Research Challenges for the Development of Interactive Spoken Language Systems[1]

Victor W. Zue

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139 USA
E-MAIL: ZUE@MIT.EDU

ABSTRACT

The last five years have witnessed unprecedented progress in human language technology (HLT) – speech recognition and language understanding capabilities are improving at a very rapid rate. 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 HLT base. This paper outlines my view on what are some of the unmet research challenges, including the need to work in *real* domains, dialogue modelling, dealing with unknown words, spoken language generation, and portability across domains and languages.

1 Introduction

In the past several years, we saw the emergence of a new breed of computer-based speech processing systems known as spoken language (or speech understanding) systems. The development of these systems is motivated by the belief that many tasks appropriate for human-computer interaction using speech fall into the realm of interactive problem solving. In these applications, whether it be searching for a restaurant or buying an airplane ticket, the solution is often built up incrementally, with the user and the computer both playing an active role in the conversation. To achieve this goal, several language-based technologies must be developed and integrated. On the input side, speech recognition must be combined with natural language processing in order to derive an *understanding* of the spoken input, often in the context of previous parts of the verbal dialogue. On the output side, some of the information that the user seeks as well as any clarification

[1]This research was supported by ARPA under Contract N00014-89-J-1332, monitored through the Office of Naval Research.

dialogue generated by the system must be converted to natural sentences and, possibly, delivered as verbal responses.

Research and development of spoken language systems has received considerable world-wide attention since the late eighties. In the United States, the Spoken Language Systems (SLS) Program sponsored by the Advanced Research Projects Agency (ARPA) of the Department of Defense provided major impetus for steady advances of the necessary technologies along many fronts. In Europe, the SUNDIAL project sponsored by the Esprit program enjoyed participation from many countries to jointly develop system that can understand several European languages in multiple domains (including train reservations and flight information) [1].

Despite significant progress in many areas, the ultimate deployment of speech-based user 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 summarize the state-of-the-art. Due to personal familiarity, I will draw primarily from my own experience in developing such systems in the United States. Interested readers are referred to the recent proceedings of the International Conference of Acoustics, Speech, and Signal Processing, the Eurospeech Conference, and the International Conference of Spoken Language Processing.

2 System Architecture and Research Issues

Figure 1 shows the major components of a typical spoken language system. The spoken input is first processed through the speech recognition component, whose goal is to convert the speech signal into a set of word hypotheses. These word hypotheses are then fed to the language understanding component. By combining syntactic and semantic constraints, this component eventually 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, the system may choose to query the user for clarification. Speech output can be generated by processing the information or clarification query through natural language generation and text-to-speech synthesis. Throughout the process, discourse information is maintained and can be fed back to the speech recognition and language understanding components. Figure 2 illustrates aspects of human-machine dialogue, including clarification and anaphoric referencing, in the MIT VOYAGER domain [2].

The requirement that the system *understand* verbal commands raises several important research issues. Perhaps the most important one is the integration of speech recognition and natural language processing technology to achieve speech understanding. Researchers in each discipline need to investigate how to exchange and utilize ideas to maximize overall system performance. In some cases, one may have to make funda-

