

Context tracking in monitoring systems: Verbmobil

Cecilia Hemming

Department of Languages, University College of Skövde
Swedish National Graduate School of Language Technology, GSLT

Introduction

Verbmobil is a speech-to-speech translation system for spontaneous dialogues. It provides mobile phone users with simultaneous dialogue translation services in English, German and Japanese for restricted topics. These topics are appointment scheduling, travel planning and remote PC maintenance. The system is special in that it mediates a dialogue by continuously monitoring and processing the participants input. Using a dialogue memory and respective domain knowledge it provides context-sensitive translations following the dialogue in any direction. In this paper I will present parts of the Verbmobil system, which together make context tracking possible. The review and study is done in the frame of the course Dialogue Systems held by Lars Ahrenberg within GSLT in the spring 2002.

The first section contains a brief system overview of the Verbmobil translation system. In section 2 issues regarding automatic speech recognition of natural language is presented and in section 3 I give an overview of Verbmobil's parsing component showing how some of the problems have been treated here. Next section deals with linguistic information representation and how the system's linguistic and dialogue components can exchange information. The 5th section treats dialogue coding and dialogue act classification methods. Section 6 deals with the translation modules and section 7 finally summarizes the topics.

1.0 System Overview

The Language Processing Modules in Verbmobil

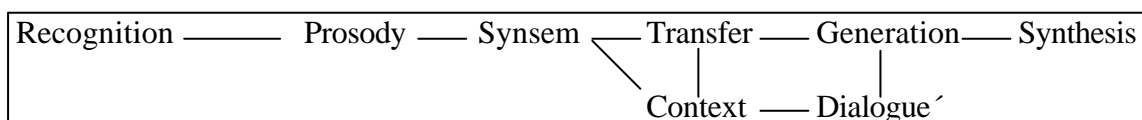


Figure 1. Overview of the language processing modules in Verbmobil

The Verbmobil architecture is based on collaborating modules handling specific tasks. As the dialogue is two-way directed, it both contains a speech recognition component that converts acoustic-phonetic parameter sequences into word strings as well as speech generation and synthesis modules. The prosody module encounters and transmits prosodic information in the word strings and so also adds information for the natural language interpretation. The Synsem module handles syntactic and semantic information. The Transfer component is working on the semantic level, using several knowledge sources. The Dialog Component deals with dialogue phases and acts and handles information of utterance context and previous dialogue.

2.0 Word hypotheses

In freely formulated speech, as a dialog goes on, thoughts are consecutively converted into words and phrases. Very often grammatical rules and syntactically right word orders are put aside to enable a more efficient communication. When it comes to analyzing spoken language, this and phenomena like hesitations, interjections, new starts and repetitions exclude many well known text parsing strategies as these often rely on wellformedness criteria. In automatic speech recognition (ASR), noise can not only be introduced by the users of the system but also by the speech recognizer itself. This two-level uncertainty makes the analysing double hard. Speech recognizers produce many erroneous word hypotheses. Most of them take a speech signal as input and produce a set of separate word hypotheses over time. A word hypothesis typically has four parts (Wermter & Weber, 1997):

- the start time
- the end time
- the word string of the hypothesis
- a plausibility of this hypothesis (based on the confidence of the speech recognizer)

The different word hypotheses, which can overlap in time, are in Verbmobil described as directed graphs. Each word hypothesis is then a separate node in this graph and it can be connected to an adjacent wordgraph if its end time is directly before the latter's start time. An example: if the word hypothesis for "am" ["on"], in the figure below, ends after 0.43 seconds and the hypothesis "sechsten" ["sixth"] starts at 0.44 seconds, these two hypotheses can successfully form a consecutive word sequence, within Verbmobil called a *word hypotheses graph* (WHG). As normally several hypotheses are produced for a phonetic string, there are also several suggestions of which word follows which in the resulting sequences. See figure 2 below for an idea of such word hypotheses sequences. At this stage the system has to be very fault tolerant to handle also speech errors, irregular word orders and wrongly recognized words (which can be due to different dialectal pronouncing, for instance).

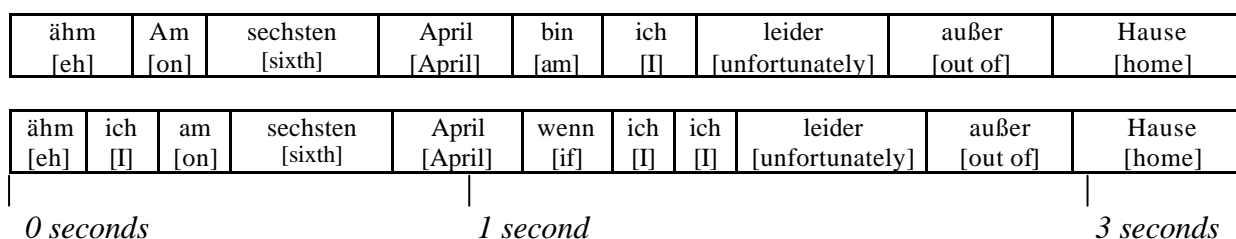


Figure 2. Two different word hypothesis sequences in a word graph. (Wermter & Weber, 1997)

