The treatment of noun phrase queries in a natural language database access system

Alexandra Klein and Johannes Matiasek and Harald Trost
Austrian Research Institute for Artificial Intelligence
Schottengasse 3
1010 Vienna
Austria
{alexandra,john,harald}@ai.univie.ac.at

Abstract

In this paper, we are going to describe some aspects of the TAMIC-P system for German, which interprets natural-language queries to databases in the social insurance domain. These natural language queries are complex NPs, consisting of clusters of NPs and PPs. The parser uses information about co-occurence and domination of linguistic elements as well as concept hierarchies in the domain to construct a parse tree and, subsequently, a derivation in quasi-logical form. This derivation can be translated into a database access query.

1 Introduction

Systems which map natural language query phrases onto database access query statements, (Trost et al., 1987), (Androutsopoulos et al., 1995), provide a natural communication environment. Yet this means that they must be able to handle vagueness, incompleteness or even ungrammaticality as these phenomena tend to be associated with language use under specific external constraints as e.g. in situations concerned with database access. In this paper., we are going to describe the natural language understanding component of the TAMIC-P[1] system for German, which interprets natural language queries addressing the databases of the Austrian social insurance institution for farmers (Sozialversicherungsanstalt der Bauern, SVB). The input queries are parsed and mapped onto a representation in quasi-logical form which serves as basis for the required database access. Simultaneously, the queries are searched for domain-specific cue words which are part of a lexical knowledge base (cf. e.g. (Christ, 1994)) This knowledge base is also accessible from the user's query interface. Additionally, many legal terms which occur in the queries may be linked to the underlying legal regulations. These regulations are available in hypertext format and allow for browsing via the user interface. Domain-specific help files will eventually also be integrated in the data to be accessed via the user interface.

An evaluation of the user requirements in this specific natural language task yielded the result that the prospective users of the system, social insurance clerks at local information days, feel most comfortable if they have the possibility to input their queries as noun phrases. For the natural language query interpretation task, this implies that complex and heterogeneous noun phrases have to be interpreted adequately.

In the following sections, we are going to describe the general aim of the TAMIC-P system, the system scenario, the kind of input the system has to be able to deal with and the parsing process. We will then give an outline of how some complex queries are treated in parsing, and finally, we are going to conclude with observations (including their implications for further work) derived from the current system setting and its application in the process of parsing NP queries in the TAMIC-P domain.

2 The TAMIC-P system for German in the Austrian scenario

The TAMIC-P system is realized in collaboration with Italian and German[2] project partners.

[1] Research in this project is funded by the European community under contract no. LE-4253. Financial support for ÖFAI is provided by the Austrian Federal Ministry of Science and Transport.

[2] DFKI GmbH, Saarbrücken, Germany, has contributed to the German language module and the associated knowledge sources. The authors would therefore like to thank Stephan Busemann for his work in this area, particularly the development of the grammar.

It consists of two language components; one for Italian in the Italian scenario and one for German in the Austrian scenario. While the interface structure (developed by the Italian partner) and the configuration of the main modules are mostly identical in the two scenarios and while both applications aim at interpreting NP queries, the two natural language components represent two distinct approaches to the query interpretation task. In this paper, we will focus on the query interpretation module as well as the required knowledge sources for dealing with German NPs.

The scenario- and language-specific part of the system consists mainly of the actual parsing component, a lexical knowledge base (LKB) representing the entities denoted in the queries and their relations to each other, the conceptual data model (CDM) specifying which entities can be found in the various databases, and the logical data model (LDM) which approximates the actual data as they occur in the domain.

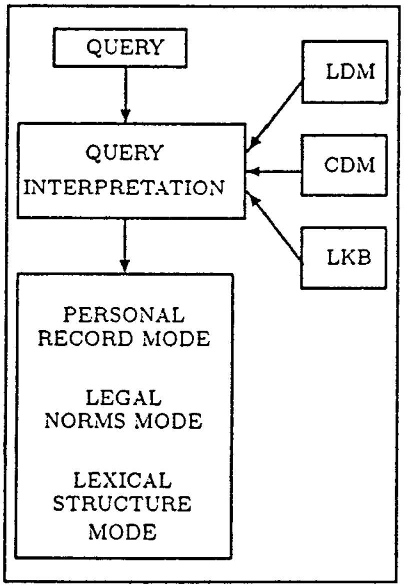

fig1: Architecture of the interpretation module for German queries in the Austrian scenario

The parsing component and the knowledge sources are connected closely. In order to construct a quasi-logical-form expression for a natural language query encountered as system input, the parsing component has to consult the relations between the denoted entities as they are represented in the database. These relations are modelled in the CDM, which is a unified and simplified version of the logical data model represented in the database. Yet it is not sufficient to construct the QLF representation: as cue words have to be identified in the queries in order to present legal norms and domain-specific lexical relations, these contexts have to be built and embedded in the framework of legal regulations and concept structures. Consequently, there are three output modes (personal information, legal texts, legal lexicon) which are presented on the interface as a card-index display. A fourth output mode (domain-specific help files) will eventually be added to the system.

