Pervasive Speech and Language Technology

Wolfgang Wahlster

German Research Center for Artificial Intelligence, DFKI GmbH
Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany
wahlster@dfki.de
http://www.dfki.de/~wahlster

Abstract. Advances in human language technology offer the promise of pervasive access to on-line information and electronic services. Since almost everyone speaks and understands a language, the development of natural language systems will allow the average person to interact with computers anytime and anywhere without special skills or training, using common devices such as a mobile telephone. The latest results and component technologies for multilingual and robust speech processing, prosodic analysis, parsing, semantic analysis, discourse understanding, translation, and speech synthesis are reviewed using the Verbmobil system as an example. Verbmobil is a speaker-independent and bidirectional speech-to-speech translation system for spontaneous dialogs in mobile situations. It recognizes spoken input, analyses and translates it, and finally utters the translation. The multilingual system handles dialogs in three business-oriented domains, with context-sensitive translation between three languages (German, English, and Japanese). We will show that the most successful current systems are based on hybrid architectures incorporating both deep and shallow processing schemes. They integrate a broad spectrum of statistical and rule-based methods and combine the results of machine learning from large corpora with linguists' hand-crafted knowledge sources to achieve an adequate level of robustness and accuracy. We argue that packed representations together with formalisms for underspecification capture the uncertainties in each processing phase, so that these uncertainties can be reduced by linguistic, discourse and domain constraints as soon as they become applicable. We show that the current core technologies for natural language and speech processing enable us to create the next generation of information extraction and summarization systems for the Web, speech-based Internet access and multimodal communication assistants combining speech and gesture.

1 Introduction

Human language technology will become pervasive in our daily lifes (see Fig. 1). When you have breakfast in the morning you can control your coffee machine by speech commands. Before you drive to a meeting you can program your car navigation system and select a music CD via voice commands. While

you are stuck in traffic, you can dictate and send an email to one of your colleagues via your WAP-enabled cell phone. In your office, you can retrieve and extract information from digital recordings of television broadcast news stored in video databases available through the Internet. In contrast to traditional TV programs, such content-based video retrieval provides information on demand. Instead of taking notes during your business meetings, you store them on your personal audio memory device. Using audio mining technology your notes are converted into searchable text that is indexed to time code on your digital audio memory. You can use your cell phone as a speech-to-speech translation device, that recognizes your spoken input, analyses and translates it to Japanese, and finally utters the translation.

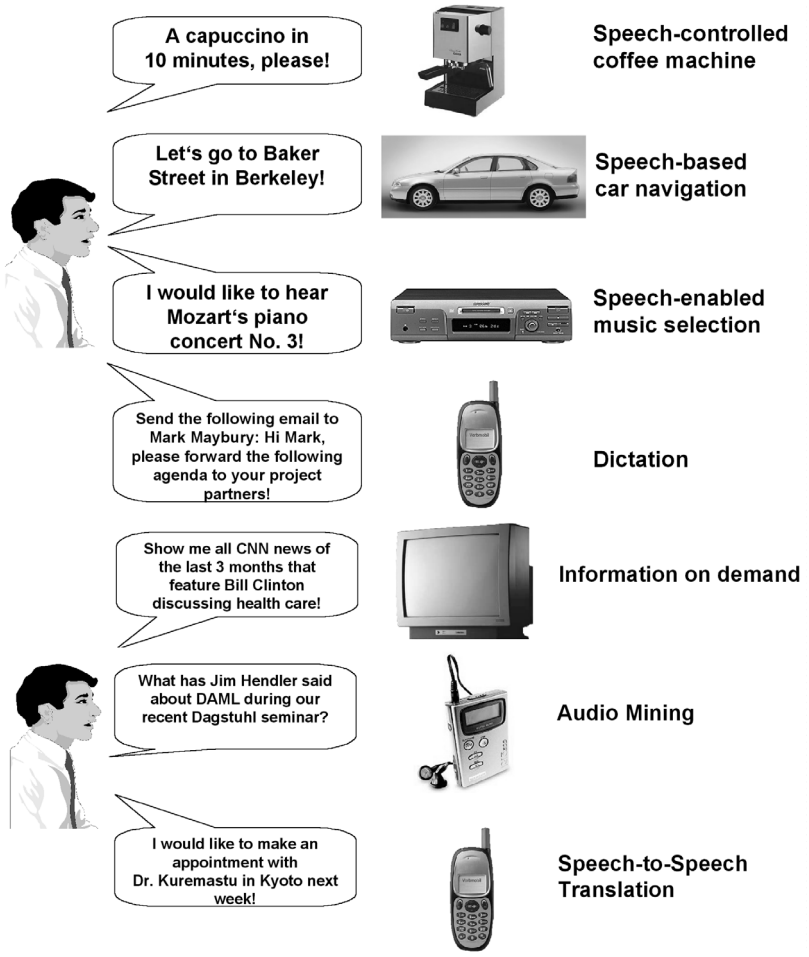

Fig. 1. Applications of Human Language Technology

Today, there exist operational demonstrators and prototypes for all the application scenarios mentioned above and some of them are being commercialized right now. This should not be misunderstood: Of course, there are many more open problems than solved ones in speech and language understanding. We will mention some of the most important open issues at the end of this paper.

Although great progress has been made in speech recognition over the past decade, the semantic level of speech analysis and the pragmatic level of speech understanding are only achieved by very few systems that work in narrowly restricted domains of discourse. Only on the the third level of language understanding all relevant ambiguities can be resolved by discourse and domain knowledge so that an unambiguous interpretation of a dialog contribution becomes possible (see Fig. 2).

