gefördert vom
Bundesministerium
für Forschung
und Technologie

Verbmobil
Verbundvorhaben

"Roger", "Sorry", "I'm still listening": Dialog Guiding Signals in Information Retrieval Dialogs

A. Kiessling, R. Kompe
H. Niemann, E. Nöth
A. Batliner

F.–A.–Universität Erlangen–Nürnberg
L.M.-Universität München

Vm

Report 31
Oktober 1994

Oktober 1994

A. Kiessling, R. Kompe
H. Niemann, E. Nöth
A. Batliner

Lehrstuhl für Mustererkennung (Inf. 5)
Friedrich–Alexander–Universität Erlangen–Nürnberg
Martensstr. 3
D–91058 Erlangen

Institut für Deutsche Philologie
Ludwig–Maximilian Universität München
Schellingstr. 3
D–80799 München

Tel.: (09131) 85 - 7799
e-mail: {kiessling}@informatik.uni-erlangen.de

Gehört zum Antragsabschnitt: 3.11, 3.12, 6.4

Das diesem Bericht zugrundeliegende Forschungsvorhaben wurde mit Mitteln des Bundesministers für Forschung und Technologie unter dem Förderkennzeichen 01 IV 102 H/0 und 01 IV 102 C 6 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei dem Autor.

From Proc. ESCA Workshop on prosody, Lund, 1993, pp. 140–143.

"Roger", "Sorry", "I'm still listening": Dialog Guiding Signals in Information Retrieval Dialogs

A. Kießling[1], R. Kompe[1], H. Niemann[1], E. Nöth[1], A. Batliner[2]

[1] Universität Erlangen-Nürnberg, Lehrstuhl für Mustererkennung (Informatik 5), Martensstr. 3, 91058 Erlangen, F.R. of Germany

[2] L.M.-Universität München, Institut für Deutsche Philologie, Schellingstr. 3, 80799 München, F.R. of Germany

ABSTRACT

During any kind of information retrieval dialog, the repetition of parts of information just given by the dialog partner can often be observed. As these repetitions are usually elliptic, the intonation is very important for determining the speakers intention. In this paper prototypically the times of day repeated by the customer in train table inquiry dialogs are investigated. A scheme is developed for the officers reactions depending on the intonation of these repetitions; it has been integrated into our speech understanding and dialog system EVAR (cf. [6]). Gaussian classifiers were trained for distinguishing the dialog guiding signals confirmation, question and feedback; recognition rates of up to 87.5% were obtained.

INTRODUCTION

Dialog systems for information retrieval are potential applications for human-machine communication. In human-human dialogs, very often (parts of) the information just given by the speaker is repeated by the partner. It can be observed e.g. in train table inquiries that the customer (henceforth C) repeats the times of arrival or departure just given by the officer (O). Very often only the intonation of this repetition of the time of day (RTD) shows the intention of C and thus governs the continuation of the dialog. In the scenario of our speech understanding and dialog system EVAR (an experimental automatic information system on train tables) the transmission of these times is a pivot point. Of course, a user-friendly system should be able to react adequately (cf. [7]). Let's e.g. consider the following dialog: O: *"... leaves Ulm at 17 23."* C: *"17 23./?"*. In the case of a rising intonation ('?') O – or the system, respectively – has to repeat the time of day, because C wants to have the time acknowledged. In the case of a falling intonation ('.') no specific reaction is necessary and the system can e.g. give the next part of information. This paper describes a corpus of *"real-life"* train table inquiry dialogs, the frequency of occurrences of RTDs, and their different intonational marking and functions. We will show how these functions can be determined automatically, and briefly how this prosodic information has been introduced into EVAR.

MATERIAL

Our investigations were based on a corpus of 107 "real-life" train table inquiry dialogs, recorded at different places, most of them conducted over the phone; for more detail cf. [1]. In most of the cases (88) the callers did not know that they were recorded. 92 dialogs concerned train schedules, the rest had other topics like fares. These 92 dialogs contained 215 utterances of C with in total 227 RTDs of arrival or departure, i.e. more than two repetitions per dialog on the average. In all but 3 cases the repetition concerned the time of day O just gave. There are two forms of time of day expressions possible in German: with or without the word *Uhr* (e.g. *"17 Uhr 23"* or *"17 23"*).

