Ceusters W, Lovis C, Rector A, Baud R. Natural language processing tools for the computerised patient record: present and future. In P. Waegemann (ed.) Toward an Electronic Health Record Europe, Proceedings, 1996;:294-300.

Natural Language Processing Tools for the Computerised Patient Record: Present and Future

W. Ceusters (1), C. Lovis (2), A. Rector (3), R. Baud (2)

(1) Office Line Engineering NV, Het Moorhof, Hazenakkerstraat 20, B-9520 Zonnegem, Belgium. Tel: +32 75 48 65 87, Fax: +32 53 62 95 55.
(2) Medicine Department,University Hospital of Geneva, Switzerland, Tel ++41-22-382-33-11, Fax+41-22-372-92-35
(3) Medical Informatics Group, Department of Computer Science, University of Manchester Manchester M13 9PL, England. Tel: +44-161-275-6133 FAX +44-161-275-6932

1. Introduction

Once upon a time - retired physicians sometimes tend to say the *good old days* - patient records - for so far used at all - were just there to serve as the practitioner's external memory. But gradually on, other purposes of the medical record where introduced. With the introduction of computer technology and certainly when the focus of *informatics* shifted to *telematics*, the requirements with respect to contents, structure and use have changed dramatically. Nowadays, the electronic patient record is recognised to be the cornerstone of the clinical practice as it permits the combination of information from different sources and provides the basis for diagnostic, therapeutic and management decisions. The more precise the information it contains, the more reliable the conclusions based upon these data will be.

Traditionally, medical record keeping is paper-based. Although its limitations have been described at various occasions, it is still recognised to be a well functioning working instrument in the hands of an experienced physician [Nygren & Henriksson 1992]. Nevertheless, automated medical record systems are supposed to outperform paper-based systems ... *by their unique potential to improve the care of both the individual patients and populations and, concurrently, to reduce waste through continuous quality improvement* [COMM 1991]. Moreover, modern multimedia technologies have the intrinsic possibilities to offer healthcare workers automated patient record systems that at the level of man-machine communication look very similar to paper-based systems, but at the same time may behave as active systems that remind, suggest and perhaps even seek out advice [Ceusters e.a. 93].

From the user's point of view, there are roughly two kinds of automated patient record systems: those in which users enter data in a very structured manner, following a well-defined classification or coding system, and those where physicians or other carers enter data in free text or near natural language.

Structured electronic healthcare records are build around structured datasets build up of administrative and medical data, a medical data-element being the smallest piece of knowledge healthcare providers can deal with: a temperature reading, the red-blood-cell count, a diagnosis, etc. As datasets conceptually are contextual aggregates of these basic elements, with highly complex interrelations, an almost unlimited number of such datasets can be created. Since many years, healthcare record systems based on such a design are thought to improve the availability, retrievability and reliability of the information they contain [Blum 1984], but their capability to store and process clinical data usually depends on "*Procrustean*" approaches to data definition and

encoding [Cristea & Mihaescu 1988]. They mostly are implemented by using controlled vocabularies, and as such, their acceptance is very limited.

In routine practice, information is collected by means of natural language, a format that is not directly suitable for computer processing. Many researchers argued however that it is necessary to conserve this approach because the description of biological variability requires the flexibility of natural language and it is generally desirable not to interfere with the traditional manner of medical recording [Wiederhold 1980]. *Free-text based healthcare record systems* have found their way to a vast majority of physicians, but the format in which the data they contain are expressed, is very difficult to use and certainly inefficient in terms of medical management, transfer of data between different systems, quality assurance and surveillance, etc. To use the information contained in free-text based electronic record systems effectively, natural language processing will become mandatory.

In this document, we present two natural-language based applications that are currently available: Lucid and Multi-TALE. They share the characteristic of being either purpose-dependent or language-dependent [Ceusters e.a. 94]. However, both systems currently exploit technology that is used to assist the development of a fully language- AND purpose independent framework that is presented at the end of this paper: GALEN.

