MEDINFO 98
B. Cesnik et al. (Eds)
Amsterdam: IOS Press

Medical Language Processing applied to extract clinical information from Dutch medical documents

Peter Spyns[a], Ngô Thanh Nhàn[b], Erik Baert[a], Naomi Sager[b], Georges De Moor[a]

[a]*Division of Medical Informatics, University Hospital Gent, Gent, Belgium*
[b]*Courant Institute of Mathematical Sciences, New York University, New York, USA*

Abstract

In this paper, we want to show how an existing morpho-syntactic analyser for Dutch (Dutch Medical Language Processor - DMLP) has been extended in order to produce output that is compatible with the language independent modules of the LSP-MLP system (Linguistic String Project - Medical Language Processor) of the New York University. The former can focus on idiosyncrasies for Dutch and take advantage of the language independent developments of the latter. This general strategy will be illustrated by a practical application, namely the extraction of clinical information from Dutch patient discharge summaries. Such an application can be of use for education, research and quality control purposes in a hospital environment.

Keywords
Medical Language Processing; Information Extraction

Introduction

At scientific congresses and in the various medical informatics journals, a lot of attention is being paid to medical language processing [1] and medical information extraction [2]. At the level of the International Medical Informatics Association, a specific working group (WG 6) has been set up to study natural-language processing and medical concept representation (next to the coding and classification of health data) [3: p.44]. Papers regarding natural language processing (NLP) presented at WG6 conferences are grouped in [4,5,6]. However, not many NLP-driven systems have actually been implemented. An overview of medical language processing systems and projects can be found in [7,8].

The following sections provide details about some aspects of NLP systems for medical English and Dutch, how information can be exchanged between them and stored in a database. Only some parts of the DMLP and LSP-MLP systems will be presented, namely those that are of importance for the experiment described below. The set-up of the test is explained and the outcomes are presented. Before a general conclusion, some ideas for discussion are provided.

Background

The Linguistic String Project - Medical Language Processor

The Linguistic String Project - Medical Language Processor (LSP-MLP) of the New York University is the first (and up until now the longest lasting) large scale project about NLP in Medicine [9,10,11]. The LSP-MLP aims at enabling physicians to extract and summarise sign-symptom information, drug dosage and response data, to identify possible side effects of medications and to highlight or flag data items [12]. In short, tasks commonly denominated by the term information extraction. Some years ago,'the LSP-MLP has also been ported to French and German, which illustrates the general applicability of its methodology and approach [13,14]. The reason for its generality lies in the use of a well defined underlying linguistic theory (String Grammar) [15], and a scientifically based sublanguage approach [16]. More recently, research focused on the automatic encoding into SNOMED codes [11] and on the extraction of information from discharge summaries for hospital management and clinical research [17,18]. With respect to the latter task, the system obtains a precision of 98.6% and a recall of 92.5% for the test samples [19]. The latest work includes the use of Standardised General Mark-up Language and World Wide Web Graphical User Interface technology to access and visualise better the requested information in a text (e.g. by highlighting words) [20].

The Dutch Medical Language Processor

The Dutch modules are typically language dependent and chiefly concern morphosyntactic analysis. A syntactic lexical database and a morphological recogniser for unknown words constitute the lexical analysis module [21]. The sentence analyser for Dutch uses Restriction Grammar (RG) as the underlying grammar formalism, which is the Prolog version of String Grammar. RG is a logic grammar formalism that combines context free rules with context sensitive information [22]. Currently, for a test sample of 35 cardiology reports consisting of 1652 sentences, a parse tree is delivered in 66% of the cases. However, not all relevant syntactic phenomena are handled (no conjunctions and complex verbal nodes) so that a corrected score probably lies around 80%. An exact account of the performance of the Dutch morphosyntactic components is given in

[23]. An exhaustive description of the DMLP can be found in [24].

