Assessing Thesaurus-Based Query Expansion Using the UMLS Metathesaurus

William Hersh, M.D., Susan Price, M.D., Larry Donohoe, M.L.I.S.
Division of Medical Informatics & Outcomes Research
Oregon Health Sciences University
Portland, Oregon, USA

Objectives: *Assess query expansion using thesaurus relationships and definitions in the UMLS Metathesaurus for improving searching performance.*
Methods: *The queries from a MEDLINE test collection (OHSUMED) were expanded using synonym, hierarchical, and related term information as well as term definitions from the UMLS Metathesaurus. Documents were retrieved from a word-statistical retrieval system and assessed for recall and precision based on relevance judgments from the test collection.*
Results: *All types of query expansion degraded aggregate retrieval performance as measured by recall and precision, although 38.6% of the queries with synonym expansion and up to 29.7% of the queries with hierarchical expansion showed improvement.*
Conclusions: *Thesaurus-based query expansion causes a decline in retrieval performance generally but improves it in specific instances. Further research must focus on identifying instances where performance improves and how it can be exploited by real users.*

Introduction

Although the major goal of the UMLS Metathesaurus is to provide linkages among different vocabularies [1], a secondary goal is to expand terminology for applications such as information retrieval (IR) systems where searchers may use different terms than writers of the documents they desire to retrieve [2]. The mismatch may be due to *synonymy*, where different terms have the same meaning (e.g., hypertension vs. high blood pressure), or *granularity*, where terms are used at different levels of detail (e.g., antibiotic therapy vs. penicillin). Thus the development of a comprehensive resource of concepts with synonym forms and semantic relationships could benefit retrieval systems.

In the past, investigators have attempted to improve information retrieval systems using the UMLS Metathesaurus. We have done extensive work attempting to provide automated indexing of concepts from the Metathesaurus in the SAPHIRE system [3, 4], showing benefit in cases of individual queries but not in the aggregate [5-7]. One benefit the Metathesaurus might provide is the ability to expand user queries to include additional terminology that is synonymous or different-grained.

Query expansion as an IR technique has been investigated for several decades. Its use has usually been in association with "word-statistical" (also called "automated indexing") techniques which were pioneered by Salton in the 1960s [8]. In these systems, queries and documents are viewed as mutli-dimensional vectors, where each term in the query and document represent a dimension in vector space. Queries are entered in "natural language," though the use of Boolean operators typical of traditional IR systems is not precluded. Retrieved documents are ordered by relevance ranking, where documents are ranked based on their similarity (usually with the vector dot product or cosine similarity). Processes like query expansion are straightforward in these types of systems, as the query is expanded with new terms (and/or reweighting of existing terms) in an attempt to find more documents that are relevant to the user's information needs.

Query expansion methods based on word-statistical approaches have been shown to be effective in the TREC experiments, which focus on government and newswire documents [9]. Significant improvements have been seen by adding terms to queries from documents ranked high by relevance ranking in the initial query [10-12]. These added terms have been shown to increase the number of relevant documents at the top of the list of retrieved documents. Srinivasan has shown effectiveness in the medical domain using a variant of this approach to expand MEDLINE queries with MeSH terms that occur in top-ranking documents as well [13].

The major goal of this study was to test whether query expansions using thesaurus relationships in the UMLS Metathesaurus could improve searching performance. Some operational IR systems already use the Metathesaurus in this capacity, although none has been formally evaluated. The PubMed

1067—5027/00/$5.00 © 2000 AMIA, Inc.

(http://ncbi.nlm.nih.gov/PubMed/) and Internet Grateful Med (http://igm.nlm.nih.gov/) systems of the National Library of Medicine (NLM) use automated explosions (expansion of the narrower-than hierarchy) of MeSH terms, while the Medical World Search Engine [14] uses both explosions and synonym expansions.

Thesauri can have three types of relationships: synonym, hierarchical, and related. The first category denotes equivalence, i.e., the different terms represent the same underlying concept, such as cancer vs. carcinoma. The second indicates a broader/narrower classification, in that the child term has an is-a relationship with the parent, e.g., Angiotensin Converting Enzyme Inhibitors vs. Captopril. The final category indicates some other type relationship that has been deemed important, e.g., hypertension vs. antihypertensive agents. Another part of most thesauri is a concept definition, and we also examined this component for query expansion.