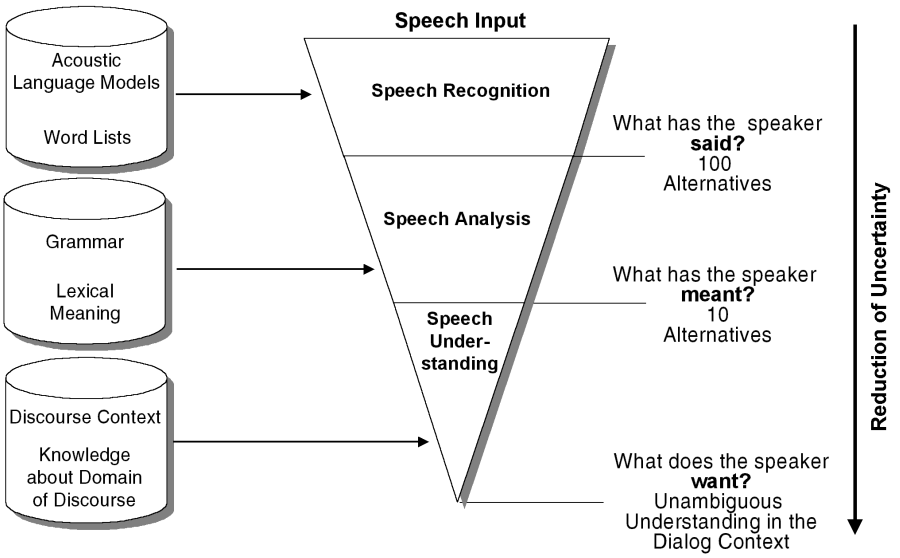

Fig. 2. Three Levels of Language Processing

The remainder of this paper is organized as follows. First, we introduce the speech-to-speech translation system Verbmobil that will serve as the concrete background of our presentation. Next, we discuss some of the grand challenges of language technology. Then we present the multi-blackboard and multi-engine approach to the robust processing of spontaneous speech. The paper ends with a discussion of open problems and conclusions.

2 The Speech-to-Speech-Translation System Verbmobil

Verbmobil is a software system that provides mobile phone users with simultaneous dialog interpretation services for restricted topics (see Wahlster, 1993, 2000). As the name Verbmobil suggests, the system supports verbal communication with foreign interlocutors in mobile situations. It recognizes spoken input, analyses and translates it, and finally utters the translation. The multilingual system handles dialogs in three business-oriented domains, with bidirectional translation between three languages (German, English, and Japanese).

In contrast to previous dialog translation systems that translate sentence-by- sentence, Verbmobil provides context-sensitive translations. Since Verbmobil must "hear between the words" — things that were communicated earlier and things about the topic being discussed — it uses an explicit dialog memory and exploits domain knowledge. Figure 3 illustrates the use of Verbmobil in the travel planning scenario.

It shows that the Verbmobil server translates the German word "nächste" into English either as "next" or "nearest" depending on a temporal or spatial question context. In Verbmobil, the dialog context is used to resolve ambiguities and to produce an adequate translation in a particular conversational situation.

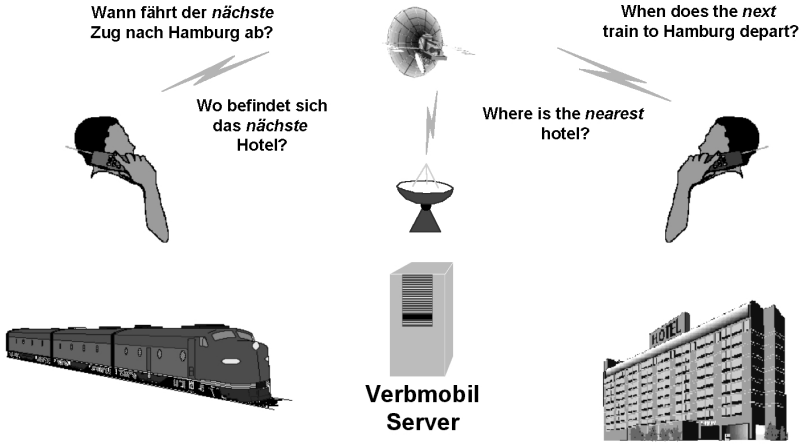

Fig. 3. Context-sensitive speech-to-speech translation

Verbmobil is the first speech-only dialog translation system. Verbmobil users can simply pick up a standard mobile phone and use voice dialing and speech commands in order to initiate a dialog translation session (see Figure 4). In contrast to previous versions of Verbmobil and other systems in the C-STAR consortium (see Woszczyna, 1999), the operation of the final Verbmobil system

is completely hands-free without any push-to-talk button. Since the Verbmobil speech translation server can be accessed by GSM mobile telephones, the system can be used anywhere and anytime. No PC, notebook or PDA must be available to access the Verbmobil translation service, just a phone for each dialog participant. In addition, no waiting time for booting computers and keyboard or mouse input to start the Verbmobil system is needed — dialog translation can begin instantaneously. Although the primary goal of Verbmobil is to support face-to-face conversations, in which all participants use their mobile phones as translation devices, Verbmobil can also be used for conversations in which the participants cannot see one another.

Fig. 4. Setting up a Verbmobil session with speech commands

Verbmobil emphasizes the robust processing of spontaneous dialogs posing difficult challenges to human language technology, that are summarized in Figure 5 and discussed in more detail below.

3 Some Grand Challenges for Speech and Language Technology

Verbmobil is the only dialog translation system to date based on an open microphone condition. It is not a "push-to-talk" system which has to be told which chunks of the sound signal represent coherent contributions by individual speakers:

Verbmobil works that out for itself from the raw input signal. The signal may be of different qualities — not necessarily from a lab-quality close-speaking microphone, for instance it can be GSM (mobile phone) quality. Thus, Verbmobil includes different speech recognizers for 16 kHz and 8 kHz sampling rates. Verbmobil deals with spontaneous speech. This does not just mean continuous speech like in current dictation systems, but speech which includes realistic disfluencies and repair phenomena, such as changes of tack in mid-sentence (or mid-word), ums and ers, and cases where short words are accidentally left out in rapid speech. For example, in the Verbmobil corpus about 20% of all dialog turns contain at least one self-correction and 3% include false starts. Verbmobil uses a combination of shallow and deep analysis methods to recognize a speaker's slips and translate what he tried to say rather than what he actually said.

