Automatic Encoding into SNOMED III: A Preliminary Investigation

Naomi Sager [1], Margaret Lyman [2], Ngô Thanh Nhàn [1], Leo J. Tick [2]

[1] Courant Institute of Mathematical Sciences, New York University, New York, NY 10012

[2] New York University Medical Center, New York, NY 10016

ABSTRACT

The Linguistic String Project (LSP) medical language processing (MLP) system converts narrative clinical reports into database tables of patient data. A procedure for mapping the output of the LSP MLP system into SNOMED III codes was developed. Preliminary results and further requirements are discussed.

INTRODUCTION

As part of the movement toward a computer-based patient record, CPR [1] (or electronic medical record system, EMRS [2]), increased attention is being paid to lexical issues: the development of the Unified Medical Language System (UMLS) [3]; the release of a much expanded version of SNOMED (SNOMED International, or SNOMED III) [4] and exploration of its use to represent clinical data [5]; new methods for maintaining controlled vocabularies [6]; integration of medical knowledge bases via an "interlingua" [7]; and the use of medical language processing (MLP) to extract clinical information from free-text patient documents [8, 9, 10], or to automatically index patient documents according to an established code [11, 12].

This paper describes initial work in using the LSP MLP system [10, 13] as the first component of an encoder that produces SNOMED III codes from free-text clinical documents. This activity seemed especially timely in light of AMIA's position paper of April 20, 1993 that named SNOMED III the "Preferred Code System" for Diagnoses (along with ICD9 which it includes); for Symptoms and Findings + Modifiers; for Anatomic location; and for Microbes and etiologies [14].

MANUAL ENCODING IN CPRI EXERCISES

To begin, we examined the results of manual SNOMED III coding that was performed for the Computer-based Patient Record Institute (CPRI) in 2 Exercises, each involving 10 case reports [15]. In the Exercises, Medical Concepts and their instantiations were manually determined from narrative case reports, which were then manually coded into a number of standard terminologies.

It was striking to observe that the medical semantic categories developed by LSP for processing clinical text were extremely similar to the Medical Concepts developed in the Exercises as a step toward manual coding. This is illustrated in Fig.1, where on the left is a list of the Medical Concepts with examples that occurred in the CPRI Exercise 1 data, and on the right the corresponding linguistic categories of the LSP MLP system. For brevity, Medical Concepts with less than 3 occurrences in the Exercise 1 data are not included in Fig. 1; the words fitted reasonably into other Medical Concepts used in the Exercises.

In the CPRI Exercises the path from patient document to code was via Medical Concepts:

Text → Medical Concepts → Coded Text

The similarity of LSP categories to Medical Concepts suggested a similar path for automatic encoding:

Text → LSP MLP System → Coded Text

To explore this possibility, an algorithm was devised (still quite preliminary) to map the output of the LSP MLP system into SNOMED III. We used texts from the CPRI Exercise 1 as input so that the results obtained by the LSP procedure could be compared with the SNOMED III codes that had been assigned manually to the texts as part of the Exercise.

MLP IN THE LSP SYSTEM

Figure 2 shows findings retrieved from the relational database table obtained from the MLP output for one case report (014A) in the CPRI Excercise 1 data. Each row contains one medical fact statement, with the words placed in semantically labeled fields. The SID field contains the sentence identifier. A sentence may contain more than one medical fact statement; the ROW field contains the row number of the medical fact statement within that sentence. The symbol # marks a "break point" between words in a given field that were separated in the sentence. Fields not queried