2. LUCID: a semi-automatic encoding tool

2.1 Coding and classification

The rationale for clinical coding is to associate concepts with precise meanings to patient data. A *concept* is defined as a unit of thought independent of any particular language [ISO 1087], a definition we take for granted in order not to be lead into discussions which would bring us back to the time of Aristotle, and perhaps even earlier. Concepts with precise meanings are to be looked for in standardised concept systems. A *system of concept* is a structured set of concepts established according to the relations between them, each concept being determined by its position in the set [ISO 1087].

Concepts in medical concept systems are represented by a term, expression or rubric, and uniquely identified by means of a code. The concept of a medical procedure by which a surgeon removes a portion of the liver is designated in ICD-9-CM by the term *partial hepatectomy*, and identified by the code *50.22*. The codes are assigned to the concepts following a coding scheme. As a result, a coding system can be viewed as a specific type of concept system in which the concepts are identified by means of a code according to a coding scheme. A *classification* is a system of concepts where the relations between the concepts are generic in nature. Usually, this hierarchical relationship is expressed in the codes themselves:

E.g.: 50. operations on the liver

50.2 local excision or destruction of liver tissue or lesion
50.21 marsupialisation of lesion of the liver
50.22 partial hepatectomy
...
...
...

Coding patient data means that a physician (or professional encoder) has to describe the patient data by means of the concepts available in the coding system to be used. The requirements to be met in order to perform the coding task adequately are [Ceusters e.a. 1994b] : 1) a perfect understanding of the meaning of the patient data (the source concepts), 2) a perfect understanding of the meaning of the concepts available in the concept system (the target concepts), 3) at least a certain level of similarity and coherence between the source concepts and target concepts, 4) facilities to search the concept system for the target concept(s) that match(es) a given source concept as closely as possible.

2.2 NLP-assisted clinical coding with Lucid

LUCID is a system developed by LNAT Associates (Switzerland), to help physicians in the assignment of more accurate codes to diagnosis expressed in natural language. The LUCID-approach is based on recommendations found in the literature [Hersh e.a. 1992, Rossi-Mori e.a. 1992] as well as on the experience in coding of the University Hospital of Geneva [Frütiger 1989] and in natural language processing [Baud e.a. 1992]. The semi-automatic tool that helps French physicians to find the good ICD-9 codes uses a large corpus of narrative French expressions that have been taken from discharge letters. Versions for other languages (Dutch, English, German, …) are available, where similar language resources have been used. The system uses full lexical (grammatical and inflexional) descriptions of words, a corpus of synonyms, equivalencies between expressions in natural language as well as partial conceptual links between words to help physicians to get the good ICD-9CM codes. When the system encounters unknown words, an intra-words processing analysis is done to try to find the meaning of the word [Lovis e.a.1995a, 1995b]. The hierarchies actually supported in the French system include ICD9-CM, ICD-10, topology in SNOMED 2 and the Swiss VESKA classification of therapeutics. All three classifications can be accessed either by narrative language or by an easy-to-use graphical hierarchy browser. For ICD hierarchies, also all rules considering exclusions have been implemented.

3. Multi-TALE: a semantic tagger for neuro-surgical procedure reports

3.1 Objectives

Multi-TALE was a small-scale project that was partly funded by the Commission of the European Union in MLAP-II contract 93-04 and co-ordinated by Office Line Engineering NV. It resulted in two syntactic-semantic taggers for English and Dutch neurosurgery reports. The function of a tagger is to perform the first and essential pre-processing stage for any natural language processing application: the labelling of words with their grammatical (sub-)categories, only taking into account a limited context. In addition to this syntactic labelling, the Multi-TALE taggers also provide

semantic tags to words and word groups found in the texts. More specifically, semantic decoration is provided on the basis of the model for surgical procedures as described in CEN ENV1828:1995 [CEN ENV1828:1995]. Although the standard was developed for the description of elementary surgical procedure expressions as they can be found in classification and coding systems, one of the hypotheses of the Multi-TALE project was that the same structure could be used to represent the particular tasks described in full text reports [Ceusters e.a. 1996].

