System Demonstration

ANTHEM

Advanced Natural Language Interface for Multilingual Text Generation in Healthcare
(LRE 62-007)

G.Deville[a], E.Herbigniaux[a], P.Mousel[b], G.Thienpont[b], M.Wéry[a]

[a] Ecole de Langues Vivantes - Facultés Universitaires de Namur
61 rue de Bruxelles - B-5000 Namur - Belgium

[b] Centre de Recherche Public - Centre Universitaire
162a avenue de la Faïencerie - L-1511 Luxembourg - Luxembourg

1. System builders and contacts

The ANTHEM project: "Advanced Natural Language Interface for Multilingual Text Generation in Healthcare" (LRE 62-007) is co-financed by the European Union within the "Linguistic Research and Engineering" program. The ANTHEM consortium is coordinated by W. Ceusters of RAMIT vzw (Ghent University Hospital) and further consists of the Institute of Modern Languages of the University of Namur (G. Deville), the IAI of the University of Saarbrücken (O. Streiter), the CRP-CU of Luxembourg (P. Mousel), the University of Liege (C. Gérardy), Datasoft Management nv — Oostende (J. Devlies) and the Military Hospital in Brussels (D. Penson).

2. System category

The ANTHEM prototype is a pilot application resulting from a 30 month European research project in the field of automatic translation and coding of medical diagnostic expressions. The aim of that project was to develop a portable, platform independent prototype of a natural language interface for the translation and encoding of medical diagnostic expressions.

3. System characteristics and Functionality description

The system was developed towards the automatic processing of utterances from the medical diagnostic sublanguage. The input must be a medical diagnosis expressed in French or in Dutch. The output can be (i) one or more translation(s) into Dutch or French, (ii) one or more semantic representation(s) containing SNOMED (Systematized Nomenclature of Medicine) codes or (iii) one or more ICD-10 (International Codification of Diseases) codes.

The performances of the system on a sample of 500 diagnoses are the following:

a) an average of 56% of the input expressions could be processed by the system;

b) an average of 84% of the produced translations were evaluated as 'at least acceptable' by a set of bilingual practitioners;

c) an average of 93% of the produced semantic representations were evaluated as 'at least acceptable' by a set of ANTHEM experts;

d) no systematic evaluation of the ICD-10 encoder was performed so far.

The system has undergone no processing time optimization so far. The majority (more than 75 %) of the processable diagnoses used for testing purposes were translated and semantically represented in less than 120 seconds each.

4. Resources

The lexicons developed in the framework of the ANTHEM project contain a total of 3.801 conceptual entries in both French and Dutch (3.638 simple items and 163 compound lexemes and multi word unit items), resulting in (i) 7.485 French and 8.281 Dutch simple words and (ii) 367 French and 328 Dutch words constituting the clex and mwu items.

The French and Dutch grammar modules used by the ANTHEM prototype are partially based on the CAT2 grammars previously developed at the IAI Saarbrücken. As the CAT2 system (i) is still in evolution and (ii) consists in a series of independent and modular morpho-syntactic elements, no precise amount of actually used rules / size can be given. Both the lexicons and the grammars are written in Prolog.

5. System internals

To achieve the above mentioned objectives, we adopted a layered approach. A medical host application uses the services of the prototype exclusively through an Application Programming Interface (API). The ANTHEM API hides all the technical details of the prototype's implementation to the medical host application. The API uses an existing MT system, called CAT2, for the translation and resorts to an expert system specifically developed for the ICD-encoding of the medical diagnostic expressions.

5.1 The ANTHEM API

We will only recall here two requirements the ANTHEM API has to meet:

1) Simpleness: In order to facilitate the integration of ANTHEM in existing or future medical applications, the API has to be as simple as possible. Therefore, we kept the number of functions of the API to a minimum.

2) Portability: In order to allow a maximum of medical applications to benefit from the services of the prototype, the API had to be portable to a wide range of platforms. Therefore we decided to implement ANTHEM in C. Moreover, as some of the prototype's components, more precisely the

CAT2 system, currently run on UNIX platforms only, we chose a client/server architecture based on Remote Procedure Calls (RPCs) for the API.

The ANTHEM architecture enables us to provide the services of the prototype to medical host applications on virtually any platform. The medical host application communicates with the API client through simple C library calls, the API client uses RPCs to communicate with the API server and, finally, the API server communicates with the MT and encoding systems through standard UNIX interprocess communication mechanisms. The role of the API is to completely hide the complexity of the underlying communication layers to the host application.

