Exploiting a Large Thesaurus for Information Retrieval

Alan R. Aronson
Thomas C. Rindflesch
Allen C. Browne

National Library of Medicine
8600 Rockville Pike
Bethesda, MD 20894

1. Background

Accuracy in information retrieval, that is, achieving both high recall and precision, is challenging because the relationship between natural language and semantic conceptual structure is not straightforward. However, effective retrieval requires that the semantic conceptual structure (or content) of both queries and documents be known. Natural language processing is one way to determine the content of a text. But, due to the complexity involved in natural language processing, various methods have been used which simulate (or approximate) representation of the content of both queries and documents.

One method of approximating the semantic content of a text is single word indexing, which can be enhanced with statistical methods, morphological processing (often stemming), and perhaps some sort of clustering to represent relationships between words. This "words-only" approach has enjoyed considerable success, especially in the vector space model (Salton 1986). However, there is a pervasive view that the method has reached the limits of its effectiveness.

Although natural language processing is difficult, its potential benefits for information retrieval have caused various researchers to investigate the use of both syntactic and semantic processing. Smeaton and van Rijsbergen (1988) and Lewis and Croft (1990), for example, report that the use of syntax in information retrieval shows promise in increasing retrieval effectiveness. In other work, Bonzi and Liddy (1988) investigate the enhancement of statistical techniques with anaphor resolution, while Sager et al. (1993) report favorably on the role of syntactic processing

in accessing medical records.[1] Some studies, however, have not been optimistic (Fagan 1987, for example).

There has also been research concentrating on semantic conceptual representation in information retrieval (Mauldin 1991, and Jacobs and Rau 1990, for example). In the area of biomedical information retrieval, a number of researchers have addressed the notion of incorporating some sort of conceptual processing. Johnson et al. (1993), for example, report on one approach using semantic processing for accessing biomedical text, while Baud et al. (1993) discuss another.

Although both syntactic and semantic processing demonstrate promise for increasing effectiveness in information retrieval, so far neither has been shown to be practical for processing unconstrained text. This is in contrast to the vector space model, which efficiently handles such text. What we propose in this paper is a retrieval methodology which takes advantage of the attractive characteristics of the vector space model, but which enhances its effectiveness through two techniques: a) underspecified syntactic analysis, which, significantly, can accommodate unconstrained text and b) the use of a large thesaurus.

While the use of a thesaurus holds a venerable position in information retrieval (Sparck Jones 1986, Salton and Lesk 1971) there has recently been a renewed interest in its application (see Evans et al. 1991, Hersh and Greenes 1990, and Hersh et al. 1994, for example). Typically, a thesaurus contains information pertaining to paradigmatic semantic relations such as synonymy and is often used for broadening the search term and thus increasing recall. Evans et al. (1991) use a thesaurus for validation of terms. In the context of biomedical information retrieval we propose mapping the text of both queries and documents to terms in the UMLS® Metathesaurus® in order to increase precision in a vector space model.

The Metathesaurus is one component of the National Library of Medicine's Unified Medical Language System® (UMLS) (See Lindberg et al. 1993). The 5th (1994) Experimental Version[2] of the Metathesaurus covers more than 150,000 concepts (including over 300,000 variants and synonyms) drawn from a variety of biomedical vocabularies, including MeSH,® ICD-9-CM, and SNOMED. The Metathesaurus indicates corresponding relationships between terms in the various vocabularies and exploits hierarchical relationships between terms as they exist within a vocabulary. The Metathesaurus provides a wealth of additional information, including the semantic type

1. See Schwartz 1990 for further discussion of syntactic processing in information retrieval.
2. The work described in this paper was based on the 4th (1993) Experimental Version of the Metathesaurus.

for each concept, definitions for many terms from *Dorland's Illustrated Medical Dictionary,* and cooccurrence with other terms in MEDLINE® citations.

