The Effect of Textual Variation on Concept Based Information Retrieval

Alan R. Aronson
National Library of Medicine
Bethesda, MD 20894

Accounting for textual variation in the documents and queries processed by information retrieval systems is considered essential for achieving good retrieval. Recent research has called into question several of the techniques used to support this endeavor. This paper reports on experiments with a concept based information retrieval system which relies on a program called MetaMap to account for textual variation in the process of mapping biomedical text such as MEDLINE® bibliographic citations to the UMLS® Metathesaurus®. The experiments confirm that the effort expended in handling textual variation is well-spent for at least one type of concept based information retrieval.

INTRODUCTION

Word based information retrieval (IR) systems have long dealt with the problem of textual variation by using a range of methods from ignoring the case of text to performing stemming in order to treat minor textual variants of the same word as a single form. Thus *Hospital* becomes *hospital*; *medicine, medicines* and *medical* become *medic*; and, unfortunately, *battery* may become *bat*. IR systems in which the focus is on concepts rather than words have an identical need to account for such textual variation. The approach to concept based IR taken here is to normalize both document text and queries, replacing text words with concepts discovered by a program called MetaMap which maps biomedical text such as MEDLINE citations to the concepts in the UMLS Metathesaurus, a knowledge base of biomedical concepts developed at NLM [1]. Actual retrieval is accomplished by processing the normalized text using a traditional statistical IR system, SMART [2]. The distinct separation of MetaMap normalization from retrieval in this approach facilitates experimentation by using the existing evaluation capabilities of SMART while allowing independent exploration of the normalization process.

Recent research has shown that universal application of traditionally accepted techniques such as ignoring case, stemming, and even the use of stop words sometimes degrades performance in IR and related systems [3,4]. MetaMap normalization can be thought of as generalized stemming. It is related to several studies which also map biomedical text to the Metathesaurus for various purposes [5-10]. The major distinguishing features of MetaMap are its use of high-level parsing, its use of knowledge based, linguistically motivated variant generation, and its principled evaluation function for ranking results. It is MetaMap's variant generation process which is directly affected by textual variation and which is the focus of this paper.

MetaMap uses three kinds of variation in the process of mapping to the Metathesaurus: acronyms and abbreviations, synonyms, and derivational variants. Spelling and inflectional variants are not considered here since the MetaMap variant generation algorithm uses a quasi-canonicalization approach which collapses all spelling and inflectional variants into a single variant. Thus spelling and inflectional variation are intrinsic to MetaMap and not susceptible to examination.

Experiments were performed by completely or partially eliminating one of the three kinds of variation from the MetaMap variant generation process and computing the effect on IR performance. The study was done using the 1995 release of the Metathesaurus on the NLM Test Collection [11] which consists of some 150 actual user queries, 3,000 MEDLINE citations, and relevancy judgements for each query.

The current study has implications for previous work using MetaMap as a fundamental component: research in semantic processing in general [12,13], ambiguity resolution in particular [14]; and, most directly, previous experimentation using all forms of MetaMap variation which produced a modest 4% increase in average precision [15].

METHODOLOGY

The process of determining the effect of textual variation on IR performance consists of mapping the entire

NLM Test Collection to the Metathesaurus for several variant generation strategies, normalizing the text of the collection based on the mapping results, and then evaluating the resulting normalized collections using SMART.

MetaMap Processing

The task of mapping from biomedical text to concepts in the Metathesaurus consists of the following five steps (see Figure 1):

