Integrating Natural Language Processing and Biomedical Domain Knowledge for Increased Information Retrieval Effectiveness

Thomas C. Rindflesch
National Library of Medicine
8600 Rockville Pike, Bethesda, MD 20894

Abstract

Underspecified semantic structures serve as the basis for indexing terms for information retrieval. Biomedical semantic types from the National Library of Medicine's Unified Medical Language System® constrain coordinate structures to increase the accuracy of the semantic representation. Preliminary experiments conducted on 3,000 MEDLINE titles and abstracts indicate that the approach contributes to increased precision.

I. INTRODUCTION

In several recent publications ([1] , [2] , [3]) we describe SPECIALIST, a system which combines natural language processing and domain knowledge to improve access to biomedical information. Although natural language processing demonstrates promise for increasing effectiveness in information retrieval, it has so far not been shown to be practical in large data bases. Two major obstacles have been the difficulty of providing a complete linguistic analysis for unrestricted text and the extensive knowledge sources required for real applications.

SPECIALIST attempts to surmount both of these problems. We finesse the difficulties associated with constructing a domain model by using an existing knowledge base, namely the Unified Medical Language System (UMLS®) [4]. We address inefficiency by looking to the notion of underspecified linguistic analysis of the sort discussed by Agarwal and Boggess [5] . An underspecified analysis of a particular structure is much simpler than a fully specified description and contributes significantly to a less complex and more efficient linguistic component. Our system attempts to use just enough linguistic analysis, both syntactic and semantic, to allow the construction of a conceptual structure which supports matching queries to documents.

After providing a brief system overview this paper discusses the treatment in SPECIALIST of a traditionally difficult linguistic structure, coordination, and concentrates on the ways in which the underspecified approach and the use of domain knowledge in the form of UMLS semantic features contribute to the analysis of coordinate structures.

II. SYSTEM OVERVIEW

The system first assigns an underspecified syntactic analysis (2) to input (1) from either a query or a document. This analysis is supported by a large lexicon [6] and the Xerox part-of-speech tagger [7], and, most importantly, identifies noun phrases for further analysis. Further analysis involves mapping noun phrases to concepts in the UMLS Metathesaurus (3), thereby providing semantic type information for further semantic analysis(4). Semantic interpretation depends on the relationships defined in the UMLS Semantic Network [8].

(1)thermography in the determination of amputation levels in ischaemic limbs

(2)minimal_syntax

[head(thermography)]

[prep(in),det(the),head(determination)]

[prep(of),mod(amputation),head(levels)]

[prep(in),mod(ischaemic),head(limbs)]

(3)Metathesaurus concepts and semantic types

“Thermography” [Diagnostic Procedure]

“Amputation” [Disease or Syndrome,Therapeutic or Preventive Procedure]

“Limbs” (“Extremities”) [Body Location or Region]

“Ischemic” (“Ischemia”) [Pathologic Function]

(4)Conceptual Structure

affects

nom([metaconc([“Ischemia”])])

theme([metaconc([“Extremities”])])

determination

instr([head([metaconc([“Thermography”])])])

theme([mod([metaconc([“Amputation”])]),head([tokens([levels])])])

has_attribute

nom([head([metaconc([“Extremities”])])])

theme([mod([metaconc([“Amputation”])]),head([tokens([levels])])])

Such semantic structures serve as the basis for matching queries to documents for information retrieval. These structures are underspecified in the sense that the system often provides only partial analyses. As such structures become more complete and accurate they support improved access to information. An inability to deal effectively with coordinate structures detracts from the accuracy of the semantic analysis.

III. COORDINATION

A. Background

Work on coordination in linguistics has proceeded under the assumption that in some sense the coordinated expressions must be similar. Schachter [9] discusses syntactic similarity in terms of constituent structure. Although Sag et al. [10] and Bouldin [11] discuss examples of the coordination of dissimilar constituents. Several designs for handling coordinate structures in NLP systems have been proposed (for example [12], [13], [14]) based on the assumption that coordination involves similar constituent structures. The approach suggested here also assumes a basic similarity in conjoined elements. It departs from tradition, however, in proposing an analysis which does not build a complete, fully specified syntactic structure. In addition, the analysis assumes that semantic information is essential to the formulation of a useful treatment of coordination.

B. General approach

In order to accommodate partial analyses, the syntactic mechanism used is closer to dependency syntax [15] than traditional constituent structure analysis. Relationships between **words** are considered to be basic to syntactic description. In coordinate structures, a coordinator signals that two words, the left and right conjuncts, are conjoined.[1] Word in this sense is a lexical entry rather than a text word; multi-word items may appear in the lexicon.

The following are defined as general constraints on coordination:

- A conjunction coordinates two elements.
- A conjunction occurs between the conjuncts it coordinates.

