A Formal Approach to Integrating Synonyms with a Reference Terminology

Harold R. Solbrig, Peter L. Elkin, M.D., Philip V. Ogren,
Christopher G. Chute, M.D., Dr.P.H
Mayo Clinic, Rochester MN

ABSTRACT

Medical terminologies continue to grow in scope, completeness and detail. The emerging generation of terminology systems define concepts in terms of their position within a categorical structure. It is still necessary, however, to access and represent the concepts using everyday spoken and written language, which introduces both lexical and semantic ambiguity. This ambiguity can have a negative impact on both selectivity and recall when it comes to associating free-form textual phrases with their coded equivalent.

Lexical ambiguity issues can often be addressed algorithmically, but semantic ambiguity presents a more difficult problem. A common solution to the semantic problem is to associate many different representational permutations with a given target concept. This approach has several drawbacks. An alternate solution is to build separate synonym tables that can serve as permuted indices into the terms representing the underlying concepts. A potential shortcoming of this approach, however, is a further reduction in the lookup selectivity.

One possible source of loss of selectivity could be "meaning drift" – the gradual change in meaning that can be introduced when following a chain of nearly synonymous words. We posited that organizing synonyms into separate "meaning clusters" might reduce this loss in precision, but the results of this study did not bear that out.

Introduction

Both the content and structure of medical terminologies continue to be refined and improved. Formal, concept-based reference terminologies are close to being reality. As coverage of these terminologies becomes more complete, the content, by necessity, is becoming more specific, comprehensive and refined as well.[1,2,3] The number of concepts represented in a terminology has evolved from the hundreds to the hundreds of thousands, and long since passed the point where they can be managed without the use of automated tools and software.[4, 5]

This emerging rigor in concept definition does not, however, preclude the necessity to represent these concepts in an everyday spoken or written language. While it may be possible to represent every individual concept with a unique term, there will always be a variety of ways in which to externally describe the concept – ways that will vary by region, specialty, intended use, etc. This collective set of alternatives is sometimes referred to as an "entry" or "colloquial" terminology[6].

Representational Variation and Synonymy

One source of representation diversity is the variation in the lexical form of words and phrases. Similar terms may differ only in word order, tense, plurality, case, etc. A significant portion of issue of lexical variation can be addressed through the use of tools like the NLM's Lexical Variant Generator (LVG)[7]. LVG can transform terms that vary solely on a lexical basis into a canonical base form that can be consistently compared and indexed.

A second source of representational variation, however, is not so readily addressed. Frequently one or more of the words in a term can be replaced by another semantically similar, but lexically quite different word or phrase without substantially altering the intended meaning. One common solution to this issue is to enumerate all of the common ways that a given concept may be represented. The February 1999 version of SNOMED-RT,[8] for example, lists *Gastrointestinal disorder* and *Disorder of the gastrointestinal tract* as alternative representations for the concept, *D5-00010: Disease of digestive tract*. This approach raises some issues, however.

The first issue with the enumeration approach is the potential combinatorial nature of word synonyms. Why shouldn't *Disease of the gastrointestinal tract, Gastrointestinal disease, GI disease, GI tract disease, GI disorder*, etc. also be listed as appropriate representations for concept *D5-00010*? A complete enumeration of all of the possible representations of each of the 110,000+ concepts currently available in SNOMED-RT could prove to be a daunting prospect.

1067—5027/00/$5.00

One potential way of solving this problem would be to create a list of synonyms and then automatically compute all possible term permutations. This approach, however, doesn't take into account the second problem with permutations, the issues of context and function. The decision as to what constitutes a synonym or near-synonym can be highly subjective. Under some circumstances it may be quite reasonable to substitute the word *disease* for the word *disorder*, while in another situation equating these two words might change the intended meaning of a statement or phrase significantly. Synonymy is context dependent. What is and isn't (sufficiently) synonymous can vary by region, specialty, institution or even sometimes by the specific individual.