5.2 The CAT2 System

CAT2 is a unification-based Machine Translation system developed at IAI Saarbrücken as a sideline of the CEC-sponsored EUROTRA program. The system is entirely written in SICStus Prolog and runs on Unix platforms in a frozen (i.e. not modifiable, but requiring a Prolog interpreter to be run) as well as in a compiled (i.e. executable, i.e. neither modifiable, nor requiring a Prolog interpreter to be run) version.

CAT2 as a rule-based system has two basic parts, the CAT2 formalism and its implementation (the software) and the CAT2 lexica and grammars (the lingware). For the project, the CAT2 formalism was used in its initial form while new lingware was developed specifically for this application. The formal properties of the system are summarized as follows: unification is its only computational mechanism, working on tree structures and on the feature structures annotated to every node of the tree. Unification may be constrained by negative, disjunctive or implicative constraints over simple and complex features.

Within ANTHEM, the CAT2 system is used to analyse Dutch and French diagnostic expressions, translating these (a) into German, Dutch and French and (b) into a semantic representation that serves as input to the Expert system. For both purposes the interlingual approach has been adopted. The interlingua consists of a set of basic word-concepts identified by their SNOMED codes and a limited set of semantic relations which link word-concepts to form larger concepts. The coupling of words to concepts takes place in the language specific CAT2 lexicons, where every conceptual entry contains a SNOMED code and its semantic type (e.g. lex='M-12000') associated with the words (e.g. string=fracture) and their morphosyntactic description.

5.3 The ICD Encoding System

One particular task in ANTHEM is the development of an automatic encoding tool for diagnoses, following the rules of the ICD-10 classification system. Although ICD has been recognised as the main nosologic system during the past century for identifying human ailments, having a well-organised and well-accepted nomenclature, it presents a number of serious drawbacks. One of these is the "gap" between the language used by physicians to register medical problems, and the limited expressiveness of ICD itself. Bridging this gap between ICD at the one hand and common medical language at the other, is the purpose of the expert system. For example, the diagnostic expression "acute inflammation of pericard" uttered by a physician, should be given the ICD-10 code 130, as this code stands for the term "acute pericarditis".

The development of the encoder has been carried out in two steps. First, a task-oriented formalism has been developed. This formalism was not meant to be used directly by an automatic reasoner for further processing, but was developed as a tool for medical knowledge engineers to make the ICD knowledge explicit. In a second phase, a set of additional tools have been worked out such as a formal syntax checker with embedded semantic constraints, a knowledge base converter to rewrite the original knowledge base in a flat file structure with intrinsic hierarchical relationships, and an encoder.

6. Hardware and software

Different technical configurations were used for validation purposes:

A twofold off-site validation was done with a client application developed for this specific purpose. The client was a small Motif application running under SunOS on a SUN SPARC LX workstation (50 Mhz, 32 Mb RAM, 1.05 Gb HD) in Namur. It used Sun's ONC/RPCs (integrated in SunOS) over the Internet to communicate with an ANTHEM server running on a SunOS machine (SUN IPX, SPARC II processor 28 MIPS, 32 Mb RAM, 80 Mb VM, 4 Gb HD) in Luxembourg. The client allowed the user to select a local file containing medical diagnostic expressions, to remotely translate, represent or encode the expressions and to save the results in a local file.

A twofold on-site validation was done at the medical informatics service of the Military Hospital in Brussels. In this case, one client application was a modified version of Datasoft's Medidoc application running under DOS at the Military Hospital. It used Sun's ONC/RPCs (through PC/NFS) over a local area network to communicate with an ANTHEM server running on a Solaris machine at the Military Hospital. Another client application was a customized version of Aurora, the computerized patient record system actually used at the Brussels Military Hospital.

7. Web Page and Email Addresses

ANTHEM Web Page: http://www.crpcu.lu/APOLLO/
W.Ceusters, Project Coordinator: werner.ceusters@rug.ac.be
G.Deville (FUNDP Namur, Belgium): gdeville@cc.fundp.ac.be
E.Herbigniaux (FUNDP Namur, Belgium): ehe@elv.fundp.ac.be
P.Mousel (CRP-CU, Luxembourg): mousel@crpcu.lu
G.Thienpont (CRP-CU, Luxembourg): thienpon@crpcu.lu
M.Wéry (FUNDP Namur, Belgium): mwery@hermes.fundp.ac.be