Automatic Analysis of Real Dialogues and Generation of Training Corpora

Jana Schwarz1, Václav Matoušek2

1 Institute of Slavistics, Technical University of Dresden, Germany
Jana.Schwarz2@mailbox.tu-dresden.de
2 Department of Computer Science, University of West Bohemia in Plzeň, Czech Republic
matousek@kiv.zcu.cz

Abstract
The development of computerized information retrieval dialogue systems communicating with the user in natural language requires the implementation of an effective training procedure with the aid of which the main modules of the dialogue system can be partly automatically developed. The presented paper describes an attempt to create the generating sentence templates automatically, using a special program package implementing an especially developed method of a quantitative linguistic analysis of transcribed real dialogues. Firstly, the program package generates a set of templates, i.e. formulas consisting of elements (terminal and non-terminal symbols) of a special grammar and describing the syntactic structure of a required subset of Czech sentences. Secondly, it generates a large corpus of unique training sentences using the sentence templates and a stochastic context-free grammar. The experimentally created corpus was used for the training of modules of a city information dialogue system.

1. Introduction
Modern computerized information retrieval dialogue systems communicating with the user in some form of natural language seem to be an effective and comfortable tool to obtain exact information on bus, train, or plane departures/arrivals, on the products on offer in hypermarkets and department stores, theater or cinema performances, interesting landmarks and many others. Such dialogue systems are at present usually developed by training their components (recognizer, linguistic analyzer, dialogue manager) using several training corpora that have to be created manually or with the aid of some artificially designed supporting means. For the creation of a corpus of training sentences we normally use some kinds of sentence generators, but a set of generating templates describing the sentence structures with corresponding probabilities of occurrence of their elements had to be created exclusively manually, on the basis of a rigorous statistic and linguistic analysis of possible sentence structures.

Therefore we would like to present some proposals which could improve the conception of automatic creation of training corpora, make them more natural and acceptable. For this purpose we used linguistic methods of interaction (conversation) analysis and an extensive corpus of authentic spoken data. Since the digital recording of real dialogues in information or shopping centres doesn’t cause difficulties, we tried to utilize recorded real dialogues from Czech tourist information centres for the automatic generation of generating templates. We decided to restrict to the task “recognition of the first turns in information demands”. Based on a smaller part of the corpus, a set of typical syntactic constructions occurring in Czech dialogue beginnings was compiled. Their elements and combination rules served to create a number of turn templates. A special program package was created, which automatically generates a training corpus of user demands.

2. Theoretical basis of linguistic analysis
Theoretical, systemic-functional and functional generative linguistics have carried important knowledge on parsing and disambiguation of sentences, on linguistic structures and functions. But since we cannot consider single sentences or written texts to be part of spoken conversation, we have to analyse natural occurring talk-in-interaction. That means we have to analyse the functioning of turn taking and co-operation between speakers, how they manage misunderstandings, which linguistic elements are used to structure and conduct the course of dialogue, which elements show the communication partner that there is something wrong, and which elements can prompt specific ways of interpreting communicative functions (illocutionary forces) of utterances by speakers.

Therefore we want to refer to the achievements of interactional linguistics and conversation analysis – two linguistic disciplines, which deal with the detailed analysis of authentic everyday talk and talk in various institutional settings. Especially important are the linguistic findings on dialogue and turn construction, because they can help us to understand how dialogue systems have to be created in order to answer the needs of the user.

Authentic, naturally occurring utterances (and consequently also demands of clients in tourist information centres) need not to be complete and/or correct in the sense of grammar and linguistics. These elliptical and other “incorrect” constructions are nevertheless fully functional, even if they occur at the beginning of the dialogue:

SB_190899_176 "wc"

01 KF: good afternoon
02 IF: good afternoon, please
03 KF: I just wanted to ask
04 bathrooms toilets

The construction of authentic talk-in-interaction follows its own rules, which we have to describe with respect to our specific corpus and the adaptability to the creation of automated dialogue systems in Czech language. It is therefore not useful to work only with terms of systemic linguistics or other linguistic theories relating to “correctness” or grammaticality of language means and use. Besides the term sentence, meaning a certain pattern (paradigm) on the systemic level of language, we should use the terms utterance (realization of a sentence), turn (part of the dialogue which is produced by one speaker
without interruption), sequence (at least two turns referring to each other and building a more complex part of the dialogue, e.g. question-answer), turn construction unit (TCU) (“the smallest linguistically possible unit ... with one or more than one TCU constituting a possible turn ...” [5], p. 14), turn fragment (unfinished stretches of talk [4], p. 3).

