UTTERANCE GENERATION FOR TRANSACTION DIALOGUES

Joris Hulstijn and Arjan van Hessen
University of Twente
PO BOX 216, 7500 AE Enschede
The Netherlands
{joris|hessen}@cs.utwente.nl

ABSTRACT

This paper discusses the utterance generation module of a spoken dialogue system for transactions. Transactions are interesting because they involve obligations of both parties: the system should provide all relevant information; the user should feel committed to the transaction once it has been concluded. Utterance generation plays a major role in this. The utterance generation module works with prosodically annotated utterance templates. An appropriate template for a given dialogue act is selected by the following parameters: *utterance type*, *body* of the template, *given* information, *wanted* and *new* information. Templates respect rules of accenting and deaccenting.

1 INTRODUCTION

Generation of appropriate system utterances for spoken dialogue systems is not trivial, especially when the dialogue involves transactions, as in systems for distant selling or ticket reservation. Dialogues that involve transactions display a more complex structure than mere inquiry or advisory dialogues. Two tasks are executed in parallel: obtaining information about the product and actually ordering the product (Jönsson, 1993). We have experimented with a dialogue system for theatre ticket reservation, called SCHISMA (Van der Hoeven et al., 1995). In our Wizard-of-Oz corpus we also find such complex behaviour: users inquire, browse and retract previous choices, for instance when tickets are too expensive. Legally, transactions like these involve obligations for both parties: the system should provide all relevant information for the user to make a fair choice. The user is bound by the transaction once it has been concluded.

Currently, we are implementing a spoken language version of SCHISMA. The quality of system prompts has a great impact on the coherence of dialogues. Wrong wording or intonation often leads to misunderstanding. Therefore the module must take the discourse effects of word-order and intonation into account (Dirksen, 1992; Van Deemter et al., 1994; Vallduví, 1990; Rats, 1996). One of the challenges is to build a system with a certain personality that is reflected in the nature of the prompts. We hope to create the impression that the system cares about the transaction. This will hopefully make users feel more committed.

The paper is structured as follows. We start with an overview of the system. In section 3 we explain the parameters that control the utterance generation module and their relation to theories of focus. In section 4 we discuss the *Fluent Dutch* text-to-speech package that is used to pronounce the templates. The paper ends with conclusions and plans for further research.

2 SYSTEM

The system is mixed initiative. There are two types of interaction: inquiry and transaction. During inquiry the user has the initiative; the system answers the user's questions. When the user has indicated that he or she wants a transaction – a reservation in our case – the system takes initiative. The system checks if all parameters needed for a transaction, like the name of the client, are known. If not, the system will ask for them. Our corpus shows that there are no separate phases in the dialogue, and no particular order. The inquiry and transaction functions of utterances are intertwined.

The system consists of three modules in a pipeline architecture. User utterances are processed by a *speech recognition* module[1]. When a keyboard-based dialogue system is adapted for speech, one of the most

[1] In the earlier keyboard-based prototype there was a corpus-based parser, preceded by a module for morphological analysis and error correction.

difficult changes that needs to be made concerns the *verification* mechanism. Current speech recognition for this type of dialogue systems, requires that correct recognition of information is verified. This can be done explicitly, by asking the user if some information was recognised correctly, or implicitly, by incorporating the recognised information in the next prompt. From user experiments we learned that the verification strategy can be greatly improved using *confidence measures* (Bouwman and Hulstijn, 1998; Rüber, 1997). When the confidence measure exceeds a threshold, verification can be suppressed. We also found that it makes sense to take the *cost* of a domain related action into account. For a crucial action, such as the closure of a transaction, information must be verified explicitly. For actions such as responses to inquiries, where recognition errors are less costly, the efficiency gain of suppressing verification prompts, outweighs the risk of misunderstanding. In the SCHISMA system, Verification prompts are distinguished by a rising intonation, indicating insecurity.

The module that decides what to do next, is called the *dialogue manager*. It maintains two datastructures: a representation of the *context* and a representation of the *plan*, the current domain-related action that the system is trying to accomplish. The dialogue manager interprets the parser output against the context. It updates the context, resolves presuppositions and anaphora and applies inference rules that constitute domain knowledge. Based on the parser output, the context, the plan and the database, the dialogue manager selects a response action. Actions are composed of basic *database actions* and basic *dialogue acts*. For example, a reservation involves a plan with sub-actions that deal with performance selection, the price and number of tickets and a final confirmation. Planning and action selection are based on a set of principles, called *dialogue rules*.

