, Massachusetts, in the USA. The system can provide information about distances, travel times, or directions between landmarks located within this area (e.g., restaurants, hotels, banks, libraries, etc.) as well as handling specific requests for information such as address, phone number or location on the map. VOYAGER served as our primary platform for developing multilingual systems, culminating in the demonstration of trilingual (in English, Italian, and Japanese) capabilities in 1994 [11].

An interesting realistic system which grew out of the ARPA ATIS effort is the PEGASUS system [45], which is connected via a modem over the phone line to a real flight reservation system. PEGASUS has knowledge of flights to and from some 220 cities worldwide. It has a fairly extensive dialogue model to help cope with difficult problems such as date restrictions imposed by discount fares or aborted flight plans due to selections being sold out. A subsequent *displayless* version of PEGASUS enables users to make flight reservations by speaking with a computer over the telephone [36].

3.2.2. GALAXY

In 1994, researchers at MIT started the development of GALAXY [14, 46], an architecture that enables universal information access using spoken dialogue. GALAXY distinguishes itself from other conversational systems in several respects. First, it is

[2] All the performance results quoted here are for the so-called "evaluable" queries, i.e., those queries that are within domain and for which an appropriate answer is available from the database.

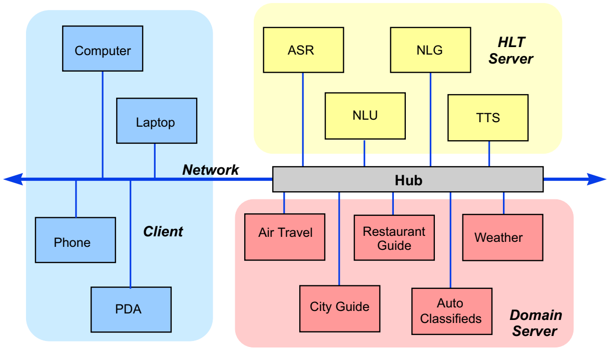


Figure 4: Architecture of GALAXY.

distributed and decentralized – GALAXY uses a client-server architecture to allow sharing of computationally expensive processes (such as large vocabulary speech recognition), as well as knowledge intensive processes. Second, it is multi-domain, intended to provide access to a wide variety of information sources and services while insulating the user from the details of database location and format. It is presently connected to many real, on-line databases, including the National Weather Services, the NYNEX Electronic Yellow Pages, and the World Wide Web. Users can query GALAXY in natural English (e.g., "what is the weather forecast for Miami tomorrow," "how many hotels are there in Boston," and "do you have any information on Switzerland," etc.), and receive verbal and visual responses. Third, it is extensible; new knowledge domain servers can be added to the system incrementally. Finally, GALAXY is mobile; it can be launched anywhere in the world using an ordinary Web browser for display and a telephone for speech input/output [21].

The GALAXY architecture has been used extensively in our group as the testbed for developing conversational interfaces and the underlying human language technology. For example, multilingual capabilities have been developed for Spanish and for Mandarin Chinese [41]. In addition, two other applications have been developed; one is an interface to a database of electronic automobile classified advertisements [23], and the other is a restaurant guide for the Boston area [37].

3.2.3. JUPITER

The most recent descendent of GALAXY is a system we call JUPITER [47]. Jupiter is a telephone-only conversational interface for weather information for more than 500 cities worldwide. The weather information is obtained from four on-line sources on the Web, and is updated several times daily. Jupiter employs GALAXY's client-server architecture, except the client is simply a telephone. It serves as a platform for investigating several research topics. First, by using the telephone as a means of accessing the information, we can empower a much larger population to access the wide range of information that is becoming available. In the scenario that we envision, a user could conduct "virtual browsing" in the information space without ever having to point or click. Second, displayless information access poses new challenges to conversational interfaces. If the information can only be conveyed verbally, the system must rely on the dialogue component to reduce the information to a digestible amount, the language generation component to express the information succinctly, and the TTS component to generate highly natural and intelligible speech. Third, channel distortions place heavy demands on the system to achieve robust speech recognition and understanding. Finally, by applying human language technologies to understanding the "content," in this case the weather forecast, we can manipulate and deliver exactly the information that the user wants, no more and no less.

For speech recognition, JUPITER makes use of the SUMMIT speech recognition system developed in our group, modified for telephone input [12]. TINA [34], a probabilistic language understanding system, parses word hypotheses and eventually creates a semantic frame representation which can be paraphrased into an SQL query. Finally, GENESIS [10] is used for generating sentence responses from the database, and the resulting sentences are played through a commercial off-the-shelf TTS system. Currently, JUPITER has a vocabulary of nearly 1500 words.