The DMLP/LSP-MLP connection

The linguistic data (= analysed sentences) are passed on from the DMLP to the LSP-MLP system via syntactic parse trees. The linguistic information of the DMLP and the LSP-MLP systems correspond in a high degree. Semantic labels, which were originally not foreseen in the Dutch lexicon, are required for further LSP-MLP semantic processing and had to be added. In addition, the Dutch grammar has been reworked to be as compatible as possible with the LSP-MLP grammar. However, some differences remain (mainly concerning the style of the grammar rules and some particularities for Dutch). For those cases, an extra conversion routine that maps the DMLP grammatical categories to the corresponding LSP-MLP labels and slightly rearranges the linguistic structure of the parse tree is integrated in the module that actually reshapes the original tree into the new format. Eventually, nearly genuine Dutch LSP-MLP trees are delivered. On the side of the LSP-MLP, some new co-occurrence patterns had to be defined for the sublanguage selection module, but no changes to the transformation and regularisation modules had to be done [1], which illustrates again the generic nature of the LSP-approach and its implemented components. The feasibility of combining both NLP systems has already been successfully demonstrated by an application involving Web-technology [25].

The Textual Information Database

At the end of the LSP-MLP processing, the sentence meaning is captured by information formats [26] that represent the semantic regularities of the medical sublanguage [27]. The FORMAT5 contains information about the patient's state (see Figure 1 - LSP-MLP Information Format (pretty print) for the sentence i "the patient has stenosis"). The sublanguage specific labels are H-PT (related to the patient) and H-DIAG (a diagnosis) (see [11] for the complete list of labels and their meaning). The fifth field contains the format identifier, and the subject occupies the 16th position followed by the verb of the sentence. The diagnosis always comes on the 20th position.

|||| SENTENCE01 | FORMAT5 |||||||||| SUBJECT = DE (H-NULL) PATIENT (H-PT) | VERB = HEEFT (H-NULL) ||| DIAG = STENOSE (H-DIAG) |||||||||||| TEXTPLUS = |

Figure 1 - LSP-MLP Information Format (pretty print) for the sentence " the patient has stenosis"

Afterwards, these formats are stored in a relational database [28]. Its columns represent semantic information templates ("sublanguage information formats") (e.g. PT [patient], MED [medication], BODYPART, DIAG [diagnosis], etc.] while the rows (the sentences of a document) contain the normalised words ("strings") and sentence parts of the document.

Extraction of information from a document stored in a table of a textual knowledge database is done by means of SQL-queries. In contrast to systems with a semantic and/or pragmatic level of analysis (e.g. [29,30]), the central notions of the textual information database are the "string" (literal) and some conceptual labels [9]. An underlying semantic lattice with interrelated concepts does not exist. Practically speaking, the various ways to express the same notion (lexical and syntactical synonymy [2]) have to be explicitly addressed in order to capture all occurrences of that notion in the actual textual information database.

The content of the database is used for information processing tasks of various nature and goal. It concerns, amongst others, automated encoding of discharge summaries [11], determination of clinical patient profiles [17], health-care quality assurance [19,28] and queries of different kinds on a patient discharge summary textual information database [18]. In principle, these tasks become available for use on Dutch medical documents as well.

Material and Methods

The corpus of nine documents contains six cardiac surgery reports and three cardiology patient discharge summaries. The sentences are of a varying length and complexity. Some minor adjustments (e.g. removal of conjunctions) have been done manually to avoid failure during the sentence analysis. More details can be found in [23]. In total 100 sentences were analysed by the Dutch components, transformed into an LSP parse tree, processed by the LSP-MLP modules and finally stored in a textual information database.

Three queries that are relevant from a clinical point of view have been defined. For each surgical deed or diagnosis mentioned in a document, the concerned location of the body must be provided. It can be important to check whether a document provides the required level of detail (with respect to the body part) for each procedure or diagnosis (for quality control procedures). These are the first two queries (diagnosis: *Q1* and surgical deed: *Q2*). The second query may only consider surgical deeds that are performed during the current hospital stay. The third query (*Q3*) aims at retrieving the reason for admission from the documents. In many cases, the reason for admission is not retained as final diagnosis. From the scientific point of view, it is interesting to find out in which cases this happens. The retrieval results for *Q2* are (partly) displayed (see Table 1 - Results for Q2 (partially truncated table)).

