DOMAIN, DIALOGUE AND SEMANTIC ANALYSIS WITHIN CITY DIALOG SYSTEM FOR CITY INFORMATION CENTRE

Roman Mouček, Kamil Ekštein

Department of Computer Science and Engineering,
University of West Bohemia,
Univerzitní 22, 306 14 Plzeň,
Czech Republic
e-mail: {moucek, kekstein}@kiv.zcu.cz

Abstract: The paper deals with a new computerized dialogue system for city information centres. The overview of dialogue system architecture is presented. The corpus of recorded sentences is shortly introduced. The basic principles of generation of sentence templates and generation of training sentences are described. The method of dialogue simulation is introduced. The main attention is paid to domain and semantic analysis. The SIL language hierarchy for semantic and knowledge representation is also mentioned.

Keywords: Dialogue system, dialogue analysis, domain analysis, semantic analysis, SIL formalism, sentence templates, semantic representation

1. INTRODUCTION

This paper deals with the domain of computerized dialogue systems as a very important and up-to-date field of artificial intelligence. The main goal of the paper is to introduce a special approach in the area of dialogue, domain and semantic analysis and semantic and knowledge representation within a proposal and development of computerized dialogue system for city information centre. This kind of system is developed at The University of West Bohemia, Czech Republic, in co-operation with Technical University in Dresden, Germany, and Pilsen City Council.

2. BASIC IDEAS

City information centres with their services are very suitable areas where the dialogue information system accessed by phone can help clients to get basic information about city, sights, public transport, local authorities, accommodation, etc. The dialogue system can serve not as the only accessible information system during then time the information centre is closed but it can also help the human operators to answer tiring and common questions. The city information system described in this paper is developed especially for Pilsen City but the basic ideas and approaches can be used during development of the similar systems.

3. BACKGROUND AND SYSTEM ARCHITECTURE

A need to build such a system for Pilsen Information Centre let to the first dialogue system architecture proposal, see Fig. 1. The architecture of the system corresponds to the architecture introduced within the Sundial project (Eckert, 1995; Ocelíková, 1998). The module of linguistic analysis is discussed in next parts.

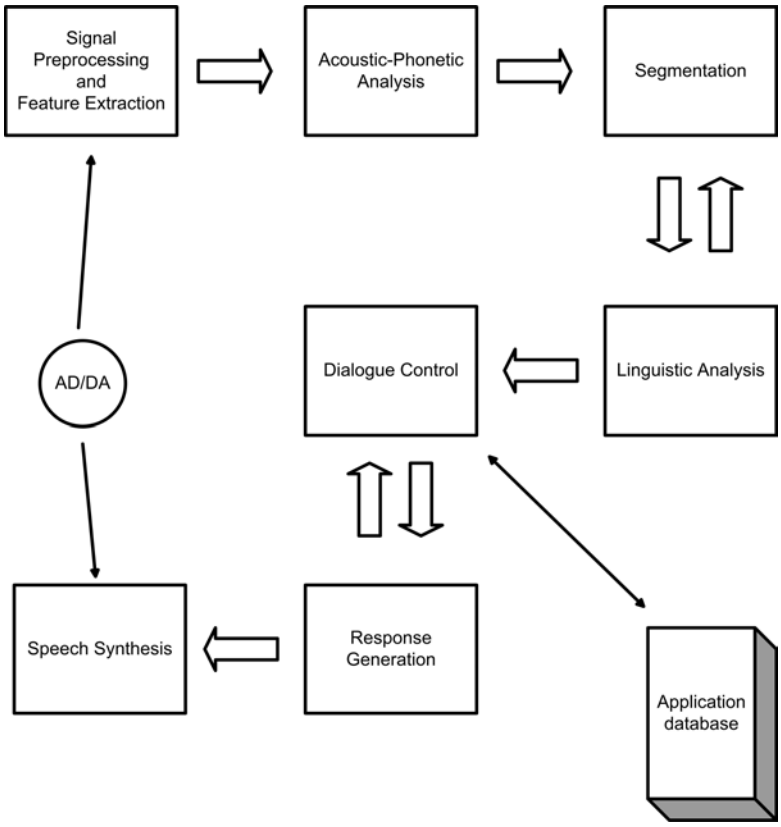

Fig. 1. System Architecture

4. BASIC RESEARCH

The architecture proposal had been preceded by a detailed analysis of the problem area (carried out by Jana Schwarz, Technical University of Dresden, Germany). There were recorded about 500 dialogues in 12 information centres in the different places of the Czech Republic. These dialogues were transcribed and analyzed to determine the course of dialogues and to asses which dialogues and turns are typical (Schwarz and Matoušek, 2001).

5. DIALOGUE ANALYSIS AND CORPUS CONSTRUCTION

5.1 Corpus of Recorded Sentences

The recorded and transcribed dialogues serve as the basis for the whole corpus of utterances possibly spoken in information centres. We eliminated all the sentences connected with the purchase and all the sentences that can be spoken only in human-human interaction.

