Semantic Processing in Information Retrieval

Thomas C. Rindflesch and Alan R. Aronson
National Library of Medicine
Bethesda, MD 20894

Intuition suggests that one way to enhance the information retrieval process would be the use of phrases to characterize the contents of text. A number of researchers, however, have noted that phrases alone do not improve retrieval effectiveness. In this paper we briefly review the use of phrases in information retrieval and then suggest extensions to this paradigm using semantic information. We claim that semantic processing, which can be viewed as expressing relations between the concepts represented by phrases, will in fact enhance retrieval effectiveness. The availability of the UMLS® domain model, which we exploit extensively, significantly contributes to the feasibility of this processing.

INTRODUCTION

Information retrieval research is keenly interested in improving retrieval effectiveness beyond that possible by using key words alone. One possibility is the use of phrases; and an array of approaches to the treatment of phrases for information retrieval can be found in the recent literature. These approaches vary most significantly with respect to the amount of linguistic analysis brought to bear in extracting phrases from free text, ranging from essentially none to a full syntactic parse. The barrier word method, for example, as demonstrated by Tersmette et al. [1] considers a phrase to be any string occurring between a list of barrier words. A more extensive analysis employs barrier words in addition to some variant generation, but attempts no (or very little) linguistic analysis [2,3,4, 5,6]. Systems which exploit linguistic structure in information retrieval may be based on a full syntactic analysis or some form of underspecified structure [7,8,9,10,11,12,13,14].

Unfortunately, the use of phrases in information retrieval—regardless of the method used to generate them—has so far not been shown to contribute significantly to retrieval effectiveness (see [9,15,3]). Due to the intuitive appeal of the use of phrases, a number of researchers have investigated enhancements to phrasal-based methods. Lewis and Croft [16], for example, discuss the use of term clustering of syntactic phrases, while Croft et al. [7] examine phrases used in conjunction with structured queries. Hersh et al. [3] suggest that adding some form of semantic processing to phrasal-based information retrieval may improve performance. In this paper, we explore an approach to semantic processing which adds to the information available about a text by specifying the relationships that exist among the concepts represented by the phrases in the text. (Also see [17, 18]). As will be seen below, we base semantic processing on an underspecified syntactic analysis and extensive variant generation, which together support the robust mapping of phrases to terms in the UMLS Metathesaurus®.

SYSTEM OVERVIEW

Our system first assigns a syntactic analysis to input from either a query or document. At the heart of the approach is a mapping of this structure to concepts in the UMLS domain model [19]. The most important information which is then available for further analysis is a semantic type for each concept which is situated in a network of such types. Further semantic processing constructs a predicate argument structure which determines how the concepts discovered in the previous phase interact within a particular linguistic structure. For example, in slightly simplified form, our syntactic component assigns the underspecified structure (1b) to the input (1a). *Thermogram* and *meningitis* map to the UMLS concepts shown in (1c), which is the semantic interpretation for this example and provides both the UMLS concepts as well as the associated semantic types. The semantic interpretation specifies the relationship which obtains between the concepts in the input phrase.

(1) a. Use of thermogram in detection of meningitis.

b. noun_phrase(
[head(use)],
[prep(of), head(thermogram)],
[prep(in), head(detection)]
[prep(of), head(meningitis)])

c. detection(
theme(head(meningitis,
concept(“Meningitis”),
semtype(‘Disease or Syndrome’))),

instr(head(thermogram,
concept(“Thermography”),
semtype(‘Diagnostic Procedure’))))

We have a running prototype which produces underspecified syntactic analysis, successfully maps noun phrases to the UMLS domain model, and then builds semantic structures.

MAPPING PHRASES IN FREE TEXT TO THE UMLS METATHESAURUS

We have attempted to combine the most effective aspects of the various approaches to using phrases in information retrieval. Our system first identifies phrases in free text using an underspecified syntactic analysis. We claim that such an approach supports the semantic representations which can effect accurate matching of queries to relevant documents but avoids the problems associated with a fully specified syntactic analysis as noted, for example, by Salton and Smith [20]. Our syntactic component is closely allied to work in underspecified syntactic analysis of the type discussed in [21] and similar in depth of coverage to the work of Evans et al [8].