These queries have been implemented in SQL and run on the textual information database. A medical doctor read the documents and provided manually the answers for the same queries. From the comparison of both sets of results, figures about the recall and precision (see Table2) have been calculated. However, these figures must be considered with the necessary caution because the extraction experiment is of a limited scale and elaboratedness (see [31,32,33,34] for reflections on and examples of extensive statistical evaluation of NLP systems in a clinical environment).

[1] The intermediary LSP-MLP components will not be discussed here due to space restrictions. A detailed account of the LSP-MLP processing chain is provided in [10,19].

[2] It must be pointed out that an intermediary module, namely the transformation component, of the LSP-MLP system specifically takes care of syntactic synonymy. Paraphrastic transformations reduce different constructions expressing the same idea to a single sentence.

Results

As a first subjective reaction, the collaborating doctor, who was up until then unaware of NLP and its potentialities, admitted to be positively impressed by the results. He had never thought that programs could be capable of achieving such results.

Medical Language Processing

A more objective measure is the number of database rows (= sentences) without any leftover strings in the TEXTPLUS field after joint DMLP/LSP-MLP processing compared to the number of sentences originally submitted for analysis. The TEXTPLUS field contains the quality assessment outcome. An empty TEXTPLUS field stands for a good analysis [19: p.149]. On a total of 100 sentences, 39 sentences have a non empty TEXTPLUS field, which means that some strings do not completely comply with the sublanguage information formats. The other strings of these sentences are put in the correct database fields, so that (some) information can be extracted. Of those 39 cases, 6 involve pronouns that have an H-NULL label (semantically void) and 5 other cases are past participles of general language words (also having the H-NULL label). As the words with an H-NULL do not participate in the "information process", these cases are not considered to be bad. Another 5 similar cases, with an "unimportant" string in the TEXTPLUS field, can be added so that finally only 23 sentences were not completely and correctly processed to allow reliable information extraction. Finally, for three sentences no analysis at all was provided so that no corresponding row in the textual knowledge database was included. A positive score of 74% for a limited and preliminary experiment can be considered as fairly good and promising.

Information Extraction

Several remarks can be made with respect to the answers to the queries that were applied to 97 of the originally 100 sentences. The 23 database rows mentioned above were processed in the regular way, entailing that the information in the TEXTPLUS field was not considered. The results are summarised in Table 2

Table 2 - information extraction score

query	recall (%)	precision (%)
Q1	85% (17/20)	94.45% (17/18)
Q2	93.75% (15/16)	83.34% (15/18)
Q3	60% (3/5)	25% (3/12)

The missing answers with respect to *Q1* are due to a wrong semantic label. Some words can be considered as a diagnosis (H-DIAG) and a disease indication (H-INDIC) depending on the clinical context. Here, a remedy could be to attribute multiple labels to more words and let LSP-MLP system determine which label applies [25]. For the second query (*Q2*), the recall is very good but the precision is a bit worse. The bad cases are surgical deeds, but from a previous (or future) hospital stay, so they should not be taken into account (rows 03 & 04 of Table 1 - Results for Q2 (partially truncated table)). A more detailed analysis taking the temporal indications into account could provide a solution. The last query (*Q3*) is the most difficult one and scores the worst. It specifically aims at the reason for the current admission. The missing sentences (2/5) are those for which no database row was provided after LSP-MLP processing. It is hoped that adjustments in the sublanguage selection module of the LSP-MLP can improve the recall percentage. The low precision score can be explained by the fact that much too many disease indicator words are retrieved. The main problem is that the relation of the disease indicator word with the admission is not clear (or lacking) in the text. This query clearly needs a conceptual approach since the admission can be expressed in various ways (admitted, admission, came for, was seen for, ...). It