FIGURE 1 - MEDICAL CONCEPTS AND CORRESPONDING LINGUISTIC CATEGORIES

MEDICAL CONCEPTS (CPRI Exercises)		LINGUISTIC CATEGORIES (LSP dictionary classes, or LSP-computed if starred)	
Diagnosis	*cancer, hypertension, anemia*	DIAG	*adenoma, diabetes, tumor, malaria*
Symptoms	*nausea, diaphoresis, discomfort*	INDIC	*abnormal, cramping, warmth*
Qualitative	*soft, vague, asymmetric, standard*	DESCR	*acinar, inactive, shotty, red*
Qualitative	*tiny shotty, within normal limits*	NORMAL	*uneventful, well, rational*
Anatomy	*thigh, ovary, hop, acetabulum*	PTPART	*decidua, ileum, sacral, ramus*
Topology	*left, superficial, bilaterally*	PTAREA	*adjacent, edge, left, ventral*
Functional Status	*normally active, appetite, oral intake*	PTFUNC	*ambulate, eat, heal, pulse*
Severity	*excessive, severe, markedly*	AMT	*impressive, mild, rare, scant*
Stage	*stage 1C, benign, 2+*	QUANT*	
Grade	*grade 2-3, grade 2, benign*	QUANT*	
Extent	*Clark's level 2, 0/3 lymph nodes*	QUANT*	
Quantitative	*.84 mm, 1.5 cm long, 5-8, 2 units*	QUANT*	
Diagnostic Exam	*Holter monitor, tilt table test,*	TXPROC	*xray, biopsy, scan, ultrasound*
Laboratory Name	*CBC, creatinine, hemoglobin, TIBC*	TXVAR	*acetone, electrolyte, FEP, WBC*
Therapeutic Procedures	*open reduction, pinning, D&C*	TTCHIR	*shunt, wired, ablation, shortened*
Treatments	*darvocet, irradiation, ibuprofen,*	TTMED	*heparin, lasix, immunization*
Treatments	*splint, Pavlik harness*	TTCOMP	*bandage, cane, catheter, mask*
Disposition	*poor prognosis, observation*	TTGEN	*care, admission, workup, visit*
Negation	*negative, without evidence of*	NEG	*absence, never, not, unable*
Reliability	*question of, rule out, possible*	MODAL	*feel, hope, likely, maybe, claim*
Chronicity	*history of, chronic, recurrent*	TMPER	*briefly, persist, sustain, usual*
Time	*nocturnal, four years, recent,*	TMLOC	*childhood, last, earlier, hence*
Time	*present for 2 years duration*	TIME*	
Physical Finding**			
motion limited, positive pivot shifts, skin folds asymmetric, visceromegaly, healing of vaginal cuff, neck supple, bleeding at biopsy site		Physiological function or anatomic site examined + Result	
internal rotation, respirations, heart rate, BP, hip motion, abduction, external rotation, flexion, range of motion, extremities		Physiological function or anatomic site examined	
prominent, nodes, incision, lymphadenopathy, less developed		Result	

* The LSP system recognizes quantitative phrases syntactically, and similarly for time expressions involving quantities (*3 days post op*).

** All examples of Physical Finding are from CPRI data, classed here into 3 groups that correspond to the LSP representation as (1) subject and predicate of a Physical Finding statement type, (2) subject only, (3) predicate only.

in the retrieval of Fig. 2 include treatment, patient management, laboratory findings and neutral descriptors, e.g. RED in sentence 14A.1.08 (Fig. 3).

The language processing causes modifiers that apply to several findings to be "distributed" so that each of the findings carries the modifier (e.g. all the negated findings in Fig. 2). Rows number 6 and 7 of SID=014A.1.02 in Fig. 2 show the effect of an unresolved syntactic ambiguity: SIGMOID CANCER AND HYPERTENSION → SIGMOID CANCER AND [SIGMOID] HYPERTENSION, in the same way as JOINT SWELLING AND REDNESS → JOINT SWELLING AND [JOINT] REDNESS. A knowledge base that identifies body parts in relation to classes of diagnoses and symptoms would resolve the ambiguity in this case.