3.2 Architecture of the Multi-TALE system for English

The English Multi-TALE prototype is a modular system consisting of several programs which operate together under Microsoft's Word for Windows, version 6.0. A typical session with the Multi-TALE prototype starts by selecting in a word processor the portion of text to be processed. The sentences selected are temporarily stored in a file that becomes the input for the syntactic tagger DILEMMA [Paulussen & Martin 1992]. This is a rule-based parts of speech tagger developed for general language. We received DILEMMA as a closed system without the possibility to improve its performance. For this reason, an intermediate module was developed to correct certain systematic errors - independent of any context within the sentence being processed - produced by DILEMMA, and to allow the processing of multi-word units (e.g. *sella tursica*, *cerebellar artery*) as one entity. The output of this module is then passed to the syntactic-semantic tagger. This part of the system is developed on a language- and domain independent mode. The linguistic and semantic knowledge it exploits is stored in separate linguistic knowledge bases (LKB). A first LKB corrects context-dependent mistakes made by DILEMMA. A second LKB contains the semantic lexicon while a third one holds the grammar. Texts are processed by the semantic tagger sentence by sentence. The analysis of any given sentence proceeds in a deterministic way, except for lexicon lookup. When for instance in a sentence two words are found that each can have three meanings or alternative syntactic characteristics, 9 possible interpretations of the sentence are processed. The last module of the Multi-TALE prototype, uses heuristic techniques to point out in such case the most likely interpretation.

As an example, the sentence "*A big fatty tumour was rapidly removed from the brain and the hole filled with pieces of artificial tissue"*, is tagged by Multi-TALE in the following way:

```
SENTENCE 001, SOLUTION 001, SEGMENT 001: remove
do      path     detnoun 4        A big fatty tumour
action  remove   papa    11a      was removed
 -      velocity adv      -       rapidly
 -       -       prep     -       from
io      anat     detnoun 4        the brain

SENTENCE 001, SOLUTION 001, SEGMENT 002: fill
do      mcr      detnoun 4        the hole
action  install  past     -       filled
 -       -       prep     -       with
m       anat     adjnoun 16mcrOf  pieces of artificial tissue
```

3.3 Results

Two types of validation have been carried out: one to test the intermediate performance of the system with fine-tuning as primary objective, and a second one to assess the results of the final modifications. Ten surgical reports (138 sentences) from the corpus, five having been used for the development of the syntactic-semantic rule-

base (training sample), and five for which this is not the case (testing sample), have been manually validated. Two human experts (physicians) were used as gold standard. Recall (number of entities retrieved and relevant over number of relevant entities in the report) and precision (number of entities retrieved and relevant over number of entities retrieved (only for second validation)) were calculated separately for testing and training sample. Calculations were based on the correct recognition of syntactic entities such as sentences, segments (surface form of the predicates), clauses (surface form of predicate arguments), simple and complex noun phrases and verb phrases, as well as on semantic information (types and semantic links correctly identified). In total 2139 syntactic units (to be mapped into 6 categories) and 857 semantic-contextual entities (to be mapped into 8 categories) were to be retrieved.

Table 1 shows the results for both validations (* indicates test not performed). The first validation revealed an acceptable performance for syntactic tagging (except for complex noun phrases) and semantic type recognition, with only very modest results for case assignment. In addition, recall in the training sample appeared to be much higher than in the test sample. Fine-tuning of the system turned out to be very effective, and led to syntactic recall for the test sample of 95.7% (precision being 95.7% also) and 89.3% for semantic recall with a precision of 94.8%. Semantic labelling still appeared to be more successful than case-assignment (recall 93.4% versus 77.4%, 60.0% and 77.8%, $p < 0,01$).