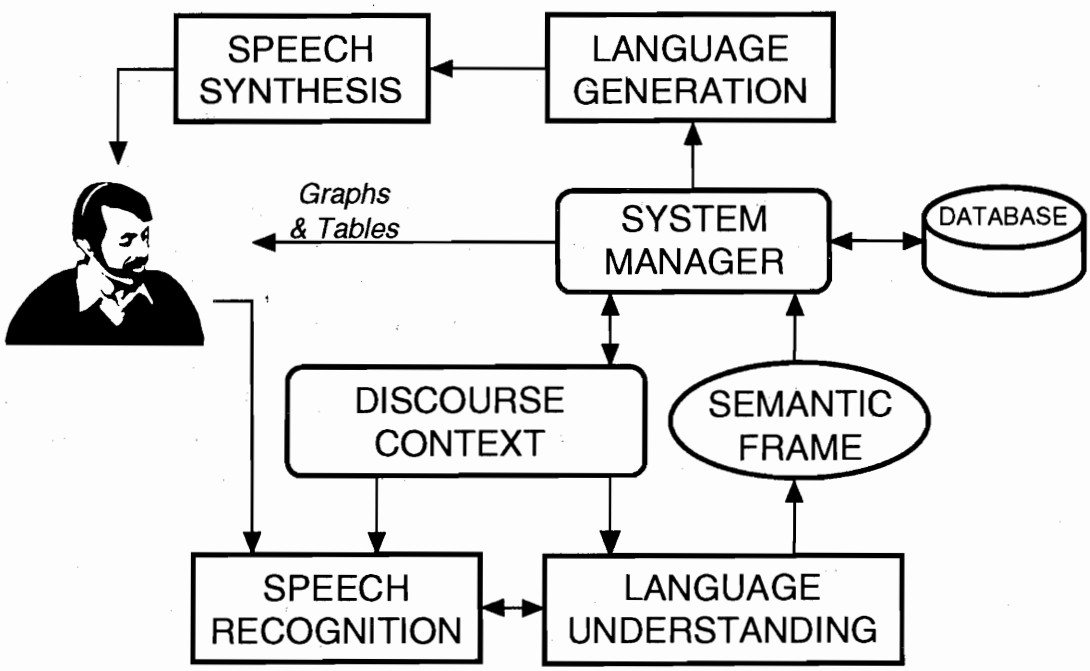


Figure 1: A generic block diagram for a typical spoken language system.

```
USER:      CAN YOU TELL ME HOW TO GET TO THE CHINESE RESTAURANT
VOYAGER:   Where are you?
USER:      I AM NEAR A BANK
VOYAGER:   I know of many banks. Can you specify which one you mean?
USER:      THE ONE ON MAIN STREET
VOYAGER:   The only bank on Main Street that I know of is the Baybank at
           226 Main Street in Cambridge.  The Chinese restaurants that I
           know of are Hong Kong and Royal East.  Can you specify which one
           you mean?
USER:      THE ONE NEAREST ME
VOYAGER:   The nearest Chinese restaurant to the Baybank at 226 Main Street
           that I know of is the Royal East.  I will try to find the best
           way to go from the Baybank at 226 Main Street to the Royal East.
           (-- Directions --)
```

Figure 2: Example dialogue in the VOYAGER domain illustrating clarification capabilities.

mental changes in the way systems are designed. For example, most natural language systems are developed with text input in mind; it is assumed that the entire word string is known with certainty. This assumption is clearly false for speech input, whereby many words are competing for the same time span, and some words may be more reliable than others because of varying signal robustness. Therefore, one may not be able to use existing natural language systems without making some modifications.

Another issue related to spoken language system development is that the system must operate in a real or realistic application domain, where domain-specific information can be utilized to translate spoken input into appropriate actions. For example, the verb "serve" conveys three entirely different meanings, depending on whether one is discussing a restaurant, the cities an airline flies to, or a tennis match. Since we are far from being able to develop systems with unconstrained capabilities, we must select applications whose domain is limited, but nevertheless useful. Realistic applications are also critical to collecting data on how people would like to use machines to access information and solve problems. The use of a constrained task also makes possible rigorous evaluations of system performance.

Finally, the system must begin to deal with interactive speech, where the computer is an active conversational participant, and where people produce speech extemporaneously. Spontaneous speech contains false starts, hesitations, and words and linguistic constructs unknown to the system. It offers significant challenges to current speech recognition and natural language systems. To make the interaction flow smoothly, the system must be able to provide feedback, especially when it is unable to fully interpret the user's query.

3 State of the Art

This section illustrates the state of the art of spoken language systems research and development by focusing on the activities in the ARPA Human Language Technology (HLT) research community in the United States in the common domain called Air Travel Information Service, or ATIS [3]. 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, periodic common evaluations have occurred at regular intervals, and steady performance improvements have been observed for all systems [4, 5, 6]. Figure 3 shows the error rates for the best ATIS systems, measured in several dimensions over the past three years. Many of the systems currently run in real-time on standard workstations with no additional hardware, although with some performance degradation.