DIALOG GUIDING SIGNALS

By repeating the time of day, C has different aims, i.e. he wants to signal O different kinds of information. Depending on the specific kind of information, mostly expressed by the intonation, the reaction of O and thus the continuation of the dialog is governed. We observed three different functional roles of the RTD: *confirmation*, *question* and *feedback* (cf. figure 2).

- Using a **confirmation**, C wants to signal O, that he got the last information, e.g. the time of arrival. Functionally, this corresponds to the word *"Roger"* in the radio traffic. Usually, the intonation ($F0$-contour) at the end of such an utterance is falling (F, cf. figure 2a). A confirmation can be frequently observed after the end of a turn of O, just at the beginning of the turn-taking by C.

  Ex: O: *You'll arrive in Munich at 5 32.*
  C: **5 32.**

- The function of a **question** is *"Sorry, please repeat"*. C signals O that he didn't understand, i.e. that he didn't get the time of day completely or that he just wants to ask O to confirm the correctness (*"correct me if I'm wrong"*). The prototypical $F0$-contour is rising (R, cf. figure 2b). These questions often occur as short interruptions during the answer phase of O.

  Ex: O: *...you'll leave Hamburg at 10 15...* *...yes, 10 15, and you'll reach...*
  C: **10 15 ?**

- By using a **feedback**, C usually wants to signal O *"I'm still listening", "I got the information"* and sometimes *"slow down, please!"* or *"just let me take down the information"*. It is normally characterized by a constant or slightly rising $F0$-contour (continuation rise, CR, cf. figure 2c) and like the question it is usually found during the answer phase of O.

  Ex: O: *...the next train leaves at 6 35...* *...and arrives in Berlin at 8 15.*
  C: **6 35 –**

Note that one has to distinguish **function** (confirmation, question, and feedback) and intonational **form** (F, R, and CR) although in prototypical cases there is an unequivocal mapping of form onto function. The dialog guiding function of a confirmation is similar to a feedback, but their intonational form is different. Normally, questions can be distinguished easily from confirmations. Feedbacks, however, are sometimes likely to be confused with questions or even with confirmations. In our material, in 100 of the 227 repetitions of C the reaction of O (confirmation of the correctness, repetition, correction or completion of the time of day) was governed by nothing but the intonation of C. In the remaining cases, there were other indicators like interrogative particles. In 64 of the 100 cases, the time of day occurred isolated; thus, the only possibility for O to interpret the intention of C correctly, is by using the intonation of these elliptic RTDs as a cue, because other grammatical indicators like word order or *Wh*-words are missing.

THE REACTION SCHEME FOR THE DIALOG SYSTEM EVAR

From the corpus we developed a scheme (cf. figure 1) showing the reactions of O depending on the intonation of the RTD of C. In the scheme it was also taken into account if the customer repeated the time of day completely, incompletely or incorrectly. The scheme was integrated into our speech understanding and dialog system EVAR [6], which in our application plays the role of O. In figure 1, for an example of an information given by the system the possible ways of repeating the last given time of day are shown in the first two columns: C can repeat the time of day not at all or completely or incompletely and correctly or incorrectly, using different intonations. In the case of the system, the word recognition module has to provide the dialog module with the analyzed word chain that has to be compared with the previous time of day given by the system. Depending on the different possible combinations, the third column shows the next dialog step, the reaction of the system. For the integration into our system a prosody module has been added to the linguistic knowledge base, a semantic network. All necessary prosodic knowledge has been specified as a set of concepts and attributes within the formalism. Since the same network contains all linguistic knowledge sources of EVAR as well, appropriate links between the prosody module on one hand and the syntactic, semantic, pragmatic and dialog module on the other hand can be easily established. For more details, cf. [6].