3.0 The Parsing Component

What if a system would cause a break for each illformed phrase going through the parser? It would not be of much use. People tend to express themselves in a very "free" way, using linguistic shortcuts, ambiguous phrases, etc. It is thus important that even suboptimal sentence hypotheses can be handled by a dialogue system. Standard top-down chart parsers can fail completely just because of a single semantic or syntactic category error. In context free grammars the construction of a complete tree demands wellformed input. But rather than having the system to continuously request repeated or reformulated phrases, one can extract the existing information also from partially erroneous sentence hypotheses. From a corrupt phrase like "I had I suggest 7 March" one could at least conclude that an agent, I, said something about the time, 7 March. This can be valuable information for the system and should not cause it to break. In Verbmobil, a combination of statistic and linguistic processing could

for instance place this utterance¹ in the dialogue act "suggest_date" with the date specification "7th Mai". In stead of trying to force a deeply structured representation onto the analysed phonetic strings, the Verbmobil solution allow three different parsers to run in parallel, each giving the best result that it can. In difficult interpretation cases, these different results are combined and presented to a selection mechanism. The selection is based on the parsers information, how it was derived, and compares this to a statistical prediction calculated from a considerable corpus of previously analysed terms. For an overview of the parsing system see figure 3.

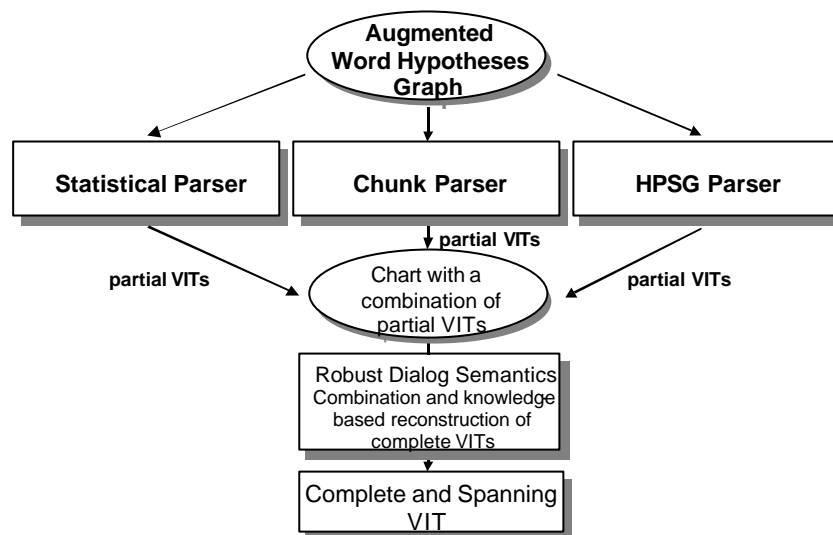


Figure 3. In Verbmobil, three parsers cooperate to create information for the semantic processing component (wahlster, 2000).

The different parsers used in Verbmobil are:

- An HPSG²-parser, that provides the most detailed and theoretically motivated analyses (most accurate but least robust).
- A chunk parser for robust, partial parsing aiming at broad data coverage (most robust but least accurate).
- A statistical LR-parser³, implementing a level of accuracy and robustness between the two other parsers.

Together, the parsers can cover a broader spectrum of robustness and accuracy. The chunk parser outputs the most robust but least accurate analyses, whereas the HPSG-parser, that can return several analyses for ambiguous inputs, produces the most accurate but least robust results. All three parsers are controlled by an integrated processing module that guides them through each word hypotheses graph by a central A* algorithm. The A* searches the best scored path according to the language model and an acoustic score calculated from the sequence's prosodic annotations (i.e. stress boundaries and intonation).

¹ An utterance corresponds to a clause and must contain a finite verb. A complex sentence with two finite verbs is considered as one utterance if one of the verbs is a complement verb and as two utterances otherwise. Within the system the following segments are also regarded as utterances: whole turns even if they do not correspond to a clause, fixed lexems or phrases, single NPs expressing certain dialogue acts, e.g. "suggest" and "clarify".

² **Head-driven Phrase Structure Grammar**, an approach to grammatical theory that tries to model human languages as systems of constraints on typed feature structures .

³ "L" = **L**eft-to-right scanning of input string, "R" = rightmost derivation in **R**everse.

Each parser uses a semantic construction component to transform its result into a semantic representation term, within Verbmobil called VIT (Verbmobil Interface Term). The different analysis results are then merged and integrated into a single chart of VITs which is further analysed by the system's semantic processing component.

