Dialogue system development for an emergency scenario

Jolanta Bachan, *Adam Mickiewicz University*
(**01.01.2007-21.06.2011**, prof. Grażyna Demenko, *Adam Mickiewicz University*)

Abstract

Dialogue systems are commonly used in call-centres and support the human telephone operators at their work. For the present work, a corpus linguistic study was performed, with the aim of finding patterns in dialogue which can be used to model typical dialogues, and of identifying stimuli which trigger alignment behaviours of particular types in humans. The main focus was set to alignment on the semantic level. A model of the human-computer dialogue is developed and implemented in a prototype dialogue system for an emergency scenario. The dialogue between the human and the computer is handled by two linked finite state automata (FSA): one for the dialogue manager and one for the map traversal.

1. Introduction

The main thesis of the Ph.D. thesis was that alignment of semantic representations is essential for successful communication in a dialogue [1]. It was stated that general function of alignment is coordination between interlocutors in order to achieve successful outcome of communication. In the present paper the steps of validation of the thesis are demonstrated. An emergency scenario and a map-task dialogue were chosen as an example and alignment of semantic representations of the map was claimed to be essential for successful communication.

The general working hypothesis was that it is possible to replace traditional emotion label sets with a generic model of the following type (which would also apply to 'emotion' in addition to 'style' if required):

TRIGGER_SITUATION → STYLE → STYLE_MANIFESTATION

The trigger situation is the particular public stress scenario which requires a certain formal or informal communication style. The style manifestation is the set of syntactic, lexical and phonological conventions which are associated with the chosen style. The specific hypothesis is that it is possible to design and implement a speech style selection module based on this model to drive synthesiser-interlocutor alignment, and to implement it in speech synthesis software. Such a module should improve the naturalness and efficiency of human-computer communication.

In order to test the thesis, a dialogue system was developed which would conform with modern dialogue theory on alignment between interlocutors. The operational goal was to develop a proof-of-concept dialogue system and the methodology in a simulated stressful emergency scenario. In the present paper, the theoretical issues are outlined and the main dialogue system development stages are presented.

2. Theory of alignment in dialogue

According to the alignment theories, alignment in dialogue takes place on semantic, syntactic and pragmatic levels. In recent years new aspects of communication have been investigated which are relevant for developing natural human-computer dialogue interaction. These include alignment of communication form and content between the interlocutors [2] and accommodation of interlocutors to each other [3]. It has been noticed that while communicating, interlocutors tend to adopt each others' behaviour such as style of speaking, vocabulary, gestures.

In the present context, alignment is meant as adaptation on the syntactic, semantic and pragmatic levels of communication between the two interlocutors, including the choice of similar lexical items and speaking style. However, it needs to be emphasised that the form, content and degree of alignment depends on the communication situation and status relations between the interlocutors. The main distinction for emergency scenarios to be made is between alignment in public and private situations. In public situations in which interlocutors do not know each other the degree of alignment of their behaviours has been found to be smaller than in face-to-face conversations between two close friends [4]. In fact, there may be deliberate non-alignment between a call-centre operator and an emotional caller, in order to calm the caller.

The dialogue system discussed in this study would have to predict the caller's knowledge and determine the common ground between the system and the caller. To do that a set of TRIGGER_SITUATIONs needs to be built containing possible situations about which the user may be talking, as well as an extensive dictionary of domain specific vocabulary containing *standard* form of relevant *colloquial* expressions.

Therefore an essential part of dialogue system design is to make it build up the extensive implicit common ground with its interlocutor. This means that the more information the system and the user share in their dialogue, the more effective conversation is and the more their situation models are aligned. From a semantic point of view, the common ground can be created by using a map, whose semantic relevance has to be negotiated in relation to reality. From a pragmatic point of view, the theory about common ground and implicit common ground assumes that there are interactive repair mechanisms using implicit and full common ground when the interlocutors' representations are not properly aligned.

3. Corpus linguistic study

For the present research, two types of dialogues were recorded in laboratory conditions: a map task and a diapix task (picture description dialogue) [5]. The map and diapix tasks are source of semi-spontaenous speech. Both dialogues were directed at crisis situations and communication in public setting, especially between people who do not know each other. As control material, neutral map task, diapix and readings were recorded. Because the diapix dialogues were not used directly for the dialogue system development, below only the map task dialogue scenario is presented.

