Argument Identification for Arterial Branching Predications Asserted in Cardiac Catheterization Reports

Thomas C. Rindflesch[†] Ph.D.
Carol A. Bean[‡] Ph.D.
Charles A. Sneiderman[†] M.D. Ph.D.
[†]National Library of Medicine, Bethesda, Maryland 20894
[‡]University of Tennessee, Knoxville, Tennessee 37996

The language describing coronary vasculature provides a suitable paradigm for research in semantic interpretation of anatomical text. As a pilot project we investigate the possibility of highly accurate retrieval of arterial branching relationships asserted in cardiac catheterization reports. Our methodology relies on the cooperation of underspecified linguistic analysis and structured domain knowledge. The satisfactory results of formal evaluation on both a training and testing set support the promise of this approach.

INTRODUCTION

Semantic interpretation maps the expressions in natural language text to a domain model, thus providing a means of normalizing the concepts and relationships encountered. Such processing can lead to advances in information management and retrieval technology. In a pedagogical context, for example, effective semantic interpretation might support innovative applications automatically linking several sources of information regarding particular body systems and associated disorders, such as clinical records, as well as textbooks on anatomy, surgery, and disease.

Comprehensive semantic interpretation of unrestrained text, however, is beyond the state of the art of natural language processing methodology. In order to scale back the complexity involved, we are focusing current research on arterial branching relationships expressed in cardiac catheterization reports.

Our decision to apply natural language processing techniques to text concerning the coronary arteries was guided not only by the medical significance of these vessels but also by their suitability in serving as a paradigm for research in semantic interpretation of anatomically-oriented text. The set of names for the coronary arteries is moderate in size; on the other hand, the names for these vessels show considerable complexity in their linguistic structure, and several of the coronary arteries have inherently ambiguous names. In addition, extensive individual variation in the specific vascular elements and their branching configurations makes it difficult to predict reliably the exact nature of the coronary vasculature exhibited by any particular patient. These factors combine to ensure that it is no trivial task, yet vitally important, to interpret correctly the language describing this aspect of human anatomy.

The text which forms the basis for this study comes from fifteen case reports (stripped of patient identification) received from the Johns Hopkins Cardiac Catheterization Laboratory. Each report is divided into explicitly marked sections which include “Indications,” “Physical Examination,” “Cardiac Catheterization,” “Hemodynamics,” “Arteriography,” and “Impression.” For this project we concentrate on the arteriography section, which describes the characteristics of the arteries observed, including size and general condition as well as the presence and location of any stenosis.

The structure of these reports is significant regarding the semantic interpretation of branching relationships in that each arteriography section is divided into labeled subsections devoted exclusively to one of the major coronary arteries. It is always the case that the branching relationships asserted in a particular subsection involve the artery named in the associated heading, as the following example illustrates.

(1) Right Coronary Artery (RCA): The right coronary artery is a large vessel which arises normally from the aorta. It gives rise to a large posterolateral as well as two posterior descending branches. There is a 90-95% stenosis in the midportion of the posterolateral branch as well as 80-90% stenosis in the proximal portion of the posterior descending vessels.

We are developing a Prolog program that reads text such as the preceding and extracts the following branching predications:

(2) Right coronary artery-BRANCH_OF-Aorta
Posterior ventricular branch of right coronary artery-BRANCH_OF-Right coronary artery
Posterior interventricular branch of right coronary artery-BRANCH_OF-Right coronary artery

Practical applications in natural language processing are directed at text which is constrained semantically, and such processors often rely extensively on characteristic linguistic patterns occurring in the genre of text being processed ([1 & 2] for example). Statistically-based methods ([3] for example) must be trained on text evincing characteristics inherent in the target input.

The research reported here takes an alternative approach. Very general, underspecified syntactic analysis is used for argument identification [4]. Rather than appeal to syntactic peculiarities found in the text being processed, we rely on general principles which exploit the structural characteristics of the input text during the semantic interpretation process. The structured domain knowledge available to the semantic processor also contributes to the accuracy of the output.

METHODS

Overview

The program we are developing initially scans the cardiac catheterization reports and determines document structure at the lowest hierarchical level. That is, the text of a subsection along with the associated heading is identified and submitted for further processing. During the subsequent phase of automatic analysis both the heading and text for each subsection are subjected to a series of steps that first identify the coronary artery terminology occurring in the sentences encountered and then combine these terms into the branching predications asserted.