Since May 1997, we have installed a toll free telephone number, so that naive users can call JUPITER and inquire about the weather. Log-files are created automatically, and the sentences are also transcribed as they come in. Thus far, we have collected over 8000 sentences from nearly 1500 calls. The data have proved to be invaluable for system development, training, and evaluation. Figure 5 is an example dialogue between a naive user and JUPITER. A comparison of this dialogue with that shown in Figure 1 suggests that JUPITER possesses some of the same characteristics as a helpful human agent.

While JUPITER has only been in existence for two months at this writing, we nevertheless have started to monitor its performance longitudinally. A recent evaluation made use of 502 sentences in 82 randomly selected calls collected from naive users using land-line, cellular, and speaker phones. Approximately 70% of the sentences contain words within the system's vocabulary. Approximately 11% of the sentences contain out-of-vocabulary words, largely uttered by curiosity seekers (e.g., "Do you have transistors between your ears?"), and a similar percentage contain no speech (laughters, hang-ups, etc.) Approximately 5% of the sentences were clipped, since JUPITER currently does not have barge-in capability, and 3% of the sentences contain mis-pronounced words. Table 1 summarizes JUPITER's performance on the in-vocabulary sentences. While the system recognizes less than 2/3 of the sentences correctly, it nevertheless understood nearly 4 out of every five sentences correctly.

U: WHAT IS THE WEATHER IN PASADENA CALIFORNIA

J: I know of the following places in California: Burbank, Fresno, Long Beach, Los Angeles, Oakland, Sacramento, San Diego, San Francisco, and San Jose. Please select one.

U: BURBANK

J: In Burbank today: areas of low clouds morning and fog, otherwise sunny. Tonight: clear. Is there something else I can do for you?

U: IS IT GOING TO RAIN IN NEW JERSEY

J: To my knowledge, the forecast calls for no rain in New Jersey today. What else would you like to know?

U: WILL IT RAIN IN WASHINGTON D C THIS WEEKEND

J: To my knowledge, the forecast calls for no rain in Washington this weekend Saturday and Sunday. What other information can I give you?

U: HOW ABOUT THE FORECAST FOR PARIS

J: I have no knowledge of the weather in Paris this weekend Saturday and Sunday. Is there anything else?

U: NO THANKS

J: Thanks for using Jupiter! Have a nice day!

Figure 5: Actual transcript of a conversation between a user (U) and Jupiter (J) over the phone. Typical conversational phenomena are annotated on the right.

CATEGORY	ERROR RATE (%)
Word Recognition	16.1
Sentence Recognition	37.0
Understanding (from text)	5.4
Understanding (from speech)	22.0

Table 1: Performance evaluation of JUPITER for the in-vocabulary sentences collected from naive users.

4. FUTURE CHALLENGES

As we can see, considerable progress has been made over the past decade in research and development of systems that can understand and respond to spoken language. To meet the challenges of developing a language-based interface to help users solve real problems, however, we must continue to improve the core technologies while expanding the scope of the underlying human language technology base. In this section, we outline some of the new research challenges that I believe have heretofore received insufficient attention.