We claim that the extensive information available in the Metathesaurus can make a significant contribution to improving retrieval effectiveness. This is disputed by Hersh et al. (1992), however, who report that mapping to the UMLS Metathesaurus provides no advantage in information retrieval. We respond to Hersh et al. by noting the importance of the effectiveness of the mapping technique. At least one other study (Yang and Chute 1993) supports the thesis that effective mapping of text to the Metathesaurus may improve results in information retrieval and suggests a statistical method (linear least squares fit) to accomplish the mapping. We agree with Yang and Chute that the effectiveness of mapping from the language of the texts to the concepts in the thesaurus is crucial for realizing the advantage of using a thesaurus. We differ from them, however, in using an approach which concentrates on symbolic processing based on linguistic analysis. We prefer this approach because it seems more likely that a symbolic method can be improved incrementally and may eventually offer a basis for advanced inferencing methods.

2. The Methodology

2.1 Overview

In the context of the SPECIALIST system (See McCray 1991 and McCray et al. 1993), we propose a method of information retrieval which enhances the vector space model and is crucially based on mapping text to concepts in the Metathesaurus. Significantly, we claim that the processing which supports this mapping is essential for effective retrieval. This processing provides intense variant generation, including abbreviation expansion, inflectional and derivational morphology, and the determination of synonymy relations, as well as a principled way of dealing with partial mappings. In addition, an important aspect is underspecified syntactic analysis, which constrains the mapping to the Metathesaurus.

Strings of text which map to Metathesaurus concepts must occur within the boundaries of a syntactic unit. The most important syntactic unit for these purposes is the simple noun phrase (that is the noun phrase without relative clauses or post-modifying prepositional phrases). An underspecified analysis which identifies simple noun phrases appears to be wholly adequate for

supporting mapping to the Metathesaurus. To employ a more elaborate analysis would be needlessly costly.

Our system shares a number of characteristics with CLARIT (Evans et al. 1991) and SAPHIRE (Hersh and Greenes 1990). However, the particular combination of characteristics is innovative. Although CLARIT uses syntactic analysis and a thesaurus, the knowledge source it uses is not as rich as the UMLS Metathesaurus. Although SAPHIRE exploits the Metathesaurus, it does not use the same mapping procedure we do, nor does it use syntactic analysis.

Figure 1 provides an overview of the way our methodology can enhance the vector space model. Input text is first processed with underspecified syntactic analysis and is then mapped to the Metathesaurus. The vector space model then accepts the resulting text, enhanced with Metathesaurus concepts. We have tested this methodology on the UMLS Test Collection (Schuyler et al. 1989) using the SMART information retrieval system (see Salton 1991) and have found that this methodology contributes to enhanced precision.

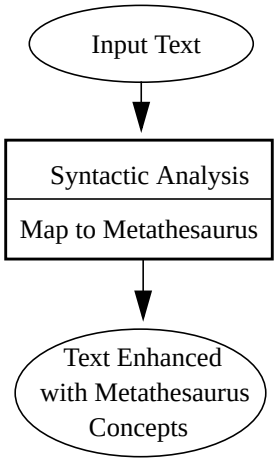

Figure 1. System Overview

For the remainder of this paper we first briefly describe the underspecified syntactic analysis we use and then discuss in some detail the methodology for mapping to the Metathesaurus. We conclude with the results of testing the system with SMART.

2.2 Syntactic analysis

Syntactic processing is supported by a large lexicon, containing over 60,000 entries with syntactic information (see Browne et al. 1993). We also rely on the Xerox stochastic part-of-speech tagger (Cutting et al. 1992). Getting the part-of-speech labels from the tagger allows the syntactic processor to be more efficient and contributes to the overall accuracy of the information retrieval process.

Our syntactic analysis concentrates on identifying simple noun phrases, that is, noun phrases in which the head is the rightmost element and which thus have no right modification.

Informally, the algorithm we use for assigning syntactic structure can be thought of as a series of filters which bring the structure into clearer and clearer focus and proceeds in two steps: a) marking simple noun phrase boundaries within a larger structure and b) applying labelling rules to identify heads and modifiers within each simple noun phrase.

In a successful syntactic analysis, heads are identified and items to the left of the head are simply labelled as "modifier". For example, the text in (1) is given the analysis in (2), where prepositional phrases (PP) and simple noun phrases are identified.