System answer: *"... In München sind Sie dann um 17 Uhr 32."*
"... You'll arrive in Munich at 5 32 p.m."

RTD		prosody–module		system reaction	
no utterance		——		——	
incorrect		——		correction	*('Nein, um 17 Uhr 32.')*
complete & correct		R	*('17 Uhr 32?')*	confirmation	*('Ja, um 17 Uhr 32.')*
		CR	*('17 Uhr 32–')*	——	
		F	*('17 Uhr 32.')*		
correct & incomplete	only minutes	R	*('32?')*	confirmation	*('Ja, um 17 Uhr 32.')*
		CR	*('32–')*	——	
		F	*('32.')*		
	only hours	R	*('17 Uhr?')*	completion	*('17 Uhr 32.')*
		CR	*('17 Uhr–')*		
		F	*('17 Uhr.')*	——	

Figure 1: *The reaction scheme for* RTD*s within the dialog system EVAR*

THE PROSODY MODULE

The task of the prosody module is to determine automatically the intonation type, i.e. F, R, CR, that are mapped onto the functional roles of the RTD, i.e. confirmation, question, and feedback. From the automatically computed $F0$-contour [5] the following 4 features are extracted: the slope of the regression line of the whole (cf. the lines in figure 2a-c) and of the final part of the $F0$-contour, and the differences between the offset (the $F0$-value of the last voiced frame) and the values of the regression lines at this offset position (related work and comparable features are e.g. reported in [7] [3] [4]). Gaussian classifiers with full covariance matrix were trained to classify into the three classes F, R, and CR and thus – prototypically – into the functional roles confirmation, question, and feedback.

DATABASES FOR THE CLASSIFIER

Two databases were recorded and digitized with 16 kHz and 14 bit: In database A one female and three male speakers (not "naive", because they are working in prosody) read 90 complete time of day utterances each (all with the word *"Uhr"*; 30 questions, confirmations, and feedbacks each). As this database was used for training, misproductions (e.g. a question was intended, but a falling $F0$-contour was produced) and erroneous $F0$-contours were discarded. Thus a total of 322 utterances could be used for training. In database B two female and two male "naive" speakers read 50 time of day expressions each. Neither misproductions nor erroneous $F0$-contours were sorted out; this database gives therefore a good impression about how the system works in real life.

EXPERIMENTAL RESULTS AND DISCUSSION

Three experiments were performed. In the first experiment database A was used for testing in a leave-one-out mode (i.e. 3 speakers were used for training, the other for testing). In the second experiment the classifier trained on database A was tested on database B. Different feature combinations (e.g. computing the slope of the 2nd regression line over the last, the last two or the last three voiced regions) were tried. The results for the best feature combination where the 2nd regression line was computed over the last two voiced regions are shown in table 1 and 2. In the leave-one-out experiment (table 1) for all 3 cases (the rows marked by R, CR, and F; number of occurrences in parentheses) good recognition rates could be achieved (average recognition rate: 87.5%). For the speaker-independent test with