Finally, each dialogue act is put into words by a general purpose *utterance generation* module. It determines the utterance-structure, wording, and prosody of each system response. General rules govern the translation of dialogue acts into the parameters that control the style of prompts. A different set of rules produces a system with a different personality. In an efficient, curt system, pronouns and implicit objects are preferred over definite NPs to express the given items, and short answers are preferred over longer answers that echo part of the question.

whq	fin verb on 2nd, wh-word on 1st	hat
decl	fin verb on 2nd, no wh-word	hat
ynq	fin verb on 1st, subject on 2nd	rising
imp	fin verb on 1st, no subject	hat
verif	verification, as decl	rising
text	longer text to be read	punct
short	short answer (PP,NP,ADV)	hat
meta	e.g. thanks, greetings, yes, no,	hat

Figure 1: Basic Utterance Types

3 PARAMETERS

The utterance generation uses a list of utterance templates. Templates contain gaps to be filled with *information items*: attribute-value pairs labelled with syntactic and lexical features. Templates are selected on the basis of five parameters: *utterance type*, the *body* of the template and possible empty lists of information items that are to be marked as *given*, *wanted* and *new*. The utterance type and body determine the word-order and the main intonation contour. The presence and number of information items in the given, wanted and new slots, as well as special features affect the actual wording and intonation of the utterance.

Utterance type The utterance type roughly corresponds to the mood of a sentence: whether it is a declarative, wh-question, yes/no question or imperative. However, many utterances do not consist of complete sentences. Often they involve short answers, commands or remarks, consisting of single VPs, NPs or PPs. Figure 1 shows a list of utterance types. It is a modified version of the utterance types developed for classification of utterances in the SCHISMA corpus (Andernach, 1996). The first column in figure 1 gives the main syntactic features; often the position of the subject and finite verb. The last column gives the main intonation contour. The regular declarative intonation is the *hat* shape: rising during the phrase and falling at the end. The regular interrogative intonation is characterised by a sharp *rising* at the end. When applied to an otherwise declarative sentence, a rising intonation indicates uncertainty. This is applied in verification prompts. For the reading of longer texts, the system assumes a *reading voice* (see below). In text, the intonation depends on punctuation marks.

Body This slot contains a label that selects the body of the template: the content that is not especially marked. It corresponds to the *tail* in the *link-focus-tail* trichotomy of Vallduví (1990). Vallduví's 'link' corresponds to the topic of the conversation, and his 'focus' corresponds to our 'new'[2]. In most templates the body is expressed by the main verb. Usually the it is deaccented, but important cue words like *niet* (not) or *maar* (but) get some extra stress.

Given This slot contains information that is to be presented as if it is given in the conversation or situation. Linguistically, this is usually reflected by the use of pronouns, definite articles or demonstratives. For most templates the given elements occupy the 'topic' position: the first position before the verb, or the position just after the verb when some new element is topicalised. With respect to intonation, given items are deaccented. Very often the 'given' items refer to the topic of the conversation (Rats, 1996), the object the conversation is currently about. But not always. The topic-comment distinction does not align with the given-new or focus-ground distinctions (Vallduví, 1990). In fact, in most utterances the global topic, the performance or the reservation, is not mentioned at all. It is implicit. In the templates the choice between pronouns, definites and implicit objects is marked by features. The decision is made according to the dialogue rules that govern the translation of dialogue acts into the five utterance parameters.

New This parameter contains something like the *focus* of the utterance. The notion of focus is not well defined for questions and imperatives. Therefore we prefer the older given-new distinction. In most templates new elements are placed towards the end of the utterance. The new field is also used for items that need contrastive stress, as in (3c) that suggests that 'Macbeth' is sold out, but no other performances.

Wanted This slot contains the type of information that is wanted from the user. This is indicated by the choice of wh-word, e.g. *why* for reasons, or by the NP in the wh-phrase, e.g. *which genre* for genres.

Examples (1 – 3) show the mechanism of template selection. Only some of the features are shown. The parameters are listed as: `utterance type; body; given; wanted; new`. Features and values are given between brackets.

(1) a. `meta:[att:sorry]; no;;;`
Nee, sorry.
No, sorry
b. `verif;;;date:tomorrow;`
Morgen?.
Tomorrow?

(2) a. `whq; want; you:[polite:+]; thing;`
Wat wilt u?
What would you like?
b. `whq; want; you:[polite:+]; performance[prep:to]; date:tomorrow`
Naar welke voorstelling wilt u morgen?
To which performance would you like to go tomorrow?

(3) a. `decl:[att:sorry]; sold_out; it;;`
Sorry, maar het is uitverkocht.
Sorry, but it is sold out.
b. `decl; sold_out; performance:[impl:+, title:'Macbeth'];;`
'Macbeth' is uitverkocht.
'Macbeth' is sold out.
c. `decl; sold_out; ;; performance:[def:+, title:'Macbeth']`
De voorstelling 'MACBETH' is uitverkocht.
The performance 'MACBETH' is sold out.