(1) Responsiveness to epidermal growth factor of human embryonic mesenchyma cells of palate by persistent rubella virus infection

(2) a. [head(responsiveness)]$_{NP}$
b. [prep(to), [mod(epidermal), mod(growth), head(factor)]$_{NP}$]$_{PP}$
c. [prep(of), [mod(human), mod(embryonic), mod(mesenchyma), head(cells)]$_{NP}$]$_{PP}$
d. [prep(of), [head(palate)]$_{NP}$]$_{PP}$
e. [prep(by), [mod(persistent), mod(rubella), mod(virus), head(infection)]$_{NP}$]$_{PP}$

The structure we impose on noun phrases is underspecified in the sense that detailed internal structure is not provided beyond the identification of the head of the structure along with all of the modifiers in the noun phrase which occur to the left of the head. This is almost exactly the approach taken by Evans et al. (1991). A similar approach is used by Greffenstette (1992) and Agarwal and Boggess (1992). Mauldin (1991) also uses an underspecified linguistic analysis, although of a somewhat different type from that used here. Other researchers use linguistic analy-

sis which is more fully specified, but yet, which provides an underspecified analysis under certain circumstances. See Strzalkowski and Vauthey 1992 and Jacobs and Rau 1990, for example.

2.3 Mapping simple noun phrases to concepts in the UMLS Metathesaurus

After all simple 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. It is important to recall that the mapping to the Metathesaurus occurs within the bounds of a noun phrase. That is, a Metathesaurus concept cannot cross a noun phrase boundary.

The process of mapping simple noun phrases to concepts in the Metathesaurus consists of generating variants of words in the phrase, finding all candidate concepts which contain a variant, computing a similarity value for each candidate, and combining one or more of the best candidates into a coherent interpretation. For example, the text *mineralize* generates variants *minerals* and *mineralization* (among others). The candidate concepts from the Metathesaurus which contain these variants are “Minerals” and “Mineralization”. The mapping algorithm determines that of these candidates “Mineralization” constitutes the best interpretation of *mineralize.*

2.3.1 Variant generation

Variant generation is determined by the knowledge available from our lexicon and knowledge bases of synonyms and derivational morphological rules. The variant generation algorithm described here is knowledge intensive and uses the following knowledge bases:[3]

- The SPECIALIST lexicon for determining spelling variations, abbreviations, acronyms, and inflectional variations;
- two knowledge bases of synonyms: one obtained by extracting synonyms from *Dorland’s Illustrated Medical Dictionary,* and an additional synonym knowledge base developed for use with SPECIALIST; and
- a knowledge base containing rules of derivational morphology.

3. Our variant generation is much simpler than that of Sparck Jones and Tait (1984). This is because our variants are only an aid to mapping from input to concepts in the domain model and are not directly used for matching queries to documents.

Variants are generated for each head and modifier as determined by syntactic analysis and include morphological variants, synonyms, acronyms and abbreviations for subsequences of words in the noun phrase. A distance value is computed to determine how much each variant deviates from the input form. Spelling and inflectional variants deviate less than synonyms, while derivational variants have the highest distance value. The results are filtered at each step using the lexicon.

For example the variants generated for input *chemical*s are given in (3) where (3a) is an inflectional variant; the first item in (3b) is an abbreviation, followed by its inflections; (3c) contains a derivational variant of *chemicals* along with its plural and (3d) and (3e) are derivational variants.

(3) a. chemical
b. chem, chems, chem's
c. chemist, chemists
d. chemically
e. chemistry

2.3.2 Metathesaurus candidates

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.

For the phrase *bone mineral density studies* the syntactic structure is given in (4) and examples of the variants are given in (5).

(4) [mod(bone), mod(mineral), mod(density),head(studies)]

(5) bone, bones, boned, boning, bony, bonier, boniest, os, ossa,
mineral, minerals, mineralisation, mineralization, mineralise, mineralize,
density, densities, dense, denseness
studies, study, studying, studious

Some candidate terms from the Metathesaurus which contain at least one of the variants are given in (6), where preferred terms are given in parentheses.

(6) “Bone Mineral Density” (“Bone Density”)
“Bone Density”
“Bone Mineralization” (“Calcification, Physiologic”)
“Bone” (“Bones”)
“Minerals”
“Mineralization”

2.3.3 Mapping between phrase and Metathesaurus terms

The final step in the mapping process combines the best candidate Metathesaurus terms to form mappings between the noun phrase and one or more Metathesaurus terms. The best candidate is determined by the degree of similarity between the noun phrase and the Metathesaurus concept, where the highest degree of similarity exists in an **exact match** in which an entire input phrase matches exactly (ignoring upper and lower case differences) to one Metathesaurus concept: *intensive care units* maps to “Intensive Care Units”. A lesser degree of similarity between a noun phrase and a 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 concept and the noun phrase are involved in the match.