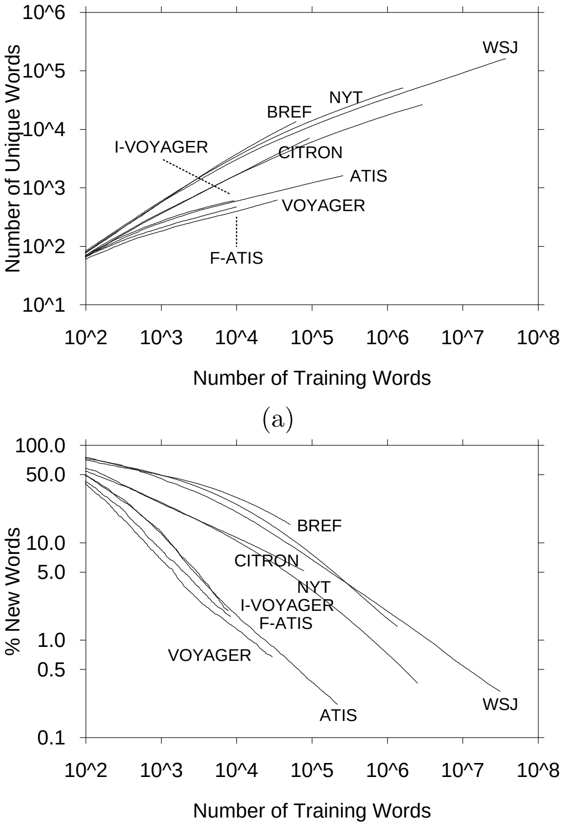
4.1. Working in Real Domains

The rapid technological progress that we are witnessing raises several timely questions. When will this technology be available for productive use? What technological barriers still exist that will prevent large-scale deployment? An effective strategy for answering these questions is to develop the underlying technologies within *real* applications, rather than relying on mock-ups, however realistic they might be. Such a strategy will force us to confront some of the critical technical issues that may otherwise elude our attention. Consider, for example, the task of accessing information in the Yellow Pages of a medium-sized metropolitan area. The vocabulary size of such a task could easily exceed 100,000, considering the names of the establishments, street and city names, and listing headings. A task involving such a huge vocabulary presents a set of new technical challenges. Among them are:

- How can adequate acoustic and language models be determined when there is little hope of obtaining a sufficient amount of domain-specific data for training?
- What search strategy would be appropriate for very large vocabulary tasks? How can natural language constraints be utilized to reduce the search space while providing adequate coverage?
- How can the application be adapted and/or customized to the specific needs of a given user?
- How can the system be efficiently ported to a different task in the same domain (e.g., changing the geographical area from one city to another), or to an entirely different domain (e.g., library information access)?

There are many other research issues that will surface when one is confronted with the need to make human language technology truly useful for solving real problems. Consider, for example, the unknown word problem. The traditional approach to spoken language recognition and understanding research and development is to define the working vocabulary based on domain-specific corpora. However, experience has shown that, no matter how large the size of the training corpora, the system will invariably encounter previously unseen words [17]. This is illustrated in Figure 6. For the ATIS task, for example, a 100,000-word training corpus will yield a vocabulary of about 1,000 words. However, the probability of the system encountering an unknown word, is about 0.002. Assuming that an average sentence contains 10 words, this would mean that approximately one in 50 sentences will contain an unknown word.

In a *real* domain such as Jupiter or the Electronic Yellow Pages, a much larger fraction of the words uttered by users will not be in the system's working vocabulary. This is unavoidable partly because it is not possible to anticipate all the words that all users are likely to use, and partly because the database is usually changing with time (e.g., new restaurants opening up). In the past, we have not paid much attention to the unknown word problem because the tasks we have chosen assume a closed vocabulary. In the limited cases where the vocabulary has been open, unknown words have accounted for a small fraction of the word tokens in the test corpus. Thus researchers could either construct generic "trash word" models and hope for the best, or ignore the unknown word problem altogether and accept a small penalty on word error rate. In real applications, however, the system must be able to cope with unknown words



(a)

(b)

Figure 6: (a) The number of unique words (i.e., task vocabulary) as a function of the size of the training corpora, for several spoken language tasks, and (b) The percentage of unknown words in previously unseen data as a function of the size of the training corpora used to determine the vocabulary empirically. The sources of the data are: F-ATIS=French ATIS; I-VOYAGER=Italian VOYAGER; BREF=French La Monde; NYT=New York Times; WSJ=Wall Street Journal; and CITRON=Directory Assistance.

simply because they will always be present, and ignoring them will not satisfy the user's needs – if a person wants to know how to go from the train station to Lucia's restaurant, they will not settle for a response such as, "I am sorry I don't understand you. Please rephrase the question." The system must be able not only to *detect* new words, taking into account acoustic, phonological, and linguistic evidence, but also to adaptively *acquire* them, both in terms of their orthography and linguistic properties. In some cases, fundamental changes in the problem formulation and search strategy may be necessary.

Aside from providing the technological impetus, however, working within real domains also has some practical benefits. While years may pass before we can develop unconstrained spoken language systems, we are fast approaching a time when systems with limited capabilities can help users interact with computers with greater ease and efficiency. Working on real applications thus has the potential benefit of shortening the interval between technology demonstration and its ultimate use. Besides, applications that can help people solve problems *will* be used by real users, thus providing us with a rich and continuing source of useful data, as we have discovered in our experience with JUPITER development.