The communicative function of a speaker’s utterance in face-to-face-talk need not to correspond with its context-free semantic interpretation. With regard to the following turn of the other speaker it is possible to interpret the communicative function of a turn correctly. For instance, queries to the location of certain objects (sights) in town are usually understood to be demands for direction giving, and utterances asking about the wish of the client are interpreted as an invitation to pose a question or demand.

According to the various, but not numerous kinds of services which are offered by tourist information centres (information on local and regional sights, institutions and so on, direction giving, sale of post cards, souvenirs ...) there is a certain spectrum of syntactical turn-constructions with similar characteristics – they differ, however, with respect to explicitness, word order or absence resp. appearance of further elements like politeness formulas. That means that many of the templates created by the program (to understand the following description see please the next paragraph) have a number of elements in common. The templates consist of a finite number and a certain kind of combination of non-terminal symbols (some of them are TCUs), but their concrete linguistic realization is quite multifarious.

For our purpose it is nevertheless necessary to go beyond the boundary of TCUs, because most of the demands are formulated using the same or very similar syntactic constructions. In order to distinguish them and to create derivation rules counting for the templates of the whole corpus, we have to define the singular components of some syntactic constructions, their role and context status. The structuring of the turns into combinable components is partly based on valency theory (the valency positions of the predicate are represented each by one non-terminal element), partly it is even deeper (numerals and determiners are treated as single components since they regularly occur at the beginning of certain types of non-terminal elements). Some examples for the structuring of user demands are:

3. Methodical approach

First of all, it is necessary to have a big amount of authentic material from the sphere of interest. Our corpus contains about 500 dialogues with an overall length of more than 13 hours. This material has to be transcibed and analysed in order to determine the course of information dialogues and to assess, which dialogues and turns are typical. In the first phase of our project, we concentrated on initial phrases containing a concrete information demand of a client.

The next step is to analyse their construction principles. As we stated above, each turn can consist of one or several turn construction units (TCUs). One important criterion of TCUs is their potential completeness, they can appear as syntactically complete sentences, noun phrases (NP), prepositional phrases (PP) or even one-word-constructions. Frequent TCUs are for instance greetings and contact formulas. In spoken dialogues, it is crucial to take into consideration all possibly constituting aspects of a TCU: syntactic construction, semantic structure, pragmatic meaning (communicative function) and prosodic characteristics, because the findings of conversation analysis in other languages let expect an interplay of all these aspects. These aspects do not necessary correspond – there are cases, in which only one aspect fulfills a distinctive function.

Furthermore, there are some particles and dialogue markers in Czech language, which are potentially polyfunctional [2]. Only their position in the dialogue, their weight (the capability to stand alone, i.e. to represent a whole turn or TCU) and their prosodic characteristics determine the several communicative functions of this linguistic elements:

The relationship between the single non-terminal symbols (realized by smaller linguistic units like noun phrase (NP), verbal phrase (VP), prepositional phrase (PP), numeral (Num), pronoun (Pron) and so on) and the dialogue unit “turn” is insofar relevant, as there are turns with only one and turns with more symbols. But only a few non-terminal symbols is really capable to serve as a stand-alone turn. In most of the cases we have to deal with several non-terminal symbols, being combined moreover in several ways. They have various syntactical weight, i.e. they can be predicative units providing valency positions, which have to be occupied, or non-predicative units serving just to occupy free valency positions (NPs, PPs). The latter can appear also alone. Furthermore, word order is quite variable.