In addition to an exact match, other types of matches can occur between a noun phrase and a Metathesaurus term. In a **simple match** the noun phrase maps to a single Metathesaurus term, although with some variation. For example, the input phrase *carotid artery* maps to “Carotid Arteries”. 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, *acidotic dogs* maps to the two terms “Acidosis” and “Dogs”.

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 (7).

(7) *synthetic organic chemical* maps to “Organic Chemicals”
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”. Mappings which do not satisfy this rule do not constitute the best mapping between noun phrase and Metathesaurus.

For candidates which do not constitute an exact match, choosing the best match is based on the degree of similarity between the noun phrase and Metathesaurus concepts. Similarity is computed by a comparison metric based on four components: centrality, variation, coverage, and cohesiveness. A normalized value between 0 and 1 is computed for each of these components. These values are combined in a weighted average in which the coverage and cohesiveness components receive twice the weight as the centrality and variation components. Each of the comparison metric components is discussed below.

The centrality value is simply 1 if the Metathesaurus concept involves the head of the phrase and 0 otherwise. For example, “Bone Mineralization”, a candidate concept for the phrase *bone mineral density studies*, receives a centrality score of 0 since it does not involve *studies*, the head of the phrase.

Variation measures the degree to which variants in the Metathesaurus concept differ from the corresponding words in the noun phrase. It is computed by first determining the “variation distance” for each variant in the Metathesaurus concept. This distance is the sum of the distance values for each step taken during variant generation. The values for each step are determined by the type of variant and are, in order of increasing distance, spelling variation, inflectional variation, synonymy and derivational variation. Of the two candidate concepts “Minerals” and “Mineralization”, the first receives a better variation score for the text *bone mineral density studies.* This is because “Mineralization”, a derivational variant, is considered to be more “distant” from *mineral* than “Minerals”, an inflectional variant.

Coverage indicates how much of the Metathesaurus concept and the noun phrase are involved in the match. For example, of the two Metathesaurus concepts “Bone Mineralization” and “Bone Density” which are candidates for mapping to the noun phrase *bone mineral density studies*, the

second gets a higher coverage value because it spans more of the input phrase (*bone mineral density*) than does the first candidate (*bone mineral*).

The cohesiveness value is similar to the coverage value but emphasizes the importance of connected components. Here, gaps in coverage are undesirable. Using the same example as above, "Bone Mineralization" receives a better cohesiveness score than does "Bone Density" as a mapping for the phrase *bone mineral density studies.* This is so because the former candidate maps to the cohesive text *bone mineral*, while the latter maps to the discontinuous text *bone ... density.*

In the final determination of the mappings between noun phrase and Metathesaurus concept, both less variation and involvement of the head contribute to a stronger match. High coverage and cohesiveness are favored, with coverage taking precedence over cohesiveness. 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 conspire to determine that of the candidate Metathesaurus terms for the phrase *bone mineral density studies* given above the best match is "Bone Mineral Density".

3. Assessing the effectiveness of the methodology

In order to determine whether underspecified syntactic analysis and mapping to the UMLS Metathesaurus could enhance the vector space model we used the SMART system and the UMLS Test Collection. We used our system to create a surrogate text from the original Test Collection. This surrogate, rather than the original text, then served as input to SMART.

3.1 The UMLS Test Collection

The UMLS Test Collection is a corpus of about 750,000 words consisting of 150 queries and 3,000 documents (approximately 25,000 major syntactic structures: sentences and complex noun phrases). The documents are MEDLINE citations (containing title, authors, abstracts, and MeSH indexing terms) in three subject categories: clinical medicine research, health sciences research, and basic science research. The queries are transcripts of requests for bibliographic information from a variety of biomedical sources and are in the language of the original requester of information. The queries range from straightforward (8) to more elaborate (9).

(8) I am looking for any information I can get on the complement system in dogs.