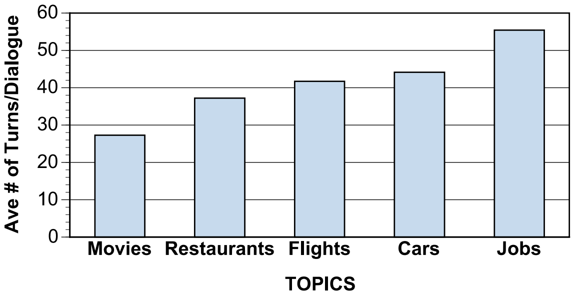


Figure 7: Averaged number of dialogue turns for several application domains.

How do we select the applications that are well matched to our present capabilities? Again, I believe the answer may lie in examining human-human data. Figure 7 displays the average number of dialogue turns per transaction for several application domains. The data are obtained from the same transcription of the 100 hours of real human-human interactions described earlier. As the data clearly show, helping a user select a movie or a restaurant is considerably less complex than helping a user to look for employment.

4.2. Spoken Language Generation

With few exceptions [11, 45, 28], current research in spoken language systems has focused on the input side, i.e., the understanding of the input queries, rather than the *conveyance* of the information.

Spoken language generation is an extremely important aspect of the human-computer interface problem, especially if the transactions are to be conducted over a telephone. Models and methods must be developed that will generate natural sentences appropriate for spoken output, across many domains and languages. In many cases, particular attention must be paid to the interaction between language generation and dialogue management – the system may have to initiate clarification dialogue to reduce the amount of information returned from the back-end, in order not to generate unwieldy verbal responses. On the speech side, recent work in synthesis based on non-uniform units has resulted in much improved synthetic speech quality [33, 18]. However, we must continue to improve speech synthesis capabilities, particularly with regard to the encoding of prosodic and paralinguistic information such as emotion. As is the case on the input side, we must also develop integration strategies for language generation and speech synthesis. Finally, evaluation methodologies for spoken language generation technology must be developed, and comparative evaluation performed.

4.3. Portability

Currently, the development of speech recognition and language understanding technologies has been domain specific, requiring a large amount of annotated training data. However, it may be costly, or even impossible, to collect a large amount of training data for certain applications, such as Yellow Pages.

Therefore, we must address the problems of producing a spoken language system in a new domain given at most a small amount of domain-specific training data. To achieve this goal, we must strive to cleanly separate the algorithmic aspects of the system from the application-specific aspects. We must also develop automatic or semi-automatic methods for acquiring the acoustic models, language models, grammars, semantic structures for language understanding, and dialogue models required by a new application. The issue of portability spans across different acoustic environments, databases, knowledge domains, and languages. Real deployment of spoken language technology cannot take place without adequately addressing this issue.

5. CONCLUDING REMARKS

In this paper, I have attempted to outline some of the important research challenges that must be addressed before spoken language technologies can be put to productive use. The timing for the development of human language technology is particularly opportune, since the world is mobilizing to develop the information highway that will be the backbone of future economic growth. Human language technology will play a central role in providing an interface that will drastically change the human-machine communication paradigm from *programming* to *conversation*. It will enable users to efficiently access, process, manipulate, and absorb a vast amount of information. While much work needs to be done, the progress made collectively by the community thus far gives us every reason to be optimistic about fielding such systems, albeit with limited capabilities, in the near future.

6. ACKNOWLEDGEMENTS

The research described in this paper is the result of collaboration with many past and current students and staff of the Spoken Language Systems Group at the MIT Laboratory for Computer Science, especially Stephanie Seneff, Jim Glass, Joe Polifroni, and Giovanni Flammia. Their contributions are gratefully acknowledged.

7. REFERENCES

[1] Asadi, A., Schwartz, R., and Makhoul, J. "Automatic Modelling for Adding New Words to a Large Vocabulary Continuous Speech Recognition System," *Proc. ICASSP*, 305-308, 1991.

[2] Bernsen, N. O., Dybkjaer, L. and Dybkjaer, H. "Cooperativity in Human-Machine and Human-Human Spoken Dialogue," *Discourse Processes*, Vol. 21, No. 2, 213-236, 1996.

[3] Blomberg, M., Carlson, R., Elenius,K., Granstrom, B., Gustafson, J., Hunnicutt, S., Lindell, R., and Neovius, L. "An Experimental Dialogue System: Waxholm," *Proc. Eurospeech*, 1867-1870, 1993.