Identifying coronary artery terminology

Initial processing to identify coronary artery terminology is based on earlier work [5] and involves sentence identification, look-up in the SPECIALIST Lexicon [6], category label ambiguity resolution [7], and an underspecified syntactic parse.

We rely on MetaMap [8] to match noun phrases to concepts in the Unified Medical Language System® (UMLS®) [9] Metathesaurus. The isolation of noun phrases referring to names for the coronary arteries in the underspecified syntactic structure sets the scene for this processing. For example, when applied to a sentence such as *The left circumflex artery gives rise to a first marginal branch,* the mapping process begins with the syntactic parse (3), in which the noun phrases referring to coronary arteries are identified and labeled.

(3) [**corart**([det(the),mod(left)],mod(circumflex),head(artery)]),
[verb(gives)],
[head(rise)],
corart([prep(to),det(a)],mod(first),mod(marginal),head(branch)])
]

Although the UMLS Metathesaurus contains a considerable number of synonyms representing its constituent concepts, we augmented these resources in several ways. To address certain variants, such as *left main* for Left coronary artery, or *LADD* for Left diagonal artery, we compiled an ancillary list of synonyms which is consulted during the mapping process. To accommodate other cases, such as *diagonal branch, diagonal vessel,* or *left circumflex vessel,* we substitute *artery* for *branch* or *vessel* in the input noun phrase before submitting it to MetaMap.

An integral component of this project is the structured domain knowledge provided by the UMLS Metathesaurus and Semantic Network. Although several constituent vocabularies in the UMLS contain extensive anatomical terminology, we consider the University of Washington Digital Anatomist (UWDA) Symbolic Knowledge Base [10] to be primary in this project. In all instances we provide the UWDA preferred term in the output, for example “Posterior interventricular branch of right coronary artery” for *posterior descending branch* in (2) above. We also rely heavily on the rich structure provided by UWDA. In addition to the traditional “is a” hierarchy, this vocabulary contains three other trees to which concepts are assigned as appropriate, namely “branch of,” “part of,” and “tributary of.”

Interpreting branching relationships

The second phase of the processing constructs a complete branching predication based on the identification of the referential vocabulary as discussed in the previous section. The two major aspects of the recognition of a semantic predication from the syntactic structure which encodes it are semantic indicator rules [11] and argument identification. The indicator rules establish a correspondence between a syntactic entity and a semantic predicate, while argument identification relies on the proper treatment of syntactic phenomena such as coordination, relativization, and anaphora.

The branching of anatomical structures is represented in the UMLS Semantic Network by two predicates:

‘branch_of’ and its inverse ‘has_branch’. These relations are encoded by several syntactic entities in the cardiac catheterization reports we have encountered: ‘branch_of’ by *is a branch of*, *arises from* (or *off of*), and *takes off from*; and ‘has_branch’ by *has branch*, *gives off*, *gives rise to*, and *bifurcates into*.

Argument identification in our system is guided by the underlying principle that the arguments must be semantically consonant with the predicate. Information ensuring this is contained in the UMLS Semantic Network, which states that both arguments of ‘branch_of’ and ‘has_branch’ must be concepts in the Metathesaurus having semantic type ‘Body Part, Organ, or Organ Component’. Argument identification is further constrained by the simple stipulation that (in the absence of other syntactic cues) when the syntactic indicator of the predicate of the relationship is a verb, the arguments of the predication are to be found on either side of the indicator. There is also a general rule that an argument from one predication cannot be “reused” for another predication, except under specified conditions (involving predicate coordination and relativization), discussed below.

The process of argument identification applies to a sentence such as (4) by taking advantage of the information in (5), where the noun phrases have been mapped to UMLS (UWDA) concepts with semantic type ‘Body Part, Organ, or Organ Component’, and indicator rules have been applied to identify the appropriate Semantic Network relationship.

(4) *The left anterior descending* artery **arises** normally **off of** *the left main*.

(5) *the left anterior descending* ⟶
“Anterior interventricular branch of left coronary artery”
arises...**off of** ⟶
‘branch_of’
the left main ⟶
“Left coronary artery”

All constraints are satisfied: the arguments are of the appropriate semantic type for the indicated relation and they occur on either side of the indicator. The final interpretation for (4) is given in (6).