---

1. Series coordination still needs to be addressed. One possible analysis is that in, for example *cats, dogs and horses, and* is involved in two coordination relationships, *cats and horses*, and *dogs and horses*. A second possibility is to treat comma (under certain circumstances) as a coordinator. Further rules are then needed in either case to insure that all the conjuncts are coordinate to each other.

- The two conjuncts have the same part of speech.

In each of the following examples, the general constraints on coordination permit the words in boldface to be coordinated by the coordinator (underlined). Note therefore, that although traditionally (5a) contains an instance of NP coordination, (5b) contains an instance of conjoined prepositional phrases, (5c) has conjoined adjectives, (5d) has coordinate verbs, and (5e) is an instance of conjoined prepositions, they are all treated similarly here.

(5) a. The **advantages** and the **limitations** of each method are discussed.

b. There were no complications of the **preparation** or of the **colonoscopy**.

c. Thermographic tests were performed during the **visual** and/or **sensory** aura.

d. Present-day therapeutic efforts only **retard** or **prevent** bone loss.

e. These samples were taken at various intervals **before** and **after** therapy.

C. Determining the consequences of a coordination relationship

Adequately dealing with coordinate structures involves two major steps. The first is the identification of the left and right conjuncts (as defined above) associated with a particular coordinator. Once the conjuncts have been found, the analysis must then consider the further linguistic consequences of the coordination relationship. If two elements are coordinate they must then have the same function (semantics permitting) in propositional structure.

For example, in (6), the coordination is identified (6b). On the basis of this information, the semantic analysis then produces propositional structure (6c).

(6) a. The effective areas of stimulation were located separately in the dorsolateral **funiculus** and in the ventrolateral **funiculus**

b. and(funiculus, funiculus)

c. has_location(effective areas of stimulation, dorsolateral funiculus)
has_location(effective areas of stimulation, ventrolateral funiculus)

For the remainder of the discussion I will focus on the first task, namely identification of the left and right conjuncts.

D. Identifying coordination relationships

The process of identifying left and right conjuncts being proposed builds on the approach in [16]. That work considers there to be a major distinction between verb coordination and other types of coordination and attempts to determine whether a given coordinator is involved in verb coordination or not. One reason to distinguish between coordination of verbs and other types of coordination concerns the location of the right conjunct in a particular coordination relationship. It is largely the case that except for verbs, the right conjunct is in the constituent immediately to the right of the coordinator.

The following is an informal statement of the algorithm for identifying coordination relations between either nouns or verbs.

- For each coordinator in a sentence, first determine whether verbs are being coordinated.
- If conditions do not exist for coordination of verbs, consider noun coordination.
- In noun coordination, the right conjunct is the head of the first NP to the right of the coordinator.
- The left conjunct in noun coordination is the head of the first NP to the left of the coordinator which is compatible with the right conjunct.

The way in which verb coordination is determined and the principles for determining compatibility of noun conjuncts are discussed in the following sections.

E. Coordination of verbs

The rules which govern verb coordination depend on the notion of an "ordination relationship." An ordination relationship is considered to obtain between the verbs in a sentence containing more than one verb. The ordination relation rules crucial for the determination of verb coordination discussed in [16] and [17] can be summarized as:

(7) a. If more than one verb occurs in a sentence, the verbs must be in an ordination relationship: coordination or subordination.

b. There is at least one verb in a sentence which is not subordinate to any other.

c. If two verbs are coordinate they are equiordinate (i.e., both are either subordinate or not).

d. A subordinator is associated with a verb to its right.

e. A verb associated with a subordinator is necessarily subordinate.

It should be noted the that information required to determine the applicability of the constraints on ordination relationships can be determined on the basis of very low level structural cues, such as the number of verbs and the occurrence of subordinators in the input string.

In the examples in (8) the verbs must be considered to be in a relationship of coordination in order to satisfy the constraints on ordination relationhips.

(8) a. Colonic bleeding lesions **were** identified in 24 of 35 patients, and hemorrhage originating proximal to the ileoceal valve **was** documented in three of these 35 patients.

b. The necrotic center of a traumatic ulcer **inhibited** measurement of an underlying inflamed base and, thus, **was** equivalent to the control in temperature.

c. We **present** and **demonstrate** the clinical model upon which the method rests.

This is so because subordination is not a possible relationship in these examples. It is thus the case that if none of the verbs in (8) can be in a relationship of subordination, each pair must be coordinate in order to satisfy all ordination relationship constraints.

Several ordination relation rules conspire to support an analysis of (9) which does not include coordinate verbs.

(9)Since alpha-atrial natriuetic peptide (ANP) plays an important role in the homeostasis of sodium **and** fluid balance, measurement of alpha-ANP concentrations might provide valuable information on the status of the critically ill.