[4] Bobrow, R., Ingria, R., and Stallard, R. "Syntactic and Semantic Knowledge in the DELPHI Unification Grammar," *Proc. DARPA Speech and Natural Language Workshop*, 230-236, 1990.

[5] Butzberger, J., Murveit, H., and Weintraub, M. "Spontaneous Speech Effects in Large Vocabulary Speech Recognition Applications," *Proc. ARPA Workshop on Speech and Natural Language*, 339-344, 1992.

[6] Carlson, R. and Hunnicutt, S. "Generic and Domain-Specific Aspects of the Waxholm NLP and Dialogue Modules," *Proc. ICSLP*, 677-680, 1996.

[7] Chow, Y., and Schwartz, R. "The N-Best Algorithm: An Efficient Procedure for Finding Top N Sentence Hypotheses," *Proc. ARPA Workshop on Speech and Natural Language*, 199-202, 1989.

[8] Dalsgaard, P., Larsen L.B., B., and Thomsen, I. (*Eds.*) *Proceedings of the ESCA Tutorial and Research Workshop on Spoken Dialogue Systems: Theory and Application*, Vigsø, Denmark, 1995

[9] Eckert, W., Kuhn, T., Niemann, H., Rieck, S., Scheuer, A., and Schukat-Talamazzini, E.G. "A Spoken Dialogue System for German Intercity Train Timetable Enquiries," *Proc. Eurospeech*, 1871-1874, 1993.

[10] Glass, J., Polifroni, J., and Seneff, S. "Multilingual Language Generation Across Multiple Domains," *Proc. ICSLP '94*. 983–976, 1994.

[11] Glass, J., Flammia, G., Goodine, D. Phillips, M., Polifroni, J. Sakai, S., Seneff, S., and Zue, V. "Multilingual Spoken-language Understanding in the MIT Voyager System," *Speech Communication*. 17 1–18, 1995.

[12] Glass, J., Chang, J., and McCandless, M. "A probabilistic framework for feature-based speech recognition," In *Proc. ICSLP*, 2277-2280, 1996.

[13] Goddeau, D. "Using Probabilistic Shift-Reduce Parsing in Speech Recognition Systems," *Proc. ICSLP*, 321–324, 1992.

[14] Goddeau, D., Brill, E., Glass, J., Pao, C., Phillips, M., Polifroni, J. Seneff, S. and Zue, V. "Galaxy: A Human-Language Interface to On-Line Travel Information," *Proc. ICSLP*, 707–710, 1994.

[15] Goodine, D., Seneff, S., Hirschman, L., Phillips, M. "Full Integration of Speech and Language Understanding in the MIT Spoken Language System," *Proc. Eurospeech*, 845–848, 1991.

[16] Grosz, B., and Sidner, C. "Plans for Discourse," in *Intentions in Communication*. MIT Press, 1990.

[17] Hetherington, I.L., and Zue, V. "New Words: Implications for Continuous Speech Recognition," *Proc. Eurospeech*, 475-931, 1991.

[18] Huang, X. "WHISTLER: A Trainable Text-toSpeech System," *Proc. ICSLP*, 2387-2390.

[19] Jackson, E., Appelt, D., Bear, J., Moore, R., and Podlozny, A. "A Template Matcher for Robust NL Interpretation," *Proc. DARPA Speech and Natural Language Workshop*, 190–194, 1991.

[20] Jurafsky, D., Wooters, C., Tajchman, G., Segal, J., Stolcke, A., Fosler, E., and Morgan, N. "The Berkeley Restaurant Project," Proc. ICSLP, 2139-2142, 1994.

[21] Lau, R., Flammia, G., Pao, C., and Zue, V. "WebGalaxy - Integrating Spoken Language and Hypertext Navigation," *These Proceedings.*

[22] Massaro, D. *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*, MIT Press, (to appear).

[23] Meng, H., Busayapongchai, S., Glass, J., Goddeau, D., Hetherington, L., Hurley, E., Pao, C., Polifroni, J., Seneff, S., and Zue, V. "A Conversational System in the Automobile Classifieds Domain," *Proc. ICSLP*, 542-545, 1996.

[24] Miller, S., Schwartz, R., Bobrow, R., and Ingria, R., "Statistical Language Processing Using Hidden Understanding Models," *Proc. ARPA Speech and Natural Language Workshop*, 278-282, 1994.

[25] Moore, R., Appelt, D., Dowding, J. Gawron, J., and Moran, D. "Combining Linguistic and Statistical Knowledge Sources in Natural-Language Processing for ATIS," *Proc. ARPA Spoken Language Systems Workshop*, 261–264, 1995.