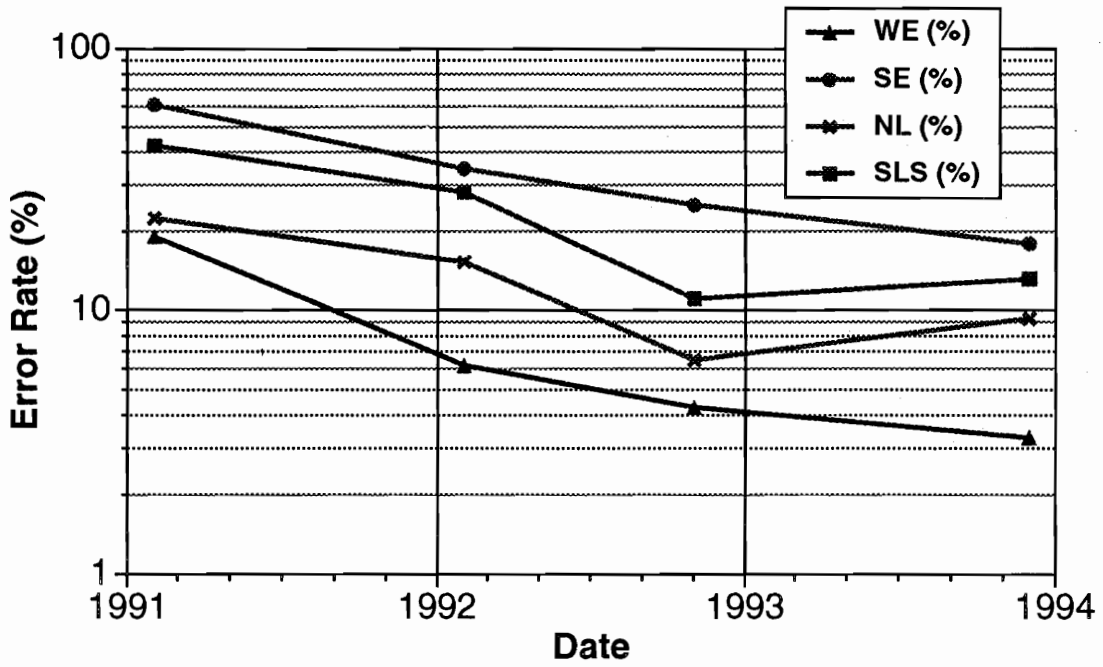


Figure 3: Best performance achieved by systems in the ATIS domain over the past three years. See text for a detailed description.

3.1 Speech Recognition

Historically, speech recognition systems have been developed with the assumption that the speech material is read from prepared text. Spoken language systems offer new challenges to speech recognition technology in that the speech is extemporaneously generated, often 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 [7, 8]. Similarly, "trash" models have been introduced to detect the presence of unknown words [9], and procedures have been devised to learn the new words once they have been detected [10].

As shown in Figure 3, the speech recognition performance has improved steadily over the past three years. Word error rate (WE) decreased by nearly six-fold while sentence error rate (SE) decreased more than three-fold in this period. In both cases, the reduction in error rate for spontaneous speech has followed the trend set forth for read speech, namely halving the error every two years. In the most recent formal evaluation of the ARPA-SLS Program in the ATIS domain, the best system achieved a word error rate of 3.3% and a sentence error rate of 18% [6]. The vocabulary size was more than 2,500 words, and the bigram and trigram language models had a perpexity of about 20 and 14, respectively. Note that all the performance results quoted in this section are for

the so-called "evaluable" queries, i.e., those queries that are within the ATIS domain and for which an appropriate answer is available from the database.

