Identification of Anatomical Terminology in Medical Text

Charles A. Sneiderman, MD, PhD
Thomas C. Rindflesch, PhD
Carol A. Bean, PhD
National Library of Medicine, Bethesda, MD 20894

We report on an experiment to use the natural language processing tools being developed in the SPECIALIST™ system to accurately identify terminology associated with the coronary arteries as expressed in coronary catheterization reports. The ultimate goal is to map from any anatomically-oriented medical text to online images, using the UMLS® as an intermediate knowledge source. We describe some of the problems encountered when processing coronary artery terminology and report on the results of a formative evaluation of a tool for addressing these problems.

INTRODUCTION

The SPECIALIST system [1] at the National Library of Medicine provides a context for research in natural language processing aimed at exploiting the Unified Medical Language System® (UMLS) [2] for improved access to biomedical information. The resources under development include the SPECIALIST Lexicon and associated lexical access programs [3] as well as a parser and a method for mapping free text to concepts in the Metathesaurus® (MetaMap [4]). Semantic interpretation, based on the Semantic Network [5], determines the relationships which obtain between the Metathesaurus concepts asserted in text. To this point the focus of research has been text-based information; development of the Visible Human image database [6] has led us to investigate the potential of using these tools for providing improved access to biomedical images as well. This involves identifying anatomical concepts and relations in biomedical text, such as the scientific literature, textbooks, and medical records, and then mapping these concepts, first to an anatomical domain model and then to anatomical images.

We see the principal challenges inherent in processing anatomically-oriented text as being two-fold: a) accurate identification of anatomical terminology and then b) determining the relationships which obtain between such concepts as asserted in the text being processed [7]. In this paper we concentrate on accurate identification of anatomical terminology in text. As a pilot project we are addressing terminology associated with the coronary arteries as found in cardiac catheterization reports.

BACKGROUND

Anatomical knowledge base

The UMLS knowledge sources serve as the domain model on which we base our processing. Of particular importance in this regard is the anatomical knowledge being added to UMLS by the University of Washington Digital Anatomist (UWDA) project [8]. The goal is to compile a fully instantiated knowledge source which integrates all concepts of human anatomy into UMLS; current coverage includes 15,345 terms representing some 9,094 concepts pertaining to the thorax. Terms representing anatomical concepts in the model attempt to reflect synonymous usage across disciplines and geographic areas. The large number of terms pertaining to the thorax alone results from a commitment to represent each discrete, visible anatomical entity explicitly as a unique concept.

Document structure

We have begun initial studies with reports of coronary angiography (stripped of patient identification) from the Johns Hopkins Cardiac Catheterization Laboratory. We would like to take advantage of the structure of these documents to help resolve certain types of indeterminacies encountered while subjecting the text to natural language processing techniques.

Exploiting the structure of documents as a guide to determining their meaning depends on textual cues which allow identification of document sections. Headings play a prominent role in this regard and can often be identified as a portion of text beginning at the left margin (and typically following a blank line). Further signature characteristics are terminal punctuation (colon, period, dash) along with lack of a finite verb. Formatting evidence includes all letters in upper case or all content words with initial upper case.

Of particular interest in processing cardiac catheterization reports are the headings which introduce sections describing particular vessels. The Johns Hopkins

reports, specifically, are divided into sections which include indication, cardiac catheterization procedure, hemodynamics, coronary arteriography (by major vessel), left ventriculography, and impression. For example, one of the reports in our study has headings indicating a major section and subsections as shown below.

(1) CORONARY ARTERIOGRAPHY:
Left Main Coronary Artery (LM):
Left Anterior Descending Coronary Artery (LAD):
Left Circumflex Coronary Artery (LCX):
Right Coronary Artery (RCA):
Left Internal Mammary Artery To The LAD:
Saphenous Vein Graft To The Circumflex Marginal:

Processing coronary artery terminology

Although research is being directed at developing general natural language techniques applicable to any medical text [9], most applications currently focus on specific areas (e.g. [10-12]). We have directed our efforts to the particular problems posed by coronary artery terminology. It is hoped that many of the specific rules being developed for the coronary arteries will generalize to the terminology for other branching structures, such as veins, nerves, and airways. The general techniques used are expected to apply to all anatomical terms.

Considered as strings, names for coronary arteries display a certain amount of complexity, as indicated in the following examples.