	First Evaluation		Second Evaluation			
	Known	Unknown	Known		Unknown	
	Recall	Recall	Recall	Precision	Recall	Precision
Sentences	100	99	100	100	98.7	100
Segments	88	86	98.0	88.3	98.5	92.3
Clauses	91	80	91.9	92.6	95.2	93.3
Simple NP's	93	85	92.8	100	94.3	100
Complex NP's	79	56	88.6	93.9	86.7	100
VP's	93	85	92.6	98.9	95.2	100
Tot. Syntax	91	82	93.3	94.4	95.7	95.7
Deeds	*	*	96.5	94.3	98.1	97.1
Anatomy	*	*	93.9	95.8	98.4	98.4
Pathology	*	*	100	88.6	94.7	100
Instrument	*	*	87.5	97.2	71.1	96.4
Tot. Sem. Types	90	71	94.6	94.2	93.4	97.8
Action	97	82	96.5	90.1	97.1	93.5
Direct Object	84	69	83.3	88.7	77.4	92.9
Indirect Object	78	61	75.0	80.0	60.0	70.6
Means	68	50	70.6	100	77.8	93.3
Tot. Semantics	83	71	91.3	92.0	89.3	94.8

Table 1: Validation results of the Multi-TALE syntactic-semantic tagger.

4. The GALEN approach

4.1 The classical trade-off

"Traditional" natural language processing applications tend to suffer from a mayor drawback: the required complexity grows if they have to operate in a broad domain, if they are to be used in a multi-lingual environment, and if they are to be designed as

generic tools that can be used for various purposes. LUCID is an example of an application that can deal with multiple languages, but that is tied to a specific purpose: finding correspondence between two natural language sentences in order to provide clinical coding facilities. Independence from language is realised by applying only "low-level" linguistic techniques, while additional semantic information is manually pre-coded. The Multi-TALE tagger is less purpose-dependent, but as an application tightly connected to the sublanguage used in neuro-surgical procedure reports. The system can be used in other languages, and in other semantic domains, if additional grammars and lexicons are developed. The latter is a well-known problem in computational linguistics, where researchers and application developers have a constant demand for large, reusable lexicons. This is exactly what - from a computational linguistics point of view - GALEN tries to solve within the medical domain. As well as quite a lot more …

4.2 The world according to GALEN

GALEN bridges the gap between the detail required in clinical practice and the abstractions needed in retrieving and coding clinical information. It provides flexible usable interfaces for clinicians of all sorts which can be tailored to local needs.

GALEN delivers terminology as *dynamic services* rather than static data sets. Terminological services are provided to application programs via a standard interface by the GALEN Terminology Servers. The terminology servers insulate applications from terminological operations and guarantee that the operations are implemented consistently [Rector e.a. 1995, Nowlan e.a. 1994].

At the heart of GALEN there is a semantically valid *model of medical concepts and terminology*, represented in a *formal language* — a model of how simple medical concepts fit together and can be re-arranged, expanded or transformed plus the related information needed for user interfaces and data entry systems. Associated with the model there is supplementary information to support different *natural languages* and different *coding schemes*. Because GALEN delivers terminology as services, individual applications need not manipulate these models directly; rather they need only ask for the information from the server [Rector e.a. 1994]. To ensure that these new technologies provide valuable benefits, their development has been guided by a series of experiments and prototype demonstrators representing a range of clinical application areas.

Currently GALEN-tools are used in a number of c*linical user interfaces and medical record systems* which use the servers to produce efficient data entry and information retrieval systems for direct use by clinicians. *Classification managers* have been developed which provide new tools for developers of coding and classification systems based on GALEN's models and servers. In addition, there are *authoring tools* for creating and maintaining decision support systems.

Even more important than the practical applications that directly serve end-users of various nature, is the generic infrastructure that has been set up in order to support those applications. There are *terminology servers* to manipulate terminology, coding systems, and medical language using models of clinical concepts. The terminology servers relieve individual applications of the need to implement complex linguistic and terminological operations. The servers transform between different coding systems and database schemata and encapsulate complex variable length descriptions

into fixed length references suitable for storage in relational databases. Implementations are available for both client-server Unix environments — the TeS — and single user PC environments — the PUPPI. In addition, developers can rely on a *model of medical concepts and terminology* — the CORE model — that is used by the Terminology Servers and that is built in a formally sound language — the GRAIL Kernel — along with associated lexicons, mappings to existing coding systems, and other auxiliary information on which the servers base their services. And finally, *methodologies and tools* for the co-operative development and maintenance of the CORE model and associated linguistic and coding information have been put in place such that efforts of the past still pay off in the future.