3.2 Language Understanding

Traditional natural language analysis is predominantly syntax-driven – a complete syntactic analysis is performed which attempts to account for *all* words in an utterance. However, it didn't take long for researchers to discover that such an approach [11, 12], while providing some linguistic constraints to the speech recognition component and a useful structure for further linguistic analysis, can break down dramatically in the presence of unknown words, novel linguistic constructs, recognition errors, and some spontaneous speech events such as false starts. Besides, spoken language tends to be quite informal – people are perfectly capable of speaking, and willing to accept, ungrammatical sentences. In contrast, other researchers have adopted a semantic-driven approach, deriving a meaning representation by spotting key words and phrases in the utterance [13]. 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. To date, almost all systems have abandoned their original goal of achieving a complete syntactic parse of every input sentence to a more robust strategy that could still answer when a full parse failed [14, 15, 16]. This can be achieved by identifying parsable phrases and clauses, along with a mechanism for gluing them together to form a complete meaning analysis [15]. Still others attempts to deal with the variabilities by deriving a semantic representation directly from the surface representation through stochastic modelling techniques [17, 18, 19].

It is difficult to objectively evaluate the performance of a natural language component, primarily because establishing the *reference* answer to a query may be difficult. For example, should the correct answer to the query, "Do you know of any Chinese restaurants?" be simply, "Yes," or a list of the restaurants that the system knows?

The ARPA-SLS community has adopted the Common Answer Specification (CAS) evaluation protocol, whereby a system's performance is determined by comparing its output, expressed as a set of database tuples, with one or more predetermined reference answers [21]. The CAS protocol has the advantage that system evaluation can be carried out automatically, once the principles for generating the reference answers have been established and a corpus has been annotated accordingly. Since direct comparison across systems can be performed relatively easily with this procedure, the community has been able to achieve cross fertilization of research ideas, leading to rapid research progress. Figure 3 shows that language understanding error rate (NL) has declined by more than two fold in the past three years.[2] This error rate is measured by passing the transcription

[2]The error rate, for both text (NL) and speech (SLS) input increased somewhat in the latest round of evaluation. This is largely due to the fact that the database has been increased from 11 cities to 46 in 1993, and some of the travel-planning scenarios used to collect the newer data were considerably more difficult.

of the spoken input, after removing partial words, through the natural language component. In the most recent formal evaluation in the ATIS domain, the best natural language system achieved an understanding error rate of 9.3% on all the "evaluable" sentences in the test set [6].

3.3 Speech Understanding

One of the critical research issues in the development of spoken language systems is the mechanism by which the speech recognition component interacts 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 [22, 23, 24], in which the recognizer can propose its best N complete sentence hypotheses[3] 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. However, it is still necessary to provide the recognizer with an inexpensive language model that can partially constrain the theories. Usually, a statistical language model such as a bigram is used, in which every word in the lexicon is assigned a probability reflecting its likelihood in following a given word.

In the N-best interface, a natural language component filters hypotheses that span the entire utterance. Frequently, many of the candidate sentences 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 control strategy, 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 are 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; any other word hypotheses are not allowed to proceed. Therefore, any theory that completes is guaranteed to parse. We have found that such a tightly coupled integration strategy can achieve higher performance than an N-best interface with a considerably smaller stack size [25, 26].

The performance of the entire spoken language systems can be assessed using the same CAS protocol for the natural language component, except with speech rather than text as input. Figure 3 shows that this "speech understanding" error rate (SLS) has fallen from 42.6% to 13.2% over the three year interval. It is interesting to note that this error rate is considerably less than the sentence recognition error rate, suggesting that a large number of sentences can be understood even though the transcription may contain errors.

[3] N is a parameter of the system that can be set arbitrarily as a compromise between accuracy and computation.

4 Future Research Challenges

As we can see, significant progress has been made over the past few years in research and development of systems that can understand 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 HLT base. In this section, I will outline some of the new research challenges that have heretofore received little 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 HLT deployment? I believe that 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, since this 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 such as Boston – a task that can be viewed as a logical extension of the VOYAGER system we developed in 1989 [2]. 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 us with 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 Boston to Washington DC), 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, some of which will be described in the remainder of this section. 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. We believe that the time is ripe for us to demonstrate the usefulness of the technology. 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.