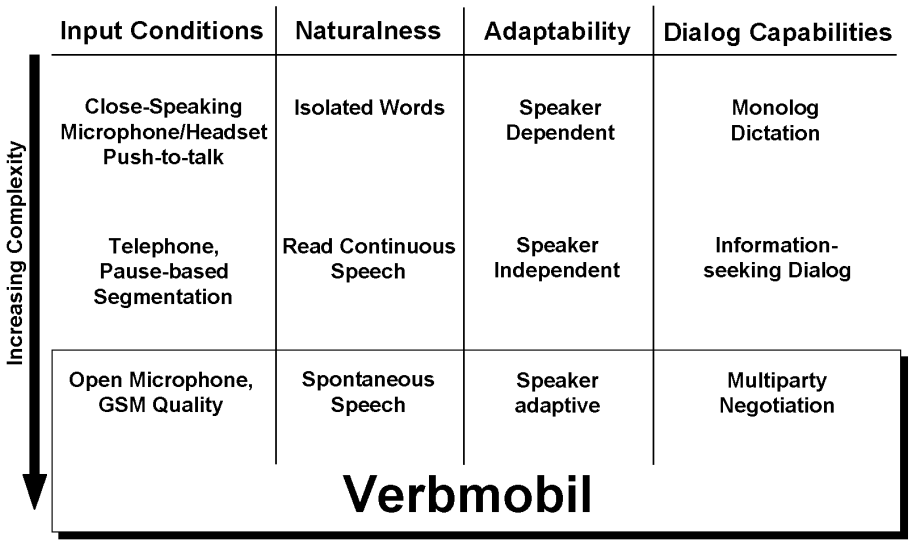

Fig. 5. Some challenges for speech and language technology

At an early processing stage prosodic cues are used to detect self-corrections. A stochastic model is used to segment the repair into the "wrong" part (the so-called reparandum) and the correction. Then the corrected input is inserted as a new hypothesis into the word hypotheses graph. Thus, Verbmobil's repair processing is a filter between speech recognition and syntactic analysis. The word lattice is augmented by an additional path that does no longer contain those parts of the utterances that the speaker tried to correct. This transformation of the word lattice is used in addition to simple disfluency filtering, that eliminates sounds like "ahh" that users often make while speaking.

In addition to this shallow statistical approach, other forms of self-corrections are also processed at a later stage on the semantic level. A rule-based repair

approach is applied during robust semantic processing to a chart containing possible semantic interpretations of the input (the so-called VIT Hypotheses Graph (VHG)).

Verbmobil applies various hand-crafted rules to detect repairs in semantic representations and to delete parts of the representation that corresponds to slips of the speaker. Verbmobil is a speaker-adaptive system, i.e. for a new speaker it starts in a speaker-independent mode and after a few words have been uttered it improves the recognition results by adaptation. A cascade of unsupervised methods, ranging from very fast adaptation during the processing of a single utterance to complex adaptation methods that analyze a longer sequence of dialog turns, is used to adjust to the acoustic characteristics of the speaker's voice, the speaking rate and pronounciation variants due to the dialectal diversity of the user community.

Verbmobil deals with mixed-initiative dialogs between human participants. Each partner has a clear interaction goal in a negotiation task like appointment scheduling or travel planning. Although these tasks encourage cooperative interaction, the participants have often conflicting goals and preferences that lead to argumentative dialogs. Therefore Verbmobil has to deal with a much richer set of dialog acts than previous systems that focused on information-seeking dialogs.

In order to ensure domain independence and scalability, Verbmobil was developed for three domains of discourse (appointment scheduling, travel planning, remote PC maintenance) with increasing size of vocabularies and ontologies.

Verbmobil is a hybrid system incorporating both deep and shallow processing schemes (see Bub et al., 1997). It integrates a broad spectrum of corpus-based and rule-based methods. Verbmobil combines the results of machine learning from large corpora with linguists' hand-crafted knowledge sources to achieve an adequate level of robustness and accuracy.

4 Verbmobil's Massive Data Collection Effort

A significant programme of data collection was performed during the Verbmobil project to extract statistical properties from large corpora of spontaneous speech. A distinguishing feature of the Verbmobil speech corpus is the multi-channel recording. The voice of each speaker was recorded in parallel using a close-speaking microphone, a room microphone, and various telephones (GSM phone, wireless DECT phone and regular phone), so that the speech recognizers could be trained on data sets with various audio signal qualities.

The so-called partitur (German word for musical score) format used for the Verbmobil speech corpora orchestrates fifteen strata of annotations: two transliteration variants, lexical orthography, canonical pronunciation, manual phonological segmentation, automatic phonological segmentation, word segmentation, prosodic segmentation, dialog acts, noises, superimposed speech, syntactic category, word category, syntactic function, and prosodic boundaries. In addition to the monolingual data, the multilingual Verbmobil corpus includes bilingual dialogs (from Wizard-of-OZ experiments, face-to-face dialogs with human in-

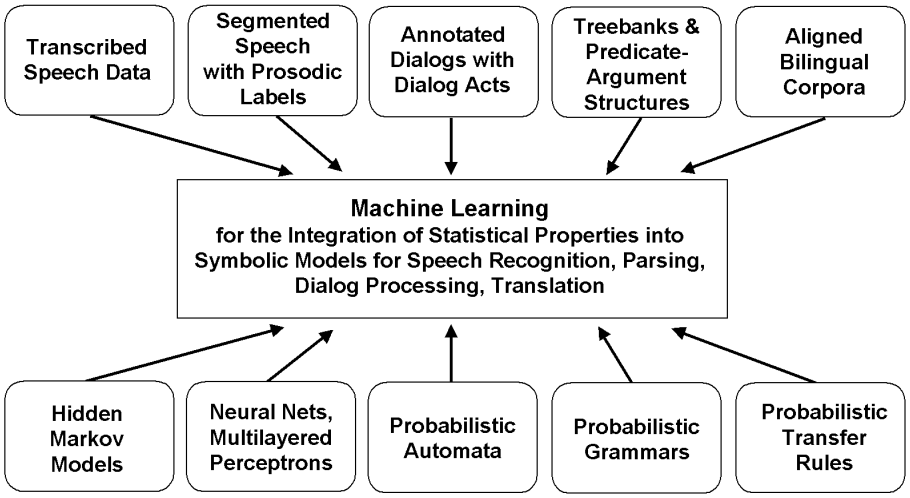

Fig. 6. Extracting statistical properties from large corpora

terpreters, or dialogs interpreted by various versions of Verbmobil) and aligned bilingual transliterations . Three treebanks for German, English and Japanese have been developed with annotations on three strata: morpho-syntax, phrase structure, and predicate-argument structure.

Various machine learning methods have been used to train Hidden Markov Models, neural nets, probabilistic automata, parsers, rule systems, translation methods and plan recognizers. The end-to-end evaluations of the various Verbmobil prototypes have shown clearly, that the robustness, coverage and accuracy of a speech-to-speech translation system for spontaneous dialogs depends critically on the quantity and quality of the training corpora.