We begin by analyzing noun phrases and prepositional phrases. In a successful syntactic analysis, heads are identified and most items to the left of the head are simply labelled as “modifier”; however, participles are singled out and labelled as such. Prepositional phrases are implicitly identified, but during the syntactic phase their attachment is not indicated.

Although the syntactic structure we produce is not fully specified, it has advantages over the unstructured phrases obtained from a barrier word approach. Most importantly, the identification of heads of noun phrases has significant consequences during the mapping of such phrases to concepts in the Metathesaurus, as will be seen in the following section. An example of the type of syntactic structure we assign is given in (2b) for the input in (2a).

(2) a. patients with sustained ventricular tachycardia treated with amiodarone

b. noun_phrase([[head(patients)],
[prep(with),mod(sustained),mod(ventricular),
head(tachycardia),pastpart(treated)],
[prep(with),head(amiodarone)]])

Note in this example that the structure is extremely flat; very little commitment is made to the internal structure of the noun phrase. For example, the past participle *treated* is not assigned a syntactic structure which directly reflects its final interpretation. At the same time, the fact that *tachycardia* is labelled as a head and distinguished from *treated* has important consequences during subsequent processing. We claim that the structural information provided by this analysis contains the optimal amount of information for further processing, namely the mapping of simple noun phrases to concepts in UMLS and the construction of a semantic interpretation.

After all noun phrases have been identified, we map these structures to concepts in the Metathesaurus using a comprehensive mapping program which employs extensive variant generation as well as a principled way of dealing with partial matches between the phrase and Metathesaurus concepts.

Variant generation is determined by the information available from our lexicon and associated knowledge bases. Variants are recursively computed by generating morphological variants, synonyms, acronyms and abbreviations for each lexical word in the input phrase. For example, all variants for the phrase *ocular complications* are listed in (3).

(3) ocular, oculars, oculus, oculi, eyepiece, eyepieces, eye, eyes, eyed, eyeing, eying, optic, optics, optical, optically, vision, ophthalmic, opthalmia, ophthalmiac, ophthalmiacs, complication, complications.

Once variants have been generated for a given phrase, candidate terms from the Metathesaurus are identified. Such candidates for a noun phrase consist of the set of all Metathesaurus terms which contain at least one of the variants computed for the phrase and which satisfy a further condition on partial matches discussed below. The candidates for *ocular complications* appear in (4), where preferred terms are given in parentheses.

(4) “Complications” (“Complication”)
“complications <1>”
“Eye”
“Optic” (“Optics”)
“Ophthalmia” (“Endophthalmitis”)
“Vision”

The final step in the mapping process combines the best candidates to form mappings between the noun phrase and one or more Metathesaurus terms. The degree of similarity between a noun phrase and a Metathesaurus concept is based on factors which take into account how much variation is used to accomplish the match, whether the head is involved, and how much of the Metathesaurus concept and the noun phrase are involved in the match.

This last criterion is based on various types of matches which can occur between a noun phrase and

a Metathesaurus term. In a **simple match** the noun phrase maps to a single Metathesaurus term. For example, the input phrase *intensive care unit* maps to "Intensive Care Units". In a **complex match** there is a partitioning of the noun phrase so that each element of the partition has a simple match to a term in the Metathesaurus. Thus, *intensive care medicine* maps to the two terms "Intensive Care" and "Medicine".

In a **partial match** the noun phrase maps to a Metathesaurus term in such a way that at least one word of either the noun phrase or the Metathesaurus term (or both) does not participate in the mapping. Some examples of partial matches are given in (5).

(5) *liquid crystal thermography* maps to "Thermography"
ambulatory monitoring maps to "Ambulatory Electrocardiographic Monitoring"
obstructive sleep apnea maps to "Obstructive Apnea"

We eliminate partial matches in which both the first and last words of the Metathesaurus term do not participate in the match. This allows *ambulatory monitoring* to map to the Metathesaurus term "Ambulatory Electrocardiographic Monitoring" above, but disallows, for example, *left ventricle* from mapping to the term "Left Ventricular Outflow Obstruction". With regard to the phrase *ocular complications*, this rule eliminates "Postoperative Complications". Mappings which do not satisfy this rule do not constitute the best mapping between noun phrase and Metathesaurus.