Table 1 - Results for Q2 (partially truncated table)

nr	9: TTCHIR (surgery)	23: BODYPART, PT-PART (LOCALISATION)
03	MET EEN ANGIOPLASTIE	ANTEROSEPTAAL PROXIMAAL OP HET DEEL LINKER
04	gedilatreerde	linker VAN DE ANTERIOR DESCENDENS TER
05	BYPASSCHIRURGIE OP DE AFDELING HEELKUNDE	CORONAIRE
15	EEN GORETEX-GREFFE INGEPLANT	FEMOROPOPLITEALE LINKS
16	EEN DOTTERDILATATIE	LINKER VAN DE ARTERIA ILIACA
17	EEN CAROTISENDARECTOMIE	RECHTS
19	VOOR INGREEP	TER HOOGTE LINKER VAN DE CAROTIS
23	PLAATSTEN EEN BIOPROTHESE	IN DE MITRALISPOSITIE
24	PLAATSTEN OOK EEN BIOPROTHESE	IN DE AORTAPOSITIE
32	DE INGREEP PLAATSEN VAN EEN GREFFE	ILIACALE LINKS
53	OPERATIEVE PROCEDURE BYPASS	CORONAIRE
54	JUMP GRAFT	VENEUZE VAN DE AORTA NAAR DE DIAGONALIS,
55	JUMP GRAFT	VENEUZE VAN DE AORTA NAAR DE EERSTE
63	OPERATIEVE PROCEDURE BYPASS	CORONAIRE
64	RECONSTRUCTIE	LINKER OP DE LAD

would be too ad hoc to enumerate in the query all the possible strings expressing the admission.

Discussion

The results of the information extraction tests show that the effort to couple the DMLP with the LSP-MLP is certainly worthwhile. Furthermore, although the syntactic processor of the DMLP still has to be optimised, the information processing modules can already be partly and satisfactorily applied as medico-administrative utilities. However, the question has to be raised if the SQL code has not become too specific for these queries and this corpus, and whether the resulting figures would present substantial discrepancies if applied to a much larger number of sentences (or database rows). Care has to be taken to only define queries that are within the power of the system. A more elaborated and extensive test involving a larger number of documents is necessary to validate these preliminary results.

Although the LSP-MLP has proved to be very valuable for medical language processing, the applied methodology does present some drawbacks. The most important one being that the textual information database effectively contains (regularised) *strings* from the original document. The same document but translated gives raise to a different set of rows (see [13: p.558]), although its medical content is the same. In a "domain modelling approach", the original and its translation lead to the same (or equivalent) representation of the knowledge.

Therefore, we strongly believe that the LSP architecture must be enhanced with a semantico-pragmatic level [35]. But we equally strongly believe that, up until now, the LSP approach still is the best method to start with. Domain modelling is not yet sufficiently well developed to deliver results and applications that are comparable to the LSP-MLP achievements. Domain modelling can profit from the LSP-experience and techniques while the performance of the LSP-based applications will certainly improve if a language independent conceptual level is integrated.

Conclusion

The material presented in this paper shows, at least in our opinion, that a language specific (Dutch) front-end (DMLP) to a domain specific (medical) information processing back-end has proven to be a workable solution. As such, it is the only existing large scale NLP system for Dutch medical text analysis. The DMLP is not yet completely finished, but its components are sufficiently well developed to implement a prototype (in combination with other existing NLP systems) for large scale medical information processing applications that can attain fairly good results. Although on the (computational-) linguistic level, several improvements are still possible, the DMLP has, at least with respect to the linguistic knowledge, passed the critical threshold beneath which positive results can be attributed to the limited size of the knowledge bases and test samples. Larger test samples ensure the results to gain in weight and importance. The modular architecture and the object oriented design methodology favour the re-usability aspect. The paper also implicitly stresses the importance of collaboration between other research groups and re-usage of each other's results.

