MetaMap Variant Generation

Alan R. Aronson

May 8, 2001

1. Overview

In order to account for textual variation in biomedical text, MetaMap computes several kinds of word variants in the process of mapping text to the UMLS Metathesaurus: spelling, inflectional and derivational variants, acronyms and abbreviations, and synonyms. The MetaMap variant generation algorithm for finding variants of a given word uses a quasi-canonicalization approach which defers consideration of all spelling and inflectional variation until the end of the process. This process is described in the next section.

2. Variant Generation

The process of mapping text to concepts in the Metathesaurus begins with the following two steps:

1. Parse the text into simple phrases and perform the remaining processing for each phrase;[1]
2. Generate the variants for the phrase where a variant essentially consists of one or more consecutive phrase words together with all of its spelling variants,[2] abbreviations, acronyms, synonyms, inflectional and derivational variants, and meaningful combinations of these.

Step 2 above begins by computing a set of variant generators for the simple phrases discovered by the parser. A variant generator is any *meaningful* subsequence of words in the phrase where a sub-

[1] Parsing is accomplished using the SPECIALIST minimal commitment parser which produces a high-level syntactic analysis rather than a full syntactic analysis. The parser optionally uses the Xerox Part-of-speech tagger which assigns syntactic labels to all textual items. The parser is very good at determining the simple noun phrases in text; and the errors it does make are normally inconsequential to MetaMap. The tagger also improves parsing results.

[2] A spelling variant of a word is just a variant having the same principal part as the word. For example, *haemorrhaged* is a spelling variant of *hemorrhaged*.

sequence is meaningful if it is either a single word or occurs in the SPECIALIST lexicon. For example, the variant generators for the 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).[1] Note the multi-word generators. Because most multi-word entries in the lexicon impart no benefit to MetaMap, a version of the lexicon containing only essential multi-word entries (such as *in vitro*) is normally used by MetaMap. A simpler example of a phrase which will be used throughout the sequel is based on the noun phrase *ocular complications*.[2] Its generators are simply *ocular* and *complications*.

The approach taken in computing variants is to compute a sequence of sets depicted in Figure 1

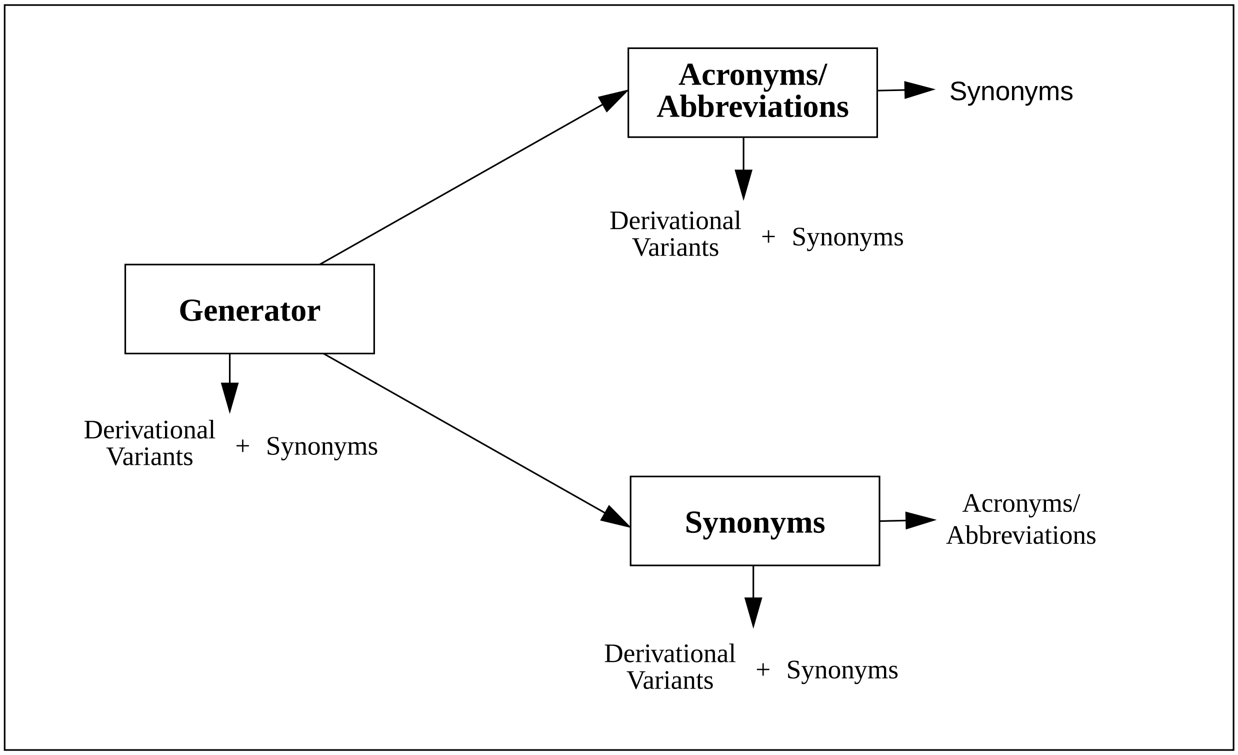

Figure 1. Variant Generation

where inflectional and spelling variation is not explicitly shown. 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 Synonyms which are highlighted with boxes in Figure 1;
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

[1] A simplified syntactic analysis for *of liquid crystal thermography* is `[prep(of), head(liquid crystal thermography)]`.

[2] A simplified syntactic analysis for *ocular complications* is `[mod(ocular), head(complications)]`.