In the final determination of the mappings between noun phrase and Metathesaurus term, both less variation and involvement of the head contribute to a stronger match. In general, a simple match represents a stronger mapping between the input phrase and the Metathesaurus term, while complex matches are less strong, and partial matches represent the weakest mapping from input to Metathesaurus. These criteria determine that of the candidate Metathesaurus terms given in (4), those listed in (6) constitute the best map to *ocular complications*.

(6) "Eye"
"Complication", "complications <1>"

RELATIONSHIPS BETWEEN PHRASES: SEMANTIC PROCESSING

Semantic interpretation indicates dependencies among the concepts identified by mapping noun phrases to concepts in the Metathesaurus. We represent these dependencies in a predicate argument structure that we call conceptual structure, which is closely related to logical form (see [22]). The arguments in conceptual structure are labelled with semantic case roles [23] in order to more clearly specify the relationships among the concepts represented. For example, we construct the (simplified) conceptual structure given in (7) to represent the semantic interpretation of *hemofiltration in digoxin overdose.* The case labels on the arguments in (7) indicate that it is digoxin overdose that is being treated through the use of hemofiltration as an instrument.

(7) treat(theme(digoxin overdose),
instr(hemofiltration))

Conceptual structures are built through the application of semantic rules which fall into two major categories. As much as possible we rely on the UMLS Semantic Network [24], since doing so diminishes the number of semantic rules we must write. When application of the Semantic Network is not possible, we appeal to rules which depend crucially on the semantic types obtained from UMLS and which are similar in spirit to those discussed in [21].

In exploiting the Semantic Network for semantic interpretation we match linguistic patterns against corresponding relationships between semantic types in the Network. As an example, consider the text *single corticospinal axons in the cat spinal cord* and note that *axons* has semantic type 'Cell Component', while *spinal cord* is of type 'Body Part, Organ, or Organ Component'. Furthermore, in the Semantic Network these two semantic types are joined by the relation 'part_of' as noted in (8).

(8) part_of('Cell Component',
'Body Part, Organ, or Organ Component')

In order to exploit these facts for semantic interpretation we need only stipulate that the preposition *in* may correspond to the Semantic Network relation 'part_of'. Then, since *single corticospinal axons in the cat spinal cord* contains the preposition *in* and since its semantic types correspond to those in (8) this relationship provides the semantic interpretation (9).

(9) part_of(nom(single corticospinal axons),
theme(cat spinal cord))

For situations in which an interpretation based on the Semantic Network does not apply, we supply rules of semantic interpretation, which crucially depend on the UMLS semantic types. In this regard, the semantic types associated with Metathesaurus concepts can be generalized. For example, *frostbite* has the semantic type 'Injury or Poisoning', while *malaria* is typed as 'Disease or Syndrome'. We collapse these types and

others referring to medically treatable conditions into the generalized type <disorder>.

An example of a domain-specific semantic rule is (10), which states that a noun phrase which is the object of the preposition *for* and whose head has any of the UMLS semantic types covered by the generalized semantic type <disorder> can modify a minimal noun phrase to the left whose head has the semantic type <therapy>. Furthermore, the rule states that in conceptual structure the relationship between the noun phrases is such that the therapy is used as an instrument to treat the disorder. Rule (10) applies to (11a) to produce (11b).

(10) [head(<therapy>)] ,[prep(for), head(<disorder>)] $\longrightarrow$ treat(theme(<disorder>),instr(<therapy>))

(11) a. Electrocoagulation for gastrointestinal hemorrhage.

b. treat(theme(gastrointestinal hemorrhage), instr(electrocoagulation))

EXPLOITING SEMANTIC STRUCTURE FOR INFORMATION RETRIEVAL

In conclusion, we suggest a method of exploiting semantic structure to improve retrieval effectiveness. The example in (12) is constructed to be paradigmatic of one problem associated with the use of either key words or phrases in information retrieval. Title (12b) is relevant to query (12a) while (12c) is not.