Experiments were done with the OHSUMED test collection that is based on 106 queries and nearly 350,000 MEDLINE references from clinical journals over the 1987-1991 time period [15]. Performance was measured by recall (the proportion of relevant documents in the collection retrieved by a query) and precision (the proportion of relevant documents returned by the query). As in many evaluations using word-statistical systems that rank retrieval output, the "11-point average" was used as a composite performance measure. This measure gives precision at fixed points of recall (<10%, 10%, 20%, etc. up to 100%) from the ranked list.

Methods

In this experiment we assessed whether expanding queries using terms from thesaurus relationships and definitions could enhance retrieval performance. All experiments were done using the OHSUMED test collection (available from ftp://medir.ohsu.edu/pub/ohsumed/) and the SMART retrieval system, version 11.0 (available from ftp://ftp.cs.cornell.edu/pub/smart/). The standard SMART recall-precision routines were used to measure performance.

In order to assess expansion of terms using Metathesaurus relationships, we developed a set of manually selected Metathesaurus concepts for each query. Queries were submitted to the SAPHIRE system for suggested Metathesaurus terms [4]. SAPHIRE breaks out phrases occurring between stop words before sending them to the concept-matching algorithm. Queries with parsimonious (i.e., complete but non-overlapping) coverage of all medical concepts by Metathesaurus terms were left as is, while those with incomplete or multiple matches were manipulated manually to obtain as close a set as possible of parsimonious concepts. The 106 queries yielded 298 terms.

An initial baseline (Run 1) was established with the OHSUMED queries using the previously established best weighting scheme [15] (see example in Figure 1). Run 2 consisted of the same run using the manually assigned Metathesaurus terms assigned as described in the previous paragraph (see Figure 2). Synonymy was assessed by adding all possible synonym variants of a term to the query. This was done by adding all words that occurred in the Metathesaurus word file (MRXW) for each Metathesaurus concept of each term, comprising run 3 (see Figure 3). Hierarchical relationships were evaluated by following the links of parent and child terms in the Metathesaurus MRREL file. Thus, terms from any vocabulary from the concept were added to the query. Expansion by adding one level of children terms was done for run 4 (see Figure 4), while expansion all the way to the "bottom" of the children hierarchy was done for run 5. Runs 6 and 7 consisted of one-level and all-level expansions for parent terms respectively (see Figure 5). Related terms were assessed by adding terms designated as related by the RO relationship in the MRREL file, comprising run 8 (see Figure 6). (We also assessed the RL relationship in this file, another related term designator in MRREL, but only 2 of the 298 terms could be expanded.) We also assessed whether text from the term definition could be used to expand queries effectively. For every concept that had a definition, its text was added to the text of the query for run 9 (see Figure 7).

The experimental measures used in this study were recall and precision. For each run, a recall-precision table was obtained that yielded precision at fixed points of recall. The performance measure for each query was the 11-point average precision at each point of recall. We also measured recall and precision at 30 documents, a standard approach in IR evaluation which gives a point value of recall and precision for a number of documents that users are willing to look at after a search. The performance measures for each run were the 11-point average precision, recall at 30 documents, and precision at 30 documents for each query in the run. Repeated measures analysis of variance was used to assess statistical significance, with Scheffe tests used to compare statistical significance between baseline and subsequent runs for the three measures.

Figure 1 – Sample query from OHSUMED test collection.

Acute tubular necrosis due to aminoglycosides, contrast dye, outcome and treatment

Figure 2 – Sample query with parsimonious Metathesaurus terms (and corresponding concepts) grouped by SAPHIRE phrase.

acute tubular necrosis due
    L0085410| Acute tubular necrosis| C0022672|
    Kidney Tubular Necrosis, Acute
aminoglycosides
    L0002556| Aminoglycosides| C0002556|
Aminoglycosides
contrast dye
    L0116993| contrast| C0110625| contrast
    L0013343| Dyes| C0013343| Dyes
outcome
treatment
    L0040807| Treatment| C0087111| Treatment <1>

Figure 3 – Sample query with synonym expansion for concept Acute Tubular Necrosis (first 10 words out of 18).

acute
atn
failure
ischemic
kidney
lesion
lower
necrosis
nephron
nephropathy

Figure 4 – Sample query with one-level child expansion for concept Aminoglycosides (first 10 terms out of 24).