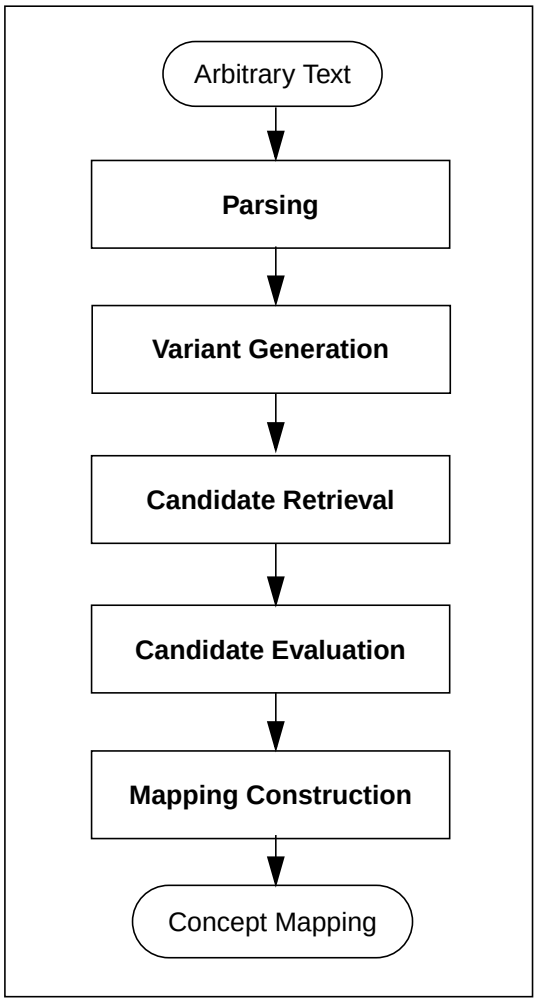

Figure 1. MetaMap processing

1. Arbitrary text is parsed into simple noun phrases; this limits the scope of further processing and thereby makes the mapping effort more tractable. Parsing is accomplished using the SPECIALIST™ minimal commitment parser [16] which produces a high-level syntactic analysis rather than a full syntactic analysis. The parser optionally uses the Xerox Part-of-speech tagger [17] which assigns syntactic labels (e.g., noun, verb) to all textual items. The parser is very good at determining the simple noun phrases in text, and the tagger improves accuracy even more.
2. For each phrase, variants are generated where a variant essentially consists of one or more phrase words together with all of its acronyms, abbreviations, synonyms, derivational variants and meaningful combinations of these;
3. The *candidate set* of all Metathesaurus strings containing at least one of the variants is retrieved;
4. Each Metathesaurus candidate is evaluated by first computing a mapping from the phrase words to the candidate's words and then calculating the strength of the mapping using a linguistically principled evaluation function consisting of a weighted average of four metrics: centrality, variation, coverage and cohesiveness. The candidates are ordered according to mapping strength; and
5. Complete mappings are constructed by combining candidates involved in disjoint parts of the phrase, and the strength of the complete mappings is computed just as for candidate mappings. The highest scoring complete mappings represent MetaMap's best interpretation of the original phrase.

An example of the mapping process is given in the next section. Details of MetaMap's algorithms can be found at http://nls3.nlm.nih.gov.

MetaMap Variant Generation

Because of its importance in determining how textual variation affects MetaMap processing, MetaMap's variant generation algorithm is described in more detail. The approach taken in computing variants is a quasi-canonicalization approach. This simply means that a variant represents not only itself but all of its inflectional and spelling variants. Collapsing inflectional and spelling variants results in significant computational savings. Actual variant generation begins by computing the set of generators for a phrase. A variant generator is any *meaningful* subsequence of words in the phrase where a subsequence is meaningful if it is either a single word or occurs in the SPECIALIST lexicon [18]. For example, the variant generators for the noun phrase *of liquid crystal thermography* are *liquid crystal thermography*, *liquid crystal*, *liquid*, *crystal* and *thermography* (prepositions, determiners, conjunctions, auxiliaries, modals, pronouns and punctuation are ignored). Note the multi-word generators. Variants are computed for each of the variant generators according to the scheme pictured in Figure 2. The computation for each generator proceeds as follows:

1. Compute all acronyms, abbreviations and synonyms of the generator. This results in the three sets Generator, Acronyms/Abbreviations, and

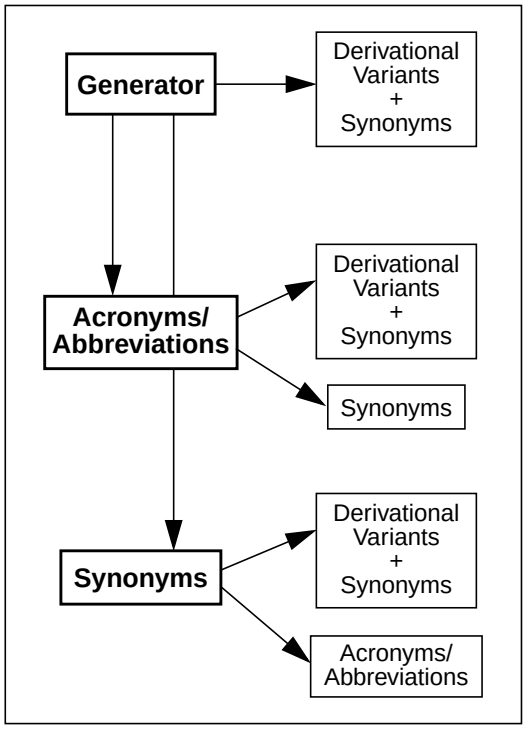

Figure 2. Variant generation

Synonyms which are highlighted with boxes in Figure 2;

2. Augment the elements of the three sets by computing their derivational variants and the synonyms of the derivational variants;
3. For each member of the Acronyms/Abbreviations set, compute synonyms; and
4. For each member of the Synonyms set, compute acronyms/abbreviations.