4 TEXT-TO-SPEECH AND PROSODY

For pronouncing the utterance templates we use the *Fluent Dutch* text-to-speech system (Dirksen and Menert, 1997). Fluent Dutch runs on top of the MBROLA diphone synthesiser (Dutoit, 1997). It uses a Dutch voice, developed at the Utrecht institute of linguistics (OTS). Fluent Dutch operates at three levels: the *grapheme* level, the *phoneme* level and a low level representation of *phones* where the length and pitch of sounds is represented. For many words, the phonetic description is taken from lexical resources of *Van Dale* dictionaries. Other prosodic information is derived by heuristic rules. We design prompts at the grapheme level. It is possible to manipulate prosody by adding punctuation at the grapheme level, by adding prosodic annotations at the phoneme level or by directly manipulating the phone level. The main operators for manipulation are the following:

question Change the usual hat-shaped intonation contour into a rising contour. This is used for verification prompts and yes/no questions.
quote Set a string of words apart from the rest of the utterance by small pauses and a lifted pitch level. This is used for names and titles.
accent Calculate the intonation centre for the accented phrase. The intonation centre is at the vowel that would receive the highest pitch, if it were pronounced in a normal declarative way. Add 10 % to the pitch level. Lengthen the vowel by 50 %. This is used for new items and important cue words.

[2] Please note that 'focus' contrasts with 'ground'. The AI notion *focus of attention* roughly corresponds to what we call 'topic'.

deaccent Calculate the intonation centre for the deaccented phrase. Level the pitch to the average of the two neighbouring intonation centres. This is used for given items.

We can also adjust other speech parameters, such as the speaking rate and the general pitch level. This influences the general impression of the voice. For reading reviews and other text fragments the system assumes a *reading voice*, characterised by a lower speaking rate and a reduced pitch range. At the end of the transaction the system concludes with a cheerful wish: *Veel plezier met <title>!* (Have fun with <title>!). It sounds cheerful, which is achieved by a higher speaking rate a slightly lifted pith level and a large pitch range.

5 CONCLUSIONS

We conclude that it is indeed possible to generate appropriate system utterances for transaction dialogues. A coherent dialogue can be achieved, among other things, by producing utterances that take the discourse effect of intonation into account. As a rule, information that is assumed to be given in the dialogue is deaccented, expressed as a pronoun, or even left out. Given information is repeated whenever the system is not confident it was recognised correctly by the speech recognition module. Such verification prompts are distinguished by a rising intonation. Information that is to be presented as new, is accented. Quoted expressions, like artist names or titles of performances, are set apart from the rest of the utterance. For reading the texts and reviews that describe the content of performances, the system assumes a 'reading voice'.

We expect to make the user feel committed by adding an extra 'handshake' at the end of the transaction. For the user it is a last opportunity to change his or her mind. It is also meant to show that the system 'cares'; it will hopefully add to the personality of the system. We hope to verify this claim by user centred evaluation studies.

Further Research User behaviour greatly depends on the quality and nature of system prompts. We will experiment with different styles of prompts. In particular, we would like to compare a brief telegram-style system, with a system that uses longer polite utterances.

The SCHISMA prototype embodies the artificial agent *Karin*, that lives in a *virtual theatre* (Nijholt et al, 1998). Visitors can walk around in a virtual copy of our local music centre and meet other agents. The environment invites an exploratory, browsing type of interaction. Karin is there to make ticket reservations and answer questions of visitors about performances and the theatre building. Karin has an artificial *talking face*. The advantage of embedding a dialogue agent in a virtual environment is that new interaction metaphors can be explored. Karin will make use of 'electronic leaflets', with photographs of the main actors and the theatre schedule in table format. The virtual theatre allows users to preview the stage from seats at different angles in the theatre hall. Future research is therefore concerned with the interaction of dialogue agents like Karin, and a virtual environment. In particular, the balance between presenting information by speech, or by other graphical means needs to be investigated.

REFERENCES

Andernach, T. (1996). A machine learning approach to the classification of dialogue utterances. In *Proceedings of NEMLAP-2*, Bilkent, Turkey.

Bouwman, G. and Hulstijn, J. (1998). Dialogue redesign with reliabiity measures. In *Proceedings of LREC*, Granada, Spain.

Van Deemter, K. et al. (1994). Generation of spoken monologues by means of templates. In *Proceedings of TWLT8*, Twente, The Netherlands

Dirksen, A. (1992). Accenting and deaccenting, a declarative approach. In *Proceedings of COLING'92*, Nantes.

Dirksen, A. and Menert, L. (1997). Fluent Dutch text-to-speech. Technical manual, Fluency Speech Technology/OTS Utrecht.

Dutoit, T. (1997). High-quality text-to-speech synthesis: An overview. *Electrical and electronics engineering*, 17(1):25–36.

Jönsson, A. (1993). *Dialogue Management for Natural Language Interfaces.* PhD thesis, Linköping University.

Van der Hoeven, G. F. et al. (1995). Schisma: a natural language accessible theatre information and booking system. In *Proceedings of NLDB*, Versailles, France.

Nijholt, A. et al. (1998). Speech and language interaction in a (virtual) theatre. In *Proceedings NLP+IA'98*, Moncton, Canada.

Rats, M. (1996). *Topic Management in Information Dialogues.* PhD thesis, Katholieke Universiteit Brabant, The Netherlands.

Rüber, B. (1997). Obtaining confidence measures from sentence probabilities. In *EUROSPEECH '97*, Rhodes.

Vallduví, E. (1990). *The Informational Component.* PhD thesis, University of Pennsylvania.