On the basis of our test corpus the syntactic and dialogic status of non-terminal symbols can be defined as follows (examples):

<table>
<thead>
<tr>
<th>S = TCU, pre-position</th>
<th>NP, TCU</th>
</tr>
</thead>
<tbody>
<tr>
<td>GREET</td>
<td></td>
</tr>
<tr>
<td>INTENT</td>
<td>VP = incomplete, S = TCU, pre-position</td>
</tr>
<tr>
<td>PLACE</td>
<td>PP</td>
</tr>
<tr>
<td>QUANT</td>
<td>Num</td>
</tr>
<tr>
<td>EXISTENCE</td>
<td>VP, incomplete</td>
</tr>
<tr>
<td>DETERMINATUM</td>
<td>Pron</td>
</tr>
<tr>
<td>INDETERMINATUM</td>
<td></td>
</tr>
<tr>
<td>POLITE</td>
<td>Particle, pre-, central or post-position independent, TCU</td>
</tr>
<tr>
<td>INSTITUT</td>
<td>NP</td>
</tr>
<tr>
<td>POSSIBILITY</td>
<td>VP, incomplete, post-position</td>
</tr>
</tbody>
</table>
There are some elements, which are capable to take a pre-(first) position and necessarily require a completion because of the regularities in spoken language. These positionally dependent elements are very important for the automatic generation of templates, because they are regularly followed by another element, which can thus be predicted. This fact has to be included in the program and can help to recognize at least the status of some unclear words or passages. Examples for positionally dependent elements (symbols) in user demands at the beginning of the dialogue (this need not count for any other dialogue position) are @DETERMIN, @INDETERMIN or @QUANT:

\[[@POLITE]50 @QUANT]50 @DETERMIN @ITEM @INTENT @INDETERMIN @ITEM @INTENT @DETERMIN @INSTITUT

Most of the expressions representing positionally dependent non-terminal symbols are facultative; especially the inflationary and redundant use of determiners in Czech spoken language is a well known fact. Elements, which are incomplete with respect to valency (@EXISTENCE, @POSSIBILITY or @ACCESS), behave analogically: They are regularly followed or preceded by @ITEM, @INSTITUT or @ACTION performed by a finite number of possible elements (terminal symbols representing more or less complex NPs with or without determiners resp. infinitives):

\[
\begin{align*}
\text{ITEM} & \quad \text{mapu zpravodaje program} \\
\text{INSTITUT} & \quad z\acute{y}v\acute{a}\acute{y}\acute{a}mek rozhledna \\
\quad & \quad "\text{japonsk\'{a} ulice" "toaleta wc"} \\
\quad & \quad "\text{vyhl\'{a}v\'{i}dkov\'{a} v\'{a} v\acute{e}\acute{v}\acute{y}v\acute{e}\acute{z}\acute{y}}" \\
\quad & \quad "\text{muzeeum um\'{e}n\'{i}}" \\
\quad & \quad "\text{po\acute{v}stv\'{o}n\'{i} banka" galerie filharmonie informace} \\
\text{ACTION} & \quad ubytovat \quad "\text{jezdit na koni" \\
\quad & \quad "\text{vym\'{e}nit pen\acute{y}v\'{i} z\acute{e}}"}
\end{align*}
\]

These elements can also – as already mentioned – occur without their syntactically bound part, the predicate. In these cases there is almost always preceding a determiner and/or an utterance standing for the non-terminal symbol @INTENT: @INTENT @INDETERMIN @ITEM @INTENT @DETERMIN @INSTITUT

At this stage of our project we have to take note of several problems concerning the structuring of units which should get the status of non-terminal symbols. These problems are reflecting the whole spectrum of different verbalizing modes and language use in spoken interaction.

4. Creation of sentence corpora

The balanced corpus of training sentences has to cover all kinds of sentence structures and multiforms of sounds, the distribution of sounds has to correspond to the speech representation. The compactness can be seen as an advantage of the corpus; it involves typically 5,000 or 10,000 training sentences which cannot be written manually (this procedure is extremely tiring and its use is disadvantageous). That means, the set of training sentences must be generated with a special sentence generator which guarantees the required statistical properties of the sentence structures. Moreover, they are composed of few sentence kinds (basic structures) containing a great number of different words and phrases.

The stochastic sentence generator [3] is used to generate a training set of Czech sentences. The input file containing original sentence templates is in our case mostly generated automatically on the basis of the above described linguistic analysis of real-life dialogues (only a small part of sentence templates has to be created manually as a learning corpus of the generator).