Derivational variants and synonyms are recursively generated since this generally produces meaningful variants. However, acronyms and abbreviations are not recursively generated since doing so almost always produces incorrect results. For example, the abbreviation *na* of *sodium* is also an acronym of both *nurse's aide* and *nuclear antigen* which are unrelated to *sodium*.

Consider the utterance *Ocular complications of Myasthenia Gravis*, a query from the NLM Test Collection. The parser detects two noun phrases, *Ocular complications* and *of Myasthenia Gravis*. A simplified syntactic analysis for *Ocular complications* is `[mod(ocular), head(complications)]`. Variants for the phrase are shown in Figure 3. The variants are arranged hierarchically according to their derivation history. Each variant is followed by its distance score from its generator and its history. For example, *ocular* and *complications* have distance score 0 and

```
ocular [0 = “”]
    oculus [3 = “d”]
    eyepiece [2 = “s”]
    eye [2 = “s”]
        optic [4 = “ss”]
            optical [7 = “ssd”]
                vision [9 = “ssds”]
            optically [10 = “ssdd”]
        ophthalmic [4 = “ss”]
            ophthalmia [7 = “ssd”]
            ophthalmiac [7 = “ssd”]
complications [0 = “”]
    complicate [3 = “d”]
```

Figure 3. Variants of *Ocular complications*

empty history because they are the generators, themselves. Similarly, *vision* has distance score 9 and history "ssds" meaning that it is a synonym of a derivational variant (*optical*) of a synonym (*optic*) of a synonym (*eye*) of *ocular*. The Metathesaurus candidates for *Ocular complications* are shown in Figure 4.

861 Complications (Complication, NOS)

861 complications <1>

777 Complicated

638 Eye

611 Optic (Optics)

588 Ophthalmia (Endophthalmitis)

579 Vision

Figure 4. Metathesaurus candidates for *Ocular complications*

The candidates are shown in order of mapping strength which has been normalized to a score between 0 and 1,000 and is displayed before the candidate. If the candidate is not the preferred name of the concept it represents, the preferred name is displayed in parentheses. The best complete mappings for the phrase consist of the Metathesaurus concept 'Eye' and either the concept 'Complication, NOS' or the concept 'complications <1>'.

Normalization

To make use of MetaMap's results, the Test Collection is normalized by replacing text with matching Metathesaurus concepts subject to the following con-

straint. In order for a concept to replace text, it must have the correct UMLS semantic type. Correctness is determined by a version of the Xerox tagger designed to handle semantic tags (semantic types) in addition to syntactic tags. Use of a semantic tagger in this way allows for choosing between competing concepts with different semantic types and also for disqualifying concepts with inappropriate semantic types.

An example will illustrate this notion of *disambiguated concept normalization*. The original text of a Test Collection query is shown below followed by its normalization. Corresponding words and concepts are underlined.

- Original text: *Adaptation of physical environment in hospitals to care for Alzheimer patients (model Alzheimer units).*
- Normalized text: *Adaptation of Physical Environment in Hospitals to Caring for Alzheimer Patients (model Alzheimer units).*

Note that no normalization occurs for *Alzheimer* since *Alzheimer* only appears in the Metathesaurus as a subpart of concepts, e.g., 'Alzheimer's Disease'. Also, even though *units* maps to the concept 'Genes' (because of an infelicitous synonymy relationship between *unit* and *gene* which is only valid in a molecular biology context), normalization does not occur since the semantic tagger chooses a tag for *units* different from that of 'Genes'.

IR Experiments

Five versions of MetaMap output differing in the type of allowable variation were used in the IR experiments:

- All Variants—the baseline version in which all types of MetaMap variation are allowed;
- No A/A—no acronym or abbreviation variants are used;
- Unique A/A—only acronyms and abbreviations with unique expansions are used. Thus, the acronym *ICU* (*Intensive Care Unit*) is used, but the abbreviation *na* (*sodium*, *nurse's aide*, *nuclear antigen*, …) is not;
- No Synonyms—no synonymy variants are used; and
- No Derivations—no derivational morphology variants are used.

Each version of MetaMap output was used to normalize the Test Collection. The resulting versions of the test collection were processed straightforwardly using SMART. The SMART weighting scheme atc (a variant of the standard term frequency-inverse document frequency scheme) was used for both queries and documents, and the 3-point average precision (i.e., averaging the precision values corresponding to recall of 0.2, 0.5 and 0.8) was used to compare results (see [19,20]).

RESULTS

The results of the experiments are shown in Table 1.