3.1 Map task emergency scenario

Each of the subjects gets a street map. (The map is presented on Fig. 1 (A) – the black circles were not marked). In this task one person has to lead the other person to get to a place in which a man with a heart attack is waiting for help. The person who is chosen to lead the other person gets a map with a marked route on it which leads among different landmarks. The other person gets a map with the landmarks only. The interlocutors cannot see each other. The task is to describe the route of how an ambulance should get to the emergency location. At the beginning one of the subjects gets a description of the situation underlining the tragic situation to invoke emotions such as fear or sadness. Additionally, the two maps the subjects get slightly differ in respect of the landmarks to create trouble in communication: on the way of the ambulance there are such obstacles as an accident, traffic jam, roadworks and a school race – these are not seen by the other person. To boost stress degree, the leading person gets 5-min time limit to perform the task.

3.2 General corpus information

Subjects: 15 males and 15 females were chosen and recorded in pairs: m:m, m:f, f:f. People who did not know each other were selected in order to assure that the dialogue would have the public character [4].

Recordings: The subjects sat alone in two quiet rooms and communicated via Skype. The recordings were performed by the MX Skype Recorder software [6] which records unlimited time audio Skype calls on two separate channels for two speakers in the stereo WAV format. The corpus contained *4h 12min* of speech, out of which *38min 45sec* were map-task emergency dialogues.

Annotation: One typical emergency map task dialogue was annotated on six tiers: (1) phones – extended SAMPA phoneme set [7], (2) syllables – (3) speech – orthographic transcription, (4) English translation of the Polish speech tier, (5) Bunt's dialogue acts main categories [8], (6) special – speech events such as filled-pauses, confirmations or hesitations.

3.3 Analysis of the selected dialogue

General measurements on one selected emergency dialogue annotated on 6 tiers were carried out (see Tab. 1). On the speech tier, 139 intervals for Speaker A were annotated and 85 intervals for Speaker B. These figures include also silences. In Speech duration column, a huge difference is to be found between the inputs given from both of the speakers. A bit more about the dialogue flow can be said when looking at the Longest silence figures. 28.87sec of Speaker's B complete silence says that the speaker let his interlocutor talk without any interruption for a long period of time. More than one dialogue act could be assigned to one stretch of speech (speech interval),

Tab. 1:
Dialogue statistics - total dialogue duration 156.49sec

Speaker	Speech tierIntervals on	Speech duration	Longest silence	Speech intervals	Dialogue acts	Syllables	Phones	Special
A	139	106.71s	4.74s	79	113	506	1231	94
B	85	42.02s	28.87s	52	70	241	579	36

therefore there is a bigger number of dialogue acts than there is the speech intervals annotated in the dialogue.

The analysis of the dialogues revealed ways of alignment of interlocutors in stress situations, ways of recovery of misalignment and general structure of a map-task dialogue. These findings were used to develop a dialogue manager FSA implemented in the dialogue system.

4. Finite-State Transducer of the map

The emergency map can be represented as a finite-state transducer (FST) where each junction corresponds to the transition node. However, not all the streets are open. Some junctions cannot be reached, because, for example, there is a traffic jam on the way or roadworks. Such junctions are not taken into account when designing the FST. Traffic on all the other streets is two-way, so turnings back are not hampered. While moving along the map, some route is followed. At a normal map, the route can be tracked thanks to the street names or the landmarks being passed on the way. In the FST, the street names and the landmarks can be replaced by Latin letters for simplification. Such an analysis of the map resulted in creation of an FST modelling the movements along the map in order to reach the goal. The design process and the FST are presented on Fig. 1. Fig. 1 (A) presents the emergency map with the marked junctions for selections. All the junctions which cannot be reached because of the obstacles on the way were not selected for the nodes of the FST. On Fig. 1 (B) there is the FST, with the enumerated transition nodes and the transitions producing Latin letters. q0 is the start node and q13 is the end node.

5. Dialogue system design

The prototype dialogue system is based on two FSA. The emergency map created and used for the map task dialogue in the dialogue corpus recording has been modelled as a transition diagram of an FSA for the map traversal. The analysis of the dialogue corpus served to create an FSA for the dialogue manager. The instruction to the human caller is to direct an ambulance from the hospital to the person with a heart attack along the streets. The human user inputs 'chat' text into the system in writing. The system communicates with the caller via audio output producing synthetic speech. In Fig. 2 the architecture of the system is presented.