FIGURE 2 - RETRIEVAL OF POSITIVE AND NEGATIVE FINDINGS FROM CASE 014A SENTENCES 1-20

SID	ROW	SIGN-SYMPTOM OR DIAG	BODY PART OR BODY FUNCTION	QUANT	TIME WORDS
014A.1.01	1	PAINFUL # RASH #	RIGHT # FACIAL # INVOLVING THE EYE		
014A.1.02	3	CORONARY ARTERY DISEASE			A PAST MEDICAL HISTORY
014A.1.02	6	# CANCER #	SIGMOID		A PAST MEDICAL HISTORY
014A.1.02	7	# HYPERTENSION #	SIGMOID		A PAST MEDICAL HISTORY
014A.1.03	1	PAINFUL RASH	DEVELOPED		APPROXIMATELY 6 DAYS P
014A.1.03	2	BLISTERS	ON THE RIGHT SIDE OF HIS FOREHEAD		
014A.1.03	3	REDNESS	ON THE RIGHT SIDE OF HIS FOREHEAD		
014A.1.03	4	REDNESS	ON THE RIGHT SIDE OF HIS # SCALP #		
014A.1.03	5	BLISTERS	ON THE RIGHT SIDE OF HIS # SCALP #		
014A.1.05	2	PAIN			
014A.1.07	1	WORSENED			GRADUALLY # OVER APPRO
014A.1.08	1	WAS SWOLLEN	RIGHT EYE #		
014A.1.08	2	WAS SWOLLEN	SURROUNDING TISSUE #		
014A.1.08	7	DIFFICULTY	VISION		
014A.1.09	2	POOR	INTAKE		
014A.1.10	4	HERPES ZOSTER		V1 DISTRIBUTION	
014A.1.11A	3	CONJUNCTIVITIS		SOME	
014A.1.13	1	HAD WORSENED	EYE		SINCE THEN
014A.1.13	3	PAINFUL	TEARING	INCREASED # VERY #	
014A.1.13	4	SWELLING			
014A.1.16A	2	PAIN			
014A.1.19	2	BE DISORIENTED			
014A.1.19	4	FEBRILE			

SID	ROW	SIGN-SYMPTOM OR DIAG	BODY PART OR BODY FUNCTION	TIME WORDS	NEGATION
014A.1.11B	1	INVOLVEMENT	CORNEAL		NO
014A.1.17	1	ASSOCIATED HEADACHE			NO
014A.1.17	2	ASSOCIATED # LOSS OF CONSCIOUS			NO
014A.1.17	3	ASSOCIATED # INJURY #	HEAD		NO
014A.1.17	4	ASSOCIATED # PAIN #	CHEST		NO
014A.1.17	5	ASSOCIATED # SHORTNESS OF BREATH	CHEST		NO
014A.1.20	1	SHINGLES	EYE	HISTORY	NO
014A.1.20	2	SHINGLES	# TO THE # FACE #	HISTORY	NO
014A.1.20	3	# OTHER INJURIES #	EYE	HISTORY	NO
014A.1.20	4	# OTHER INJURIES #	# TO THE # FACE #	HISTORY	NO

AUTOMATIC ENCODING USING MLP

Figures 3, 4, 5 and 6 contain results of the automatic encoding for sentences 08, 10, 11 and 20 of the text of report 014A. Part A contains the CPRI Exercise 1 data. Part B shows the results of the LSP automatic encoding.

The input to the encoding algorithm consists of the words in a row of the relational database (excluding the TIME field for now). Four types of search are employed. "Full search" constructs a text string of all the words in the row in their order of occurrence in the sentence, and searches SNOMED strings for a complete or partial match. "Field search" does the same for each nonempty field in the row. "Text flow break search" does the same for the substrings within a field (strings set off by break marks "#"). Finally, as a last resort, "Single word search" treats every word individually.

The "best match" is the set of codes that covers the most words in the least number of codes (Wingert strategy [16]). However, by displaying all the non-null results of searches, as is done in Figs. 3-6, including those that cover the same word string by different codings (Fig. 3, Part B, Row 1), the automatic encoder also functions as a means of discovering alternative codings within SNOMED.

DISCUSSION OF PRELIMINARY RESULTS

Because the encoding algorithm is at a preliminary stage of development, no attempt has been made to compare automatic vs. manual codings quantitatively. A fully developed encoder will require auxiliary knowledge about word relations (synonomy, hyponomy, hypernomy), methods for treating inflectional differences (singular/plural, adjective/noun, etc.), analysis of compound medical words into semantic roots [17] and a representation of correspondences (*cardi* ↔ *heart*). This experiment used an ad hoc synonym list for all term equivalences. For some word relations, SNOMED itself, while the target of encoding, is also a tool. Synonomous variants are often included under the same code number. The text word SWOLLEN in Fig. 3 was manually coded to M-02570 Swelling, NOS, the preferred term (class 01). However, 5 other terms with the same numerical code are Bulge, Tumefaction, Tumescence, Swollen, Bulging, having other class numbers. The value of bringing together variants under one code number is seen in Figs. 4 and 6. In Fig. 4 (text sentence 10) we learn of the diagnosis Herpes Zoster (DE-32400), and ten sentences later the report of no history of Shingles (DE-32400).