Aminoglycosides
Amikacin
Amikacin Sulfate
Butirosin Sulfate
Framycetin
Genticin
Gentamicins
Hygromycin B
Kanamycin
Kantrex

Figure 5 – Sample query with one-level parent expansion for concepts Contrast and Dye.

contrast
Dyes
Indicators and Reagents
Miscellaneous Drugs and Agents
Industrial product NOS

Figure 6 – Sample query with related term expansion for concept Treatment (first 10 terms out of 35).

Treatment
Aftercare
Ambulatory Care
Cost Control
Counseling
Crisis Intervention
Day Care
Delivery of Health Care
Euthanasia
Preventive Medicine

Figure 7 – Sample query with definition expansion for concept Acute Tubular Necrosis.

Acute kidney failure resulting from destruction of tubular epithelial cells. It is commonly attributed to exposure to toxic agents or renal ischemia following severe trauma.

Results

Table 1 shows the results from runs 1-9. The first column of numbers lists the 11-point average. The next two columns list recall and precision for the top 30 documents retrieved by each query. These values give an operational assessment of the average performance for each run. The final column of numbers lists the number of queries where the expansion method improved the 11-point average over baseline.

While the manually designated Metathesaurus terms (run 2) performed comparably to the baseline, every expansion degraded aggregate performance. All of the differences were statistically significant with the exception of word expansions. All hierarchical expansions worsened performance, with all-level expansion poorer than single-level. Table 1 also shows, however, that each expansion resulted in some queries with improved performance. The highest number of improved queries came from word expansion, with over 40% of queries showing improved recall-precision. Both one-level and all-level children expansions as well as the one-level parent and related term expansions showed benefit about one-quarter of the time.

Discussion

Our results show that expanding queries with Metathesaurus terms did not confer any aggregate advantage as measured by recall and precision. On the average, retrieval performance did not improve from adding synonymous, hierarchical, or related terms or term definitions. However, in some instances, such as word expansion, improvement was seen over the baseline in 40% of queries. Clearly there may be a role for expansion when applied in certain instances. Further research with real users must delineate those instances.

There are several limitations to this study that indicate further investigation is warranted. The first limitation is that we focused on only one test collection which has only one type of document, the MEDLINE reference. Since MEDLINE records already contain MeSH terms, a slight amount of thesaurus-based expansion is already present, in that MeSH terms contain different expressions of concepts in the title or abstract. As noted already, it has been shown that the presence of these MeSH terms leads to a 10% improvement in performance as measured by average precision [15]. Therefore other test collections, particularly those derived from full-text databases, may have a better chance of accruing performance benefits than bibliographic databases such as MEDLINE.

A second limitation is that we assessed only one type of IR system, the word-statistical system. It is possible that a Boolean system would yield different results than we achieved here, although there is no compelling reason to believe so, since these systems would be using the same terms.

The final limitation was the batch nature of our experiments. We attempted to systematically assess benefit from different expansion techniques on the collection as a whole. It is possible that searchers could add terms in an individualized manner leading to better performance. Indeed, experienced searchers for years have touted the benefits of explosions in traditional MEDLINE systems. However, it should be noted that no formal evaluation of this technique has been performed and in fact, a previous study showed that experienced searchers using this test collection derive little performance benefit using all MEDLINE features (including MeSH terms and explosions) over use of text words only [16]. While the explode function makes sense for human-indexed systems that instruct indexers to assign terms at the most specific level of detail, it may be less desirable for systems that use thesauri in automated ways.

The results of this study show that thesaurus-based automated query expansion does not necessarily improve searching performance. This is in contrast to automated query expansion based on words [10-12] and indexing terms [13] from the document. These results therefore question whether operational systems such as PubMed, Internet Grateful Med, and others should use automated hierarchy explosions. These results also parallel those of our previous work, which show that use of thesauri tools such as the UMLS Metathesaurus can provide benefit for IR tasks such as automated indexing and query expansion in a substantial minority of queries but not in the aggregate. The next major challenge is to determine how to apply these benefits in instances where they are likely to succeed and avoid them when they are prone to failure. Further improvements in thesauri, indexing approaches, and other language technologies should help, especially if augmented by more systematic study of users and their information needs.