4.2 Dialogue Modelling

Human verbal communication is a two-way process involving multiple, active participants. Mutual understanding is through direct and indirect speech acts, turn taking, clarification, and pragmatic considerations. Our experience with the development of the PEGASUS system for on-line travel planning has convinced us that an effective spoken language 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. There has been some theoretical work on the structure of human-human dialogue [28], but this has not yet led to effective insights for building human-machine interactive systems. The importance of continuing research in human/computer dialogue modelling cannot be overstated.

An example of the beginning of dialogue-based research using a real application backend is the MIT PEGASUS system. PEGASUS is an outgrowth of the ARPA-SLS ATIS common task development effort. It is a spoken language interface connected to an on-line flight database – the American Airlines EAASY SABRE system with over 700,000 active subscribers. The system leverages off our spoken language technology development in the ATIS domain, and enables users to book real flights using EAASY SABRE. The input query is transformed by the speech understanding system to a semantic frame representation that captures its meaning. The System Manager transforms the semantic representation into an EAASY SABRE command, transmits the command to the application backend, formats and interprets the resulting information, and generally manages the dialogue with the user. Figures 4 and 5 contains an example of the log of an actual round-trip booking to illustrate the system's capability.

The system is fully functional in the sense that members of our group have been able to use it to make actual travel arrangements since last spring, using naturally spoken English. In a recent evaluation [27], the system was able to complete 7 out of 10 actual bookings. For experienced users, a round-trip transaction can be completed using 10 queries in about 5 minutes of elapsed time.

4.3 The New Word Problem

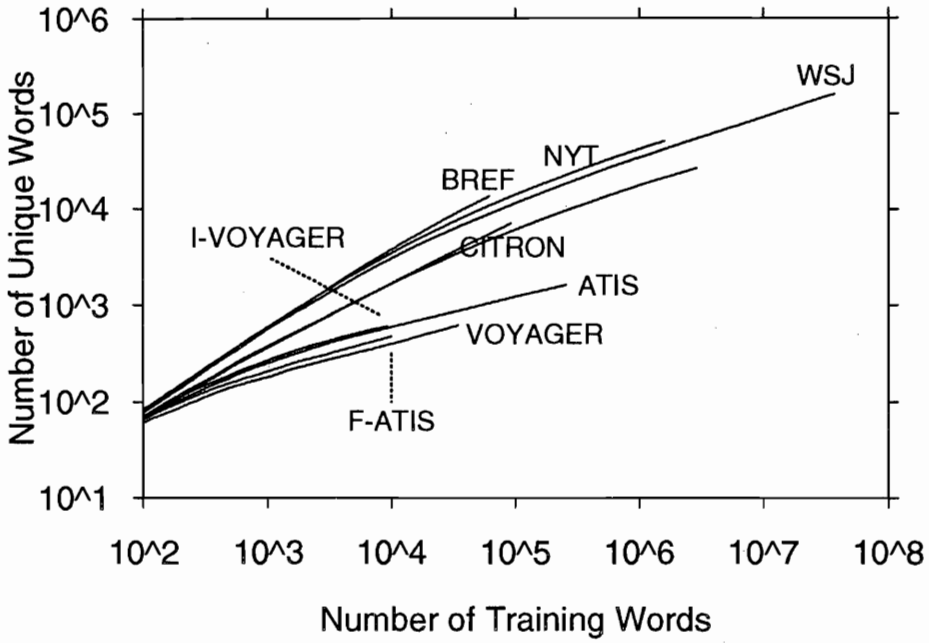
Yet another important issue concerns unknown words. The traditional approach to spoken language recognition and understanding research and development is to define the working vocabulary based on domain-specific corpora [29]. However, experience has shown that, no matter how large the size of the training corpora, the system will invariably encounter previously unseen words. This is illustrated in Figure 6. For the

```
USER:     I WANT TO GO FROM BOSTON TO SAN FRANCISCO
PEGASUS:  What date will you be travelling on?
USER:     OCTOBER TWENTIETH
PEGASUS:  Here are the flights from Boston to San Francisco on Wednesday October
          twentieth.
              FLIGHT   FROM  TO   DEPARTURE    ARRIVAL  NSTOPS
           1  DL  565  BOS   SFO   7:00 A.M.  11:55 A.M.  1
           2  US  31   BOS   SFO   7:00 A.M.  12:03 P.M.  1
           3  NW  181  BOS   SFO   7:30 A.M.  10:58 A.M.  0
           4  UA  31   BOS   SFO   8:10 A.M.  11:23 A.M.  0
           5  NW  353  BOS   SFO   8:15 A.M.   1:18 P.M.  1
           6  UA  59   BOS   SFO   9:05 A.M.  12:34 P.M.  0
            ...
          You can ask for more flights.
USER:     I'D LIKE A NONSTOP UNITED FLIGHT
PEGASUS:  Here are the non-stop United flights from Boston to San Francisco on
          Wednesday October twentieth.
              FLIGHT   FROM  TO   DEPARTURE    ARRIVAL  NSTOPS
           1  UA  31   BOS   SFO   8:10 A.M.  11:23 A.M.  0
           2  UA  59   BOS   SFO   9:05 A.M.  12:34 P.M.  0
           3  UA  93   BOS   SFO   5:10 P.M.   8:23 P.M.  0
           ...
USER:     BOOK THE THIRD FLIGHT PLEASE
PEGASUS:  Please wait while I determine what fares are available for this flight.
          The following fares are available.  Which fare class would you like?
              ONE WAY   FARE CLASS      MIN-STAY  MAX-STAY  RETURN REQ'D
           1    $ 214.  THRIFT DISCOUNT   YES        30         YES
           2    $ 234.  COACH DISCOUNT    YES        30         YES
           3    $ 294.  THRIFT DISCOUNT   YES        NO         YES
           4    $ 304.  COACH DISCOUNT    YES        30         YES
           ...
USER:     COACH DISCOUNT
PEGASUS:  The following fares are available.  Which fare class would you like?
              ONE WAY   FARE CLASS      MIN-STAY  MAX-STAY  RETURN REQ'D
           1    $ 234.  COACH DISCOUNT    YES        30         YES
           2    $ 304.  COACH DISCOUNT    YES        30         YES
           3    $ 336.  COACH DISCOUNT    31         NO         YES
           ...
```