It has to be adapted to the universal generator structure, i.e. all Czech orthographic symbols have to be represented in the PCS notation, e.g. special characters as \( \acute{d} = \{\{a\},f=\{\{i\},f=\{\{r\},\acute{a}=\{\text{accent 23u}} \)

The generator produces sentences with the aid of a context-free grammar. This grammar involves in addition to the usual standard start symbol a set of special “non-visible” non-terminal symbols \( TEMPL_1 \ldots TEMPL_m \) corresponding to the respective question/answer types. These symbols represent “empty” places that will be only expanded; they are named sentence templates and they occur in the format

\[
TEMPL_i \rightarrow [\text{@N}_i | \text{@P}_1 | \text{@P}_2 | \text{@P}_3 | \ldots
\]

where symbols \( N_i, i = 1, 2, \ldots \) represent the names of non-terminal symbols and numbers \( p_j, j = 1, 2, \ldots \) the occurrence probability of the joint non-terminal symbol; production rules of the grammar get the form

\[
N_i \rightarrow [\text{@P}_1 | \text{@P}_2 | \text{@P}_3 | T_1 | T_2 \ldots
\]

where symbols on the right hand side of the rule separated by vertical lines represent alternatives of the rule, \( N_i \in V_N, T_j \in V_T \), the terminal symbols consisting of more words must be closed in quotation marks, e.g. “muzeeum um\'{e}n\'{i}”. But only one of them can be selected for the sentence generation (all these alternatives occur there with the same probability). Alternatives in brackets supplemented by an integer number \( p \) represent the optionally occurring part of the sentence (named by the non-terminal symbol \( N_i \)) with the corresponding occurrence probabilities. The following example shows a small part of the created file of templates and grammar rules:

Example 1: A part of a template file

\[
\begin{align*}
\text{[@POLITE]50 @QUANT]50 @DETERMIN @ITEM. @INTENT]33 @INTERROG]50 @POSSIBILITY @ACTION @INSTITUT]25? @GREET]25 @INTENT @ITEM]50 @QUANT]25. @INFORMATION]50 @PLACE @ACTION @CONJUNCT]50 @INTENT @INTERROG. @GREET]50 @HESITATION]50 @INTENT]50 @INTERROG]50 @EXIST @INSTITUT]50 @INSTITUT. @GREET]25 @INTERROG]25
\end{align*}
\]

\[
\%
\]

\[
\text{INTENT } \"[\{a\} sem se } \text{\textbackslash v\{s\}la jenom zeptat" \\
\text{chci se takhle na to pod\'{i}v\'{i}vat"} \\
\text{\"[\{a\} sem si p\'{v}\{r\}\{s\}la tadydle pro"} \\
\text{\"[\{a\} m\'{v}\{a\}m takovej dotaz"} \\
\text{\"cht\{v\}\{e\}l smie se zeptat"} \\
\text{\"[\{a\} sem se ch\{t\}\{v\}\{e\}l z\acute{e}ptat"} \\
\text{\"m\'{o}v\{u\}v\{z\}u se zeptat"}
\]
\]

\[
\text{POLITE pros\'{i}v\{i\}m } \\text{\"[\{a\}s} \\
\text{p\{v\}\{e\}kn\{v\}\{e\} } \\text{\"prosim } \\text{\textbackslash v\{a\}s} \\
\text{\"pros\'{i}v\{i\}m } \\text{\textbackslash v\{a\}s} \]
\]

\[
\%
\]

\[
\text{INTERROG kde } \"\text{\textbackslash s ker\'{v}\{y\} strany" kdepak jak}
\]

The following facts can be seen as the advantages of the presented grammar definition:
- a great number of training sentences can be generated;
- the word dictionary can be stored in one place and it can be later easier modified;
- different word group significance can be considered.

On the other hand, the impossible supervising for the sentence meaning and for the duplicate occurrence of some terminal sym-
bols, e.g., local names, can be seen as the disadvantage of this concept. Due to this disadvantage a small number of sentences may be generated syntactically correct, but meaningless. But the occurrence probability of such sentences can be reduced by an adequately great variety of template variations.