4. For each member of the Synonyms set, compute acronyms/abbreviations.

The sequence of sets corresponding to the above description is:

- **G**—a generator;
  GSPs—the spelling (same-part) variants of G;
  GIs—the inflections of G;
- **GDs**—the derivational variants of G;
  GDSIs—the synonyms and their inflections of GDs;
- **GAAs**—the acronyms/abbreviations of G;
  GAASPs—the spelling variants of GAAs;
  GAAIs—the inflections of GAAs;
- **GAADs**—the derivational variants of GAAs;
  GAADSIs—the synonyms and their inflections of GAADs;
- **GSs**—the synonyms of G;
  GSSPs—the spelling variants of GSs;
  GSIs—the inflections of GSs;
- **GSDs**—the derivational variants of GSs;
  GSDSIs—the synonyms and their inflections of GSDs;
- **GAASs**—the synonyms of GAAs;
  GAASIs—the inflections of GAASs;
- **GSAAs**—the acronyms/abbreviations of GSs; and
  GSAAIs—the inflections of GSAAs.

Sets shown in bold do not involve spelling variation or inflection. The issue of whether to recursively generate variants of a given type is handled as follows:

- Acronyms and abbreviations are not recursively generated since doing so almost always produces incorrect results. For example, the abbreviation *na* of *sodium* has expansions *nurse's aide* and *nuclear antigen* which are unrelated to *sodium*; and

- Derivational variants and synonyms are recursively generated since this often produces meaningful variants.

The variants computed for the generator *ocular* are shown in Figure 2. Following each variant is its variant distance score, a rough measure of how much it varies from its generator. Each step of

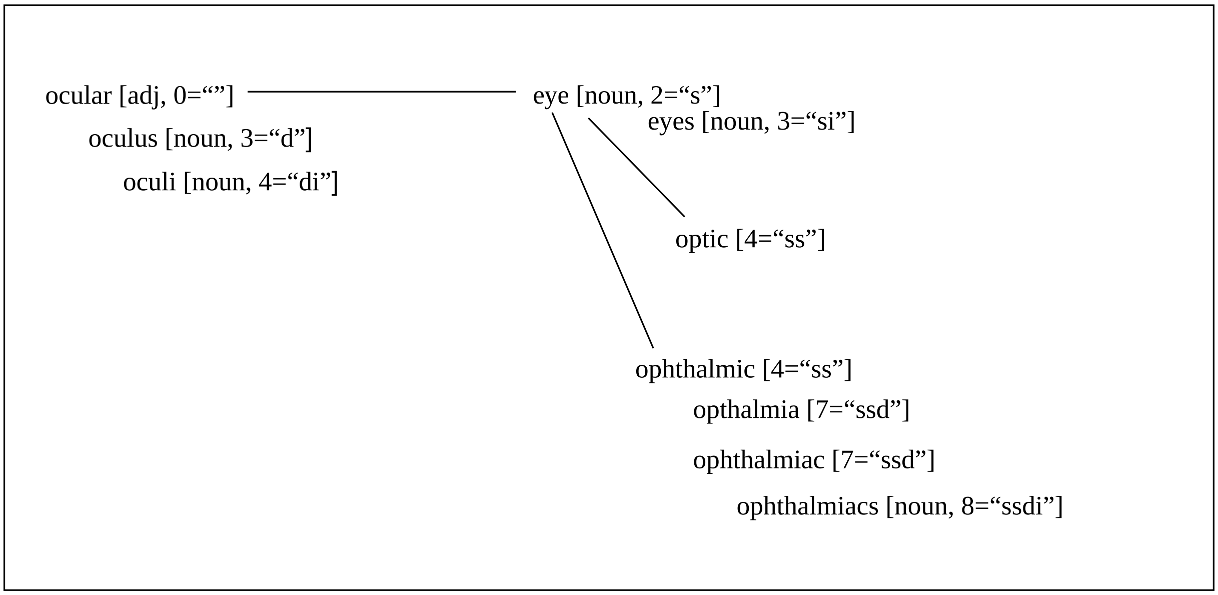

Figure 2. Variants for the generator *ocular*

the generation process adds a history element and a variant distance score according to Table 1.

Variant Type	**Distance Value**
spelling (`p`)	0
inflectional (`i`)	1
synonym (`s`) or acronym/abbreviation (`a`, `e`)	2
derivational (`d`)	3

Table 1. Variant Distances

For example,

- *oculus* (with variant distance 3 and history “d”) is simply a derivational variant of the generator *ocular*;
- *eyes* (with variant distance 3 and history “si”) is an inflectional variant of a synonym (*eye*) of *ocular*; and
- *ophthalmiacs* (with variant distance 8 and history “ssdi”) is an inflection of a derivational variant (*ophthalmiac*) of a synonym (*ophthalmic*) of a synonym (*eye*) of *ocular.*

The following MetaMap options have an effect on the variant generation process:

- `-a --no_acros_abbrs`, `-u --unique_acros_abbrs_only`, `-d --no_derivational_variants`, and `-D --an_derivational_variants` affect which kinds of variation are allowed. The first two options prohibit acronym/abbreviation variants entirely or restrict them to those cases with unique expansions. The last two options apply similarly to derivational variants, the last option restricting derivational variation to that between an adjective and a noun;

- `-z --term_processing` affects parsing and, therefore, has an indirect effect on variant generation. Text which would normally be processed as separate phrases is handled monolithically; and
- `-8 --dynamic_variant_generation` causes MetaMap to employ the algorithm described here instead of using tables of pre-computed variants which have also been filtered so that only variants actually occurring in the Metathesaurus are produced.