(9) We have a most interesting patient who has Hodgkin's disease and has presented with a liver abscess due to Nocardia species! Request search for papers detailing infections, specifically liver abscesses, in patients with Hodgkin's disease; spectrum of clinical illness infections due to Nocardia sp.; infections on patients with Chronic Granulomatous disease.

The 150 queries are divided into approximately 50 queries for each of the three subject areas noted. The collection was created by an expert searcher translating the original user's request into a formal Boolean query composed primarily of the key words from the MeSH vocabulary and then conducting a search on the MeSH indexing terms for a subset of MEDLINE citations. The precision of these searches was determined to be about 65% by a domain expert who examined (and marked as relevant or nonrelevant) the citations retrieved by each of the 150 searches.

3.2 The surrogate text

For each query and citation in the Test Collection we produced a surrogate text by replacing phrases (or parts of phrases) in the original text with their corresponding Metathesaurus concept. Any phrase or phrase component which did not have a mapping was left in the text. (A similar method is used in Hersh et al. 1992.) An example surrogate text is given in (11) for the input text (10), which is a query. (Metathesaurus concepts are capitalized and underlined in the surrogate).

(10) Input text—Please do literature search for any relationship between chloroquine and low blood pressure in people with pre-existing hypertension. Also possible interaction with diuretics to exaggerate hypotensive effect.

(11) Surrogate text—Please do Literature search for any relationship between Chloroquine and Hypotension in people with pre-existing Hypertension. Also possible interaction with Diuretics to exaggerate Hypotension effect.

Note especially that two phrases in the input text, *low blood pressure* and *hypotensive* map to the concept "Hypotension". This fact indicates one way in which mapping to the Metathesaurus contributes to increased precision. When SMART processes (11) as a query it will not consider documents pertaining to blood poisoning or blood culture relevant, as it would have done when processing (10) as a query.

The texts in (12) and (13) provide a further example of the positive effect of Metathesaurus synonyms.

(12) Input text—Plasma cell dyscrasias.

(13) Surrogate text— Paraproteinemias.

In addition to “Plasma Cell Dyscrasias”, other synonyms of “Paraproteinemias” are “Monoclonal Gammopathies” and “Paraimmunoglobulinemias”. Consequently any of these terms occuring in text would map to “Paraproteinemias” and thus documents containing any of these terms would be retrieved by a query containing any other.

3.3 Results

When the parts of the input text corresponding to Metathesaurus concepts have been replaced with the concepts, SMART operates normally on the new surrogate text. The benefits of having both text and Metathesaurus concepts can be seen in Figure 1 which shows recall/precision curves

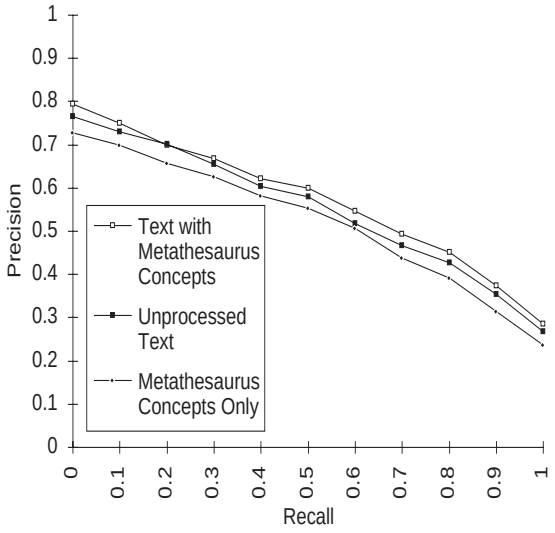

Figure 2. Recall/Precision Curves for the NLM Test Collection

produced by SMART running on three versions of the Test Collection. In addition to the original text and the text enhanced with Metathesaurus concepts the third text contains only the Meta-

thesaurus concepts which resulted from the mapping process (but not the original text which failed to map).

The curve labelled “Metathesaurus Concepts Only” refers to a surrogate text which excludes the original text failing to map to the Metathesaurus. Average precision for that text is lower than that for the unprocessed text. This contrasts with the curve for the surrogate text which includes both text and Metathesaurus concepts (labelled “Text with Metathesaurus Concepts”), which is better than the unprocessed text.

The increase of average precision for our method over use of the plain text is about 4%. While the average precision figure which we achieve is so far not dramatically better than that attainable with unprocessed text, the results are nevertheless promising.