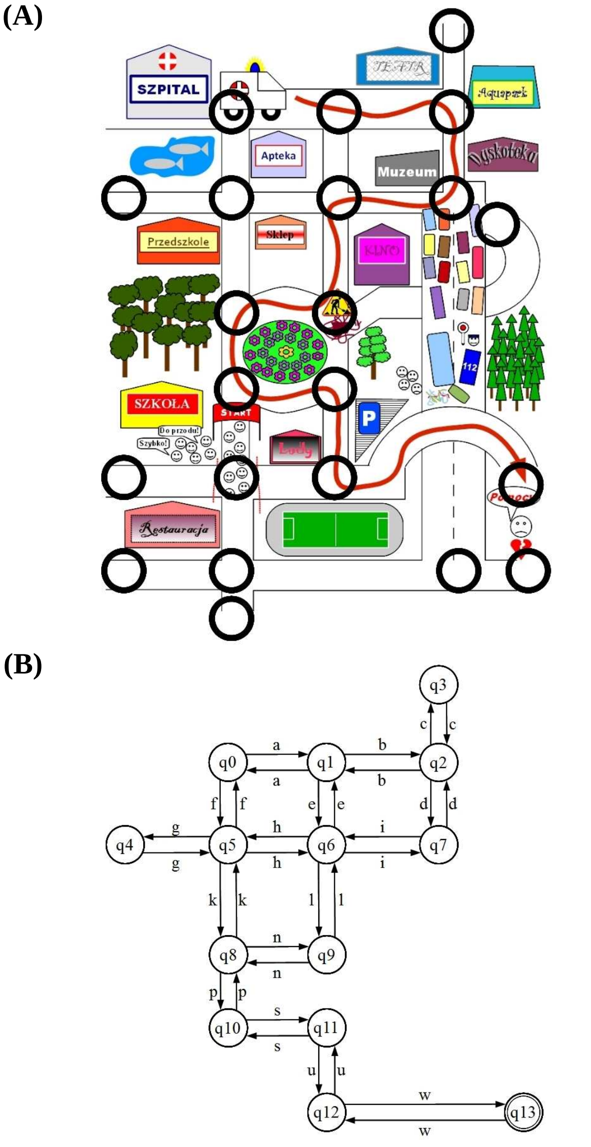

Fig. 1: (A) Emergency map with marked junctions; (B) Emergency dialogue automaton with the nodes representing the selected reachable junctions.

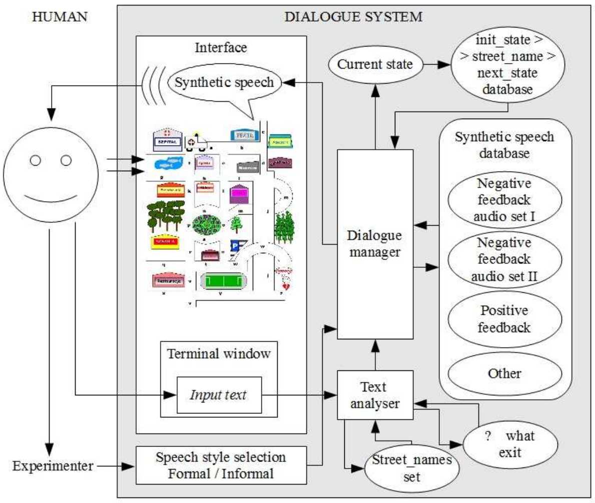

Fig. 2: Dialogue system architecture.

The interface in human-computer communication is based on text input from the caller into the terminal window. As already mentioned, the computer communicates with the user via synthetic speech played via loudspeakers. Additionally, the caller is provided with a street map on a computer screen. Based on personal characteristics of the user or randomly either formal or informal speech style is selected by an experimenter for the dialogue.

6. Dialogue system implementation

The prototype dialogue system was implemented as a command line application. The written input from the user is entered on the command line and the synthetic speech output from the system comes from the loudspeakers.