5 The Main Components of Verbmobil

The screenshot of Verbmobil's control panel provides an overview of the main components of the system. The overall control and data flow is indicated by arrows pointing upwards on the left side of the screenshot, from left to right in the middle and downwards on the right side. On the bottom various input devices can be selected (Europeans call a cell phone a "handy"). Since Verbmobil is a multilingual system it incorporates three speech recognizers and three speech synthesizers for German, English and Japanese.

A distinguishing feature of Verbmobil is its multi-engine parsing architecture. Three parsers based on different syntactic knowledge sources are used to process the word hypotheses graphs (WHG) that are augmented by prosodic information extracted by the prosody module (see Section 5 below). All parsers use the multistratal VIT representation as an output format. Since in most cases the parsers produce only fragmentary analyses, their results are combined in a chart

of VIT structures. A chart parser and a statistical LR parser are combined in a package that is visualized in the screenshot as "integrated processing". These shallow parsers produce trees that are transformed into VIT structures by a module called semantic construction (see Figure 7). This syntax-semantics interface is primarily lexically driven . The module with the label "deep analysis" is based on a HPSG parser for deep linguistic processing in the Verbmobil system.

Verbmobil is the only completely operational speech-to-speech translation system that is based on a wide-coverage unification grammar and tries to preserve the theoretical clarity and elegance of linguistic analyses in a very efficient implementation. The parser for the HPSG grammars processes the n best paths produced by the integrated processing module. It is implemented as a bidirectional bottom-up active chart parser.

The statistical translation module starts with the single best sentence hypothesis of the speech recognizer. Prosodic information about phrase boundaries and sentence mode are utilized by the statistical translation module. The output of this module is a sequence of words in the target language together with a confidence measure that is used by the selection module (not shown in the control panel) for the final choice of a translation result. Verbmobil includes two components for case-based translation. Substring-based translation is a method for incremental synchronous interpretation, that is based on machine learning methods applied to a sentence-aligned bilingual corpus. Substrings of the input for which a contiguous piece of translation can be found in the corpus are the basic processing units. Substring pairs are combined with patterns for word order switching and word cluster information in an incremental translation algorithm for a sequence of input segments. The other component for case-based translation is based on 30000 translation templates learned from a sentence-aligned corpus. Date, time and naming expressions are recognized by definite clause grammars (DCGs) and marked in the WHG. An A* search explores the cross-product graph of the WHG with the subphrase tags and the template graph. A DCG-based generator is used to produce target language output from the interlingual representation of the recognized date, time and naming expressions. These subphrases are used to instantiate the target language parts of translation templates.

Dialog-act based translation includes the statistical classification of 19 dialog acts and a cascade of more than 300 finite-state transducers that extract the main propositional content of an utterance. The statistical dialog classifier is based on n-grams and takes the previous dialog history into account. The recognized dialog act, the topic and propositional content are represented by a simplistic frame notation including 49 nested objects with 95 possible attributes covering the appointment scheduling and travel planning tasks. A template-based approach to generation is used to transform these interlingual terms into the corresponding target language. The shallow interlingual representation of an utterance is stored together with topic and focus information as well as a deep semantic representation encoded as a VIT in the dialog memory for further processing by the dialog and context evaluation component.

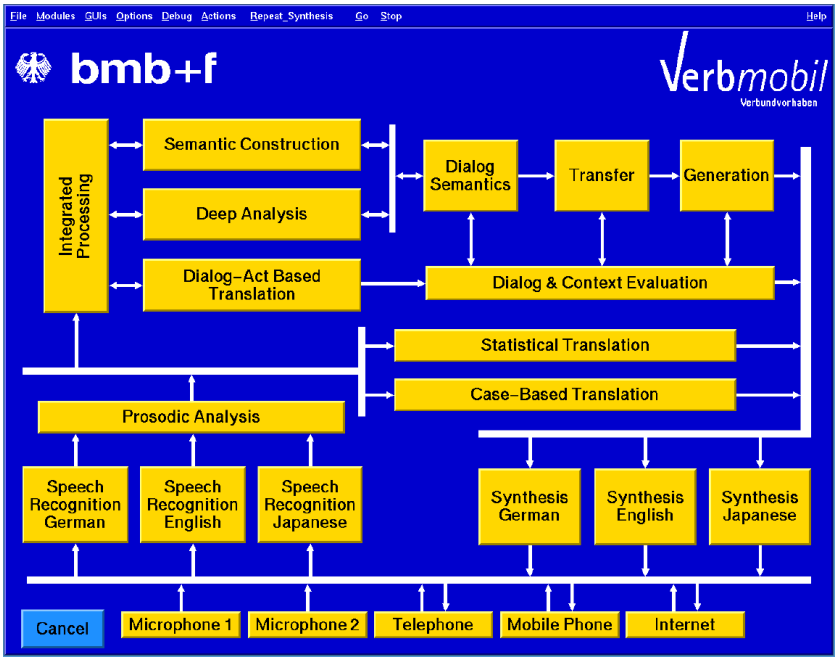

Fig. 7. A snapshot of Verbmobil's control panel

The dialog component includes a plan processor, that structures an ongoing dialog hierarchically in different dialog phases, games and moves. Dialog acts are the terminal nodes of the tree structure that represents the dialog structure. Information about the dialog phase is used eg. during the semantic-based transfer for disambiguation tasks. In addition, inference services are provided by the dialog and context component eg. for the completion of underspecified temporal expressions and the resolution of anaphora or ellipsis. Temporal reasoning is used for example to transform expressions like "two hours later" or "next week" into fully specified times and dates stored in the dialog memory for summarizing the results of a negotiation. The transfer module triggers contextual reasoning process only in cases where a disambiguation or resolution is necessary for a given translation task. For example the German noun "Essen" can be translated into "lunch" or "dinner" depending on the time of day, which can be derived by contextual reasoning. Disambiguation and resolution on demand is typical for Verbmobil's approach to translation, since various forms of underspecification and ambiguity can be carried over into target language, so that the hearer can resolve them.