3.4 Improving the methodology

Precision could no doubt be increased by correcting mapping errors. For those simple noun phrases which map, at least partially, to a Metathesaurus concept, the strategy just described chooses the correct concept around 90% of the time. By correct concept we mean that the concept chosen is an appropriate mapping of the text in the given context. In this sample, incorrect mappings to Metathesaurus concepts fall into two general categories: those caused by variant generation and those caused by our failure to resolve ambiguity.

Problems in variant generation can be due to morphology, acronym expansion, or synonym expansion. For example, in the mapping for the phrase *hard physical work* morphological variant generation causes *hard* to match the concept “Hardness”.

Errors due to acronym expansion often stem from an analysis which is incorrect for other reasons. For example, in the phrase *Le Fort I osteotomy,* the term *Le Fort* does not occur in the Metathesaurus, nor does it occur in any of our knowledge bases. We therefore treat it as two tokens *le* and *fort*. *Le* occurs in our acronym knowledge base as an abbreviation for *lupus erythematosus*, which maps to the corresponding Metathesaurus concept “Lupus Erythematosus”. Although this can be solved by adding *Le Fort* to our lexicon, it is unlikely that we will ever have complete lists. A general solution to such problems is needed.

While generally valuable, our robust generation of synonyms occasionally leads to error. For example, *ventricle* as such does not occur in the Metathesaurus. In one of our synonym lists *ventricle* is listed as a synonym of *ventriculus,* which is also a synonym of *stomach*. We thus map the string *ventricle* to the concept “Stomach” regardless of the context in which it occurs.

There are a number of terms in the Metathesaurus, such as “Ventilation”, which are ambiguous. We so far do not disambiguate these terms given the context in which they appear. Other terms are not ambiguous in the Metathesaurus but map to words which are ambiguous in English. We will thus have the wrong concept when such terms occur in contexts other than the one specified in the Metathesaurus. An example of such a term is “Conditioning”, which has only the psychology denotation in the Metathesaurus.

In order to resolve the infelicitous and ambiguous mappings under discussion, we are currently pursuing research based on distribution patterns of semantic types which occur in text. Semantic types in UMLS are features such as ‘Disease or Syndrome’, ‘Diagnostic Procedure’, and ‘Anatomical Structure’ which indicate the semantic content of each Metathesaurus concept with which they are associated. Preliminary results indicate that these distribution patterns can be exploited with statistical techniques to solve at least some of the problems mentioned.

The general way in which such a solution might work can be seen by referring to the example noted above in which the *le* of *Le Fort I Osteotomy* mapped infelicitously to “Lupus Erythematosus” (which has semantic type ‘Disease or Syndrome’). If the proposed mapping were carried out, “Osteotomy” (with semantic type ‘Therapeutic or Preventive Procedure’) would occur as the head of a noun phrase having a modifier with semantic type ‘Disease or Syndrome’. If statistical methods can determine that this pattern rarely (or never) occurs, then the mapping would be disallowed, and *Le Fort* would be left as is in the text.

4. Conclusion

The results obtained by submitting a moderately large test collection to the SMART system indicate that underspecified syntactic processing and effective mapping of text to concepts in the UMLS Metathesaurus have a positive effect on the vector space model. Although a complete semantic conceptual representation would be ideal for representing the content of text for information retrieval, it is not currently possible to provide such a representation. The surrogate text enhanced with Metathesaurus concepts which our system produces appears to provide a representation closer to the ideal than is possible with unprocessed text.

There are several additional reasons why mapping to the Metathesaurus is significant for information retrieval. There is a great deal more information available in the Metathesaurus than is available in traditional thesauri, which typically concentrate on synonymy information. Once

mapping to the Metathesaurus has been accomplished this information can be exploited. For example, hierarchical relationships between concepts (such as “isa”) are provided for many concepts. Strzalkowski and Vauthey (1992) explore the advantages these can provide for information retrieval and they describe a method of computing them from text. This information is available directly from the Metathesaurus. With regard to the cooccurrence of terms with other terms in MEDLINE citations, Harbourt et al. (1993) describe a system which exploits this information. We have suggested above that the semantic types may be valuable in resolving mapping ambiguities.

