Object-Oriented Controlled-Vocabulary Translator Using TRANSOFT+HyperPAD

G. William Moore, M.D., Ph.D.
Jules J. Berman, Ph.D., M.D.

Laboratory Service, Veterans Affairs Medical Center, Baltimore, Maryland; Department of Pathology, University of Maryland School of Medicine, Baltimore, Maryland; and Department of Pathology, The Johns Hopkins Medical Institutions, Baltimore,Maryland.

Abstract

Automated coding of surgical pathology reports is demonstrated. This public-domain translation software operates on surgical pathology files, extracting diagnoses and assigning codes in a controlled medical vocabulary, such as SNOMED. Context-sensitive translation algorithms are employed, and syntactically correct diagnostic items are produced that are matched with controlled vocabulary. English-language surgical pathology reports, accessioned over one year at the Baltimore Veterans Affairs Medical Center, were translated. With an interface to a larger hospital information system, all natural language pathology reports are automatically rendered as topography and morphology codes. This translator frees the pathologist from the time-intensive task of personally coding each report, and may be used to flag certain diagnostic categories that require specific quality assurance actions.

Introduction

Coding diagnoses with SNOMED numbers is one of the most time-consuming computer-related activities performed by pathologists [1]. One must review each specimen's diagnosis on each report, find a SNOMED diagnosis that is a close match for the diagnosis, as well as related diagnostic terms that apply to a specific diagnosis. For instance, a squamous carcinoma of skin might have an *in situ* component (coded as squamous cell carcinoma, *in situ*), an invasive component (coded as squamous cell carcinoma, keratinizing), and a yeast superinfection (coded as *Candida*, or as candidiasis). It may arise in the larynx ('larynx' topography), and it may have a metastasis to a neck lymph node (topography 'neck' or 'lymph node' or 'cervical'). The metastatic lesion would also require an additional morphology code (squamous cell carcinoma, metastatic). It is easy to see how a single diagnostic report can give rise to a rather large list of coded items. Currently, many anatomic pathology information systems provide a code when the pathologist inputs a diagnostic entity from the controlled vocabulary list. This saves the pathologist from looking up each code number, but does not eliminate the need to manually extract diagnostic entities from reports.

The professional time allocated to coding reports can easily require one or more hours each day, and the benefits of coding (dependable retrieval of reports) rests largely on the care expended in the coding process. Friedman has shown that the retrieval of reports by SNOMED code may be less efficient than often assumed [2].

Automatic coding is now available in several commercial systems [3]. The efficacy of the few available automatic coding software products has not been subjected to scientific scrutiny and validation. It is likely that instances of unsatisfactory searches on SNOMED-coded reports result largely from a reluctance or inability of human coders to adequately encode pathology reports [4,5].

TRANSOFT is a table-driven, public-domain computer translation shell [6,7], using the file structure of the Veterans Affairs (VA) File Manager (FileMan), the core database management system used in 169 Veterans Affairs Medical Centers [8]. The user supplies the dictionary and a grammar in the augmented-transition-network (ATN) style common to many computer translators [9-11]. This demonstration is an object-oriented version of the TRANSOFT computer-translation-shell, applied to one year of English-language surgical pathology reports collected at the Baltimore VA Medical Center, using HyperPAD ((C) 1989, Brightbill-Roberts & Co., Ltd. [12]) as the user interface, into a controlled medical vocabulary as the target language. Words or phrases in the edited reports are pointed to topography and morphology terms [13]. TRANSOFT was used to rearrange the natural language word order used in pathology reports into a standard format for target terms.

0195-4210/91/$5.00

HyperPAD

HyperPAD is an object-oriented scripting environment for the IBM PC or compatibles, akin to HyperCard for the Macintosh computer [14-17]. The public-domain edition of TRANSOFT is supported by a run-time-only version of HyperPAD, called BROWSER, which gives the user access to translation and word processing functions, but does not allow reprogramming. HyperPAD objects consist of buttons, fields, pages, backgrounds, and pads. Pages are limited to 30 kilobytes apiece, but the number of pages is limited only by available disk storage. Operations on data contained in fields are controlled by 'scripts' (programs) written in the PADtalk language. Input surgical pathology reports reside in the 'input' field; the English-to-target lexicon resides in the 'lexicon' field; and the grammatical parser (word rearrangement formulas) resides in the 'grammar' field.