5.2 Corpus of Generated Sentences

The transcribed dialogues were analyzed to automatically create the sentence templates. The special program package used as analyser working in two modes (generation of templates and generation of training sentences) implemented especially developed method of a quantitative linguistic analysis (Schwarz and Matoušek, 2001). However, the whole corpus of generated sentences covers too wide area to be elaborated in the module of linguistic analysis. Thus the corpus has been divided by another program tool into several domains. The basic ideas of domain analysis are described in Section 6, the other information can be found in (Mouček and Taušer, 2002). The corpus of generated sentences was completed by sentences obtained by simulation of dialogue between user and computer in two selected domains, see Section 5.3.

5.3 Dialogue Simulation and Corpus of Simulated Sentences

The corpus of real and generated sentences has been completed by sentences obtained by simulation of dialogue between user and computer. Nearly all users of computerized dialogue system change their dialogue strategy if they speak to computer instead of human being. Many people also reject to speak to computer and require switching to the human operator immediately. Account on that we decided to simulate this kind of dialogue and to complete the whole corpus. The process of corpus construction is shown in Fig. 2.

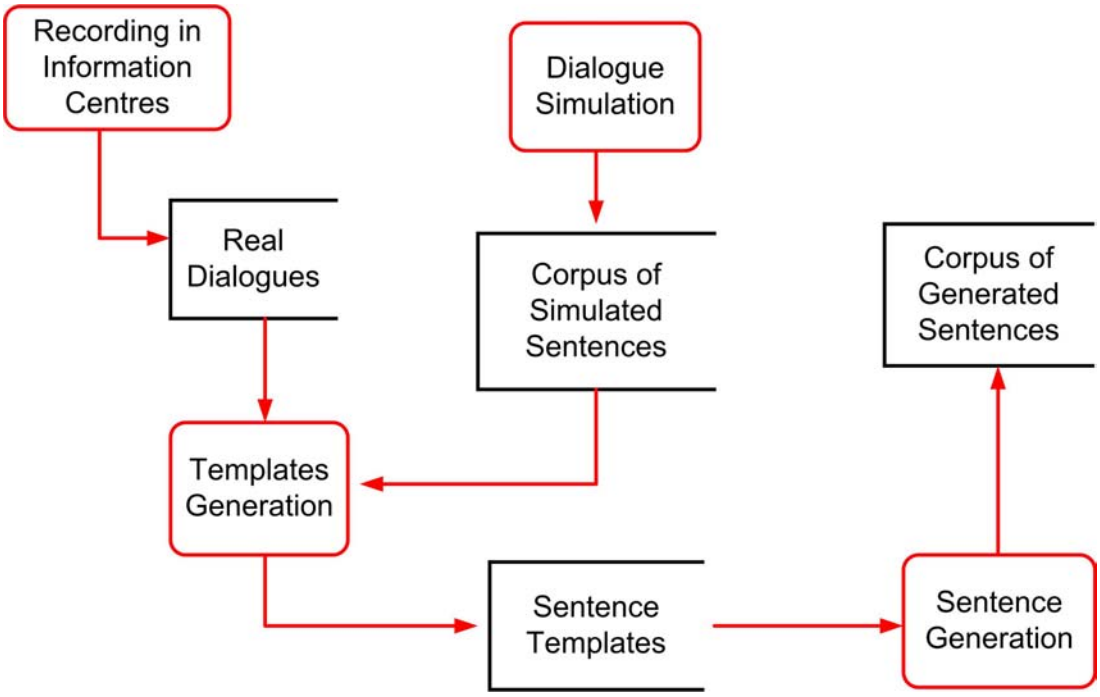

Fig. 2. Corpus Construction

5.4 Usage of Corpus

The whole corpus is used for training of the dialogue system modules:

- a speech recognition module, see (Ekštein, 2001),
- a module of linguistic analysis.

6. DOMAIN ANALYSIS

The whole corpus of sentences has been further analyzed. The sentences were divided into domains by another program tool working on the principle of word meaning evaluation.

The keyword file the program works with consists of:

- number and description of domain,
- list of keywords with their weights,
- list of keyword collocations with their weights.

The domain rules and restraints were established. As a result we obtained a set of domains and the rate of sentences (corpus of sentences generated from sentence templates and obtained by dialogue simulation) corresponding to each domain. The domain overview is presented in Table 1. A total number of 4.236 sentences were analyzed, 21.3% sentences cannot be classed with a domain.

Table 1. Domain Overview

Domain	**Sentence rate [%]**
Transport	8.6
Accommodation	13.2
Food Services	2.4
Institutions	12.2
Culture	11.3
Companies	9.6
Sport Facilities	7.5
Tourism	14.8

6.1 Domain Selection

Two domains presented in Table 1 were selected to be elaborated in a new dialogue system - transport and institution domain. The transport domain has been selected because of the similarity to the existing kind of problems solved during the development of the dialogue system within the SUNDIAL project. The institution domain is the domain quite different from the transport domain.