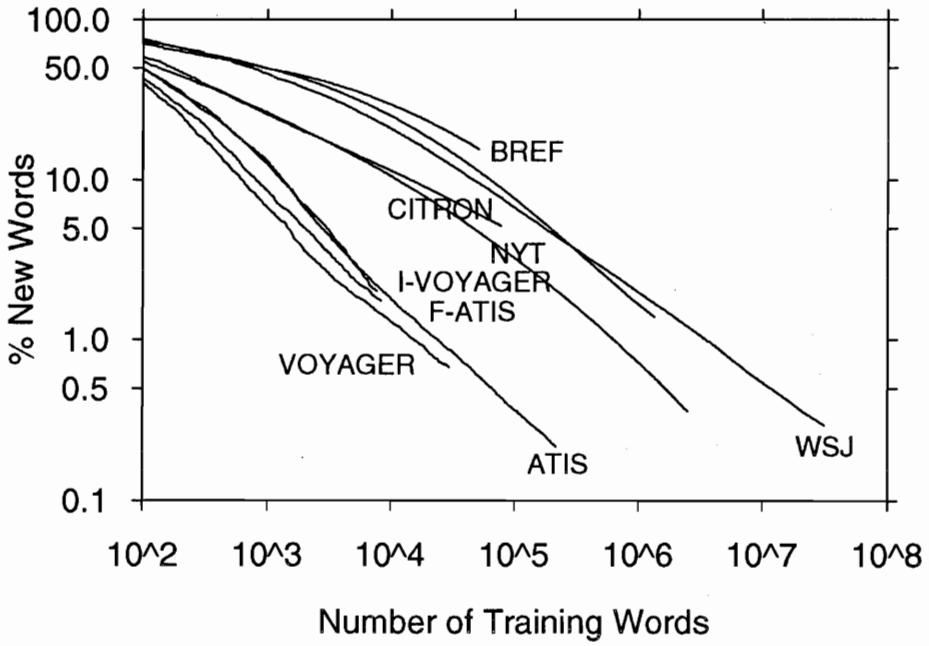
Figure 4: An example of an actual verbal booking dialogue using PEGASUS. Due to space limitations, irrelevant parts of the system's responses have been omitted.

```
USER:     THE CHEAPEST ONE
PEGASUS:  I'll book United flight 93 from Boston to San Francisco on Wednesday
          October twentieth for you.  Your reservation has been entered.
          Now I'll help you with the return flight.
          What date will you be returning on?
USER:     ON SUNDAY
PEGASUS:  Here are the United flights from San Francisco to Boston on Sunday
          October twenty-fourth.
              FLIGHT  FROM  TO   DEPARTURE    ARRIVAL  NSTOPS
           1  UA  92   SFO   BOS   7:45 A.M.   4:06 P.M.  0
           2  UA  20   SFO   BOS   1:30 P.M.   9:54 P.M.  0
           3  UA  90   SFO   BOS  10:10 P.M.   6:32 A.M.  0
           ...
USER:     SHOW ME NONSTOP FLIGHTS LEAVING IN THE MORNING
PEGASUS:  Here is United flight 92 from San Francisco to Boston on Sunday October
          twenty-fourth.
              FLIGHT   FROM  TO   DEPARTURE    ARRIVAL  NSTOPS
           1  UA  92   SFO   BOS   7:45 A.M.   4:06 P.M.  0
USER:     I'LL TAKE THIS FLIGHT
PEGASUS:  Please wait while I determine what fares are available for this flight.
          I'll book United flight 92 from San Francisco to Boston on Sunday
          October twenty-fourth for you.  Your reservation has been entered.
          Give me a moment to confirm your reservation.
          Here's your ticket summary.
          You saved a total of 894 dollars over regular coach fare.
          Based on fares in effect today, your airline tickets will cost:
          Each adult:    488.00 USD
          Fare codes:    QE7ONQ, QE7PNQ
          Adults:        1
          Total fare:    488.00 USD
          ...  Please wait while I retrieve your itinerary...
               STEPHANIE SENEFF                              Price:   488.00
                                                             Total:   488.00
          ------------------------------------------------------------------
              Date      Flight      Class  From    To      Leave    Arrive
           1  OCT20     UA  93        Q    BOS     SFO      510P     823P
           2  OCT24     UA  92        Q    SFO     BOS      745A     406P
          ------------------------------------------------------------------
          Can I help you with something else?
```