Practically speaking, in the future we would like to add extra semantic labels to all the Dutch dictionary entries [1], preferably in as automated a way as possible. This would allow to apply in a broad way some application programs already available for the English and French medical sublanguages (e.g., see [18,19,28]) to the medical Dutch as well. The feasibility of actually applying NLP in medicine in a clinical environment has already been convincingly proved [33,36]. The present work is to be seen as a step towards the same goal, but for Dutch.

References

[1] Baud R, Rassinoux AM, & Scherrer JR. Natural Language Processing and Semantical Representation of Medical Texts, *Meth. Inf. Med.* 1992: 31: 117 - 125.

[2] Hersh W. *Information Retrieval, a Health Care Perspective*, Springer-Verlag, New York, 1996.

[3] van Bemmel J, & McCray A, eds. *Yearbook of Medical Informatics*, Schattauer, 1996.

[4] Scherrer JR, Coté R, & Mandil S, eds. *Computerized Natural Medical Language Processing for Knowledge Representation*, North Holland, 1989.

[5] McCray A, Safran C, Chute C, & Scherrer JR. Natural Language and Medical Concept Representation, *Meth. Inf. Med.* 1995: 34 1/2 (special issue)

[6] Chute C, ed. *Preprints of the IMIA WG6 Conference on Natural Language and Medical Concept Representation*, Jacksonville, Florida, 1997.

[7] Friedman C, & Johnson S. Medical Text Processing: Past achievements, future directions. In: Ball M, & Collen M, eds. *Aspects of the Computer-based Patient Record*, Springer - Verlag, 1992; pp. 212 - 228

[8] Spyns P. Natural Processing in Medicine: An Overview, *Meth. of Inform. in Medicine* 1996: 35: 285 - 301

[9] Sager N. *Natural Language Information Processing: a computer grammar of English and its applications*, Addison-Wesley, Reading MA, 1981.

[10] Sager N, Friedman C, & Lyman M. *Medical Language Processing: Computer Management of Narrative Data*, Addison Wesley, Reading, MA, 1987.

[11] Sager N, Lyman M, Nhan NT, & Tick L. Medical Language Processing: Applications to Patient Data Representation and Automatic Encoding. In: [5] pp. 140 - 146

[12] Lyman M, Sager N, Friedman C, & Chi E. Computer-structured Narrative in Ambulatory Care: Its Use in Longitudinal Review of Clinical Data. In: *Proceedings of SCAMC 85*, 1985; pp. 82 - 86

[13] Nhàn NT, Sager N, Lyman M, Tick L, Borst F, & Su Y. A Medical Language Processor for Two Indo-European Lan-

[1] A smaller special purpose dictionary was used for the test.

guages. In: *Proc. of SCAMC 89*, 1989; pp. 554 - 558

[14] Oliver N. *A sublanguage based medical language processing system for German*, New York University [Ph.D. thesis], 1992

[15] Harris Z. *String Analysis of Sentence Structures*, Mouton, The Hague, 1962

[16] Grishman R, & Kittredge R. *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*, Lawrence Erlbaum Ass., Hillsdale, NJ, 1986.

[17] Borst F, Lyman M, Nhan NT, Tick L, Sager N, & Scherrer JR. TEXTINFO: A Tool for automatic Determination of Patient Clinical Profiles Using Text Analysis. In: *Proc. of SCAMC 91*, 1991; pp. 63 - 67

[18] Sager N, Lyman M, Nhàn NT, Tick L, Borst F, & Scherrer JR. Clinical knowledge bases from natural language patient documents. In: *Proceedings of MEDINFO 92*, 1992; pp. 1374 - 1381

[19] Sager N, Lyman M, Bucknall C, Nhàn NT, & Tick L. Natural Language Processing and the Representation of Clinical Data, *JAMIA*, 1994: 1: 142 - 160

[20] Sager N, Nhàn NT, Lyman M, & Tick L, (1996), Medical Language Processing with SGML display. In: *Proceedings of SCAMC 96*, 1996; pp. 547 - 551