The user needs to insert street names which do not violate the street arrangement of the map to make the system move from one state to the next state. Additionally, the user has the following function words to enter:

1. "`what`" – repeats the last audio output;
2. "`?`"– the current state at which the ambulance is is printed as text (e.g. "`The ambulance is at state: q4!`") and a special map with transition states is displayed;
3. "`exit`" – exits the dialogue loop and moves to the farewell section of the program.

The dialogue manager (DM) finite-state automaton with example utterances is presented in Fig. 3. The figure presents a semi-coupled DM automaton of exemplar dialogue acts, in which the automaton on the top is the dialogue system (DS) automaton and the automaton at the bottom is the expected human user (HU) input automaton. The transitions between the interlocutors, i.e. the turn change, is marked with dashed arrows. The red dotted-dashed arrows (also local loops) show the dialogue flow where the caller input does not match the expected input. For example, the dialogue system will not move to the requested dialogue act until it does not get the positive feedback from the caller that he knows where the hospital is. In general, there is no turning back from one state to the other on the individual automata, i.e. caller and dialogue sequence automata. The local loops are implemented, but no backward arrows are designed. However, when working together, the turnings back are expected. These backward transitions are from DS:q6 to HU:q3 and HU:q4 to DS:q5. At this part of the dialogue, the DM automaton is connected with the map traversal automaton presented in Fig. 1 (B). As long as the caller moves through the dialogue, he also moves through the map.

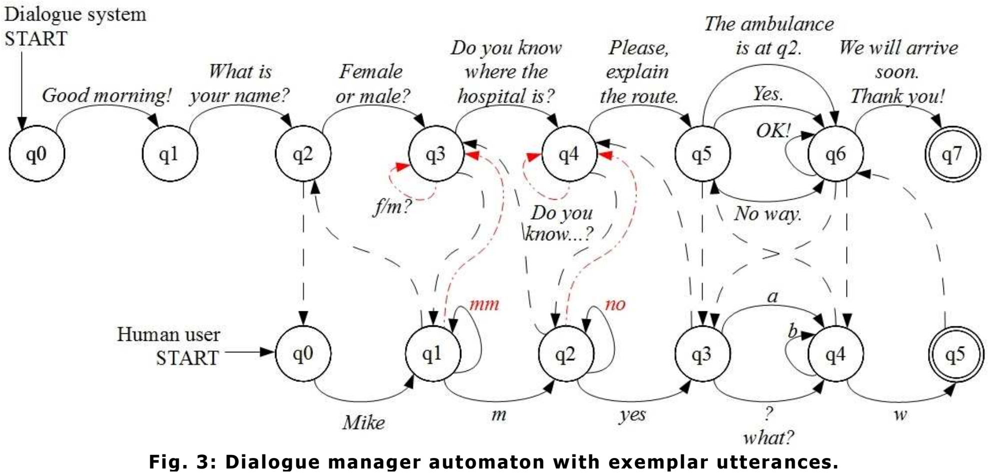

Fig. 3: Dialogue manager automaton with exemplar utterances.

The dialogue is divided into three sections: opening, direction description in the while-loop and the farewell.

In the opening section, the system greets the caller, asks for the name and sex and whether it is clear where the hospital is. If the answer is "yes", it asks to explain the route.

The negotiation of the route is processed by one while-loop. The caller inputs the street names or any function word in order to carry out the task. If the suggested street name was correct and the system made a move on its map, positive feedback is produced. If the suggested street is impassable or it violates the street arrangement on the map, negative feedback from audio set II is produced. When the text input is neither the street name nor the function word, it is recognised as a non-existent street and negative feedback from audio set I is produced. At any time, the user can insert one of the tree function words.

After the system gets to the state q13, the farewell is produced. The system promises that the emergency service will do everything they can to arrive quickly and informs that if they have any trouble to get to the place, they will call again. At the end section, the inputs provided by the caller are printed to the TXT log file.

6.1 Implemented utterances

The observations made on the alignment phenomenon between the speakers in the dialogue corpus, especially results of the communication in the emergency scenario, made it possible to select utterances which will match best the requirements of the aligned or cooperatively non-aligned dialogue system. Those requirements (cf. [9]) are: informativeness, relevance, briefness, politeness. The first three requirements correspond to Gricean Maxims of Cooperation. The fourth requirement results from the observations of how the dialogues in the dialogue corpus were carried out.