An important feature of the method used is that each individual medical fact statement within a complex sentence is encoded as a unit. This is especially

FIGURE 3 - ENCODING OUTPUT, SENTENCE 8

014A.1.08 THE RIGHT EYE, PARTICULARLY THE SURROUNDING TISSUE, WAS SWOLLEN, RED, AND TEARING AND HE HAD DIFFICULTY WITH HIS VISION.

A. CPRI Exercise 1 Data
Medical Concept

<Symptoms>	swollen
<Symptoms>	red
<Anatomy>	eye
<Topology>	right
<Symptoms>	difficulty with vision
<Symptoms>	tearing

Manual SNOMED Coding from Medical Concepts

M-02570	01	Swelling, NOS
M-04040	01	Red color
T-AA000	01	Eye, NOS
G-A100	01	Right
DA-74900	02	Decreased vision, NOS
T-AA970	01	Tears

B. Automatic Encoding from MLP Output

Row 1:	T-AA010	01Right eye
	G-A100	01Right
	T-AA000	01Eye, NOS I C69.9
	M-02570	05Swollen
Row 2	G-A125	02Surrounding
	T-D0050	02Tissue, NOS
	M-02570	05Swollen
Row 4:	T-AA010	01Right eye
	T-AA970	01Tears
	G-A100	01Right
	T-AA000	01Eye, NOS I C69.9
Row 6:	T-AA970	01Tears
	G-A125	02Surrounding
	T-D0050	02Tissue, NOS
Row 7:	F-F0000	01Vision, NOS

---

important when a sentence contains both positive and negative findings (Fig. 5).

A final point regarding Fig. 4, where the difference between manual coding and automatic coding for the text *V1 distribution* is striking. Whereas the manual coder recognized that *V1 distribution* referred to the Ophthalmic division of the fifth cranial nerve, SNOMED code T-A8210, the automatic encoder matched the text *V1* with SNOMED V1 under "TNM classification of malignant tumors after operation". Further, the MLP system placed V1 DISTRIBUTION

FIGURE 4 - ENCODING OUTPUT, SENTENCE 10

014A.1.10 HE WAS EVALUATED AT THE EMERGENCY ROOM HERE AND FELT TO HAVE HERPES ZOSTER OF THE V1 DISTRIBUTION .

A. CPRI Exercise 1 Data
Medical Concept

<Diagnosis>	herpes zoster
<Anatomy>	V1 distribution

Manual SNOMED Coding from Medical Concepts

DE-32400	01	Herpes zoster, NOS
T-A8210	02	Ophthalmic division of fifth cranial nerve

B. Automatic Encoding from MLP Output

Row 4:	DE-32400	01Herpes zoster, NOS I (L-36401) I 053.9
	G-F231	01V1

FIGURE 5 -ENCODING OUTPUT - SENTENCE 11

014A.1.11 HE HAD BEEN SEEN BY OPHTHALMOLOGY AND THEY ALSO THOUGHT THAT THERE WAS SOME ASSOCIATED CONJUNCTIVITIS BUT NO CORNEAL INVOLVEMENT .

A. CPRI Exercise 1 Data
Medical Concept

<Diagnosis>	conjunctivitis
<Negation>	no
<Anatomy>	corneal involvement

Manual SNOMED Coding from Medical Concepts

DA-75605	01	Conjunctivitis, NOS
G-A201	01	Negative
G-A658	02	Not involving
T-AA200	01	Cornea, NOS

B. Automatic Encoding from MLP Output

Row 3:	DA-75605	01Conjunctivitis, NOS I 372.30
	G-A606	01Some
Row 1:	G-A657	02Involving
	T-AA200	01Cornea, NOS I C69.1
	G-A201	01Negative

---

in the QUANT, not BODYPART, field, which only shows that mathematicians and linguists should not be let loose on clinical vocabulary without proper medical supervision.