We would also claim that our method, which employs linguistic analysis along with mapping to the Metathesaurus provides an advantage over methods which do not involve linguistic analysis. It seems quite likely that semantic conceptual structure, based on linguistic processing, will eventually be needed to gain a significantly deeper understanding of free text.[4] This deeper understanding is almost certainly required in order to support advanced processing such as inferencing and question answering.

Finally, we comment on the relationship between statistical methods and symbolic processing in information retrieval. We do not see any antagonism between the two approaches. Rather, we would like to suggest that a valuable symbiosis is possible and desirable between them. Specifically, we claim that a surrogate text enhanced with concepts can improve any of the statistical methods. We have tested this hypothesis with a traditional vector space model (SMART) and have found that the enhanced text does in fact achieve better results than the plain text alone. While we have not so far tested this method with other statistical models it seems reasonable to assume that using a probabilistic model (see Belkin and Croft 1992) or latent semantic indexing model (see Deerwester et al. 1990) on our enhanced text would also produce results better than those achieved with the statistical model and plain text.

Acknowledgments

We would like to acknowledge Alexa T. McCray, Amir Razi, and Suresh Srinivasan for their contributions to this research.

4. In Rindflesch and Aronson 1993 we discuss an extension to the system described here which produces semantic conceptual structure.

References

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

Baud, Robert; Christian Lovis; Laurence Alpay; Anne-Marie Rassinoux; Jean-Raoul Sherrer; Anthony Nowlan; and Alan Rector (1993). “Modelling for natural language understanding”. Charles Safran (ed.) *Proceedings of the 17th Annual Symposium on Computer Applications in Medical Care*, 289-293.

Belkin, Nicholas J., and W. Bruce Croft (1992). “Information Filtering and Information Retrieval: Two Sides of the Same Coin?” *Communications of the ACM* 35(12):29-38.

Bonzi, Susan, and Elizabeth Liddy (1988). “The use of anaphoric resolution for document descriptions in information retrieval”. Yves Chiaramella (ed.) *11th International Conference on Research and Development in Information Retrieval*, 53-65.

Browne, Allen C.; Alexa T. McCray; and Suresh Srinivasan (1993). *The SPECIALIST Lexicon*. National Library of Medicine, Report No. NLM-LHC-93-01. (Available from NTIS, Springfield VA: PB93-217248).

Cutting, D.; J. Kupiec; J. Pedersen; and P. Sibun (1992). “A practical part-of-speech tagger”. *Proceedings of the Third Conference on Applied Natural Language Processing*.

Deerwester, Scott; Susan T. Dumais; George W. Furnas; Thomas K. Landauer; and Richard Harshman (1990). “Indexing by latent semantic analysis”. *Journal of the American Society for Information Science* 41:391-407.

Evans, David A.; Kimberly Ginther-Webster; Mary Hart; Robert G. Lefferts; and Ira A. Monarch (1991). “Automatic indexing using selective NLP and first-order thesauri”. *RIAO 91*, 624-44.

Fagan, Joel L. (1987). *Experiments in Automated Phrase Indexing for Document Retrieval: A Comparison of Syntactic and Non-Syntactic Methods*. Cornell University doctoral dissertation.

Greffenstette, Gregory. 1992. “Use of syntactic context to produce term association lists for text retrieval”. Nicholas Belkin, Peter Ingwersen, and Annelise Mark Pejtersen (eds) *Proceed-*

ings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 89-97.

Harbourt, Anna M.; Edmund J. Syed; William T. Hole; and Lawrence C. Kingsland, III (1993). "The ranking algorithm of the Coach browser for the UMLS Metathesaurus". Charles Safran (ed.) *Proceedings of the 17th Annual Symposium on Computer Applications in Medical Care*, 720-724.

Hersh, William R., and Robert A. Greenes (1990). "SAPHIRE—An information retrieval system featuring concept matching, automatic indexing, probabilistic retrieval, and hierarchical relationships". *Computers and Biomedical Research* 23:410-425.

Hersh, William R.; David H. Hickam; R. Brian Haynes; and K. Ann McKibbon (1994). "A performance and failure analysis of SAPHIRE with a MEDLINE test collection". *Journal of the American Medical Informatics Association* 1:51-60.