4.0. Verbmobil Interface Term, VIT

The VITs play a central role in the Verbmobil system: they carry all linguistic information of an utterance that is needed for the translation but they are also important in that they provide an interface representation used for information exchange between linguistic and dialogue components.

A VIT carries multi-layered linguistic information encoded in a record-like data structure following the Discourse Representation Theory (DRT) of Kamp and Reyle from 1993 (Dorna, 1996). Fields within the VIT-structure contain variable-free lists of non-recursive terms, known as "flat" (as opposite to *deep*) set representations. The data give for instance information about *semantics, scope, sorte, morphosyntax, prosody and discourse*, which correspond to the information needed for translation, see figure 4 below. Beside "VIT id" and "Index" all slots are encoded as variable free term lists. Lists are advantageous as they are easy to manipulate. In for instance Lisp and Prolog they are built-in already, and they can easily be ported to other programming languages.

Slot name	Description
VIT id	Combines a unique tag for a turn segment described by the current VIT and the word lattice path used in its linguistic analysis.
Index	A triple consisting of the entry points for traversing the VIT representation.
Conditions	Labelled conditions describing the possibly underspecified semantic content of an utterance.
Constraints	Scope and grouping constraints, used for underspecified quantifier and operator scope representation.
Sorts	Sortal specifications for instance variables introduced in labelled conditions.
Discourse	Additional semantic and pragmatic information, e.g. discourse roles for individual instances.
Syntax	Morpho-syntactic features, e.g. number and gender of individual instances.
Tense and Aspect	Morpho-syntactic tense combined with aspects and sentence mood information, e.g. used for computing surface tense.
Prosody	Prosodic information such as accenting and sentence mood.

Figure 4. Overview of the Verbmobil VIT, a consistent representation for linguistic information exchange among the system's linguistic modules (Schielen & al. 2000).

Within the VIT, the slots with information can be seen as analysis layers that collect different types of linguistic information from various Verbmobil-components. They can be accessed and manipulated separately depending on each component's functionality. The information between and within the different layers are linked together by constant symbols, so called *labels* (formulas or VIT conditions arguments), *instances* (discourse referents) and *holes* (underspecified scope, resembling parts of formulas with unknown content). These constants can be seen as skolemized⁴ logical variables each denoting a node in a graph. With skolemized input unwanted input-variable unification is avoided during the matching of individual transfer rules against the semantic representation.

So, the VIT representation combines information from the different levels of linguistic analysis and this is how the system is able to deal with ungrammatical input. E.g. if an argument is missing for a verbal semantic predicate, there is still information about its case and sort and what syntactic subcategorization frame it would be placed in. The overall content of a VIT corresponds to a single (segmented) utterance in a dialog turn⁵, which enables the different linguistic parts to work incrementally.

⁴ Skolemization in this regard corresponds to replacing existentially quantified variables with a fully new constant that stands for the object that is asserted to exist.

⁵ The definition of a turn within Verbmobil is one contribution of a dialogue participant that can be further divided into utterances. For the definition of utterances, see footnote 1, page 3.

The goal of the linguistic analysis components is to produce the most accurate representation of a spoken input within the time available (Pinkal et al. 2000). Whereas the results from the three parsers are combined and merged into a single chart, the statistical selection module picks out only one of the resulting translation threads to be used as input for the translation component. The most probable of several possible and concurrent translation candidates is thus used for generating the target language (TL) phrase. The input to the transfer component is one single VIT analysis representation of a source language (SL) utterance, and the output is a corresponding VIT for the TL-synthesis.

The VIT-representation also constitutes the critical part when it comes to generating dialogue on-demand minutes and also dialogue summaries. Relevant data are for this selected from the dialogue memory and converted into sequences of VITs. The generation module is used to produce textual documents which can be translated by the transfer module.

5.0 Dialogue Coding

The two levels of dialogue coding in Verbmobil are: *dialogue acts* superordinated by *dialogue phases*. Each of the phases is optional and do not have to be ordered in a certain way. In negotiation dialogues the system distinguishes between five different dialogue phases:

Hello - participants greet each other and introduce themselves.

Opening - the negotiating topic is introduced.

Negotiation - the negotiation phase between Opening and Closing.

Closing - negotiation is finished, agreement is met and the topic is maybe summarized.

God_Bye - participants say good bye to each other.