4.3 GALEN at work

At least seven families of applications will in the future make use of the GALEN Terminology Server and CORE model:

- Medical records, clinical user interfaces and clinical information systems
- Natural language understanding and translation systems
- Clinical Decision Support Systems
- Management of, and conversion amongst, coding and classification schemes
- Bibliographic retrieval
- Information retrieval, intelligent querying, and epidemiological analysis
- Mediation between the semantic content of heterogeneous information systems

These applications will bring major benefits to clinicians and other health care professionals through the availability of more intuitive interfaces. Health service managers and planners will benefit from the more accurate cost effective information captured by clinicians in the course of patient care. Systems Developers, Integrators, and vendors will be able to bring on the market clinical applications that are easier to develop, maintain and integrate. Information providers and developers of decision support systems will more easily penetrate the market by being able to offer better linkage to medical records and to existing sources of information such as bibliographical resources.

5. Conclusion

Medical information systems are sufficiently large and varied such that no one vendor can expect to provide all of the systems needed in even a single hospital, let alone for the health service as a whole. Yet many of the benefits of information systems derive primarily from integration and sharing of information. A wide variety of specialist systems need to share information and common interfaces. Many of these varied systems would benefit from natural language interfaces and some, such as automatic linkage to abstracts of the literature, are even impractical without it. Generic multilingual solutions are required if the range of services to be built is to meet the demand. Furthermore, it is essential that the natural language processing components share the underlying concept structure used by the various applications.

Electronic patient record systems are no exception to this. A wealth of knowledge is needed to enter information in those systems consistently and to use the information afterwards for various purposes. Provided that a highly acceptable system can be designed in a specific environment, then developers surely will want to make it

available to other users. Whilst much re-use of system components is feasible within a given market segment, there are significant costs associated with the 'localisation' of systems to the needs of other markets. Perhaps the most important of these costs is the localisation to the linguistic needs of each national market in Europe.

Medicine is a descriptive, language intensive activity, and the costs of developing, and perhaps more importantly maintaining, the linguistic resources needed to localise clinical systems are clearly high. This presents a genuine barrier to the development of systems for use in Europe. Any practical approach to the management and exploitation of linguistic resources in large scale clinical information systems must be based on common methods and internal representations for linguistic information. This information must be reusable across a wide range of systems and local variants of those systems, and the cost of maintaining that information must be separable from those of maintaining the rest of the system.

It is in this spirit that GALEN is being developed. For sure, it is a slow process, and applications such as Lucid and Multi-TALE will keep their value for many years. But in the long run, only systems that are based on GALEN-technology will provide a secure and stable basis both for customers and suppliers.

6. References

[Baud e.a. 1992]

Baud RH, Rassinoux AM, Scherrer JR. Natural Language Processing and Semantical Representation of Medical Texts. Methods of Information in Medicine, 1992, 31: 2.

[Blum 1984]

Blum B.I. (Ed.): Information systems for patient care. New-York, Springer-Verlag, 1984.

[Ceusters e.a. 1993]

Ceusters W, Bonneu R, De Moor G, Lapeer R, Thienpont G. The Challenge of the Nineties: Bringing Multimedia Healthcare Records to Life. In: Reichert A, Sadan BA, Bengtsson S, Bryant J, Piccolo U (eds.), Proceedings of MIE'93. Freund Publishing House, England, 1993; 594-599.

[Ceusters e.a. 1994]

Ceusters W, Deville G, Buekens Ph. The Chimera of Purpose- and Language Independent Concept Systems in Health Care. In Barahona P, Veloso M, Bryant J (eds.) Proceedings of the XIIth International Congress of EFMI, 208-212, 1994.

[Ceusters e.a. 1994b]

Ceusters W, Mommaerts JL, Devlies J. Terminological Systems and Formalisms for Medical NLP-applications. ANTHEM Deliverable D4-1, 1994.

[Ceusters e.a. 1996]

Ceusters W, Deville G, De Moor G. Automated extraction of neurosurgical procedure expressions from full text reports: the Multi-TALE experience. In: Brender J. et al. (eds.) Medical Informatics Europe ’96, 154 - 158, IOS Press 1996.

[CEN ENV1828:1995]

CEN ENV 1828:1995. Medical Informatics - Structure for classification and coding of surgical procedures.