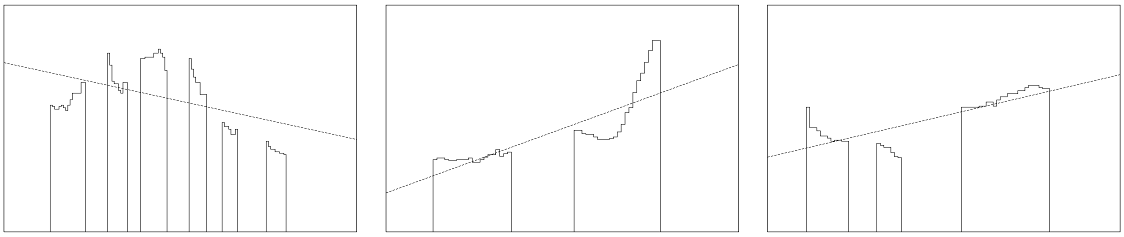
Table 1: *Results for leave-one-out*

	R	CR	F
R (97)	81.4	18.6	0.0
CR (107)	7.5	87.9	4.7
F (118)	1.7	5.1	93.2

Table 2: *Results for database* B

	R	CR	F
R (70)	87.1	7.1	5.7
CR (64)	21.9	37.5	40.6
F (66)	3.0	7.6	89.4

the naive speakers (table 2) we obtained an average recognition rate of 71.3%. The decrease in performance is due to the fact that no utterances were discarded and that the naive speakers obviously had enormous difficulties in the controlled production of a continuation rise: whereas questions and confirmations were recognized with approximately the same recognition rate (88%) as in the first experiment, it was much more difficult to classify the feedbacks correctly. As a final experiment the classifier trained on database A was tested on a subset of the above mentioned "real-life" material. Due to the sometimes very noisy telephone quality, only 32 isolated RTD's could be used for classification. Their reference type (R, F, CR) was determined by auditory tests. For classification, the same features as described above were extracted from the digitized signal. All the 10 confirmations, all the 5 questions and 7 of the 17 feedbacks were classified correctly.



2a: F, Confirmation **2b:** R, Question **2c:** CR, Feedback

Figure 2: *Prototyp. F0-contours (*F, R, CR*), their functional roles, and regression lines*

FINAL REMARKS

In [2] we show that the prosodic marking of sentence modality is more distinct in elliptic utterances than in non-elliptical utterances. Therefore we expect our modeling of question, confirmation, and feedback with R, F, and CR to work reasonably well not only with RTDs in train table dialogs, but also within other scenarios, where short elliptic utterances in clarification dialogs are used (like e.g. prizes in fare dialogs). However, our modeling is not exhaustive, because if e.g. in a confirmation a contrastive accent is positioned on the last syllable, or vice versa, in a question on the first syllable, our model will possibly not work adequately. Moreover, RTDs might not be purely isolated. They do often occur together with additional particles (like *"yes", "no")* or with repetitions of city names. In future, we plan to take into account the other possibilities of accentuation as well as non-isolated RTDs.

Acknowledgements: This work was supported by the German Ministry for Research and Technology (*BMFT*) in the joint research project ASL/VERBMOBIL. Only the authors are responsible for the contents of this paper.

References

[1] A. Batliner, A. Kießling, R. Kompe, E. Nöth, and B. Raithel. *Wann geht der Sonderzug nach Pankow? (Uhrzeitangaben und ihre prosodische Markierung in der Mensch-Mensch– und in der Mensch-Maschine-Kommunikation).* In *Proc. DAGA '92*, volume B, pages 541–544, Berlin, 1992.

[2] A. Batliner, C. Weiand, A. Kießling, and E. Nöth. *Why sentence modality in spontaneous speech is more difficult to classify and why this fact is not too bad for prosody.* In this volume.

[3] N. Daly and V. Zue. *Acoustic, Perceptual, and Linguistic Analyses of Intonation Contours in Human/Maschine Dialogues.* In *Int. Conf. on Spoken Language Processing*, pages 497–500, Kobe, 1990.

[4] N. Daly and V. Zue. *Statistical and Linguistic Analyses of F0 in Read and Spontaneous Speech.* In *Int. Conf. on Spoken Language Processing*, volume 1, pages 763–766, Banff, Canada, 1992.

[5] A. Kießling, R. Kompe, H. Niemann, E. Nöth, and A. Batliner. *DP-Based Determination of F0* contours from speech signals. In *Proc. ICASSP*, volume 2, pages II–17–II–20, San Francisco, 1992.

[6] R. Kompe, A. Kießling, T. Kuhn, M. Mast, H. Niemann, E. Nöth, K. Ott, and A. Batliner. *Prosody takes over: A prosodically guided dialog system.* To appear in: *Proc. Eurospeech93*, Berlin, Sept. 1993.

[7] A. Waibel. *Prosody and Speech Recognition.* Morgan Kaufmann Publishers Inc., San Mateo, California, 1988.