The subordinator *since* is associated with *plays* (7d), rendering that verb necessarily subordinate (7e). Therefore if *plays* is coordinate with *might provide* both verbs are subordinate (7c). Since these are the only two verbs in the sentence, however, such a construal violates (7b). Therefore the only possible analysis of (9) is one which considers the verbs to be in a relationship of subordination and which assumes that the *and* coordinates something other than verbs.

F. Coordination of nouns

The general constaint on coordination which requires the two conjuncts to be the same part of speech needs to be further constrained. At least in the biomedical research literature, a large number of coordination relationships between nouns occurs when the two conjuncts are either both

relational nouns or have compatible semantic types.[1] Relational nouns are those which can take arguments (other than possessive *of* phrase). This includes nominalizations, but also words like *size* (as in *the size of the desk*).

I exploit this generalization by stipulating the following rule:

(10)A nominal coordination is allowed only if the two conjuncts are compatible. They are compatible if they are both relational nouns or if they have consonant semantic types.

In each sentence in the following examples (11) this rule allows only the word in bold on the left of the coordinator to be the left conjunct of the coordination relationship. In each case that word is a relational noun matching the relational noun in bold on the right of the coordinator.

(11) a. There was a significant **increase** in plasma calcium and a significant **decrease** in plasma phosphate.

b. Nucleotide sequence **analysis** of the spacer regions flanking the rat rRNA transcription unit and **identification** of repetitive elements.

c. We determined the bone mineral **density** of the lumbar spine and the **strength** of back extensors in 68 healthy postmenopausal Caucasian women.

With regard to consonant semantic types, in (12), *circle of Willis*, an anatomical term, must be coordinated with the consonant term *artery. Communication* is a relational noun, and *blood*, while an anatomical term, is not a head.

(12)Blood supply was provided by communication between a tortuous megadolichobasilar **artery** and the **circle of Willis** through enlarged posterior communicating arteries

Similarly in (13), the drug *gentamicin* can only be coordinated with another drug *tobramycin*. Both *effect* and *clearance* are relational nouns, and *hemofiltration* denotes a therapeutic procedure.

(13)The effect of continuous arteriovenous hemofiltration on the clearance of either **tobramycin** or **gentamicin** was studied in eight critically ill patients.

---

1. [Agarwal and Boggess 1992[5]] also exploit semantic type compatibility in their treatment of coordination.

G. Deficiencies

The current treatment of nominal coordination is deficient in requiring that the two conjuncts must be either relational nouns or have consonant semantic types. In some instances this requirement will leave coordination undetermined (14), since it is not the case in this example that all the conjuncts are relational nouns, nor do they have consonant semantic types.

(14)Immunologic effects of blood transfusion upon **renal transplantation**, tumor **operations**, and bacterial **infections**.

Leaving coordination undetermined is considered to be less serious than assigning an incorrect analysis, as will happen in (15).

(15)The reliability of thermography for the ~~determination~~ of the level of amputation for an ischaemic lower limb was compared with that of the **doppler flowmeter** and the clinical **judgement** of an expericenced surgeon.

Although *judgement* is a relational noun, it is actually coordinated with *flowmeter*, rather than with the relational noun *determination*. It appears to be the case that such examples are not common.

IV. PRELIMINARY TESTING

Intuitively it would seem that translating a text into semantic conceptual structure should contribute to retrieval effectiveness, since the conceptual structure regularizes and canonicalizes text. The treatment of coordination is important in this regard in that a correct analysis of coordination contributes to a more accurate conceptual structure. We have begun to test our conceptual structures with respect to increasing information retrieval precision using the vector-space statistical model of information retrieval (SMART, [18]).

Using this word-based model to represent semantic predicational structure forces us to give up some of the information inherent in the hierarchical semantic structure. However, we feel that the advantage gained, namely ready access to an efficient means of comparing semantic structures offsets the disadvantage.

We have based our testing on the UMLS Test Collection [19]. This collection contains 150 natural language queries and 3,078 documents consisting of titles and abstracts of journal articles from the biomedical literature. The entire textual content of the collection comprises approximately 730,000 words in approximately 25,000 sentences or complex noun phrases.

We ran SMART on the test collection twice, once on the plain text and a second time on a surrogate text transformed by adding the corresponding conceptual structures to each major linguistic structure. For example, the raw text (16) with conceptual structure (17) was transformed into surrogate text (18). This was done for all queries and documents in the test collection and then this new surrogate text was presented to SMART.

(16)the late effect of subtotal thyroidectomy and radioactive iodine therapy on calcitonin secretion and bone mineral density in women treated for graves disease.