(6) Anterior interventricular branch of left coronary artery-BRANCH_OF-Left coronary artery

The basic strategy for argument identification is augmented slightly in order to accommodate coordination and relativization. In noun phrase coordination, such as in (7) between the phrases *a left circumflex* and *left anterior descending branches*, the concept mappings for the two coordinate phrases must each appear analogously in predications which are otherwise identical.

(7) The left main gives off *a left circumflex* **and** *left anterior descending branches*.

This condition on the semantic interpretation of coordinate noun phrases is satisfied by the branching relationships in (8), which the system extracted from (7).

(8) Anterior interventricular branch of left coronary artery-BRANCH_OF-Left coronary artery
Circumflex branch of left coronary artery-BRANCH_OF-Left coronary artery

The sentence (9) provides an example of predicate coordination (between *arises from* and *gives rise to*).

(9) The right coronary artery *arises* normally *from* the aorta **and** *gives rise to* a posterior descending artery.

This phenomenon allows sharing (or reuse) of an argument. In this instance the concept mapping for *the right coronary artery* can be shared between the semantic predicates based on *arises from* and *gives rise to.* The branching predications for (9) given in (10) illustrate this shared argument (Right coronary artery). (‘has_branch’ is later normalized to ‘branch_of’.)

(10) Right coronary artery-BRANCH_OF-Aorta
Right coronary artery-HAS_BRANCH-Posterior interventricular branch of right coronary artery

A syntactic construction which involves a relative clause modifying the subject of the preceding clause is seen in (11), where *which gives rise to a single posterior descending branch* modifies the preceding subject, namely *the RCA.*

(11) The RCA *is* a moderate sized short vessel **which** *gives rise to* a single posterior descending branch and is without significant disease.

Subject relativization is treated by our system similarly to predicate coordination, in that a single argument is allowed to be shared between two predications. In this instance, the two predications are based on *is* and *gives rise to.* The predication based on *is* is not a branching relationship and is suppressed by the system for now. The branching relationship expressed in (11) is given in (12).

(12) Right coronary artery-HAS_BRANCH-Posterior interventricular branch of right coronary artery

We also employ a limited form of anaphora resolution based on [12]. The implementation applies only to the pronoun *it* occurring in subject position. In such instances, this pronoun is considered to corefer with the subject of the immediately preceding sentence.

For example, in (13), *it* in the second sentence is replaced by the subject of the preceding sentence, namely *the LAD.*

(13) **The LAD** is a large vessel which proceeds around the apex. **It** gives rise to a first diagonal branch which is a large vessel and contains a 30% ostial stenosis.

Once this substitution has resolved the anaphora, processing proceeds normally, and the interpretation for the second sentence is the predication given in (14).

(14) Left diagonal artery-BRANCH_OF-Anterior interventricular branch of left coronary artery

The final step in semantic interpretation is to validate the proposed output against the UWDA domain knowledge available for the subsection being processed. This procedure disallows a branching predication that does not occur in the UWDA branching hierarchy. For example, syntactic argument identification incorrectly generated the branching relationship (16) from the sentence (15); however, validating against domain knowledge eliminated this error from the final output.

(15) The right coronary artery arises normally off of the aorta, is diffusely, mildly diseased and gives off collateral flow to the LAD.

(16) -FP->Anterior interventricular branch of left coronary artery-BRANCH_OF-Right coronary artery

The etiology of this error is that *the LAD* rather than *collateral flow* was taken to be the object of *gives off*. Such syntactic errors are normal in natural language processing technology. An appeal to domain knowledge, guided by the structure of the documents being processed, can eliminate at least some of these errors.

RESULTS

In order to formally evaluate our methodology, we divided the fifteen reports into two groups: CCR1 (reports 1 through 7) and CCR2 (reports 8 through 15). CCR1 contains 212 sentences, while CCR2 has 319. For the arteriography section from each report, the sentences asserting a branching relationship (40 in CCR1 and 44 in CCR2) were identified and the correct relationship referred to was inserted by hand (CAS). These marked sentences served as the standard against which we calibrated the program under development. The group from CCR1 served as a training set, while those from CCR2 were used for testing.

During development, the program was honed to fit the training set (CCR1) by modification guided by error analysis iteratively performed on the output from successive versions of the program. The final version of the program was then run on the testing set (CCR2). The program was not modified to accommodate the structural peculiarities of CCR2; however, synonyms not encountered previously (such as *LADD* for Left diagonal artery) were added to our supplemental lists.