3 The TAMIC-P scenario

Queries which have to be interpreted in the TAMIC-P domain concern all areas of social security, i.e. pension, health and accident insurance. In order to provide the required information, several databases have to be consulted. As social insurance employees in advisory dialogues have to work under extreme cognitive pressure due to limited time and other situational constraints, it is the aim of the TAMIC-P project to simplify this advisory process by providing one interface for the various different tasks which have to be performed:

- consult citizen's insurance data in several databases
- consult the relevant laws and regulations
- consult a legal glossary for related terms and concepts

In TAMIC-P, these tasks are based on the interpretation of natural language queries and the interaction with the interface. For the user, this verbal and graphical interaction faciliates obtaining the requested results from the databases as well as the norms and the compilation of legal terms. At the same time, clients using the system have to rely on great robustness in the query interpretation task so that queries do not result in the distribution of false, inadequate or incomplete information.

4 The corpus

As it has already been mentioned, queries entered into TAMIC-P by SVB clerks concern personal data as well as legal affairs. A fairly typical query would be for instance

Ersatzzeiten wegen Kindererziehung
'Exemption times because of child raising'

At the present state of the system, this query has three dimensions with regard to its interpretation: First, it concerns the personal insurance records stored for a specific persion (this person is determined by the context) who may or may not have acquired the requested type of insurance months. Second, it refers to a special official status of insurance months which is defined in legal texts describing the benefits which can be derived from different kinds of insurance times. Third, the query and its underlying concepts have to be compared to related queries and concepts: I.e. in a wordnet-type structure, *Ersatzzeiten* ('exemption times') is in this specific use synonymous to *Ersatzmonate* ('exemption months') and belongs to the category of *Versicherungszeiten* ('insurance times'). Furthermore, there are several different types of specific 'exemption times because of child raising' (e.g. raising adopted children, grandchildren etc.) which have to be considered in an evaluation of the insurance records, particularly if a citizen applies for retirement pension.

An evaluation of user queries has shown that the SVB clerks were reluctant to form full, grammatically elaborate sentences; instead, they preferred to rely on noun phrases denoting the requested concepts from the social insurance domain. Regarding the example, this does not come as a surprise as - apart from the additional cognitive effort which is required in using a complete sentence - using an NP seems to be a natural way of including the three dimensions of the specific personal information from the insurance records, the legal norms, and the lexical knowledge base in the noun phrase quoted above. In contrast, three complete sentences have to be formed to refer to the same dimensions if noun phrases are avoided. The corresponding English phrases for the German full examples are:

- *Which exemption times because of child raising are stored for Mrs/Mr x?*
- *What are the legal regulations concerning exemption times because of child raising?*
- *What are the relevant lexical properties associated with exemption times because of child raising?*

Note that the second and the third example use some kind of metalanguage to link 'exemption times because of child raising' to the required dimension. In contrast to this, the noun phrase 'exemption times because of child raising' needs no metalanguage und points elliptically to all three dimensions. Therefore, if only noun phrases are used, their inherent vagueness and incompleteness at least provide means to consult all the relavant information sources simultaneously.

For the NP analysis, this implies that much effort has to be invested in modelling the entities and relations denoted by the linguistic expressions, and certainly also the mapping between the linguistic and the conceptual levels. Tighly packed linguistic structures refer to complex conceptual structures. The dependencies in the conceptual model are mirrored in the linguistic expressions. As we are dealing with noun phrase queries, it is not possible to regard verbs as assigning the key structural dependencies. Yet it can be said that certain conceptually pivotal nouns 'subcategorize' (to use this term loosely for illustration purposes) for specific linguistic objects which are determined by the underlying conceptual data model as well as certain conventions of use. These 'subcategorization relations' are often confirmed by prepositions with weak semantics:

Ersatzzeiten **wegen** *Kindererziehung*
'Exemption times **because of** child raising.'