(2) Anterior interventricular coronary artery
Left sinuatrial nodal artery
Third posterior ventricular branch of circumflex coronary artery
Large posterior septal branch of posterior interventricular artery
Obtuse marginal branch of circumflex branch of left coronary artery

However, linguistic processing can considerably reduce this complexity. All terms for arteries can be considered to be simple or complex depending on how many simple noun phrases they contain. "Obtuse marginal branch of circumflex branch of left coronary artery," for example, which contains three simple noun phrases, is complex, while "Left sinuatrial nodal artery" is simple. The general rule for the identification of simple terms for arteries is that (in certain semantic contexts) if the head of a noun phrase (the right-most noun in the phrase) belongs to a certain small set of nouns (e.g. artery, branch, ramus, vessel) then that noun phrase is a simple term referring to an artery. Any number of simple artery terms contiguously joined by prepositions constitutes a complex artery term.

Even though the UWDA vocabulary has broad coverage, there are still a number of differences between the anatomical terms found in clinical reports and the anatomy vocabulary in the UMLS Metathesaurus. Some of these differences will simply have to be listed, for example, "Main nodal artery" for "Crista terminalis branch of sinoatrial nodal artery." However, the relationship between "Left circumflex coronary artery" and "Circumflex branch of left coronary artery" is more systematic and can be handled with general procedures currently being devised.

Abbreviatory devices and underspecified terms

The use of abbreviatory devices with regard to names for coronary arteries is seen in examples such as "Conus branch" as a synonym of "Conus branch of right coronary artery," or in "Posterior descending artery" as a synonym for "Posterior descending branch of right coronary artery." Synonymy due to abbreviatory devices is not always listed in the Metathesaurus, and thus its identification by natural language processing techniques is important for accurate identification of anatomical terminology.

Shortened expressions, such as those given above, fall into essentially two patterns, one resulting from truncation and the other from gapping. The first example given, "Conus branch," is the result of truncating all but the first simple term in a complex term ("Conus branch of right coronary artery"). "Posterior descending artery" is similar to its longer synonym, but the shorter term has an internal gap which is filled by the string *branch of right coronary* in the full expression. Several variations on the gapping paradigm are seen in variant names for coronary arteries as illustrated in (3) through (5). In each instance, the (a) and (b) examples are synonymous.

(3) a. Anterior interventricular branch of left coronary artery
b. Anterior interventricular coronary artery

(4) a. Diagonal branch of left coronary artery
b. Left diagonal artery

(5) a. Atrioventricular nodal branch of circumflex coronary artery
b. Artery of atrioventricular node

(3b) differs from the paradigm example for gapped abbreviations given above in that only *branch of left* has been omitted. (4b) is missing an internal gap in

comparison to (4a); however, in addition, the modifier *left* has been shifted to the front of the term. In the transition from (5a) to (5b) two strings *branch* and *circumflex coronary* have been removed, and the head of the term, *artery* has been moved to the front.

Regardless of the type of gapping used to create a shorter synonymous variant from a full expression referring to a coronary artery, such variants are generally listed in the UWDA Symbolic Knowledge Base (and thus in UMLS). All the examples involving gapping listed above appear as synonyms in UMLS. However, abbreviated variants involving truncation are not generally listed, and thus it is important to process them with the techniques being developed in order to accurately identify them as coronary artery terms.

Abbreviated variants such as those just described are a type of underspecified expression in comparison to the full term. Such underspecificity underlies ambiguity which must be resolved in a particular textual context. For example, "Atrial branch" can be an abbreviated variant of either "Anterior atrial branch of right coronary artery" or "Atrial branch of circumflex coronary artery." If textbooks or clinical records employ the shorter, ambiguous variant, they typically do so in the context of a discussion which is clearly marked as describing the characteristics of either the right or left coronary artery.

Another type of underspecified (and hence ambiguous) term is seen in the following example from a cardiac catheterization report. *It gives rise to a moderate to large sized first marginal branch...* Taken out of context, the term *first marginal branch* is ambiguous in that it might refer to a branch of either the left or right coronary artery. However, this example occurs in a section of the report clearly marked with the heading "Left Circumflex Coronary Artery (LCX)," making it clear that the left marginal branch is intended. Similar to these truncated expressions are single words such as *vessel*, *collateral*, *branch*, and *marginal*, which are highly ambiguous when considered in isolation. This ambiguity is also susceptible to resolution given the context in which these terms appear. Finally, pronominal anaphora (*it*, for example, in the sentence above) can be thought of as a type of underspecificity which becomes tractable given a specific context.

Syntactic processing and semantic interpretation

As noted earlier, effectively extracting information from anatomically oriented text in order to provide access to biomedical images depends on two phases of processing: a) syntactic processing for accurate identification of anatomical terminology and b) subsequent semantic interpretation for determining the relationships which exist between the concepts identified in the earlier phase. The research being conducted here can be viewed as focusing on the syntactic processing which supports accurate identification of coronary artery terminology. Complementary research [7] concentrates on interpreting the particular semantic relations based on this terminology as expressed in coronary catheterization reports.