[26] Oviatt, S. L. "Multimodal Interfaces for Dynamic Interactive Maps", *Proc. Conference on Human Factors in Computing Systems: CHI '96*, 95-102, 1996.

[27] Pallett, D., Fiscus, J., Fisher, W., and Garafolo, J., Lund, B., Martin, A., and Pryzbocki, M. "Benchmark Tests for the ARPA Spoken Language Program," *Proc. ARPA Spoken Language Systems Technology Workshop*, 5-36, 1995.

[28] Pan, S. and McKeown, K. "Spoken Language Generation in a Multimedia System," *Proc. ICSLP*, 374-377, 1996.

[29] Peckham J. "A New Generation of Spoken Dialogue Systems: Results and Lessons from the SUNDIAL Project," *Proc. Eurospeech*, 33-40, 1992.

[30] Pieraccini, R., Levin, E. and Lee, C.H. "Stochastic Representation of Conceptual Structure in the Atis Task," *Proc. DARPA Speech and Natural Language Workshop*. 121-124, 1992.

[31] Price, P. "Evaluation of Spoken Language Systems: the Atis Domain," *Proc. DARPA Speech and Natural Language Workshop*, 91-95, 1990.

[32] Rudnicky, A., Lunati, J-M., and Franz, A. "Spoken Language Recognition in an Office Management Domain," *Proc. ICASSP*, 829–832, 1991.

[33] Sagisaka, Y., Kaiki, N., Iwahashi, N., and Mimura, K. "ATR ν-Talk Speech Synthesis System," *Proc. ICSLP*, 483-486, 1992.

[34] Seneff, S. "Tina: A Natural Language System for Spoken Language Applications," *Computational Linguistics*, Vol. 18, No. 1, 61-86, 1992.

[35] Seneff, S. "Robust Parsing for Spoken Language Systems," *Proc. ICASSP*. 189-192, 1992.

[36] Seneff, S., Zue, V. Polifroni, J., Pao, C., Hetherington, L., Goddeau, D., and Glass, J. "The Preliminary Development of a Displayless Pegasus System," *Proc. ARPA Spoken Language Technology Workshop*, 212-217, 1995.

[37] Seneff, S. and Polifroni, J. "A New Restaurant Guide Conversational System: Issues in Rapid Prototyping for Specialized Domain," *Proc. ICSLP*, 665-668, 1996.

[38] Seneff, S. Goddeau, D., Pao, C., and Polifroni, J. "Multimodal Discourse Modelling in a Multi-User Multi-Domain Environment," *Proc. ICSLP*, 188-191, 1996.

[39] Stallard, D. and Bobrow, R. "Fragment Processing in the DELPHI System," *Proc. DARPA Speech and Natural Language Workshop*. 305-310, 1992.

[40] Sutton, S., Kaiser, E., Cronk, A., and Cole,R. "Bringing Spoken Language Systems to the Classroom," *These Proceedings.*

[41] Wang, C. Glass, J., Meng, H., Polifroni, J., Seneff, S., and Zue, V. "Yinhe: A Mandarin Chinese Version of the Galaxy System'" *These Proceedings.*

[42] Ward, W. "Modelling Non-Verbal Sounds for Speech Recognition," *Proc. DARPA Workshop on Speech and Natural Language*, 47-50, 1989.

[43] Ward, W. "The CMU Air Travel Information Service: Understanding Spontaneous Speech," *Proc. ARPA Workshop on Speech and Natural Language*, 127-129, 1990.

[44] Ward, W. "Integrating Semantic Constraints into the SPHINX-II Recognition Search," *Proc. ICASSP*, II-17-20, 1994.

[45] Zue, V., Seneff, S., Polifroni, J., Phillips, M., Pao, C., Goddeau, D., Glass, J., and Brill, E. "Pegasus: A Spoken Language Interface for On-Line Air Travel Planning," *Speech Communication*, 15, 331-340, 1994.

[46] "Navigating the Information Superhighway Using Spoken Language Interfaces," *IEEE Expert*, vol. 10, no. 5, 39-43, 1995.

[47] Zue, V. , Seneff, S., Glass, J., Hetherington, L., Hurley, E., Meng, H., Pao, C., Polifroni, J., Schloming, R., and Schmid, P. "From Interface to Content: Translingual Access and Delivery of On- Line Information," *These Proceedings.*