Figure 5: Continuation of the example shown in Figure 4.



(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.

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 contain 10 words, this would mean that approximately one in 50 sentences will contain an unknown word.

In a *real* domain such as Electronic Yellow Pages, a 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 is open, unknown words account for a small fraction of the word tokens in the test corpus. Thus researchers can 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 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 MIT 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.

4.4 Spoken Language Generation

With few exceptions [2, 30, 27], 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. While the systems may not be entirely deaf, they are certainly mute!

Spoken language generation is an extremely important aspect of the human-computer interface problem, especially if the transactions were to be conducted over a telephone. It is also crucial for the task of speech-to-speech translation. We must develop models and methods that will generate natural sentences appropriate for spoken output, across many domains and languages [31]. In many cases, we must pay particular attention to the interaction between language generation and dialogue management – the system may have to initiate clairifcation dialogue to reduce the amount of information returned from the backend, in order not to generate unwieldy verbal responses. On the speech side, we must continue to improve speech synthesis capabilities, particularly with regard to the encoding of prosodic and paralinguistic information such as emotion and mood. 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.

4.5 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 issues that must be addressed before spoken language technology 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 enable users to efficiently access, process, and manipulate 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 future.

6 References

[1] Peckham J., "A New Generation of Spoken Dialogue Systems: Results and Lessons from the SUNDIAL Project," *Proc. Eurospeech*, 33–40, September 1994.