The gold standard for CCR1 contains 66 branching relationships. The program identified 59 such predications in CCR1, all of which were correct. Recall and precision for the results of the training set were thus 89% and 100%, respectively. The gold standard for the testing set, CCR2, expresses 71 branching relationships. When the program ran on CCR2 it recovered 59 of these assertions, all of which were correct, thus yielding 83% recall and 100% precision.

DISCUSSION

A failure analysis revealed several predictable sources for the errors observed. The majority of the false negatives generated while processing CCR1 are ultimately due to a single coronary artery term that does not map to a concept in the UMLS Metathesaurus, namely *ramus intermedius*, as seen in the following examples.

(17) a. The left main coronary artery arises normally off the aorta and gives rise to the left circumflex, the **ramus intermedius** and left anterior descending branches.

b. The left circumflex arises normally from the left main and proceeds through the AV groove giving rise to a proximal **ramus intermedius** which is a moderate-sized vessel, a moderate-sized first marginal branch, and a larger second marginal branch.

Note that the term is used ambiguously to refer either to a branch of the Left coronary artery or of the Circumflex branch of the left coronary artery.

The branching predications missed while processing CCR2 are largely due to a variety of linguistic phenomena that do not occur in CCR1. Several false negatives result from the recurrence of a syntactic structure like that seen in (18).

(18) The left main coronary artery arises normally from the **aorta,** gives rise to an LAD and left circumflex systems.

The structure of this sentence is nonstandard in that there should be a coordinator instead of a comma after *aorta*. The program was not able to recognize that *gives rise to* is intended to be coordinate with *arises*, and thus was not able to find a subject for *gives rise to.*

The anaphora inherent in the phrase *its distal portion* in (19) was not recognized by the program, and hence the branching relationship asserted in this sentence was not retrieved.

(19) Its distal portion gives rise to a small posterolateral branch.

Finally, the program was not able to map the noun phrase *a large second RV free wall branch* in (20) to a concept in UMLS. Due to this omission, the program did not recognize any branching relationships in this sentence.

(20) The RCA is a large vessel which gives rise to a moderate sized first and **a large second RV free wall branch**, a large posterior descending artery and a small right posterolateral branch.

Significantly, the perfect precision achieved for both the training and testing sets was due largely to the fact that we validated the output of the program against the UWDA branching hierarchy. Any erroneous output produced by the program was eliminated by this process.

CONCLUSION

The results of the formal evaluation of our methodology for extracting arterial branching relationships from cardiac catheterization reports demonstrate the effectiveness of combining underspecified natural language processing techniques with the structured domain knowledge provided by the UMLS. We are confident of the feasibility of extending this methodology to a more comprehensive normalization of the semantic content of anatomically-oriented text. Further extensions to this research include investigating ways in which semantic interpretation interacts with associated cognitive structures such as those representing spatial phenomena typically found in anatomical discouse ([13]).

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 for this study.

References

1. 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 AMIA Annual Fall Symposium*, 1997:829-833.

2. Fiszman M, Chapman WW, Evans SR, and Haug PJ. Automatic identification of pneumonia related concepts on chest x-ray reports. In Lorenzi NM (ed.) *Proceedings of the AMIA Annual Symposium*, 1999:67-71.

3. Taira RK and Soderland SG. A statistical natural language processor for medical reports. In Lorenzi NM (ed.) *Proceedings of the AMIA Annual Symposium*, 1999:970-4.

4. Rindflesch TC, Rajan JV, and Hunter L. Extracting molecular binding relationships from biomedical text. *Language Technology Joint Conference ANLP-NAACL*, 2000.

5. Sneiderman CA, Rindflesch TC, and Bean CA. Identification of anatomical terminology in medical text. In Chute CG (ed.) *Proceedings of the AMIA Annual Symposium*, 1998:428-32.

6. 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.

7. 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.

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

9. 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.

10. Rosse C, Mejino JL, Modayur BR, Jakobovits R, Hinshaw KP, and 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.

11. Bean CA, Rindflesch TC, and Sneiderman CA. Automatic semantic interpretation of anatomic spatial relations in clinical text. In Chute CG (ed.) *Proceedings of the AMIA Annual Symposium*, 1998:897-901.

12. Lappin S and Leass HJ. An algorithm for pronominal anaphora resolution. *Computational Linguistics* 1994:20(4):535-62.

13. Bean CA. Formative evaluation of a frame-based model of locative relationships in human anatomy. In Masys DR (ed.) *Proceedings of the AMIA Annual Fall Symposium*, 1997:625-9.