Within these two domain possible user questions and domain range have been widely analyzed.

6.2 Transport Domain

Within domain of transport the dialogue system should be able to:

- find a connection between two stops of city transport,
- find a connection between two suburbs of the city,
- find a next connection from the stop determined by a user,
- provide a basic information about bus and railway stations,
- provide a basic information about bus and train connections, etc.

7. SEMANTIC ANALYSIS

The corpus of sentences is used in the process of semantic representation and interpretation. The SIL formalism (McGlashan, 1991; McGlashan, 1994) developed within the SUNDIAL project is used as a language suitable for knowledge representation.

7.1 SIL Concepts

SIL consists of basic formalism CoreSIL, representations of partial descriptions of utterance called UFOs and methods of hierarchical inference. It provides a simple semantic representation of utterances at two different levels of detail:

- a linguistically oriented level,
- a task oriented level.

The utterance is described in terms of concepts or type-feature structures at both levels. At the linguistically oriented level, the final structure corresponds to a compositional semantic representation. This level of the SIL language can theoretically represent all structures on the basis of linguistic theories. At the task-oriented level, the final structure corresponds to a back-end application.

7.2 SIL Extension

We have made an extension of the CoreSIL definitions in the case of transport domain. We have also tried to extend CoreSIL hierarchy in the case of institution domain. It seems now that it will be not necessary to define a completely new SIL hierarchy for this domain. During completing of a CoreSIL hierarchy we try to find another domain-independent hierarchical part, which can be reused for further domain extension.

7.3 Main Sentence Concept and Local Semantic Concepts

The corpus of sentences for transport domain is divided into several groups and all the sentences in one group are described by specific semantic concept. This main sentence concept is determined by the existence of key structures in the sentence or it can be derived from the local semantic concepts found in the sentence. The main sentence concept can be also introduced as a composition of local semantic concepts keeping in all semantic information, which is passed to the module of semantic interpretation. The type of main sentence concept also determines the possible system answers and final task concept used as the interface to the application database.

The local semantic concepts describe the certain parts of the sentence and compose the main sentence concept. The length of the sentence part described by local semantic concept is various and it follows several rules:

- the meaning of the sentence part is unique within the domain area,
- the composition of local semantic concepts in the interpretation phase is unique in the domain area,
- the determination of local semantic concepts covering the sentence is as universal as possible.

We can see that these rules can be contradictory in many cases. We need to use the universal local semantic concepts to reuse them in the other main sentence concepts and in the other domains. On the other side, the phase of semantic interpretation fails in the case of very small and universal local semantic concepts.

8. CONCLUSION AND FUTURE WORK

This paper presents the basic proposal of city information centre dialogue system and especially the part of dialogue, domain and semantic analysis. The phase of dialogue and domain analysis are now nearly completed. The work will be finished by solving the problems of domain overlapping. In the future especially the module of linguistic analysis and dialogue manager module will be worked on. We consider to propose and to implement a module extracting semantic information from the user utterance for two domains, to extend an existing SIL hierarchy and to build eventually a new SIL hierarchy, to propose and to implement a part of dialogue manager (belief module, dialogue module) and to determine and to implement a domain-dependent and domain-independent part of a parser and SIL hierarchy influenced by processing of two different domains.

ACKNOWLEDGEMENT

This research is supported by the Grant of Czech Republic Ministry of Education, MSM 235200005 Information Systems and Technologies.

REFERENCES

Eckert W. (1995). Gesprochener Mensch-Maschine-Dialogue, *PhD Thesis*, Universität Erlangen-Nürnberg, Germany.

Ekštein K. (2001). SpART I – An Experimental Automatic Speech Recognizer, In *Proceedings of SPECOM 2001*, Moscow, Russia.

McGlashan S. (1991). A Proposal for SIL, Sundial WP6 (unpublished).

McGlashan S. (1994). The Role of Semantics in Spoken Dialogue Translation Systems, In *Proceedings of the Second International Conference on Maschine Translation*, Cranfield, England.

Mouček R. and K. Taušer (2001). Dialogue System for City Information Center, In *Proceedings of the 6th World MultiConference on Systemics, Cybernetics and Informatics SCI 2002*, Orlando, USA, 2002, volume XII, pp. 563 - 567.

Ocelíková J. (1998). Zpracování významu spontánních promluv při dialogu člověka s počítačem (Semantic Representation and Interpretation of Spontaneous Utterances in Human Computer Interaction), *PhD Thesis*, Plzeň, Czech Republic.

Schwarz J. and V. Matoušek (2001). Creation of a Corpus of Training Sentences Based on Automated Dialogue Analysis, In *Proceedings of the 4th International Conference TSD 2001*, Železná Ruda, Czech Republic.