Hersh, William R.; David D. Hickam; and T. J. Leone (1992). "Words, concepts, or both: Optimal indexing units for automated information retrieval". Mark E. Frisse (ed.) *Proceedings of the 16th Annual Symposium on Computer Applications in Medical Care*, 644-648.

Jacobs, Paul S., and Lisa F. Rau (1990). "SCISOR: Extracting information from on-line news". *Communications of the ACM* 33(11):88-97.

Johnson, Stephen B.; Anthony Aguirre; Ping Peng; and James Cimino (1993). "Interpreting natural language queries using the UMLS". Charles Safran (ed.) *Proceedings of the 17th Annual Symposium on Computer Applications in Medical Care*, 294-298.

Lewis, David D., and W. Bruce Croft (1990). "Term clustering of syntactic phrases". Jean-Luc Vicick (ed.) *13th International Conference on Research and Development in Information Retrieval*, 385-404.

Lindberg, Donald A. B.; Betsy L. Humphreys; and Alexa T. McCray (1993). "The Unified Medical Language System". *Methods of Information in Medicine* 32:281-291.

Mauldin, Michael L. (1991). *Conceptual Information Retrieval: A Case Study in Adaptive Partial Parsing*. Boston: Kluwer Academic Publishers.

McCray, Alexa T. (1991). “Extending a Natural Language Parser with UMLS Knowledge”. Paul D. Clayton (ed.) *Proceedings of the 15th Annual Symposium on Computer Applications in Medical Care*, 194-198.

McCray, Alexa T.; Alan R. Aronson; Allen C. Browne; Thomas C. Rindflesch; Amir Razi; and Suresh Srinivasan (1993). “UMLS knowledge for biomedical language processing”. *Bulletin of the Medical Library Association* 81:184-194.

Rindflesch, Thomas C., and Alan R. Aronson. (1993). “Semantic processing in information retrieval.” Charles Safran (ed.) *Proceedings of the 17th Annual Symposium on Computer Applications in Medical Care*, 611-615.

Sager, Naomi; Margaret Lyman; Leo J. Tick. Ngo Than Nhan; and Christine E. Bucknall (1993). Charles Safran (ed.) *Proceedings of the 17th Annual Symposium on Computer Applications in Medical Care*, 265-268.

Salton, Gerard (1991). “Development in automatic text retrieval”. *Science* 253:974-980.

Salton, Gerard (1986). “Recent trends in automatic information retrieval”. F. Rabitti (ed.) *Proceedings of the ACM Conference on Research and Development in Information Retrieval*, 1-10.

Salton, Gerard, and C. Buckley (1991). Global text matching for information retrieval. *Science* 253:1012-1015.

Salton, Gerard, and M. E. Lesk (1971). “Information analysis and dictionary construction”. Gerard Salton (ed.) *The SMART retrieval system: Experiments in automatic document processing*, 115-142. Englewood Cliffs, NH: Prentice-Hall, Inc.

Schuyler, Peri L.; Alexa T. McCray; and Harold M. Schoolman (1989). “A test collection for experimentation in bibliographic retrieval”. B. Barber, D. Cao, D. Qin, G. Wagner (eds.) *MEDINFO 89*, 810-912. Amsterdam: North-Holland.

Schwartz, Christoph (1990). “Automatic syntactic analysis of free text”. *Journal of the American Society for Information Science* 41:408-417.

Smeaton, A. F., and C. J. van Rijsbergen (1988). “Experiments on incorporating syntactic processing of user queries into a document retrieval strategy”. *Proceedings, 11th International Conference on Research & Development in Information Retrieval*, 31-51.

Sparck Jones, Karen (1986). *Synonymy and semantic classification*. Edinburgh University Press.

Sparck Jones, K., and J.I. Tait (1984). “Automatic Search Term Variant Generation”. *Journal of Documentation* 40:50-66.

Strzalkowski, Tomek, and Barbara Vauthey (1992). “Information retrieval using robust natural language processing”. *Proceedings, 30th Annual Meeting of the Association for Computational Linguistics*, 104-111.

Yang, Yiming, and Christopher G. Chute (1993). “Words or concepts: The features of indexing units and their optimal use in information retrieval”. Charles Safran (ed.) *Proceedings of the 17th Annual Symposium on Computer Applications in Medical Care*, 685-689.