(12) a. Query: Intra-carotid injection of drugs for the treatment of malignant gliomas

b. Title1: Intra-carotid BCNU chemotherapy for malignant gliomas

c. Title2: Association of internal carotid aneurysms and temporal glioma

A reasonable translation of (12a) into a Boolean query might be (carotid AND glioma). It is not advisable to include *injection* in the Boolean query, since the concept represented by *intra-carotid injection* could well be represented in text by some form of *infusion, perfusion,* or *chemotherapy,* at least. Given this query it is not possible to reject the nonrelevant (12c).

The use of phrases does not solve the problem, and in fact makes it worse. The Boolean translation of the query using phrases would probably be (intra-carotid injections AND gliomas). This rejects the nonrelevant title, but also rejects the relevant title.

The use of terms from the Metathesaurus alone also does not help. “Injections” (the term for *intra-carotid injection*) does not match “Chemotherapy” (the term for *intra-carotid BCNU chemotherapy*).

A solution based on semantic processing depends on the partial (and simplified) conceptual structures for the query and texts given in (13).

(13) a. treat(theme(malignant gliomas), instr(intra-carotid injection))

b. treat(theme(malignant gliomas), instr(intra-carotid BCNU chemotherapy))

c. co-occurs_with(cotheme(internal carotid aneurysm), cotheme(temporal glioma))

The most important aspect of (13) relevant to the problem under discussion is that the query and the relevant title involve the predicate *treat*. In (13a) and (13b) a concept with the semantic type ‘Disease or Syndrome’ is treated by a concept with semantic type ‘Therapeutic or Preventive Procedure’. A quite different semantic structure has been assigned to the nonrelevant title (13c), in which a ‘Disease or Syndrome’ co-occurs with an ‘Acquired Abnormality’.

These facts based on semantic conceptual structure can be used to improve retrieval precision by including a stipulation on the retrieval mechanism which states that in order for a query to match text, the main predicate in the semantic structure of the query must match the main predicate in the conceptual structure of the text. This requirement eliminates the nonrelevant title (13c) above as a possible match to the query. Once such a requirement has been met, the normal Boolean query can be issued to retrieve the relevant title (13b).

In so far as semantics is able to identify relationships between phrases and thus more precisely represent the content of text, we see this type of processing as showing considerable promise for being able to enhance existing information retrieval techniques based on phrases, whether the phrases are directly identified in text or result from mapping to a controlled vocabulary.

ACKNOWLEDGMENTS

We would like to acknowledge Allen C. Browne, Alexa T. McCray, Amir Razi, and Suresh Srinivasan for their contributions to this project.

References

1. Tersmette KWF, Scott AF, Moore GW, Matheson NW and Miller RE. “Barrier word method for detecting molecular biology multiple word terms.” In Greenes RA (ed) *Proceedings of the 12th Annual SCAMC*, 207-211, 1988.

2. Elkin PL, Cimino JJ, Lowe HJ, Aronow DB, Payne TH, Pincetl PS and Barnett GO. “Mapping to MeSH: The art of trapping MeSH equivalence from within narrative text.” In Greenes RA (ed) *Proceedings of the 12th Annual SCAMC*, 185-190, 1988.

3. Hersh WR, Hickam DH and Leone TJ. “Words, concepts, or both: Optimal indexing units for automated information retrieval.” In Frisse ME (ed) *Proceedings of the 16th Annual SCAMC*, 644-648, 1992.

4. Lin R, Lenert L, Middleton B and Shiffman S. “A free-text processing system to capture physical findings: Canonical phrase identification system (CAPIS).” In Clayton PD (ed) *Proceedings of the 15th Annual SCAMC*, 168-172, 1991.

5. Miller RA, Gieszczykiewicz FM, Vries JK and Cooper GF. “CHARTLINE: Providing bibliographic references relevant to patient charts using the UMLS Metathesaurus knowledge sources.” In Frisse ME (ed) *Proceedings of the 16th Annual SCAMC*, 86-90, 1992.