[COMM 1991]

Committee on Improving the Patient Record in Response to Increasing Functional Requirements and Technological Advances. The Computer-Based Patient Record: an Essential Technology for Health Care. Dick and Steen (eds), National Academy Press, Washington DC, 1991.

[Cristea & Mihaescu 1988]

Cristea D., Mihaescu T. Combining menus with natural language processing in recording medical data. Journal of Clinical Computing 1998; XVI,5-6, 156-166.

[Frütiger 1989]

P. Frütiger. Prospective encoding, transcoding and pragmatic representation of medical language. In : Computerised Natural Medical Language Processing for Knowledge Engineering (Ed. JR. Scherrer, RA. Côté, SH. Mandil) Elsevier, Amsterdam, 83-94, 1989

[Hersh e.a. 1992]

Hersh WR, Evans DA, Monarch IA, Lefferts RG, Handerson SK, Gorman PN. Indexing Effectiveness of linguistic and non-linguistic Approaches to automated Indexing. In : Proceedings MEDINFO 92 (Ed. Lun KC, Degoulet P, Piemme TE, Rienhoff O), North-Holland, Amsterdam, 1992, pp. 1402-1408

[ISO 1087]

International Standards Organisation. Terminology - Vocabulary (ISO - 1087), Geneva, ISO, 1987.

[Lovis e.a. 1995a]

Lovis C, Michel PA, Baud R, Scherrer JR. Use of a semi-automatic conceptual ICD-9 encoding system in an Hospital Environment. In P. Barahona et al. Eds, Lecture Notes in Artificial Intelligence, Springer-Verlag, 1995

[Lovis e.a. 1995b]

Lovis C, Michel PA, Baud R, Scherrer JR. Word Segmentation Processing : A way to exponentially extend medical dictionaries. In RA. Greenes & al. Eds, proceedings of MEDINFO 95, 1995

[Nowlan e.a. 1994]

Nowlan W, Rector A, Rush T, Solomon W. From Terminology to Terminology Services. 18th Annual Symposium on Computer Applications in Medical Care (SCAMC-94). Washington, DC: , 1994: 150-154.

[Nygren & Henriksson 1992]

Nygren E, Henriksson P. Reading the Medical Record, analysis of physicians' ways of reading the medical record. Computer Methods and Programs in Biomedicine, 1992;39: 1-12.

[Paulussen & Martin 1992]

Paulussen, H and Martin W. DILEMMA-2: a lemmatizer-tagger for medical abstracts, in Proceedings of the Third Conference on Applied Natural Language Processing (ACL), Trento, 141-146, (1992).

[Rector e.a. 1994]

Rector A, Gangemi A, Galeazzi E, Glowinski A, Rossi-Mori A. The GALEN CORE Model Schemata for Anatomy: Towards a re-usable application-independent model of medical concepts. In: Barahona P, Veloso M, Bryant J, (ed). Twelfth International Congress of the European Federation for Medical Informatics, MIE-94. Lisbon, Portugal: , 1994: 229-233.

[Rector e.a. 1995]

Rector A, Solomon W, Nowlan W, Rush T. A Terminology Server for Medical Language and Medical Information Systems. *Methods of Information in Medicine* 1995;34:147-157.

[Rossi-Mori e.a. 1992]

Rossi Mori A, Bernauer J, Pakarinen V, Rector AL, Robbè P, Ceusters W, Hurlen P, Ogonowski A, Olesen H. Models for Representation of Terminologies and Coding Systems in Medicine. In : Proceedings of 'Opportunities for European and US cooperation in standardization in Health care Informatics', 1992, Geneva

[Wiederhold 1980]

Wiederhold G. Databases in healthcare. Stanford University, Computer Science Department, Report No. STAN-CS-80-790, 1980.

Short Biography

Dr. W. Ceusters is general manager of Office Line Engineering NV, a small company specialised in Medical Language Engineering. He obtained degrees in neuropsychiatry, informatics and knowledge technology. He is involved in several European projects in the domain of medical natural language processing and together with the co-authors involved in the GALEN-project.