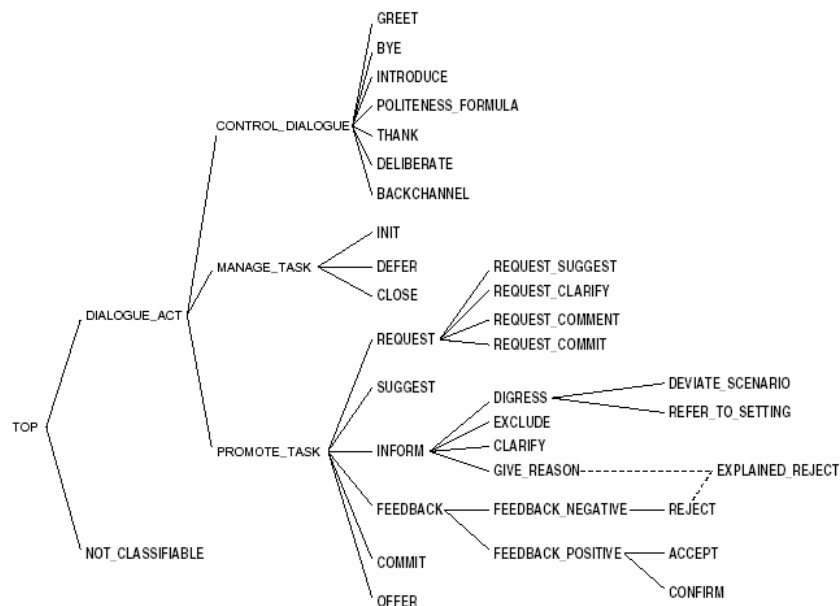


Fig. 5. The dialogue act hierarchy in Verbmobil (Alexandersson, 1998).

To translate a word properly it can be crucial to know in what phase of the dialogue it was uttered. With this information it is for instance possible to disambiguate the German word *wiederholen*. The sense of its English translation *recapitulate* is normally used to give a summary of a discussion, thus it can be found at the discussion end, in the closing-part. The word in its sense *repeat* can instead occur at any place in the dialogue. The dialogue phases are further divided into different Dialogue Acts. An

utterance can be assigned one, two or even more dialogue acts if one single definition does not cover the utterance illocution (e.g. a counter-suggestion can be viewed as being a REJECT and a SUGGEST at the same time).

The Verbmobil approach to utterance representation is based on the assumption that the dialogue act (henceforth DA) plus its propositional content forms the meaning of an utterance, this in accordance with Stephen C. Levinson's claim that "the illocutionary force and the propositional content of utterances are detachable elements of meaning" (Alexandersson et al. 2000:449). All the four sentences below are normally used with different illocutionary forces (i.e. performing different speech acts), nevertheless, they express the same propositional content (that the addressee will go home):

I predict that you will go home. Go home! Are you going to go home? I advise you to go home.
(Alexandersson et al., 1998)

The number of propositions co-occurring with a DA is constrained and this information can be used when annotating DAs to utterances. Each DA has a set of possible propositions, e.g. ACCEPT can contain the accepted proposition's anaphoric or explicit reference (a date or duration, a location, a transportation or accommodation, an action). The distinction between the DAs CONFIRM and CLOSE can be made in that the former must contain propositional content and the latter must not.

Compared to other DA annotation schemes the Verbmobil coding schema is rather large and complex. The system is special though in that it has to consider features about the dialogue as well as the different translations. A closer look on other comparable schemes does not fit in this paper, but a survey of different coding schemes can be found at the MATE project web site. The MATE project, funded by the European Commission, aimed to work out standards in this area. Markup schemes used so far makes it difficult to reuse both corpora and annotation tools, furthermore different systems use different levels of annotation (prosody, morphosyntax, coreference, dialogue acts), see <http://www.dfki.de/mate/d11/chap4.html>.

5.1 Classification Methods

Analyses of speech acts in especially English and German tutorial dialogues have suggested that the first three words of a speech act is a significant cue for DA classification. The previous two turns in a dialogue are also known to be moderately predictive of the next DA. Presume that an utterance in a negotiation is recognised as an offer. The following utterance to process can then be searched for sequences that signal appropriate speech acts, like for instance a REJECTION which could be represented by the string "I'm sorry!". Marineau and her colleagues suggest in (Marineau et al., 2000) that likewise features could be very efficient input units in a neural network system. They also stress the view that neural networks might outperform symbolic systems in classifying ambiguous speech acts, this by the net's ability to weigh conflicting information.

Other useful features that can be used for the classification is prosodic information. Results in (Shriberg et al. 1998) claims that DAs are redundantly marked in natural conversation and that a variety of automatically extractable prosodic features could aid dialogue processing. In Verbmobil, prosodic information like duration, pitch, energy and pauses, is used to 1) extract information about boundaries (e.g. phrase and word boundaries), 2) for DA segmentation and recognition and 3) for sentence mood designation helping the transfer based translation. 4) It also gives information about word accents used by the generation component and 5) about prosodic features that plays a role for speaker-adaptation in the speech synthesis. (Batliner et al. 2000:107-108) claim that Verbmobil would be the only complete speech understanding system that really uses prosodic information and they point to some major difficulties. According to them a main problem of using prosodic information is that it is not clear how many prosodic classes one can distinguish. In addition to that, different prosodic

information influences each other, different functions realized with the same parameters interfere with each other and there is also a trading relation amongst some of them: a smaller value of one parameter can be compensated with a greater value of another.