[21] Spyns P, & De Moor G. A Dutch Medical Language Processor, *Int. J. of Bio-Med. Comp.* 1996: 41:181-205

[22] Hirschman L, & Dowding J. Restriction Grammar: a logic grammar. In: Saint-Dizier P, & Szpakowicz S. eds. *Logic and Logic Grammars for Language Processing*, Ellis Horwood, 1990; pp. 141 - 167

[23] Spyns P & De Moor G. A Dutch Medical Language Processor: Part II Evaluation, *International Journal of Medical Informatics*, (in revision)

[24] Spyns P. *Natural Language Processing in the Bio-medical Area: Design and Implementation of an Analyser for Dutch*, K.U. Leuven (Ph.D. thesis), 1996

[25] Spyns P, Nhàn NT, Baert E, Sager N, & De Moor G. Dutch Sublanguage Semantic Tagging combined with Mark-Up Technology. In: *Proc. of the 5th Conf. on Applied Natural Language Processing*, 1997; pp. 182 - 189

[26] Chi E, Lyman M, Sager N, Friedman C, & Macleod C. A Database of Computer-structured Narrative: Methods of computing complex relations, In: *Proceedings of SCAMC 85*, 1985; pp. 221 - 226

[27] Friedman C. Automatic Structuring of Sublanguage Information: Application to Medical Narrative. In: [16], pp. 85 - 102

[28] Hirschman L, Story G, Marsh E, Lyman M, & Sager N. An Experiment in Automated Health Care Evaluation from Narrative Medical Records, *Computers and Biomedical Research* 1981: 14: 447 - 463.

[29] Zweigenbaum P. MENELAS, Coding and Information Retrieval from Natural Language Patient Discharge Summaries. In: Laires M, Ladeira M, & Christensen J, eds. *Health in the New Communication Age*, IOS Press, Amsterdam, 1995 pp. 82 - 89

[30] Rassinoux AM, Juge C, Michel PA, Baud R, Lemaître D, Jean F, Degoulet P, & Scherrer JR. Analysis of Medical Jargon: the RECIT system. In: *Proceedings of AIME 95*, 1995; pp. 42 - 52

[31] Friedman C, & Hripcsak G. Evaluating Natural Language Processors in the Clinical Domain. In: [6]: pp. 41-52

[32] Zweigenbaum P, Bouaud J, Bachimont B, Charlet J, & Boisvieux JF. Evaluating a Normalized Conceptual Representation Produced from Natural Language Patient Discharge Summaries. In: *Proceedings of the 1997 AMIA Annual Fall Symposium*, pp. 590 - 594

[33] Hripcsak G, Friedman C, Alderson P, DuMouchel W, Johnson S, & Clayton P. Unlocking Clinical Data from Narrative Reports: A Study of Natural Language Processing, *Annals of Internal Medicine* 1995: 9: 681 - 688

[34] Gundersen M, Haug P, Pryor A, van Bree R, Koehler S, Bauer K, Clemons B. Development and Evaluation of a Computerized Admission Diagnoses Encoding System, *Computer and Biomedical Research*, 1996, 29: 351 - 372

[35] Hahn U, & Romacker M. Text Structures in Medical Text Processing: Empirical Evidence and a Text Understanding Prototype, In: *Proc. of the 1997 AMIA Annual Fall Symposium*, 1997; pp. 819-823

[36] Haug P, Christensen L, Gunderson M, Clemons B, Koehler S, & Bauer K. A Natural Language Parsing System for Encoding Admitting Diagnoses. In: *Proc. of the 1997 AMIA Annual Fall Symposium*, 1997; pp. 814 - 818

Address for correspondence

Division of Medical Informatics
University Hospital Gent
De Pintelaan 185 (5K3), B-9000
Gent
Belgium
e-mail: Peter.Spyns@rug.ac.be
URL: http://allserv.rug.ac.be/~pspyns