The transfer component is basically a rewriting system for underspecified semantic representations using Verbmobil's VIT formalism. Semantic-based transfer receives a VIT of a source language utterance and transforms it into a VIT for the target language synthesis. This means that the transfer module abstracts

away from morphological and syntactic analysis results. The final Verbmobil system includes more than 20000 transfer rules. These rules include conditions that can trigger inferences in the dialog and context evaluation module to resolve ambiguities and deal with translation mismatches, whenever necessary. The transfer component uses cascaded rule systems, first for the phrasal transfer of idioms and other non-compositional expressions and then for the lexical transfer. The translation of spatial and temporal prepositions is based on an interlingual representation in order to cut down the number of specific transfer rules. Semantic-based transfer is extremely fast and consumes on the average less than 1% of the overall processing time for an utterance.

Verbmobil's multilingual generator includes a constraint-based microplanning component and a syntactic realization module that is based on the formalism of lexicalized tree-adjoining grammars. The input to the microplanning component are VITs produced by the transfer module. A sentence plan is generated that consists basically of lexical items and semantic roles linking them together. The microplanner decides about subordination, aggregation, focus and theme control as well as anaphora generation. The syntactic realization component can either use LTAG grammars that are compiled from the HPSG grammars used for deep analysis or a hand-written LTAG generation grammar. For English and Japanese the grammars that were designed for analysis are usable for generation after an offline-compilation step.

The speech synthesizer for German and American English follows a concatenative approach based on a large corpus of annotated speech data. The word is the basic unit of concatenation, so that subword units are only used if a word is not available in the database.

The synthesizer applies a graph-based unit selection procedure to choose the best available synthesis segments matching the segmental and prosodic constraints of the input. Whenever possible the synthesizer exploits the syntactic, prosodic and discourse information provided by previous processing stages. Thus for the deep processing stream it provides concept-to-speech synthesis, whereas for the shallow translation threads it operates more like a traditional text-to-speech system resulting in a lower quality of its output.

6 Using Prosodic Information at all Processing Stages

Verbmobil is the first spoken-dialog interpretation system that uses prosodic information systematically at all processing stages. The results of Verbmobil's multilingual prosody module are used for parsing, dialog understanding, translation, generation and speech synthesis (see figure 8). This means that prosodic information in the source utterance is passed even through the translation process to improve the generation and synthesis of the target utterance. Prosodic differences in one language can correspond to lexical or syntactic differences in another; for instance, a German utterance beginning "wir haben noch ..." may be translated by Verbmobil into English either "as we still have ..." or as "we have another ..." depending whether noch is stressed. Although prosody is used

in some other recent speech recognition systems, the exploitation of prosodic information is extremely limited in these approaches. For example, the ATR Matrix system (see Takezawa et al., 1998) uses prosody only to identify sentence mood (declarative vs. question). We believe that Verbmobil is the first system to make significant use of prosodic aspects of speech. The prosody module of Verbmobil uses the speech signal and the word hypotheses graph (WHG) produced by the speech recognizer as an input and outputs an annotated WHG with prosodic information for each recognized word. The system extracts duration, pitch, energy, and pause features and uses them to classify phrase and clause boundaries, accented words and sentence mood. A combination of a multilayer perceptron and a polygram-based statistical language model annotate the WHG with probabilities for the classified prosodic events.

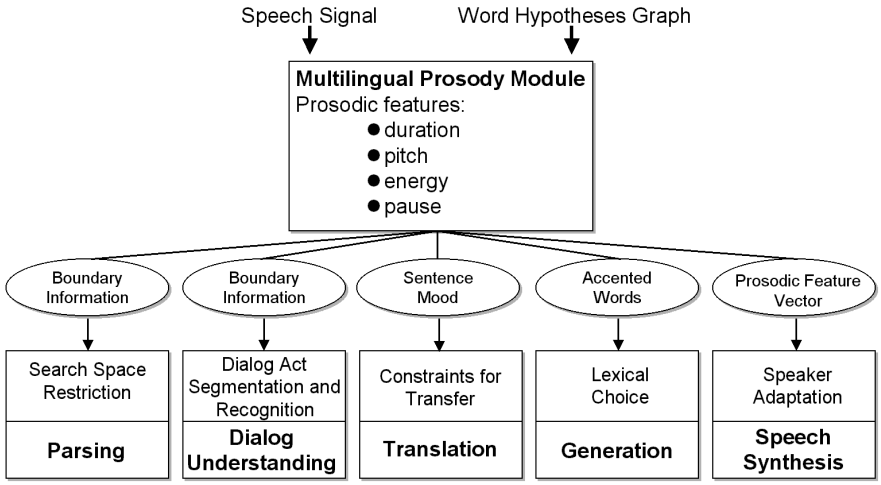

Fig. 8. The role of prosodic information in Verbmobil

Verbmobil uses the probabilistic prosodic information about clause boundaries to reduce the search space for syntactic analysis dramatically. During parsing, the clause boundary marks that are inserted into the WHG by the prosody module play the role of punctuation marks in written language. Dialog act segmentation and recognition is also based on the boundary information provided by the prosody module. Prosodic cues about sentence mood is often used in Verbmobil's translation modules to constrain transfer results, if there is not enough syntactic or semantic evidence for a certain mood (e.g. question). The information about word accent is used to guide lexical choice in the generation process. Finally, during speech synthesis the extracted prosodic features are used for speaker-adaptation.

7 The Multi-Blackboard Architecture of Verbmobil

The final Verbmobil system consists of 69 highly interactive modules. The transformation of speech input in a source language into speech output in a target language requires a tremendous amount of communication between all these modules. Since Verbmobil has to translate under real-time conditions it exploits parallel processing schemes whenever possible. The non-sequential nature of the Verbmobil architecture implies that not only inputs and results are exchanged between modules but also top-down expectations, constraints, backtracking signals, alternate hypotheses, additional parameters, probabilities, and confidence values.

198 blackboards are used for the necessary information exchange between modules. A module typically subscribes to various blackboards. Modules can have several instances, e.g. in a multiparty conversation there may be two German speakers, so that two instances of the German speech recognition module are needed.