FIGURE 6 - ENCODING OUTPUT, SENTENCE 20

014A.1.20 THE PATIENT HAS NO HISTORY OF SHINGLES OR OTHER INJURIES TO THE EYE OR FACE.

A. CPRI Exercise 1 Data
Medical Concept

<Diagnosis> shingles
<Diagnosis> injuries
<Anatomy> eye
<Anatomy> face
<Negation> no

Manual SNOMED Coding from Medical Concepts

DE-32400	02	Shingles
M-10000	01	Injury, NOS
T-AA000	01	Eye, NOS
T-D1200	01	Face, NOS
G-A201	01	Negative

B. Automatic Encoding from MLP Output

Row 1: DE-32400 02Shingles | (L-36401) | 053.9
T-AA000 01Eye, NOS | C69.9
G-A201 01Negative
Row 2: DE-32400 02Shingles | (L-36401) | 053.9
T-D1200 01Face, NOS | C76.0
G-A201 01Negative
Row 3: T-AA000 01Eye, NOS | C69.9
G-A201 01Negative
G-A609 01Other
M-10000 01Injury, NOS | 959.-
Row 4: T-D1200 01Face, NOS | C76.0
G-A201 01Negative
G-A609 01Other
M-10000 01Injury, NOS | 959.-

References

[1] Dick RS, Steen EB, eds. *The Computer-Based Patient Record. An Essential Technology for Health Care*. Wash DC: Nat'l Acad. Press, 1991.

[2] RFA: Applied Research Relevant to an Electronic Medical Record. *NIH Guide for Grants and Contracts*, Vol 23:5, Feb 4, 1994, pp 5-8.

[3] Humphreys BL, Lindberg DAB. The UMLS Project:: a distributed experiment in improving access to biomedical information. North-Holland, Amsterdam: *MEDINFO 1992*;265-8.

[4] Coté RA, Rothwell DJ, Beckette R, Palotay J, eds. *SNOMED International*. Northfield, IL: College of American Pathologists, 1993.

[5] Campbell KE, Das AK, Musen MA. A logical foundation for representation of clinical data. *J Am Med Informatics Assoc.* 1994;1:218-232.

[6] Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge-based approaches to the maintenance of a large controlled medical terminology. *JAMIA.* 1994;1:35-50.

[7] Masarie FE, Miller RA, Bouhaddou O, Giuse NB, Warner HR. An Interlingua for electronic interchange of medical information: using frames to map between clinical vocabularies. *Comput Biomed Res* 1991;24:379-400.

[8] Zingmond D, Lenert LA. Monitoring free-text data using medical language processing. *Comput Biomed Res.* 1993;26:467-81.

[9] Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. *JAMIA.* 1994;1:161-74.

[10] Sager N, Lyman M, Bucknall C, Nhàn N, Tick LJ. Natural language processing and the representation of clinical data. *JAMIA.* 1994;1:142-60.

[11] Satomura Y, Do Amaral MB. Automated diagnostic indexing by natural language processing. *Med Inf (Lond)* 1992;17:149-63.

[12] Sager N, Lyman M, Nhàn NT, Tick LJ. Medical Language Processing: Applications to Patient Data Representation and Automatic Encoding. *Meth Inform Med*, to be published.

[13] Sager N, Friedman C, Lyman MS. *Medical Language Processing: Computer Management of Narrative Data.* Addison-Wesley, 1987.

[14] Board of Directors of the American Medical Informatics Association. Position Paper: Standards for Medical Identifiers, Codes and Messages Needed to Create an Efficient Computer-Stored Medical Record. *J Am Med Informatics Assoc.* 1994;1:1-7.

[15] Case histories and concept identification obtained from Computer-based Patient Record Institute (CPRI) exercises presented at *Annual Symposium on Computer Applications in Medical Care*, Wash DC, Nov 1993.

[16] Wingert F. An Indexing System for SNOMED. *Meth Inform Med.* 1986;25:22-30.

[17] Wingert F. Morphologic Analysis of Compound Words. *Meth Inform Med.* 1985; 24:155-62.