Problems of synonymy, abbreviatory expressions, ambiguity, and underspecificity must be addressed before noun phrases in text can be successfully mapped to concepts in UMLS, and a final semantic interpretation is effected. Given the importance of syntactic processing in this regard we conducted a formative evaluation to test the effectiveness of the techniques being developed.

METHODS

Of the catheterization reports received from Johns Hopkins, fifteen describe diagnostic angiography with a section typically containing at least two coronary artery terms. In order to provide a standard against which to test our methodology, the physician author marked in the diagnostic reports any words or phrases referring to coronary arterial structures as coronary entities (CE). Implied references to coronary structures, such as *the vessel*, *a branch*, and *no left collaterals*, were also marked. (6) is an example of a sentence from a report with coronary artery terms marked.

(6) The ||CE|left anterior descending artery|CE|| has a 50-70% stenosis in the ||CE|proximal section|CE|| prior to the takeoff of any of the ||CE|diagonals|CE||.

We then drew on the resources of the SPECIALIST system to provide a syntactic analysis which could serve as the basis for the focused processing pertinent to the particular problems inherent in coronary artery terminology. An underspecified syntactic structure, based on information in the Lexicon and a stochastic tagger [13] for resolution of part-of-speech ambiguities, was produced for each sentence in our test set of cardiac catheterization reports, as in the following example.

(7) a. The left anterior descending artery arises from the left main coronary artery and gives rise to two moderate sized diagonal vessels.

b. [[det(the), mod(left), mod(anterior), mod(descending), head(artery)

[verb(arise)]
[prep(from), det(the), mod(left), mod(main), mod(coronary), head(artery)]
[conj(and)]
[verb(give)]
[head(rise)]
[prep(to), mod(two), mod(moderate), mod(sized), mod(diagonal), head(vessel)]]

Note that the syntactic structure is underspecified in that no commitment is made to the internal structure of noun phrases. It is our experience that such an analysis is adequate as a basis for the further processing being considered here. Subsequent analysis then proceeded as discussed in the preceding sections, and in the ideal case, the program determined exactly those noun phrases (8b) marked by hand (8a) in the records.

(8) a. The **left anterior descending artery** arises from the **left main coronary artery** and gives rise to two moderate sized **diagonal vessels**.

b. left anterior descending artery
left main coronary artery
diagonal vessels

RESULTS

The total number of coronary artery terms marked by hand in the records processed was 213, while the total number of such terms found by our automatic procedures was 199. Recall as a partial measure of effectiveness was thus 83%. Of the 199 terms identified, 176 were correct, and thus precision was 88%.

Although the reports were marked independently of the analysis, the development of the rules underlying the analysis was not conducted in ignorance of the markings. Due to this bias, quantitative results reported for this training set are probably better than those achievable for a test set; however, the general trend is likely to persist.

As indicated by the comparatively weak recall score, most of our errors were false negatives. The following section addresses some of the problems we encountered.

DISCUSSION

Certain errors resulted from deficiencies in the underlying syntactic analysis. One such problem is due to an inadequate treatment of coordinated structures. For example, the human judge had marked *mid and distal LAD* in (9a) as being a coronary artery term. However, syntactic processing was not able to determine the correct structure for this phrases, and hence only *distal LAD* (9b) was recognized.

(9) a. Otherwise, the **mid and distal LAD** are free of significant stenosis.

b. distal LAD

Other syntactic errors were due to failure of the tagger to correctly resolve part-of-speech ambiguities. For example in (10) *proceeds* functions as a verb. However, the tagger erroneously analyzed this word as a noun, and subsequent syntactic processing thus considered *the LAD proceeds* to be a single noun phrase rather than a noun phrase (*the LAD*) followed by a verb (*proceeds*). Since *proceeds* as the head of this phrase is not one of the words allowed as the head of a coronary artery term, the program ignored the entire phrase as containing a relevant term.

(10) a. The **LAD** proceeds around the apex and is a **large vesse**l.

b. large vessel

The largest proportion of errors were due simply to lexical issues, which can be addressed by expanding our lists of terms pertinent to the coronary artery identification. Previously unidentified acronyms, such as *PDA*, *PDA I*, and *RCA* occurred throughout the records and were missed by the program. Some of these are defined locally in the text, and we are pursuing a program to identify and expand such cases.

In addition to the names of the coronary arteries, we also attempted to identify reference to part-whole relationships, such as *proximal portion* in (11).

(11) a. This **marginal** has serial 70-80% stenoses in its **proximal portion** prior to bifurcating.

b. marginal
proximal portion

However, due to an incomplete list of terms referring to such phenomena, some expressions of this type, such as *proximal section* in (12) were missed.