6. Wagner MM. “An automatic indexing method for medical documents.” In Clayton PD (ed) *Proceedings of the 15th Annual SCAMC*, 1011-1017, 1991.

7. Croft WB, Turtle HR and Lewis DD. “The use of phrases and structured queries in information retrieval.” In Bookstein A, Chiaramella Y, Salton G and Raghavan VV (eds.) *Proceedings of the 14th Annual International ACM/SIGIR Conference*, 32-45, 1991.

8. Evans DA, Ginther-Webster K, Hart M, Lefferts RG and Monarch IA. “Automatic indexing using selective NLP and first-order thesauri.” *RIAO 91*, Autonoma University of Barcelona, April 2-5, 624-44, 1991.

9. Fagan JL. *Experiments in Automatic Phrase Indexing for Document Retrieval: A Comparison of Syntactic and Non-Syntactic Methods*. Cornell University Doctoral Dissertation, 1987.

10. Mauldin ML. *Conceptual Information Retrieval: A Case Study in Adaptive Partial Parsing*. Boston: Kluwer Academic Publishers, 1991.

11. McCray AT. “Extending a Natural Language Parser with UMLS Knowledge.” In Clayton PD (ed.) *Proceedings of the 15th Annual SCAMC*, 194-198, 1991.

12. Metzler DP and Haas SW. “The constituent object parser: Syntactic structure matching for information retrieval.” In Belkin NJ and van Rijsbergen CJ (eds.) *Proceedings of the 12th Annual International ACM/SIGIR Conference*, 117-126, 1989.

13. Smeaton AF and van Rijsbergen CJ. “Experiments on incorporating syntactic processing of user queries into a document retrieval strategy.” In Chiaramella Y (ed.) *Proceedings of the 11th International ACM/SIGIR Conference*, 31-51, 1988.

14. Sparck Jones K and Tait JI. “Automatic search term variant generation.” *Journal of Documentation* 40:50-66, 1984.

15. Croft WB and Das R. “Experiments with query acquisition and use in document retrieval systems.” *Proceedings of the 13th Annual International ACM/SIGIR Conference*, 349-368, 1990.

16. Lewis DD and Croft WB. “Term clustering of syntactic phrases.” In Vidick J-L (ed.) *Proceedings of the 13th Annual International ACM/SIGIR Conference*, 385-404, 1990.

17. Miller PL, Smith P, Morrow JS, Riely CL and Powsner SM. “Semantic relationships and MeSH.” In Greenes RA (ed) *Proceedings of the 12th Annual SCAMC*, 174-179, 1988.

18. Wendlandt EB and Driscoll JR. “Incorporating a semantic analysis into a document retrieval strategy.” In Bookstein A, Chiaramella Y, Salton G and Raghavan VV (eds.) *Proceedings of the 14th Annual International ACM/SIGIR Conference*, 270-279, 1991.

19. Humphreys BL and Lindberg DAB. “The Unified Medical Language System Project: A Distributed Experiment in Improving Access to Biomedical Information.” In Lun KC, Degoulet P, Piemme T and Rienhoff O (eds.) *Proceedings of MEDINFO 92*, 1496-1500, 1992.

20. Salton G and Smith M. “On the application of syntactic methodologies in automatic text analysis.” In Belkin NJ and van Rijsbergen CJ (eds.) *Proceedings of the 12th Annual International ACM/SIGIR Conference*, 137-150, 1989.

21. Agarwal R and Boggess L. “A simple but useful approach to conjunct identification.” *Proceedings, 30th Annual Meeting of the Association for Computational Linguistics*, 15-21, 1992.

22. Allen J. *Natural Language Understanding.* Menlo Park, CA: Benjamin/Cummings, 1987.

23. Fillmore CJ. “The case for case.” In Bach E and Harms R (eds.) *Universals in Linguistic Theory.* New York: Holt, Rinehart, and Winston, 1-90, 1968.

24. McCray AT and Hole WT. “The Scope and Structure of the First Version of the UMLS Semantic Network.” In Miller RA (ed.) *Proceedings of the 14th Annual SCAMC*, 126-130, 1990.