The conversion of TRANSOFT from MUMPS/FileMan was straightforward. Every string manipulation, database operation, and DOS interface could be substituted with corresponding HyperPAD operations. However, for large-scale sorting procedures, MUMPS was significantly faster than HyperPAD. Furthermore, HyperPAD requires some planning in order to partition a large lexicon or grammar into pages size 30 kilobytes or less.

TRANSOFT

TRANSOFT is a computer translation shell for translating medical statements into formal codes. Defining features of the source language (English) and target language (topography and morphology terms) reside in databases, not in program code. By contrast, commercially available translation programs are 'turnkey' or 'hands off' systems, where the user is dependent on the vendor to install modifications [3]. Part-of-speech designators for TRANSOFT include a combination of punctuation, traditional parts-of-speech, and medical axes as follows:

[=start-sentence
]=end-sentence
,=comma
T=topography
M=morphology
A=adjective
C=conjunction
N=noun
R=preposition
U=undetermined
Y=negation

A single pathology report forms a 'TRANSOFT paragraph', in which it is expected that every medically significant ambiguity can be resolved from the contents of that paragraph. Translations are obtained by pointing English words or phrases in the lexicon to the corresponding target term, and rearranging word order according to a 'parsing formula', as for example:

```
[ prostatic   adenocarcinoma .
[      T          M         ]
1      2          3         4
```

In this case, we have a four-word paragraph. The first word is '[', the start-sentence marker. The second word is 'prostatic', with part-of-speech T. The third word is 'adenocarcinoma', with part-of-speech M. The fourth word is '.', with part-of-speech], the end-sentence marker. The part-of-speech pattern for the entire paragraph is '[TM]'. In the grammar field, [TM] points to the parsing formula '1[2t3m4]', which means: put [(=start-sentence) into target position 1; put t (=topography code) into target position 2; put m (=morphology code) into target position 3; put] (=end-sentence) into target position 4. The final translation is: [PROSTATE ADENOCARCINOMA].

In a more complex example:

```
[ adenocarcinoma of prostate with stromal-hyperplasia.
[         M     R  T     C              M          ]
1         2     3  4     5              6          7
```

This is an seven-word paragraph, with '[' as the first part-of-speech, etc. The part-of-speech pattern for the entire paragraph is '[MRTCM]'. The parsing formula for this translation is '1+5[3m0r2+6t0c7m4+8]', which means: put [(=start-sentence) into target positions 1 and 5; put the first m (=adenocarcinoma) into target position 3; put r (=preposition) into target position 0, i.e., delete; put t (=topography code) into target positions 2 and 6; put c (=conjunction) into target position 0, i.e., delete; put the second m (=stromal-hyperplasia) into target position 7; put] (=end-sentence) into target positions 4 and 8. The final translation is: [PROSTATE ADENOCARCINOMA][PROSTATE STROMAL-HYPERPLASIA].

Discussion

Despite the enormous expense of medical record keeping, the textual part of medical records is not routinely translated into controlled vocabularies for quality assurance (QA) reviews. Commercial translators perform poorly with long sentences, or do not produce

standardized outputs that are portable to existing hospital information systems [3]. TRANSOFT is a public-domain translator, portable to MS-DOS-based or UNIX-based microcomputers, as well as to a range of minicomputer and mainframe operating systems. TRANSOFT uses the file structure of the VA hospital information system, with the largest userbase worldwide.

HyperPAD provides an intuitive user interface. All files (input text, lexicon, grammar) may be imported into or exported from HyperPAD as plain ASCII files. Object-oriented script languages have been dubbed the 'BASIC of the 1990s', since it is relatively easy to write serviceable programs for many common applications [16]. Compared to BASIC, scripting languages execute more slowly, but this disadvantage is more than compensated by their rapid prototyping capabilities. The input, lexicon, and grammar windows provided by HyperPAD are more than just attractive attachments; they help the developer and user understand how the data tables interact to obtain the final translation.