The original utterances from the exemplar dialogue were annotated on the phone level and then synthesised using the automatic close copy speech (ACCS) synthesis [10] with the PL2 MBROLA male voice created for this prototype dialogue system. The selected utterances were then left untouched or were slightly modified in order to meet formal and informal criteria of the speaking style. Some utterances were invented to match the dialogue scenario. All the utterances were synthesised using MBROLA [11] and exported to the WAV audio format and integrated with the dialogue system. All the feedback utterances are chosen randomly from the audio sets at a runtime of the dialogue system. The other utterances at the dialogue opening and closure are generated according to the scenario.

6.2 Evaluation

The dialogue system underwent thorough evaluation according to the EAGLES standards [12]. First, diagnostic evaluation was performed to check whether the system runs without failures.

Having underwent successfully the diagnostic evaluation, the dialogue system faced functional testing and judgement testing with the human users in laboratory setting. The setting of the evaluation is presented on Fig. 4. 52 people took part in the evaluation. In the evaluation, mainly young students took part: 19-23 years old. However, there were also a few older people in their late 20's and one 52-year-old man whose voice was used for creating the PL2 MBROLA voice.

The functional testing was based on evaluation of the dialogue between the human and the computer. The dialogue model was tested, the scenario and the successfulness of communication by the means of the actual conversation between the test participant and the computer. After the dialogue was finished, the test participant was asked to assess different domains of the system on the 5-point rating scale, where 1 was the lowest grade and 5 was the highest grade. The system's 4 domains were evaluated: speech style selection module, speech synthesis, dialogue manager and system design. For simplification, the test participants were asked to evaluate 7 categories called: friendliness, speech quality, speech intelligibility, dialogue, dialogue naturalness, system attractiveness and ease of usage.

Last but not least, the system was tested in field conditions at the Researchers' Night 2011 in Poznań. Around 80 people conversed with the system in order to direct the ambulance from the hospital to the person with the heart attack. They were informed that the time of their dialogue was being measured and if they carried out the task in a very short time, they would get a little prize.

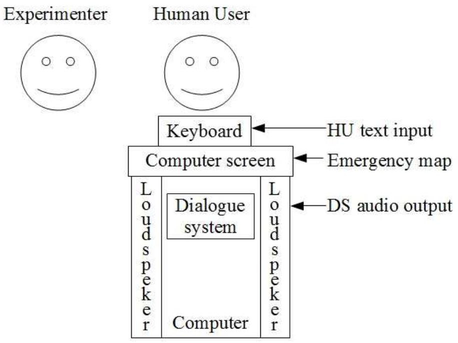

Fig. 4: Dialogue system evaluation setting.

6.3 Results

In the laboratory test 14 females and 12 males took part to evaluate each of the two scenarios: formal and informal. Altogether 52 people took part in the laboratory evaluation. The duration time of all the dialogues in formal and informal scenarios lasted about 75min 34sec and 76min 48sec respectively. The number of inputs inserted during one dialogue is almost the same and equals 20.54 inputs for the formal and 20.26 inputs for informal scenarios. The same equality applies to the average length of the path and the number of inputs of the "`what`" function word. The "?" word appeared much more frequent in dialogues with the informal speech styles and suggests that the informal speech was less intelligible.

The basic statistics of the functional testing show that both, the formal and informal speech styles provided similar conversational circumstan-ces to the dialogue participants, with the tendency of the informal speech to be less intelligible.

All the subjects accomplished the communication with the computer successfully, which means that they were aligned at least on the essential semantic level. According to the semantic level interpretation adopted for the map-task scenario, the human user as well as the computer, i.e. the dialogue system, must have had the same or at least a compatible semantic representation of the map to move along the map and assure success of communication by getting to the end point. Misalignments happened, but the dialogue system effectively recovered from those misalignments.

The results of the judgement testing of the system were very high: 4.11 for the formal dialogue scenario and 4.30 for the informal scenario, where 5 was the highest grade.

When it comes to the field testing at the Researchers' Night, all the people carried out the task successfully, regardless the crowd and noise around and even despite the fact that some little children had problems finding letters on the keyboard. The age of the youngest child was 5 years old. The best time was 55sec and it was reached by a boy who tried to beat the record a few times. This shows that familiarisation with the system makes the usage easier and more effective. The second time was 1min 4sec by a girl who watched her friend doing the task before her. Then there were a few good times around 1min 30sec and the other times varied a lot, up to 4min 30sec.