For automatic classification of DAs in larger corpora, there are mainly three basic methods used: *statistical methods using language models* (e.g. Reithinger und Klesen, 1997), *neural networks* (NN) (e.g. Kipp, 1998) and *transformation based learning* (TBL) (e.g. Samuel et al. 1998).

In the classification by TBL Samuel and his colleagues uses a modified version of Brill's TBL-method. To get round the sparse data problem they let the ir system consider several utterance features. These features are: cue phrases, word n-grams, speaker information, punctuation marks, number of words, the entire utterance for 1-, 2- and 3-word utterances, DA of preceding and following utterances and they also use what they call *dialogue act cues*, which they claim are more effective than solely cue phrases and word n-grams. There are three groups of these dialogue act cues: *traditional cues*, as "but" and "so"; *potential cues*, as "thanks" and "See you!"; and finally *domain cues* belonging to a certain domain, as "what time?" and "busy" in appointment scheduling corpora. Given a set of rule templates (restricting the range of patterns to consider) a sequence of symbolic rules is devised from a correctly tagged training corpus, this using a Monte Carlo algorithm⁶. The rules are then one after the other applied to every utterance to be classified. An example: The first rule in the learned sequence, labels every utterance with the dialogue act SUGGEST. Then the second rule is applied, which could be for instance: *IF an utterance \underline{u} contains the word \underline{w} AND the tag on the utterance preceding \underline{u} is \underline{X} THEN change \underline{u} 's tag to \underline{Y}* . After this the third, the fourth and the rest of the rules are applied in the order they occur. This means that an utterance can change tag several times before all rules in the list have been applied and the final DA classification has taken place. This method has reached results close to statistical ones, see below.

Kipp classified speech acts with neural networks, using Elman's simple recurrent network⁷. The network contained a group of 18 modular networks, representing 18 different DAs. Each of these networks had one YES and one NO neuron as output unit. The resulting yes/no outputs were neither probability values nor comparable amongst each other but they could tell which net's value was the highest (i.e. which of the associated DAs that was the best guess for a certain utterance). The output of a single modular network \underline{d} for an utterance $\underline{u} = \{w_0, \dots, w_n\}$ with output functions \underline{yes}_d and \underline{no}_d was computed by the formula:

$$\frac{1}{n} \sum_{k=0}^n (\underline{yes}_d - \underline{no}_d)$$

The weighted outputs of the modular networks formed the input to a selector network, eventually determining the correct speech act category, see figure below. The training was done on 467 German dialogues taken from the Verbmobil-corpus and Kipp's networks yielded a recall of 60.45 % on the test set. A further evaluation of the method consisted of combining statistics and neural networks, replacing the original speech representation by a vector with statistical information. Each word was here represented by a vector with the length set to 18, one component for each DA, and thus reflecting the words probability distribution over all DAs: $P(d|w)$, the probability that a word \underline{w} indicates the DA annotation of \underline{d} . Applying this enhanced (hybrid-) method to the same data increased the recall to 66.31 %.

⁶ Monte Carlo methods can roughly be described as statistical simulation methods that use sequences of random numbers to perform a simulation. The name "Monte Carlo" comes from the similarity to games of chance.

⁷ A recurrent neural network also contains connections from output nodes back to hidden layer and/or input nodes and allow interconnections between nodes of the same layer. The output of a node is thus dependent of other nodes at the previous instant (Mehrotra et al 1997).

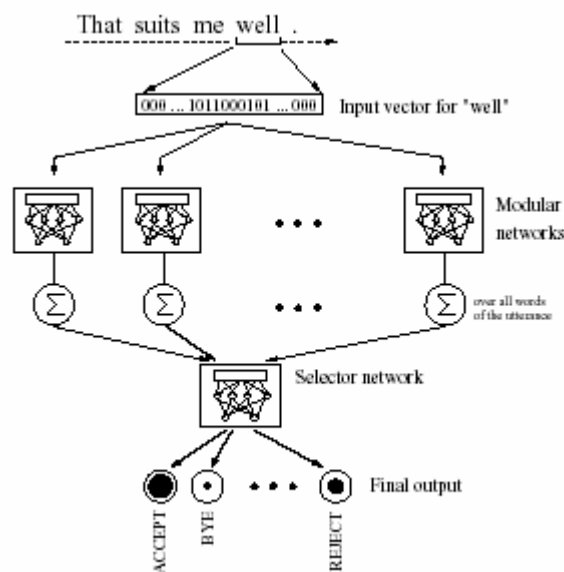


Figure 6. An example of Kipp's neural net system for classification of dialogue acts (Kipp, 1998).