(12) a. The **left anterior descending artery** has a 50-70% stenosis in **the proximal section** prior to the takeoff of any of the **diagonals**.

b. left anterior descending artery
diagonals

Related terminology, currently missed, is indicated in the examples in (13), where only the pertinent terms are highlighted.

(13) a. The circumflex coronary artery is subtotally occluded just at its **takeoff** from the left main coronary artery.

b. Immediately prior to the **bifurcation** and following the bifurcation the first vessels are free of significant stenosis.

c. It has multiple irregularities in its **proximal course** before giving rise to a large first septal perforator and then a moderately large first diagonal.

Other false negatives (highlighted) due essentially to lexical issues are given in (14).

(14) a. There is an approximately 2-cm gap between **the total right** and the collaterals which fill via left to right to the distal RCA.

b. It bifurcates into an LAD and **left circumflex system**.

CONCLUSION

We view the encouraging results of our evaluation, particularly regarding precision, as being indicative of the feasibility of using existing natural language processing techniques to provide enhanced access to anatomical text and images. Subsequent processing, built on current results, needs to successfully access anatomical concepts in the UWDA knowledge source and correctly interpret the relationships (especially spatial) expressed in clinical text such as cardiac catheterization reports. If this can be accomplished it will, for example, be possible to indicate in an online image the location of the occlusion described in text such as:

(15) The circumflex coronary artery is subtotally occluded just at its takeoff from the left main coronary artery.

This ability could support advanced pedagogical tools designed for an electronic curriculum. In addition, problems due to synonymy and abbreviatory expressions encountered in processing biomedical text are just those likely to impede user access to an image database like the Visible Human. The techniques being developed here will be valuable in developing improved interfaces to such databases.

Acknowledgments

We are grateful to W. Lowell Maughan, M.D. and Mrs. Jamie Riley of Johns Hopkins Medical Institutions, Baltimore for the cardiac catheterization reports which form the basis of this study.

References

1. McCray AT, Aronson AR, Browne AC, Rindflesch TC, Razi A and Srinivasan S. UMLS knowledge for biomedical language processing. *Bulletin of the Medical Library Association* 81, 1993, 184-194.

2. Humphreys BL, Lindberg DAB, Schoolman HM, and Barnett GO. The Unified Medical language System: An informatics research collaboration. *Journal of the American Medical Informatics Association* 1998:5(1):1-13.

3. McCray AT, Srinivasan S and Browne AC. Lexical methods for managing variation in biomedical terminologies. In Ozbolt JG (ed.) *Proceedings of the 18th Annual SCAMC*, 1994, 235-239.

4. Aronson AR, Rindflesch TC, and Browne AC. Exploiting a large thesaurus for information retrieval. *Proceedings of RIAO* 94, 1994:197-216.

5. Rindflesch TC and Aronson AR. Semantic processing in information retrieval. In Safran C (ed.) *Proceedings of the 17th Annual SCAMC*, 1993:611-615.

6. National Library of Medicine (U.S.) Board of Regents. *Electronic imaging: Report of the Board of Regents*. U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health, 1990. NIH Publication 90-2197.

7. Bean, CA, Rindflesch TC, and Sneiderman CA. Automatic semantic interpretation of anatomic spatial relationships in clinical text. Submitted to 1998 AMIA Annual Fall Symposium.

8. Rosse C, Mejino JL, Modayur BR, Jakobovits R, Hinshaw KP, Brinkley JF. Motivation and organizational principles for anatomical knowledge representation: The Digital Anatomist Symbolic Knowledge Base. *Journal of the American Medical Informatics Association* 1998:5(1):17-40.

9. Friedman C. Towards a comprehensive medical language processing system: methods and issues. In Masys DR (ed.) *Proceedings of the 1997 AMIA Annual Fall Symposium*, 1997:595-599.

10. Sager N, Lyman M, Tick LJ, Nhan NT, and Bucknall CE. Natural language processing of asthma discharge summaries for the monitoring of patient care. In Safran C (ed.) *Proceedings of the 17th Annual SCAMC*, 1993:265-268.

11. Hripcsak G, Friedman C, Alderson PO, DuMouchel W, Johnson SB, Clayton PD. Unlocking clinical data from narrative reports: a study of natural language processing. *Ann Intern Med*, 1995; 122(9):681-8.

12. Jain NL and Friedman C. Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports. In Masys DR (ed.) *Proceedings of the 1997 AMIA Annual Fall Symposium*, 1997:829-833.

13. Cutting D, Kupiec J, Pedersen J and Sibun P. A practical part-of-speech tagger. In Proceedings of the Third Conference on Applied Natural Language Processing, 1992.