Acknowledgements

This work was supported by grants DE-FG03-94ER61918 from the Department of Energy and LM06311 from the National Library of Medicine.

Table 1 – Results of runs 1-9 with 11-point average, recall and precision at 30 documents retrieved, and number of queries where expansion improves over baseline. (* p < .001 with comparison to baseline, ** p < .0001 with comparison to baseline)

Run	R-P average (11 pt.)	Recall at 30 documents	Precision at 30 documents	Number of queries (%) improved over baseline
1. Baseline	.2592	.3480	.1960	NA
2. Parsimonious terms	.2598	.3469	.1967	39 (38.6%)
3. Word expansion	.2410	.3470	.1696	42 (41.6%)
4. Child expansion – one level	.1680*	.2385	.1422	24 (23.8%)
5. Child expansion – all levels	.1540**	.2100	.1244	24 (23.8%)
6. Parent expansion – one level	.1719**	.2749	.1508	29 (28.7%)
7. Parent expansion – all levels	.0255**	.0360	.0234	3 (2.9%)
8. Related term expansion	.1570**	.2336	.1366	28 (27.7%)
9. Definition expansion	.1236**	.2007	.1059	11 (10.9%)

References

1. Lindberg DAB, Humphreys BL, and McCray AT, The Unified Medical Language System project. *Methods of Information in Medicine*, 1993. 32: 281-291.
2. McCray AT, Srinivasan S, and Browne AC. Lexical methods for managing variation in biomedical terminologies. In *Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care*. 1994. Washington, DC: Hanley & Belfus 235-239.
3. Hersh WR, Evaluation of Meta-1 for a concept-based approach to the automated indexing and retrieval of bibliographic and full-text databases. *Medical Decision Making*, 1991. 11: S120-S124.
4. Hersh WR and Leone TJ. The SAPHIRE server: a new algorithm and implementation. In *Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care*. 1995. New Orleans, LA: Hanley-Belfus 858-862.
5. Hersh WR, Hickam DH, and Leone TJ. Word, concepts, or both: optimal indexing units for automated information retrieval. In *Proceedings of the 16th Annual Symposium on Computer Applications in Medical Care*. 1992. Baltimore: McGraw-Hill 644-648.
6. Hersh WR, et al., A performance and failure analysis of SAPHIRE with a MEDLINE test collection. *Journal of the American Medical Informatics Association*, 1994. 1: 51-60.
7. Hersh WR and Hickam DH, An evaluation of interactive Boolean and natural language searching with an on-line medical textbook. *Journal of the American Society for Information Science*, 1995. 46: 478-489.
8. Salton G, Developments in automatic text retrieval. *Science*, 1991. 253: 974-980.
9. Voorhees EM and Harman D, Overview of the Sixth Text REtrieval Conference (TREC). *Information Processing and Management*, 2000. 36: 3-36.
10. Buckley C, et al. Automatic query expansion using SMART: TREC 3. In *Overview of the Third Text REtrieval Conference (TREC-3)*. 1994. Gaithersburg, MD: NIST 69-80.
11. Evans DA and Lefferts RG. Design and evaluation of the CLARIT TREC-2 system. In *The Second Text REtrieval Conference (TREC-2)*. 1993. Gaithersburg, MD: NIST 137-150.
12. Broglio J, et al. Document retrieval and routing using the INQUERY system. In *Overview of the Third Text REtrieval Conference (TREC-3)*. 1994. Gaithersburg, MD: NIST 29-38.
13. Srinivasan P, Retrieval Feedback in MEDLINE. *Journal of the American Medical Informatics Association*, 1996. 3: 157-168.
14. Suarez HH, Hao X, and Chang IF. Searching for information on the Internet using the UMLS and Medical World Search. In *Proceedings of the 1997 Annual AMIA Fall Symposium*. 1997. Nashville, TN: Hanley & Belfus 824-828.
15. Hersh WR, et al. OHSUMED: an interactive retrieval evaluation and new large test collection for research. In *Proceedings of the 17th Annual International ACM Special Interest Group in Information Retrieval*. 1994. Dublin: Springer-Verlag 192-201.
16. Hersh WR and Hickam DH, The use of a multi-application computer workstation in a clinical setting. *Bulletin of the Medical Library Association*, 1994. 82: 382-389.