The statistical approach finally is represented by the classification using n-gram language models implemented in the Verbmobil system (Reithinger und Klesen, 1997). This approach takes a word string as input. It decides which dialogue act \underline{D} describes the illocution of a distinct word string \underline{W} with statistical models for each \underline{D} , under the condition of the \underline{W} . As experiments showed that the number of correct classified acts was increased by 3% when taking the dialogue history into account, the classification formula used is the following:

$$D = \operatorname{argmax}_{D'} P(W|D') P(D'|H)$$

1505 dialogues from the Verbmobil data collection were annotated with 19 of the DAs in figure 5 above. From the original 32 acts, seven were not specified as they together covered less than 1% of the annotated utterances and the other six were removed as they often were confused with other closely related acts. All utterances from each dialogue act in the manually annotated corpus were collected and the relative word and transition frequencies calculated, giving the respective a-priori probabilities.

6.0 Translation in context

In contrast to other MT systems presenting several translations to choose among for each input, Verbmobil provides one single and, in respect to its pragmatic and communicative aspects, best approximated translation for each utterance. The conflicting goals of aiming towards at the same time quick, robust, efficient and reliable speech-to-speech translation have been approached by using both deep and shallow processing. On the one hand statistical translation gives quick-and-dirty results and is robust against speech recognition errors, on the other hand semantic transfer is less robust, demands more computer power but then produces higher quality translations. The system includes five translation engines: *statistical translation*, *case-based translation*, *substring-based translation*, *dialog-act based translation*, and *semantic transfer*.

A central repository for the Verbmobil dialogue data, which can be used by the different components, is the *dialogue memory*. It receives data from different modules and provides data on request. Every

module belongs to a special translation track and the dialogue turn is segmented in its own way. This is why the dialogue memory stores distinct lists of segmented objects for each turn and track. A segment object can contain DA, dialogue phase, topic, content expression and the corresponding VIT. It is uniquely defined by a distinct turn number as well as the begin and end time of the translation track.

The statistical translation module uses information from both a translation model and a language model. These models have been trained on a corpus of segmented and transliterated dialogue turns translated by humans. The translation model contains a stochastic lexicon and word position parameters. The component receives the single best sentence hypothesis from the ASR, added with prosodic information of sentence mood, accentuated syllables and phrase boundaries. Alignment templates are used to find word group dependencies in the SL and TL. The output of the statistical translation module is a sequence of TL-words with confidence measure which is presented to the selection module.

The case-based translation module uses translation templates learned from a sentence-aligned corpus. The problem of uncertainty, caused by the non perfect ASR, can to some extent be avoided by allowing a word lattice with scored alternatives. The translation component can then decide what the input probably was. The first step in the process is to identify edge sequences of the lattice that are candidates for embedded subphrases corresponding to certain template variables. Definite clause grammars are used to detect and mark date, time and name expressions. These grammars allow a comparison and alignment of the SL and TL expressions based on the semantic content. An A* search algorithm is used to find the best matching SL template for the word lattice. Not only exact matches but also mismatches and transitions for word omissions or additional words in the input are allowed. In the cost calculation for the search algorithm these “errors” acquire an appropriate penalty. The recognition confidence of the ASR module for the lattice’s matched path is also considered in the cost calculation. So if a solution is found, the confidence value of the path is accordingly affected by these phenomena. After the SL template matching, SL phrases are bound to variables in the translation templates and the interlingual (IL) representation of these dates, times and names are transformed to TL instances by generation grammars. Finally the target language utterance is generated by instantiating the TL parts of the matched templates with the translated subphrases. (Auerswald, 2000). The Substring-Based Translation approach is an attempt to provide a translation for any piece of input (words or sequences of words) as soon as it is available from the ASR. This, even if it means that the first translation turns out to be wrong after that further segments have been processed. This is automatically signalled by the system by insertion of a marker “*I mean...*” before the translation correction. In the training statistical word alignment is combined with preprocessed translation chunks and contextual clustering of word classes.

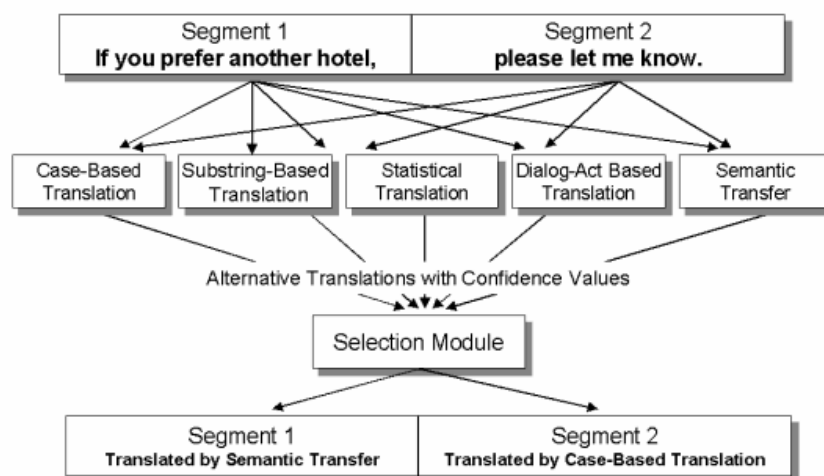


Fig. 5. The five translation engines of Verbmobil (Wahlster, 2000).