(17) a. effect
nom([head([metaconc([“Subtotal thyroidectomy”])])])
nom([mod([metaconc([“Radioactivity”])]),mod([metaconc([“Iodine”])]),
head([metaconc([“therapy”])])])
theme([head([metaconc([“secretion”])])])
theme([head([metaconc([“Bone Density”])])])
modArg([metaconc([“Late”])])

b. secretion
theme([metaconc([“Calcitonin”])])

c. treat
theme([head([metaconc([“Graves’ Disease”])])])
instr([mod([metaconc([“Radioactivity”])]),mod([metaconc([“Iodine”])]),
head([metaconc([“therapy”])])])
patn([head([metaconc([“Women”])])])

(18) a. effect nom Subtotal thyroidectomy nom Radioactivity Iodine therapy
theme secretion theme Bone Density modArg Late

b. secretion theme Calcitonin

c. treat theme Graves’ Disease instr Radioactivity Iodine therapy patn Women

The results of this testing show a 4% increase in average precision using the surrogate text containing conceptual structure over those obtainable on the basis of unprocessed text. These

results, while modest, nonetheless indicate that the approach to NLP, including the treatment of coordination, show promise for continued research.

V. REFERENCES

[1] A. T. McCray et al., "UMLS knowledge for biomedical language processing," *Bulletin of the Medical Library Association,* vol. 81, pp184-194, April 1993.

[2] T. C. Rindflesch and A. R. Aronson, "Semantic processing in information retrieval," *Proceedings, Seventeenth Annual Symposium on Computer Applications in Medical Care,* Washington, D. C., October 1993, pp. 611-615.

[3] A. R. Aronson, T. C. Rindflesch, and A. C. Browne, "Exploiting a large thesaurus for information retrieval," *RIAO 94 Conference Proceedings,* New York, October 1994, pp. 197-216.

[4] D. A. B. Lindberg, B. L. Humphreys, and A. T. McCray, "The Unified Medical Language System," *Methods of Information in Medicine,* vol. 32, pp. 281-291, 1993.

[5] R. Agarwal and L. Boggess, "A simple but useful approach to conjunct identification," *Proceedings, 30th Annual Meeting of the Association for Computational Linguistics*, Newark, DE, June 1992, pp. 15-21.

[6] A. C. Browne, A. T. McCray, and S. Srinivasan, *The SPECIALIST Lexicon,* National Library of Medicine, Report No. NLM-LHC-93-01. (Available from NTIS, Springfield, VA: PB93-217248), 1993.

[7] D. Cutting, J. Kupiec, J. Pedersen, and P. Sibun, "A practical part-of-speech tagger," *Proceedings, Third Conference on Applied Natural Language Processing,* Trento, Italy, April 1992, pp.

[8] A. T. McCray and W. T. Hole, "The scope and structure of the first version of the UMLS Semantic Network," *Proceedings, Fourteenth Annual Symposium on Computer Applications in Medical Care,* Washington, D. C., November 1990, pp. 126-130.

[9] P. Schachter, "Constraints on coordination," *Language,* vol. 53, pp. 86-103, March 1977.

[10] I. A. Sag, G. Gazdar, T. Wasow, and S. Weisler, "Coordination and how to distinguish categories," *Natural Language and Linguistic Theory,* vol. 3, pp. 117-171, 1985.

[11] J. M. Bouldin, "The Discourse Condition on Coordination," Fifteenth Annual Minnesota Conference on Language and Linguistics, Minneapolis, October 1989.

[12] V. Dahl and M. C. McCord, “Treating coordination in logic grammars,” *American Journal of Computational Linguistics*, vol. 9, pp. 69-91, April 1983.

[13] L. Hirschman, “Conjunction in Meta-Restriction Grammar,” *The Journal of Logic Programming*, vol. 3, pp. 299-328, December 1986.

[14] W. A. Woods, “An experimental parsing system for transition network grammars,” in R. Rustin (ed.), *Natural Language Processing*, New York: Algorithmics Press, 1973, pp.145-149.

[15] R. A. Hudson. *Word Grammar*, Chicago: University of Chicago Press, 1984.

[16] M. B. Kac and T. C. Rindflesch, “Coordination in reconnaisance-attack parsing,” *Proceedings, Twelfth International Conference on Computational Linguistics*, Budapest, August 1988, pp. 285-290.

[17] T. C. Rindflesch, *Linguistic aspects of natural language processing*, University of Minnesota doctoral dissertation, 1990.

[18] G. Salton, “Development in automatic text retrieval,” *Science*, vol. 253, pp. 974-980, 1991.

[19] P. L. Schuyler, A. T. McCray, and H. M. Schoolman, “A test collection for experimentation in bibliographic retrieval,” *MEDINFO 89*, pp. 810-912, 1989.