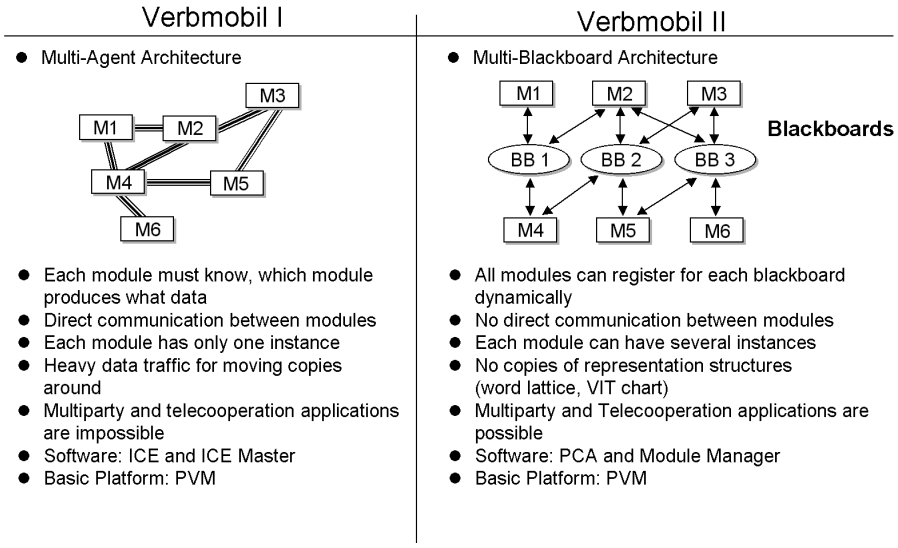

Fig. 9. A comparison of the architecture of Verbmobil I and II

The final Verbmobil system is based on a multi-blackboard architecture that pools processing modules around blackboards representing intermediate results at each processing stage. It turned out that such a multi-blackboard approach is much more efficient than the more general multi-agent architecture used in the first Verbmobil prototype. Due to the huge amount of interaction between modules a multi-agent architecture with direct communication among module agents would imply 2380 different interfaces for message exchanges between the

69 agents. Figure 9 summarizes the advantages of the multi-blackboard approach vs. the multi-agent approach for the Verbmobil architecture.

In a multi-blackboard architecture based on packed representations at all processing stages (speech recognition, parsing, semantic processing, translation, generation, speech synthesis) using charts with underspecified representations the results of concurrent processing threads can be combined in an incremental fashion. All results of concurrent processing modules come with a confidence value, so that selection modules can choose the most promising results at each processing stage or delay the decision until more information becomes available. Packed representations such as the WHG (Word Hypotheses Graph) and VHG (VIT Hypotheses Graph) together with formalisms for underspecification capture the non-determinism in each processing phase, so that the remaining uncertainties can be reduced by linguistic, discourse and domain constraints as soon as they become applicable.

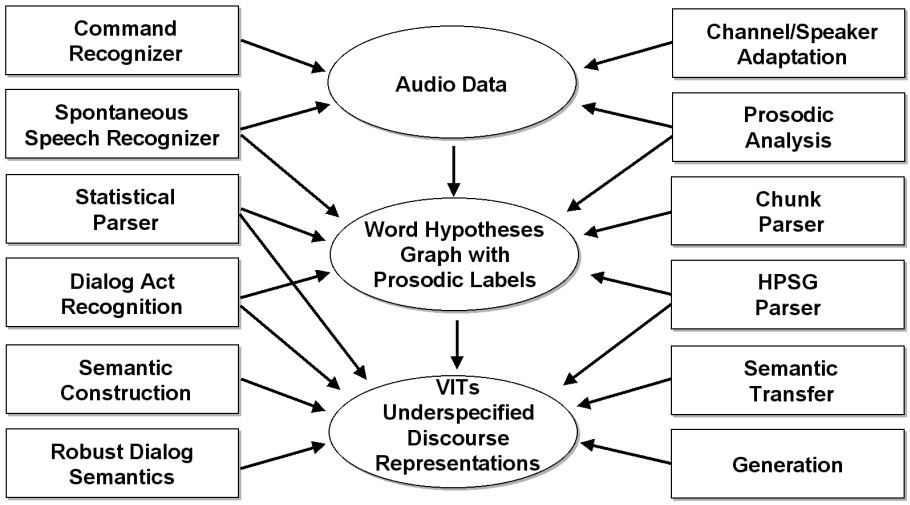

Fig.10. Some key blackboards with their subscribing modules

VITs (Verbmobil Interface Terms) are used as a multi-stratal semantic representation by the central blackboards for the deep processing threads in Verbmobil. The semantic representation in a VIT is augmented by various features concerning morpho-syntax, tense, aspect, prosody, sortal restrictions and discourse information. VITs form the input and output of the modules for robust semantic processing and semantic-based transfer. The initial design of the VIT representation language was inspired by underspecified discourse representation structures. VITs provide a compact representation of lexical and structural ambiguities and scope underspecification of quantifiers, negations and adverbs. Figure

10 illustrates the role of VITs as a common semantic representation language for the blackboards of Verbmobil.

8 Verbmobil's Multi-Engine Approach

Verbmobil performs language identification, parsing and translation with several engines simultaneously. Whereas the multi-engine parsing results are combined and merged into a single chart, a statistical selection module chooses between the alternate results of the concurrent translation threads, so that only a single translation is used for generating the system's output.