The dialogue-act based translation is realized from the statistical dialogue act classifier and a cascade of finite-state transducers that extract the utterance's main propositional content. The content together with the topic and the recognized dialogue act is represented by a simplistic frame notation of nested objects with possible attributes. This interlingual representation is transformed into the TL by template based generation. To be available for further processing, it is then stored in the dialogue memory together with topic and focus information as well as the utterance's VIT containing the deep semantic representation.

The semantic-based transfer component, finally, combines Beaven's Shake-and-Bake method with a semantic transfer approach. The translation component receives a SL VIT as input. If necessary it retrieves information and inferences from the evaluation component (containing dialogue processing and contextual disambiguation) and then act as a rewriting system constructing a TL output set by applying bilingual transfer rules. It uses a cascaded rule system, first for the transfer of idioms and other non-dividable expressions and then for lexical items. The component finds translation equivalences to sets of semantic entities rather than morpho-syntactic lexical items as in the original Shake-and-Bake method. To reduce the number of transfer rules, spatial and temporal prepositions are translated by means of IL. This translation component is very fast, using approximately 1% of the overall utterance processing time. (Emele et al. 2000)

6.1 Keeping track of the context

As mentioned above, in Verbmobil, ambiguities are represented through underspecification, and when the ambiguity has to be resolved (for the translation) the resolution is triggered on demand. The context module performs several types of disambiguation and uses several knowledge sources in conjunction to resolve anaphora and lexical ambiguities. Examples of knowledge sources are: morphological as well as syntactic and semantic information, knowledge about dialogue states or world knowledge. The dialog component provides all the information about the utterance context and the previous dialogue.

In Verbmobil's semantics module, the final processing step is called Discourse and Dialog Semantics (DaD). DaD focuses on resolving ambiguities that require context information as prosody, DAs and dialogue history. The component gets VITs (the ones selected for further processing) as input and adds to them prosodic information calculated from the word hypotheses graph, that is: word-accent, word-emphasis and prosodic sentence mood (whether the sequence is declarative or a question) (Bos and Heine, 2000:337). The dialogue history is built up from the semantic representations by continuously saving a stack of the last five disambiguated VITs. In addition to that, all possible antecedents for anaphora and ellipsis resolution from the two precedent turns can be accessed from a second stack. It keeps track of the dialogue structure, storing information about the users' goals and intention. It keeps information about the thematic structure, i.e. stores relationships between the dialogue suggestions and finally it contains data about the referential structure, which represents the conceptual and linguistic data for the different utterances.

6.2 Contextual Disambiguation

To generate appropriate translations one has to keep track of the context. In a dialogue system with near-realtime translation ambitions, it is especially important to optimize the disambiguation processes. Some ambiguities can be resolved by using information found in the same utterance, in these cases the transfer module solves the problem itself within the VIT currently processed. For other resolutions, though, this local information is not sufficient. The transfer formalism allows the use of both local and non-local information in the translation rules' conditions, so information from external

sources is also available. A dialogue history is an important knowledge source for contextual information. The German utterance "*Geht es bei Ihnen?*" can be interpreted as "*Does it suit you?*" when discussing about times and dates, but at the same time as "*How about your place?*" when negotiating about a place to meet. Consider also the following phrases:

Wie wäre es um eins? Wir könnten gemeinsam *Essen* gehen.
How about one o'clock? We could go for *lunch* together.

Wie wäre es abends? Wir könnten gemeinsam *Essen* gehen.
How about in the evening? We could go for *dinner* together.

(examples from Emele et al. 2000: 367)

It is not possible to disambiguate the German word *Essen* with the help of information found within the sentences. We need knowledge to specify the time of the day and connect this with the appropriate translation. Inferences and information that guide context sensitive translations are taken from the evaluation components, which handles contextual disambiguation and dialogue processing.

7.0 Summary

In this paper I have looked into some components in Verbmobil that together make it possible to translate and make summaries with the context of each utterance taken into account.

The speech recognition component of Verbmobil takes as input a spoken source language utterance in German, English or Japanese and gives as output a word hypotheses graph, containing alternative string sequences annotated with probability scores. There are three parsers running in parallel: a chunk parser, which gives a very robust result, a HPSG-parser, which is much more accurate than robust and finally a statistical parser. The system's prosody component complete the word lattice produced with prosodic information, like stress boundaries and intonations. The word lattices are record-like data structures, so called VITs, encoding multi-layered linguistic information. The VITs carry all needed linguistic information of utterances and also provide an interface representation used for communication between components within the system.