	Average Precision	Improvement over Unprocessed
All Variants	0.5968	5.0%
No A/A	0.5907	3.9%
Unique A/A	0.5887	3.5%
No Synonyms	0.5887	3.5%
No Derivations	0.5822	2.4%

Table 1. Average precision and improvement over unprocessed

They are expressed both as raw average precision values and as percentage improvement over the average precision obtained without processing the text at all (0.5686). The best results are obtained when all variation is allowed.

CONCLUSION

The experiments described above show that all forms of variation used by MetaMap enhance retrieval performance albeit in varying degrees. The contribution to retrieval performance of each type of variation can be seen by comparing its average precision to that of the All Variants case: the greater the difference, the greater the contribution. Thus derivational variation contributes the most, followed by synonymy and then acronyms and abbreviations. Restricting acronyms and abbreviations to unique ones is slightly better than allowing them all.

The most interesting result is the one favoring unique acronyms and abbreviations over arbitrary ones. This confirms an intuition that while all forms of variation are good in general, it is nevertheless useful to try to discover practical ways of limiting variation to increase accuracy. Further experiments such as exploring the recursive depth allowed during variant generation or applying semantic restrictions to the process may lead to a better understanding of textual variation in order to enhance information retrieval performance.

Acknowledgments

I want to thank Tom Rindflesch and Zoë Stavri for many helpful discussions during the writing of this paper.

References

1. Lindberg DAB, Humphreys BL and McCray AT. "The Unified Medical Language System." *Methods of Information in Medicine* 32:281-291, 1993.

2. Salton G (ed). *The SMART retrieval system: Experiments in automatic document processing*. Englewood Cliffs, NH: Prentice-Hall, Inc, 1971.

3. Church KW. "One Term or Two?" In Fox E, Ingwersen P, and Fidel R (eds) *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 310-318, 1995.

4. Riloff E. "Little Words Can Make a Big Difference for Text Classification." In Fox E, Ingwersen P, and Fidel R (eds) *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 130-136, 1995.

5. Canfield K, Bray B, Huff S and Warner H. "Database capture of natural language echocardiographic reports: a Unified Medical Language approach." In Kingsland LC, III (ed.) *Proceedings of the 13th Annual SCAMC*, 559-563, 1989.

6. Chute CG, Yang Y, Tuttle MS, Sherertz DD, Olson NE and Erlbaum MS. "A preliminary evaluation of the UMLS Metathesaurus for patient record classification." In Miller RA (ed.) *Proceedings of the 14th Annual SCAMC*,161-165, 1990.

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

8. 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.

9. Wagner MM. "An automatic indexing method for medical documents." In Clayton PD (ed.) *Proceedings of the 15th Annual SCAMC*, 1011-1017, 1992.

10. 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.

11. Schuyler PL, McCray AT and Schoolman HM. "A test collection for experimentation in bibliographic retrieval." Barber B, Cao D, Qin D and Wagner G (eds.) *MEDINFO 89*, Amsterdam: North-Holland, 810-912, 1989.

12. Rindflesch TC and Aronson AR. "Semantic processing in information retrieval." In Safran C (ed.) *Proceedings of the 17th Annual SCAMC*, 611-615, 1993.

13. Sneiderman C, Rindflesch TC and Aronson AR. "Extracting Physical Findings from Free-Text Patient Records." *1995 AMIA (American Medical Informatics Association) Spring Congress*, Cambridge, Massachusetts, June 25-28, 1995.

14. Rindflesch TC and Aronson AR. "Ambiguity Resolution while Mapping Free Text to the UMLS Metathesaurus." In Ozbolt JG (ed.) *Proceedings of the 18th Annual SCAMC*, 240-244, 1994.

15. Aronson AR, Rindflesch TC and Browne AC. "Exploiting a Large Thesaurus for Information Retrieval." In *RIAO (Computer aided information retrieval) 94 Conference Proceedings*, 197-216, 1994.

16. McCray AT, Aronson AR, Browne AC, Rindflesch TC, Razi A and Srinivasan S. "UMLS knowledge for biomedical language processing." *Bulletin of the Medical Library Association* 81:184-194, 1993.

17. Cutting D, Kupiec J, Pedersen J and Sibun P. "A practical part-of-speech tagger." In *Proceedings of the Third Conference on Applied Natural Language Processing*, 1992.

18. McCray AT, Srinivasan S and Browne AC. "Lexical methods for managing variation in biomedical terminologies." In Ozbolt JG (ed.) *Proceedings of the 18th Annual SCAMC*, 235-239, 1994.

19. Tague JM. "The pragmatics of information retrieval experimentation." In Jones KS (ed.) Information Retrieval Experiment, 59-102, 1981.

20. Salton G. "The Smart environment for retrieval system evaluation—advantages and problem areas." In Jones KS (ed.) Information Retrieval Experiment, 316-329, 1981.