Synonymy may also depend upon the function to which the terminology is being applied. If the primary intent is to present an end user with a list of possible selections, recall may take precedence over selectivity – meaning that partial synonyms are useful. If, however, the purpose is automated coding, selectivity becomes paramount and only nearly exact synonyms should be considered. Pre-computing term permutations takes none of these factors into account.

Synonyms in a Separate Terminological Space

A possible solution to both of the above issues would be to maintain synonyms in a completely separate "space" from the terms that represent the concepts. This would simplify the task of maintaining the reference terminology as only the preferred term and non-computable variations such as eponyms and metonyms would need to be entered. The synonyms would function as a permuted index into the primary, allowing the decision of what constitutes "close enough" to be postponed until context and function could be determined.

This above approach has been used to cross-reference web pages[9]. When it comes to medical web sources, however, the application of synonyms didn't necessarily lead to improved performance. In the example cited above, synonyms adversely impacted selectivity without yielding the expected improvement in recall. While these results apply specifically to web pages rather than reference terminologies, there is no reason to believe that the issues would not be similar in both situations. Given this, we began by asking what might be done to improve the behavior and characteristics of synonym based indexing applied to a reference terminology. How might we recognize our primary goal, improvement in recall, without suffering a corresponding loss of selectivity? Among the potential causes of the selectivity problem, two possibilities stand out:

1) **The set of synonyms used in some of the previous experiments was quite general.** The University of Arizona studies, in particular, had demonstrated that better result could be achieved when the set of synonyms was constrained to the medical domain rather than representing the whole of the English language[9]. We conjecture that further constraining the scope of the synonyms might continue to yield additional improvements in the system behavior.
2) **The synonyms may have been subject to unacceptable "meaning drift"** – WordNet[10] and other synonym collections[11,12] embody the notion of synonym sets – a focus point around which words with similar meanings can be clustered. Unless, however, this clustering is maintained throughout the *entire* indexing process, it becomes possible to introduce an undesirable "meaning drift". As an example, the word *birth* could be used as a near synonym for *congenital* in one situation and for *childbirth* in another. This should not, however, result in the word *congenital* indexing terms containing the word *childbirth*.

The rest of this document focuses on the second problem – an attempt to study the impact of "meaning drift" on synonym indexing. For the purposes of the study, we chose to use a reasonably small and focused synonym set, under the hopes that minimizing the potential impact of the first problem might more clearly focus the second.

Approach

We started with a formal model of the characteristics that we believed necessary for a separate synonym cross-reference. The model was implemented as a set of relational tables and Java services deployed in the framework of the Mayo terminology services[13] - distributed component architecture. The tables were then populated with a selected set of clinically specific synonyms, acronyms and abbreviations.

We processed all of the active terms in the February 1999 beta release of SNOMED-RT. Each term was converted to a "canonical" lexical form, split into words and each word was looked up in the synonym table. If a matching synonym was found, we made a reference to the term using the corresponding meaning cluster identifier. This index provided a means to go from a word to a set of one or more

meaning clusters and then to locate all terms that had the same or similar words within the clusters.

The resulting index was examined to determine the degree of completeness, coverage and degree of concept "blurring". As a final step, we utilized a set of test terms to determine the impact of the index on the precision and recall of concept retrieval.