The Verbmobil-strategy is to keep ambiguities ambiguous as far as possible in the process, this is realised by representing ambiguous parts through under-specification. Then when a resolution is needed it is done on demand. The context module uses several knowledge sources in conjunction to resolve anaphora and lexical ambiguities.

There is a conflict in aiming for translations that at the same time should be as quick, robust, efficient AND reliable as possible. The approach used in Verbmobil, that of combining information of both deep and shallow processing, is one that seems to bring a consecutive translation system like this as far as it can get with existing techniques. Quickly performed statistical translation, that are robust against speech recognition errors, opposed to less robust semantic transfer, demanding more computer power but resulting in higher quality translations, make the two opposite and complementing approaches. There are five translation engines in all, performing: *statistical translation, case-based translation, substring-based translation, dialog-act based translation, and semantic transfer.*

References

Alexandersson, J., Buschbeck-Wolf, B., Fujinami, T., Kipp, M., Koch, S., Maier, E., Reithinger, N., Schmitz, B., Siegel, M. 1998. Dialogue Acts in VERBMOBIL-2, Report 226.

Alexandersson, J., Engel, R., Kipp, M., Koch, S., Küssner, U., Reithinger, N. and Stede, M. Modeling Negotiation Dialogs. In Wahlster, W. 2000. *VerbMobil: Foundations of Speech-to-speech Translation*. Springer-Verlag, Berlin-Heidelberg, Germany.

Alexandersson, J., Reithinger, N., Maier, E. 1997. Insights into the Dialogue Processing of VERBMOBIL, Report 191.

Auerswald, M. Example-Based Machine Translation with templates. In Wahlster, W. 2000. *VerbMobil: Foundations of Speech-to-speech Translation*. Springer-Verlag, Berlin-Heidelberg, Germany.

Batliner, A., Buckow, J., Niemann, H., Nöth, E., Warnke, V. The Prosody Module. In Wahlster, W. 2000. *VerbMobil: Foundations of Speech-to-speech Translation*. Springer-Verlag, Berlin-Heidelberg, Germany.

Bos, J. and Heine, J. Discourse and Dialog Semantics for Translation. In Wahlster, W. 2000. *VerbMobil: Foundations of Speech-to-speech Translation*. Springer-Verlag, Berlin-Heidelberg, Germany.

Dorna, M. 1996. *The ADT-Package for the VerbMobil Interface Term*. VerbMobil Report 104, IMS, University of Stuttgart, Germany.

Emele, M., C., Dorna, M. Lüdeling, A., Zinsmeister, H. and Rohrer, C. Semantic-based Transfer. In Wahlster, W. 2000. *VerbMobil: Foundations of Speech-to-speech Translation*. Springer-Verlag, Berlin-Heidelberg, Germany.

Kipp, M. 1998. The neural path to Dialogue Acts. In *Proceedings of the European Conference on Artificial Intelligence*, 175-179.

Marineau, J., Wiemer-Hastings, P., Harter, D., Olde, B., Chipman, P., Karnavat, A., Pomeroy, V., Rajan, S., Graesser, A. and the Tutoring Research Group. 2000. *Classification of Speech Acts in Tutorial Dialog*. The Department of Psychology, University of Memphis.

Mehrotra, K., Mohan, C. K., Ranka, S. 1997. *Elements of Artificial Neural Networks*. MIT Press, Massachusetts.

Pinkal, M., Rupp, C. J. and Worm, K. Robust Semantic Processing of Spoken Language. In Wahlster, W. 2000. *VerbMobil: Foundations of Speech-to-speech Translation*. Springer-Verlag, Berlin-Heidelberg, Germany.

Reithinger, N. and Klesen, M. 1997. Dialogue Act Classification Using Language Models. In *Proceedings of Eurospeech -97*, 2235-2238.

Samuel, K., Carberry, S., Vijay-Shanker, K. 1998. Computing Dialogue Acts from Features with Transformation-Based Learning. In *Proceedings of the American Association for Artificial Intelligence*, 90-97.

Schielen, M., Bos, J., Dorna, M. VerbMobil Interface Terms (VITs). In Wahlster, W. 2000. *VerbMobil: Foundations of Speech-to-speech Translation*. Springer-Verlag, Berlin-Heidelberg, Germany.

Shriberg, E., Bates, R., Stolcke, A., Taylor, P., Jurafsky, D., Ries, K., N., Martin, R. Meteer, M., Van Ess-Dykema C. Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech? 1998. in *LANGUAGE AND SPEECH Special Issue on Prosody and Conversation*.

Wahlster, W. An Overview of the Final Verbmobil System. In Wahlster, W. 2000. *Verbmobil: Foundations of Speech-to-speech Translation*. Springer-Verlag, Berlin-Heidelberg, Germany.

Wermter, S. and Weber, V. 1997. *SCREEN: Learning a Flat Syntactic and Semantic Spoken Language Analysis Using Artificial Neural Networks*. Verbmobil-Report 190. University of Hamburg, Germany