7. Conclusions

In the present paper, a prototype dialogue system creation for use in the emergency scenario was presented. Analysis of dialogue corpora of a map task emergency scenario was used to create a finite-state automaton modelling the dialogue in a prototype dialogue system. Finally, a prototype dialogue system was developed and evaluated with human users in laboratory and field tests. The prototype dialogue system combined text input with speech output and its core was based on two linked finite state automata: one for the dialogue manager and one for map traversal. The laboratory and field settings of the evaluation task demonstrated alignment of the semantic representation of the map, as all the human users finished the task successfully.

Additionally, the alignment of the dialogue system was based on speech style selections: formal and informal. The speech style selection demonstrated the claim that the traditional emotion label sets used in general speech synthesis may be replaced by the speech *style* in the dialogue systems. Such a system not only takes into account the communication in public situations, but also aligns with the user on the same levels as the human users would do, so not emotional, but semantic, syntactic and pragmatic.

Bibliography

[1] Bachan, J. 2011. Communicative Alignment of Synthetic Speech. Ph.D. Thesis. Institute of Linguistics. Adam Mickiewicz University in Poznań, Poland.

[2] Pickering, M.J. & Garrod, S. 2004. Toward a mechanistic psychology of dialogue. In: *Behavioral and Brain Sciences 27,* pp. 169-225

[3] Giles, H., Coupland, N., & Coupland, J. 1992. Accomodation theory: Communication, context and consequences. In H. Giles, J. Coupland, & N. Coupland (Eds.), *Contexts of accommodation* (pp. 1-68). Cambridge: Cambridge University Press

[4] Batliner, A., Steidl, S., Hacker, Ch., Nöth, E. 2008. Private emotions versus social interaction: a data-driven approach towards analysing emotion in speech. In:*User Modelling and User-Adapted Interaction - The Journal of Personalization Research 18,* pp. 175-206

[5] Bradlow, A. R., Baker, R. E., Choi, A., Kim, M. and van Engen, K. J. 2007. The Wildcat Corpus of Native- and Foreign-Accented English. In: *Journal of the Acoustical Society of America*, 121(5), Pt.2, p. 3072

[6] MX Skype Recorder v4.3.0. Copyright © 2006-2010

[7] Demenko, G., Wypych, M. & Baranowska, E. 2003. Implementation of Grapheme-to-Phoneme Rules and Extended SAMPA Alphabet in Polish Text-to-Speech Synthesis. In: G. Demenko & M. Karpiński (Eds.) *SLT, Vol. 7.* Poznań: PTFon, pp. 79-95

[8] Bunt, H. 2000. Dialogue pragmatics and context specification. In: H. Bunt & W. Black, (Eds.) *Abduction, Belief and Context in Dialogue. Studies in Computational Prag-matics.* Amsterdam: J. Benjamins, pp. 81–150

[9] Grice, H.P. 1975. Logic and conversation. In: Cole, P. & Morgan, J. (Eds.) *Syntax and Semantics, Vol. 3.* New York: Academic Press. pp. 41-58

[10] Bachan, J. 2007. Automatic Close Copy Speech Synthesis. In: *Speech and Language Technology. Vol. 9/10.* Ed. G. Demenko. Poznań: PTFon, pp. 107-121

[11] Dutoit, T., Pagel, V., Pierret, N., Bataille, F. & van der Vrecken O. 1996. The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes. In: *Proceedings of ICSLP 96, Vol. 3.* Philadelphia, pp.1393-1396

[12] Gibbon, D., Mertins, I. & Moore, R. 2000. *Handbook of Multimodal and Spoken Dialogue Systems: Terminology, Resources and Product Evaluation.* New York: Kluwer Academic Publishers

This work was partly funded by the research grant No. N N104 119838 titled "Communicative alignment of synthetic speech"

Author:

Jolanta Bachan, Ph.D.
Institute of Linguistics
Adam Mickiewicz University
Al. Niepodległości 4
61-874 Poznań, Poland
tel. +48 502 544 279
fax +48 61-829-36-62
email: jolabachan@gmail.com