The structure of the generator template set is developed on a basis of the linguistic analysis of the corpus of transcribed real dialogues recorded in approximately 20 information offices and centers of the Czech Republic. The collected real queries were generalized and “standardized” from the viewpoint of the semantic information and 29 sentence templates were obtained in the first attempt.

5. Creation of sentence templates

The automated creation of training sentence templates requires the manual creation of a small set of template patterns further used as a “learning material” for the program initialization. The implemented program package reads the template patterns, creates an internal representation of the concrete sentence grammar and analyzes the corpus of transcribed “real-life” dialogues in the last step. Based on the “learned” fundamental (manually written) sentence templates it creates the template to each analyzed dialogue request. In cases, in which the linguistic-analysis-module cannot classify some part or the input sentence, it asks for help of the user (from this point of view the creation of sentence templates doesn’t run automatic, but only automated) for the determination of the “unknown” part of the sentence and “learns” the user’s decision (new input).

The program package consists of seven program modules written in C++ programming language, in which the sentence structures are implemented as a simple chained dynamic lists. The grammar rules “learned” from the manually prepared template patterns are represented by multi-level chained lists interconnected by especially developed data structure. The calls of program modules are controlled by PERL control routine.

6. Experiments and results

When creating the templates for the sentence generation, we were confronted with several linguistic problems:

First of all it is sometimes difficult to decide, which units belong together or could be also interpreted as separate units. This strongly depends on the word order (which is in Czech language quite variable) and the relatively bound position of enclitics. That means, we have to deal with similar constructions representing the same non-terminal symbol, but having different syntactic properties, for instance: máte tady (‘do you have here’, first position in a request), tady máte (‘here you have’, position after an interrogative).

The second, quite serious problem is the divergence between Czech standard language and colloquial Czech. In everyday spoken communication there is an overwhelming prevalence in the use of colloquial Czech, which has, of course, a lot of regional specifics. Nevertheless, the use of one of these codes depends on a range of situation-determined aspects including the relationship between the communication partners. In our corpus, we find a wide range of variations and combinations of both codes (code-switching is an extremely widespread phenomenon in Czech quotidian communication). Even though the clients and the information specialists don’t know each other, they communicate very often in an informal, relaxed way, i.e. using colloquial Czech. Nonetheless, there are people talking exclusively standard Czech (esp. older clients). What we do not know is: Which code will people use in order to communicate with the automatic dialogue system, which is more anonymous and represents a fairly new kind of communication? At this time, we know less about the regularities in Czech human-machine communication. That is why we have to take into account all problems and variations of code-switching between standard and colloquial Czech.

A third – well known and often cited – problem is, that we are concerned with a highly flective language. The compatibility of and congruence between terminal symbols is strongly influenced by this fact. Furthermore, there are more combination possibilities of adjectives and other linguistic means because of certain unification processes in spoken Czech (some endings are for instance enforced, others are disappearing).

A corpus of 2,000 training sentences was experimentally created using a set of automated prepared sentence templates. The probability distribution of sentence structures was tested; it corresponds to the presumptions. The corpus will be used to train a program module of the word recognizer of a new developed tourist information dialogue system.

7. Conclusion

During the first phase, we analyzed only the first requests of clients, their first utterances after entering the tourist information office. About 60 percent of our corpus of real dialogues could be used for the automatic generation of sentence templates, but we expect a much higher percentage of usable dialogues for the generation of other sentence templates (further questions and requests of clients, problem solving, misunderstandings and so on), because recording troubles and missing passages occurred mostly at the beginning of dialogues. Nevertheless, our corpus of real dialogues in Czech tourist information centres is big enough and represents various regions of the Czech Republic.

With regard to authentic occurring data and the results of sentence generation we have to state, that there are constructions which have to be valued as ungrammatical from the point of view of standard grammar, but which are fully functional and acceptable in spoken communication. They do not (or rarely) cause misunderstandings, rather they are looked at as economic linguistic means in order to solve problems quickly. The training corpus has to cover all these cases.

8. Acknowledgement

The work presented in this paper was partly supported by the Bilateral Cooperation Project CZE-00-034, partly by the Czech Ministry of Education under contract MSM 2352 00005, and by the German-Czech Fund of Future under contract No. A/IV/146.

9. References