Hall and Lemoine describe their experience in examining 2500 consecutive SNOMED-coded reports at each of two London hospitals [4]. They found error rates of 10% and 16%. Other authors have obtained similar results [5]. Among the erroneous codes in the Hall and Lemoine study, about three-quarters were 'irretrievable', i.e., errors 'which would preclude the case being found by a reasonably diligent researcher in the future'. The authors conclude that 'many of the errors seem to be due to laziness in coding, with failure to consult the appropriate manual and reliance on memory for common codes.... It was apparent that many staff...failed to appreciate the implications of poor or incomplete coding, and considered it to be a burdensome task.'

The most important reason for recovering records coded in controlled medical vocabularies is QA [15]. In our laboratory, QA consists of examining the sequence of events in each patient's medical history and flagging exceptions, say, a patient with a suspicious biopsy and no followup. In order for QA to be cost-effective, computer translation must be fully automated and have a seamless interface from routine reports into the larger information system. Furthermore, the false-negative error rate must be low, or else rare events, such as QA-exceptions, may not be reliably detected. TRANSOFT/HyperPAD forms an environment in which natural language medical documents are translated into controlled medical vocabularies, and can serve in the recovery of primary medical records in hospital information systems. This technology can lead to better quality assurance in routine medical practice.

References

[1] Dudrey ET, Watts MT: A dBase III surgical pathology reporting and encoding microcomputer system. Am J Clin Pathol 93:91-97, 1990.

[2] Friedman BA: The impact of new features of laboratory information systems on quality assurance in anatomic pathology. Arch Pathol Lab Med 112:1189-1191, 1988.

[3] Weilert M, Aller RD, Pasia OG: System selection learned the hard way. CAP Today 5:38-49, 1991.

[4] Hall PA, Lemoine NR: Comparison of manual data coding errors in two hospitals. J Clin Pathol 39:622-626, 1986.

[5] Enlander D: Computer data processing of medical diagnoses in pathology. Am J Clin Pathol 63:538-544, 1975.

[6] Moore GW, Riede UN, Polacsek RA, Miller RE, and Hutchins GM: Automated translation of German to English medical text. Am J Med 81:103-111, 1986.

[7] Moore GW, Wakai I, Satomura Y, and Giere W: TRANSOFT: Medical translation expert system. Artif Intell Med 1:149-157, 1989.

[8] Davis RG: FileMan: A User Manual. Bethesda, MD: National Association of VA Physicians, 1987.

[9] Woods W: Transition network grammars for natural language analysis. Commun Assn Comp Mach 13:591-606, 1970.

[10] Vasconcellos M and Leon M: SPANAM and ENGSPAN: Machine translation at the Pan American Health Organization. Comput Linguist 11:122-136, 1985.

[11] Hutchins WJ: Machine Translation: Past, Present, Future. Chichester: Ellis Horwood Ltd, 1986.

[12] Brightbill Roberts & Co., Ltd., 120 E. Washington Street, Suite 421, Syracuse, NY 13202, TEL: 1-315-474-3400.

[13] Wingert F, Rothwell D, and Cote R: Automated Indexing into SNOMED and ICD. In, Scherrer JR, Cote RA, and Mandil SH (eds.), Computerized Natural Medical Language Processing for Knowledge Representation. Amsterdam: Elsevier Science Publishers B.V. pp. 201-239.

[14] Winkler D and Kamin S: Hypertalk 2.0: The Book. Bantam Books, New York, 1990.

[15] Berman JJ: Solving Quality Assurance Problems with Object Scripting Languages. Artif Intell Med 3:161-171, 1991.

[16] Wood L: Script Languages. The BASIC of the 1990s? BYTE 16(8):244-250, 1991.

[17] Allen D, Crabb D, Loeb L, Malloy R, Nance B, Rosh W, Sheldon K, Wagner P: What is a programming language? Scripting tools are real programming languages. BYTE 16(8):103-104, 1991.