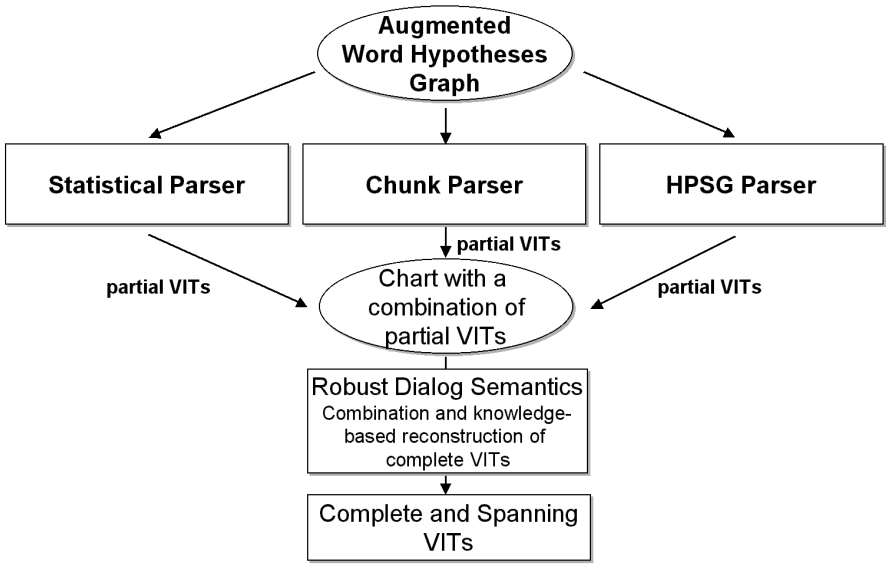

Fig. 11. Verbmobil's multi-engine parsing approach

Verbmobil uses three parallel parsing threads: an incremental chunk parser, a probabilistic LR parser and a HPSG parser. These parsers cover a broad spectrum with regard to their robustness and accuracy. The chunk parser produces the most robust but least accurate results, whereas the HPSG parser delivers the most accurate but least robust analysis. All parsers process the same word hypotheses graph with its prosodic annotations. The search for the best scored path (according to the acoustic score and the language model) is controlled by a central A* algorithm that guides the three parsers through the word hypotheses graph. The HPSG parser may return more than one analysis for ambiguous inputs, whereas the chunk parser and statistical parser return always only one result. Each parser uses a semantic construction component to transform its analysis results into a semantic representation term. Even partial results of the

different parsing engines are integrated into a chart of VITs, that is further analyzed by the robust semantic processing component.

The final Verbmobil system includes five translation engines: statistical translation, case-based translation, substring-based translation, dialog-act based translation, and semantic transfer. These engines cover a wide spectrum of translation methods. While statistical translation is very robust against speech recognition problems and produces quick-and-dirty results, semantic transfer is computationally more expensive and less robust but produces higher quality translations (see Figure 12). However, it is one of the fundamental insights gained from the Verbmobil project, that the problem of robust, efficient and reliable speech-to-speech translation can only be cracked by the combined muscle of deep and shallow processing approaches.

The language identification component of Verbmobil uses also a multi-engine approach to identify each user's input language. The three instances of the multilingual speech recognizer for German, English, and Japanese run concurrently for the three first seconds of speech input. A confidence measure is used to decide which language is spoken by a particular dialog participant. The language identification component switches to the selected recognizer that produces a word hypotheses graph for the full utterance. Verbmobil's error rate for this type of language identification task is only 7.3% .

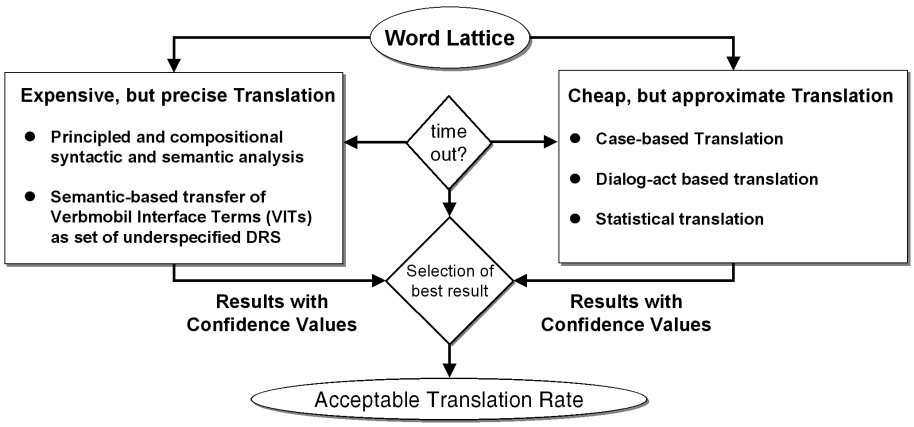

Fig. 12. Competing strategies for robust speech translation

9 Summarizing Dialogs

Another novel functional feature of Verbmobil is the ability to generate dialog summaries. Suppose that two speakers negotiate a travel plan: one can ask the system either to specify the final agreement, omitting the negotiating steps, or to summarize the steps of argument while leaving out irrelevant details of wording.

A dialog summary can be produced on demand after the end of a conversation (see Figure 13). The summaries are based on the semantic representation of all dialog turns stored in the dialog memory of Verbmobil. It is interesting to note that dialog summaries are mainly a bye-product of the deep processing thread and the dialog processor of Verbmobil. The most specific accepted negotiation results are selected from the dialog memory. The semantic-based transfer component and the natural language generators for German and English are used for the production of multilingual summaries. This means that after a conversation over a cell phone the participants can ask for a written summary of the dialog in their own language. The dialog summary can be sent as a HTML document using email. In the context of business negotiations Verbmobil's ability to produce written dialog summaries of a phone conversation is an important valued-added service.

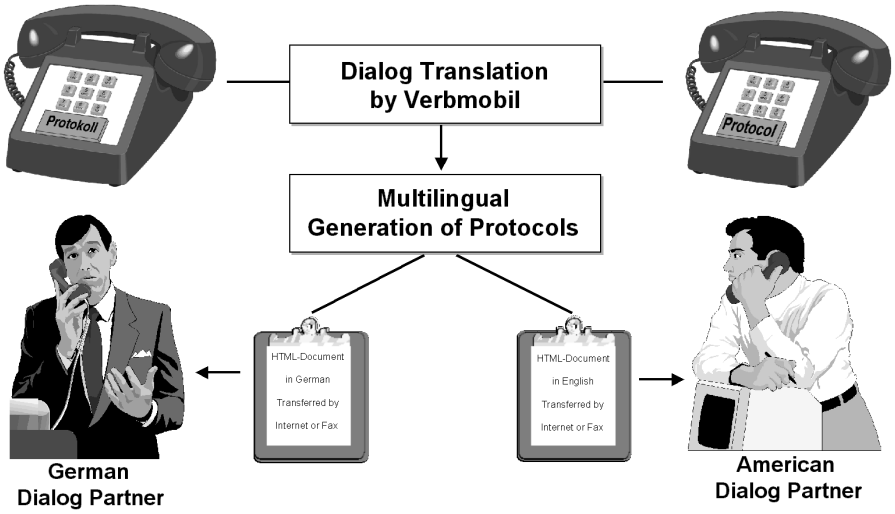

Fig. 13. The generation of a multilingual dialog summaries

10 Conclusion: Lessons Learned, Open Problems, and Future Impact

One of the main lessons learned from the Verbmobil project is that the problem of speech-to-speech translation of spontaneous dialogs can only be cracked by the combined muscle of deep and shallow processing approaches:

- deep processing can be used for merging, completing and repairing the results of shallow processing strategies
- shallow methods can be used to guide the search in deep processing
- statistical methods must be augmented by symbolic models to achieve higher accuracy and broader coverage
- statistical methods can be used to learn operators or selection strategies for symbolic processes

The final Verbmobil architecture supports large and robust dialog systems and maximizes the necessary interaction between processing modules:

- in a multi-blackboard and multi-engine architecture, that is based on packed representations on all processing levels and uses charts with underspecified multi-stratal representations, the results of concurrent processing threads can be combined in an incremental fashion
- all results of concurrent and competing processing modules come with a confidence value, so that statistically trained selection modules can choose the most promising result at each stage, if demanded by a following processing step-packed representations together with formalisms for underspecification capture the uncertainties in each processing phase, so that the uncertainties can be reduced by linguistic, discourse and domain constraints as soon as they become applicable.

10.1 Open Problems

Although the Verbmobil project has successfully met all project goals, many open problems in language technology remain that we must solve in the next decade. Current language technology relies heavily on machine learning approaches. But there are three major problems with the current corpus-based learning methods:

- data collection is very expensive
- the training data sets are cognitively unrealistic
- the data sparseness causes problems for important, but infrequent words

In addition, there are various problems with hand-crafted knowledge sources that are used in hybrid processing schemes to complement the statistical models:

- the methods are still quite brittle
- the knowledge sources are often domain dependent
- the scalability of hand-crafted knowledge sources is limited

Although enormous progress has been made during the past decade, most operational systems work only in restricted domains, with limited vocabularies, and for a single language only. Scaling up, rapid porting to new discourse domains, and achieving full multilinguality are key challenges for future research in language technology.

Fig. 14 surveys the mayor application areas for multilingual language technology for the next decade. In addition to dialog translation, multilingual access to huge video and audio archives has a great application potential. Speech-based web access will become increasingly important in mobile and hands-free situations. Finally, the synergistic use of speech and gesture recognition will lead to more intuitive user interfaces to advanced e-services. In SmartKom, the follow-up project to Verbmobil, we are working on such an intelligent multimodal interface agent.

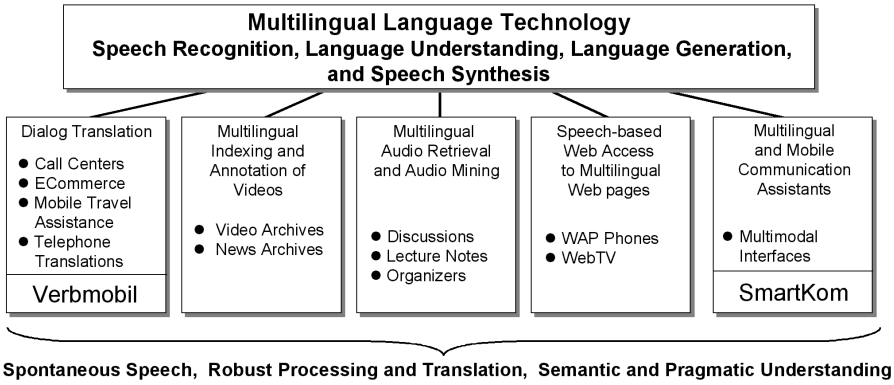

Fig. 14. International research trends in multilingual systems

10.2 Future Impact

In the next decade, speech technology will allow us to store retrieve and process spoken language like we are doing it today for written texts. Thus in many situations, writing and reading may be substituted by speaking and hearing. Meeting minutes, transcribed interviews, or lecture notes contain less information than the original spoken contributions, since the emotional colouring, the disambiguation and the focusing effects of prosody cannot be covered by the transcription. Thus important information about the speaker's affective state, the situative context and the speaker's intention are lost in textual transcriptions. When intelligent audio mining methods will allow us to easily retrieve every utterance that we have produced or heard during our lifetime, we may return to a more oral society like during the thousands of years before Gutenberg. However, these early oral societies had no mass storage for audio information, no automatic

processing and no retrieval tools for spoken language. Today in our textual society, we pass news and knowledge mainly textually, since we have digital mass storage for texts and can easily process and retrieve texts on our computers. Let us conclude with a speculative claim: Human language technology will allow us to return from a textual knowledge society to a more oral knowledge society in about fifty years, when the digital storage, processing and retrieval of spoken language will be as easy, fast and widely available as it is today for written language. After all, it is well known that the human cognitive system is more adapted to speaking and hearing than to writing and reading.

References

1. Bub, T., Wahlster, W., and Waibel, A. *Verbmobil: The Combination of Deep and Shallow Processing for Spontaneous Speech Translation.* In Proceedings of the International IEEE Conference on Acoustics, Speech and Signal Processing, München, Germany (1997) 71–74
2. Cole,R., Mariani, J., Uszkoreit, H., Zaenen, A, Zue, V. (eds.) *Survey of the State of the Art in Human Language Technology.* Cambridge: Univ. Press (1998)
3. Jurafsky, D., Martin, J.H. Speech and Language Processing *An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* New Jersey: Prentice-Hall (2000)
4. Manning, C.D., Schütze, H.: *Foundations of Statistical Natural Language Processing.* Cambridge: MIT Press (1999)
5. Maybury, M, Wahlster, W. (eds.): *Readings in Intelligent User Interfaces.* San Francisco: Morgan Kaufmann (1998)
6. Takezawa, T., Morimoto, T., Sagisaka, Y., Campbell, N., Iida, H., Sugaya, F., Yokoo, A., and Yamamoto, S. *A Japanese-to-English Speech Translation System: ATR-MATRIX.* In Proceedings of the ICSLP (1998) 957–960
7. Wahlster, W. *Verbmobil: Translation of Face-to-Face Dialogs.* In Proceedings of the Fourth Machine Translation Summit, Kobe, Japan (1993) 128–135
8. Wahlster, W. (ed.): *Verbmobil: Foundations of Speech-to-Speech Translation.* Berlin, Heidelberg, New York: Springer (2000)
9. Woszczyna, M. (ed.) *Proceedings of the C-STAR Workshop.* Schwetzingen, Germany (1999)