The function of the preposition is to indicate the relation. In German, these types of noun phrase are often turned into a compound with the meaning remaining unchanged:

Kindererziehungsersatzzeiten
'child raising exemption times'

These remarks already describe the NP typology encountered in queries: complex NP clusters, NP-PP clusters (with faded prepositional semantics) and complex compounds. Of course, any combinations of the three types may also occur.

5 Parsing NP queries

5.1 Compositionality and non-compositionality

Generally, the parsing process in the TAMIC-P query interpretation component for the Austrian scenario relies strongly on the hypothesis that the semantics of a phrase (represented for example in quasi-logical form) can be derived by composition of the QLFs of the parts the phrase is composed of. This implies that linguistic paraphrases which denote the same object or set of objects have to eventually end up in identical representations. As we are dealing with a limited domain, which usually restricts the number of options available for interpretation, this approach is feasible. Yet so far, we have no possibility of dealing with queries beyond the field of social-insurance. The domain also ensues that the conceptual data model, representing entities and relations in the actual database, often contains simple objects, attributes or attribute values which are referred to by complex query elements on the natural-language level. In order to tackle this problem, a filter mechanism is used which

- treats all natural language utterances in a compositional manner, possibly involving QLF prediates (originating from the lexicon) that do not denote a CDM object and
- applies a set of substitutions to the QLF resulting from the parse that transform the complex description into the simplistic one contained in the CDM.

Since the number of such 'noncompositional' objects in the CDM is limited and their 'compositional' meaning in terms of a QLF is unique, not very many substitution definitions are required, and it is not difficult to come up with them. At the same time, we are aware of the fact that a larger domain might require a broader analysis approach (Rayner, 1993).

5.2 Lexical resources

As far as lexical resources are concerned, the query interpretation module for German uses two resources of lexical data: a small morphological lexicon, which can be employed for morphological annotation, and a repository of semantic and 'subcategorization' information included in the lexical knowledge base (LKB). In the LKB, each entry represented as a synset in a wordnet-style structure contains slots for synonyms, hyponyms and hypernyms, as well as a specification of the quasi-logical form as it can be derived from the conceptual data model. Returning to the example of *Ersatzzeit wegen Kindererziehung* ('exemption time because of child raising'), from the LKB we obtain basically the following structure:

```
ERSATZZEIT WEGEN KINDERERZIEHUNG

CATEGORY: n-filler
SYNONYMS: Ersatzmonat wegen
Kindererziehung
HYPERNYMS: Ersatzzeit, Kindererziehung
HYPONYMS:
QUASI-LOGICAL FORM: :ARG (BEITRAG ?X),
:RESTR (HAS-QUAL ?X <K>), :VAR (?X)
DESCRIPTION: Typ von Ersatzzeit, der
aufgrund von Kindererziehungszeiten
angerechnet werden kann
```

CATEGORY indicates whether we are dealing with an object, an attribute or an attribute value ('n-filler' indicates that this is an attribute value). SYNONYMS, HYPERNYMS and HYPONYMS each list related lexical concepts according to the conceptual data model. The QUASI-LOGICAL FORM describes the combination of the elements which are needed for mapping the query representation onto the appropriate database access statement, and DESCRIPTION documents a legal explanation of the term.

5.3 The parser

The parser uses the information stored in the lexical resources for assigning a structure to the linguistic expression. This information is matched and combined according to grammatical knowledge (encoded in the grammar and the morphological lexicon) as well as domain knowledge inherent in the lexical knowledge base. From the parser's point of view, it is desirable to obtain a single feature structure containing both subcategorization information and logical form from the lexicon lookup. How semantic information has to be combined during parsing may not only depend on the QLFs in the CDM for the signs to be combined, but also on lexical features of the head and the syntactic role that is filled by the argument (or modifier). In the sentence

John beats Paul

for example, the meaning depends crucially on whether *John* or *Paul* fills the subject slot. Although these differences occur much more frequently with verbs, which are rather irrelevant in the TAMIC-P corpus, this behaviour can also be observed with nouns.

Therefore, the following decisions for the encoding of semantic and subcategorization information have been taken:

1. All substantive categories (nouns, adjectives, verbs) provide subcategorization information for both arguments and modifiers they may take (modifiers/adjuncts are viewed as kind of optional arguments).
2. For each element subcategorized for, syntactic and semantic information is provided, restricting the possible fillers.
3. In order to explicitly specify the semantic relationship of the head and the argument in the CDM, 'glue' predicates may be specified which bind the quasi-logical forms of the head and the argument.