[2] Zue, V., Glass, J., Goodine, D., Leung, H., Phillips, M., Polifroni, J. and Seneff, S. "The VOYAGER Speech Understanding System: A Progress Report," *Proc. DARPA Speech and Natural Language Workshop*, 160–167, October 1989.

[3] Price, P., "Evaluation of Spoken Language Systems: the ATIS Domain," *Proc. DARPA Speech and Natural Language Workshop*, 91–95, June 1990.

[4] Pallett, D., Dahlgren N., Fiscus, J., Fisher, W., Garafolo, J., and Tjaden, B., "February 1992 ATIS Benchmark Test Results," *Proc. DARPA Speech and Natural Language Workshop*, 15–27, February 1992.

[5] Pallett, D., Fiscus, J., Fisher, W., and Garafolo, J., "Benchmark Tests for the DARPA Spoken Language Program," *Proc. ARPA Speech and Natural Language Workshop*, 7–18, March 1993.

[6] Pallett, D., Fiscus, J., Fisher, W., and Garafolo, J., Lund, B., and Pryzbocki, M., "1993 Benchmark Tests for the ARPA Spoken Language Program," *Proc. ARPA Speech and Natural Language Workshop*, March 1994.

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

[8] 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, February 1992.

[9] Asadi, A. and Schwartz, R., "Automatic Detection of New Words in a Large Vocabulary Continuous Speech Recognition System," *Proc. ARPA Workshop on Speech and Natural Language*, 263–265, October 1989.

[10] 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, May 1991.

[11] 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, June 1990.

[12] Seneff, S., "TINA: A Natural Language System for Spoken Language Applications," *Computational Linguistics*, Vol. 18, No. 1, 61–86, March 1992.

[13] Ward, W., "The CMU Air Travel Information Service: Understanding Spontaneous Speech,", *Proc. DARPA Speech and Natural Language Workshop*, 127–129, June, 1990.

[14] 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, February, 1991.

[15] Seneff, S., "Robust Parsing for Spoken Language Systems," *Proc. ICASSP*, 189–192, March 1992.

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

[17] 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, February, 1991.

[18] Kuhn, R. and De Mori, R., "Learning Speech Semantics with Keyword Classification Trees," *Proc. ICASSP*, 55–58, April 1994.

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

[20] Seneff, S., Meng, H., and Zue, V., "Language Modelling for Recognition and Understanding Using Layered Bigrams," *Proc. International Conference on Spoken Language Processing*, 317–320, October, 1992.

[21] Bates, L., Boisen, S., and Makhoul, J., "Developing an Evaluation Methodology for Spoken Language Systems," *Proc. ARPA Workshop on Speech and Natural Language*, 102–108, June 1990.

[22] 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, October 1989.

[23] Soong, F. and Huang, E., "A Tree-Trellis Based Fast Search for Finding the N-best Sentence Hypotheses in Continuous Speech Recognition", *Proc. ARPA Workshop on Speech and Natural Language*, 199–202, June 1990.

[24] Zue, V., Glass, J., Goddeau, D., Goodine, D., Leung, H., McCandless, M., Phillips, M., Polfroni, J., Seneff, S., and Whitney, D., "Recent Progress on the MIT VOYAGER Spoken Language System," *Proc. International Conference on Spoken Language Processing*, 1317–1320, November 1990.

[25] 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, September 1991.

[26] Goddeau, D., "Using Probabilistic Shift-Reduce Parsing in Speech Recognition Systems," *Proc. International Conference on Spoken Language Processing*, 321–324 October 1992.

[27] 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," To appear in *Speech Communication*, 1994.

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

[29] Hetherington, I.L., and Zue, V. "New Words: Implications for Continuous Speech Recognition," *Proceedings of the European Conference on Speech Communication and Technology*, Berlin, Germany, September 1993.

[30] Glass, J., Goodine, D., Phillips, M., Sakai, S., Seneff, S., and Zue, V. "A Bilingual VOYAGER System," *Proc. Eurospeech*, 2063–2066, September 1993.

[31] Glass, J., Polifroni, J., and Seneff, S. "Multilingual Language Generation Across Multiple Domains," *These Proceedings*, September, 1994.