5.4 Grammar

Based on the analysis of a small corpus, grammatical rules were derived. At this point, the grammar consists of 9 context-free rules. They cover to whole range of queries which have been encountered so far. In the first version, complex prenominal modifiers and relative clauses are not implemented, but the corpus suggests that these are added benefits rather than urgent needs.

For the expression *Ersatzzeit wegen Kindererziehung* , we use the following rules:

```
GRAMMAR FRAGMENT:

TOP -> DetP AP N NP PP*
NP -> NP1 Narg
NP1 -> DetP A* N
PP -> P NP
```

Categories in curly brackets are optional. The asterisk represents the Kleene star. Underlined categories are heads. PP can be a temporal modifier. DetP can be a complex determiner, corresponding e.g. to *between three and six* , *up to five*, *more than six*, *at most three*, or a simple article like *the* or *a*. Lexical categories other than determiners include N, A, P.

The derivation with the above grammar fragment is as follows:

Ersatzzeiten wegen Kindererziehung
'Exemption times because of child raising'

```
DERIVATION:

TOP
(NP (NP1 (N (Ersatzzeiten) )),
PP (P (wegen), NP (NP1 (N
Kindererziehung))))
```

Lookup in the LKB resulted in a structure which has already been described in the previous chapter:

```
ERSATZZEIT WEGEN KINDERERZIEHUNG

CATEGORY: n-filler
SYNONYMS: Ersatzmonat wegen
Kindererziehung
HYPERNYMS: Ersatzzeit, Kindererziehung
HYPONYMS:
QUASI-LOGICAL FORM: :ARG (BEITRAG ?X),
:RESTR (HAS-QUAL ?X <K>), :VAR (?X)
DESCRIPTION: Typ von Ersatzzeit, der
aufgrund von Kindererziehungszeiten
angerechnet werden kann
```

By combining the grammar rules with the information from the LKB as well as the CDM, we obtain the following result:

```
RESULT OF PARSING (QLF):

((:ARG (CDM::BEITRAG ?FS18561)) (:VAR
?FS18561) (:SEL)
(:RESTR (CDM::BEITRAGQUAL ?FS18561
?FS18379) (QUAL-VALUE ?FS18379)
(IN ?FS18379 (43
E3))(CDM::PERSON_KONTEXT ?FS18561
?FS19601)))
```

During the parsing process, it was correctly recognized that the query concerns a type of BEITRAG ('contribution'), with a certain qualification associated with the contribution. The codes 43 and E3 indicate which categories are assigned to the qualifications of the contributions in the database. From the parser output, a database access query can be constructed, which retrieves the required element from the person's insurance records in the database. The person is identified in the quasi-logical form with the PERSON_KONTEXT predicate.

6 Dealing with complex queries in parsing

One example for a more complex query is the phrase

Kindererziehungszeiten der Person mit der Versicherungsnummer 1001050610
'child raising times of the person with social insurance number 1001050610'

This query is asking more generally for any insurance times associated with child raising. In the Austrian social insurance system, citizens are often identified by their insurance number, as it is the case in this query. The syntactic analysis according to the grammar yields the following derivation where the social insurance number has been omitted for simplification:

```
DERIVATION:

TOP
(NP (NP1 (N Kindererziehungszeiten))),
(NP (NP1 (DetP der), (N Person)),
(PP (P mit), NP (NP1 (DetP der), (N
Versicherungsnummer ...))))
```

The parser uses co-occurence rules to grasp the phenomenon of nominal 'subcategorization'. In the example cited above, there is an entry which specifies that 'insurance number' tends to occur (sometimes connected by *von*, 'of') with a 'person' who has an insurance number and with the actual 'number'. The SUBCAT entry looks as follows:

```
SUBCAT:

(def_subcat (cdm::PERSONVSNR ?x ?y)
((:syn ((cat .  (np pp)) (prep .
"von")) :sem (cdm::PERSON ?x))
(:syn ((cat .  (cardinal string))) :sem
(= ?y ?z)))
```

These pieces of information are used to build up the parse tree:

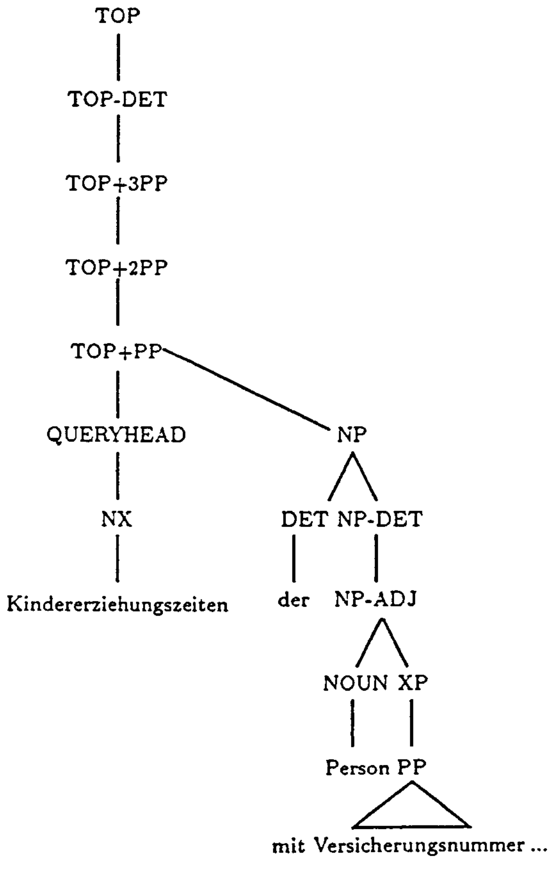

fig2: (Simplified) parse tree

This parse tree is then mapped onto the quasi-logical form:

```
RESULT OF PARSING (QLF):
```

```
((:ARG (CDM::BEITRAG ?FS18561)) (:VAR
?FS18561) (:SEL)
(:RESTR (CDM::BEITRAGQUAL ?FS18561
?FS18379) (QUAL-VALUE ?FS18379)
(IN ?FS18379 (43 51 87 A8 A9 E1 E3 E5
E6 E7 E8 E9))
(CDM::PERSON ?FS19601) (CDM::PERSONVSNR
?FS19601 ?FS18419)
(= ?FS18419 "1001050610")
(CDM::PERSON_KONTEXT ?FS18561
?FS19601)))
```

As *Kindererziehungszeiten* is a hypernym for *Ersatzzeiten*, there are more different types of qualification available.

The compositional aspect of this approach also allows the analysis of phrases like *Kindererziehungszeiten der Person mit der Versicherungsnummer 1001050610 nach 1940* ('child raising times of the person with insurance number 1001050610 after 1940'), *Kindererziehungszeiten der Person mit dem Nachnamen Huber nach 1940* ('child raising times of person with last name Huber after 1940') etc. The analyses for these complex phrases are carried out in the same manner as for the simpler example described above.

7 Conclusion and further work

So far, this approach seems to work efficiently in the limited domain of social insurance information. The integrated treatment of NP 'subcategorization' allows for an analysis of syntactic attachment as well as the interpretation of semantic structure. The mapping of parse trees onto quasi-logical form also combines syntactic and semantic representation. This way, information needed for constructing the database access query does not get lost in the analysis process. The analysis also provides information which can be used in determining the appropriate legal norms and the concept hierarchy as it is represented in the LKB.

As it was mentioned above, a problem which has not yet been tackled is the treatment of queries from outside the domain. As the users of the system are insurance clerks, they tend to be very cooperative. Therefore we have not encountered uncooperative behaviour in our user studies. Within the actual domain, work on covering the whole range of user queries is being carried out. There are some possible types of extremely complex relations for farmers cultivating property owned by someone else. These relationships have be be modelled in the database for the computation of social insurance contributions and may lead to complex queries. Additionally, it would be useful to include an analysis of compounds, which – as it is well-known – can be rather complex in German. From the evaluation of these types of possible queries, however, we have gained the impression that they can be fully analyzed in the compositional approach described in this paper.

References

Ion Androutsopoulos, Graeme D. Ritchie, and Peter Thanisch. 1995. Natural Language Interfaces to Databases - An Introduction. *Natural Language Engineering*, 1(1):29–81. DAI RP-709.

Oliver Christ. 1994. A Modular and Flexible Architecture for an Integrated Corpus Query System. In *COMPLEX'94*.

Manny Rayner. 1993. *Abductive Equivalential Translation and its Application to Natural Language Database Interfacing*. Ph.D. thesis, University of Stockholm.

Harald Trost, Ernst Buchberger, Wolfgang Heinz, Christian Hoertnagel, and Johannes Matiasek. 1987. 'Datenbank-DIALOG' - a German language interface for relational databases. *Applied Artificial Intelligence*, 1(2):181–203.