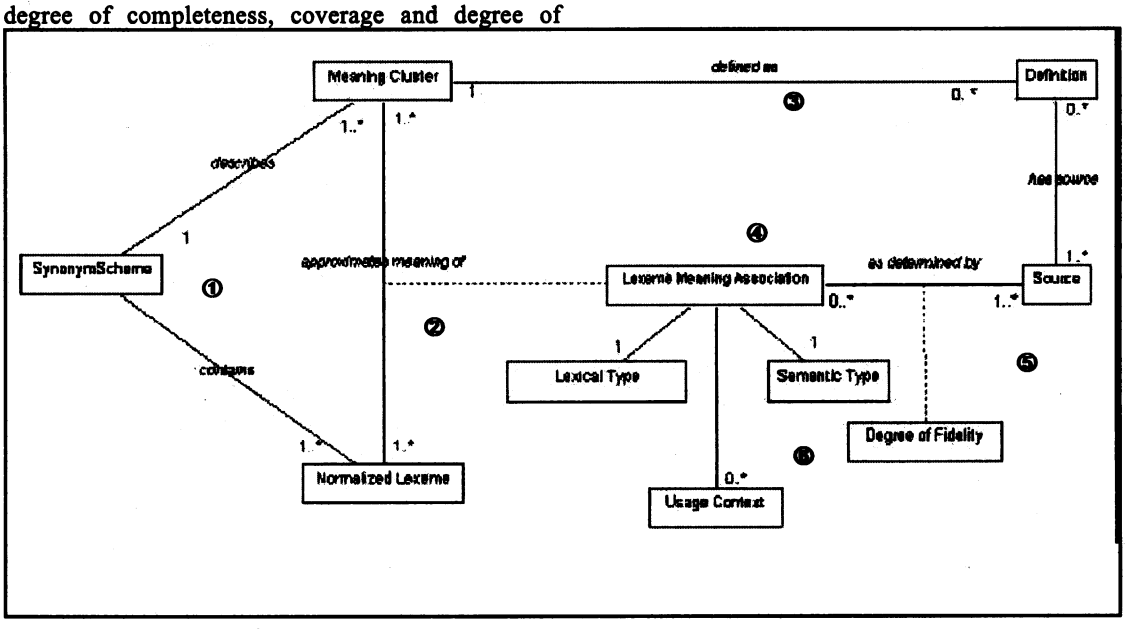

Figure 1 - Synonym Scheme Model

Results

The Model

The model of the logical characteristics of a synonym scheme is shown in **Figure 1** In this model, a *Synonym Scheme* represents the overarching entity that describes both a set of *Meaning Clusters* and a set of *Normalized Lexemes* (words, acronyms, abbreviations, etc. which have been "normalized" - converted into their lexical "base" form) ❶. Each *Normalized Lexeme* is then associated at least one *Meaning Cluster* ❷. *Meaning Clusters* may be accompanied by *Definitions* ❸, the purpose of which is to fix and describe the intent of the cluster.

Each association between a *Normalized Lexeme* and a *Meaning Cluster*, a *Lexeme Meaning Association* ❹ is asserted by one or more authorities or *Sources*. Each *Source* can independently state their own perception of relative "fidelity" (exact, broader, narrower, approximate) of the association ❺. Each association is further described by a *Lexical Type*, a *Semantic Type* and a list of *Contexts* ❻ in which the association is deemed to be applicable. Typical lexical types include *acronym, abbreviation, prefix, suffix, etc.* while a *Semantic Type* might be *synonym, antonym, metonym,* etc.

Populating the Model

The model was implemented as a service within the context of the Mayo terminology services[13] – distributed component architecture. It was populated with a set of about 2800 synonyms, which had gathered by the Mayo Clinic to aid in manual coding of clinical problems. The content and clustering and content was reviewed and edited by a Mayo clinician.

We first attempted to determine how much additional information this synonym set added to the beta version of SNOMED-RT. In theory, if SNOMED-RT defined a concept with a meaning similar to that of a synonym cluster, concept composition should already provide an access mechanism to a similar set of target concepts. As an example, if the SNOMED-RT beta already contained a concept *heart*, similar results should be realized by adding *"cardiac"* to the list of terms associated with the *heart* concept as would be by creating meaning cluster for the terms *"cardiac"* and *"heart"*.

To answer this question, we selected all of the concepts in the SNOMED-RT beta with single word preferred names. An analysis of the associated terms gave us a reasonable approximation of the overlap, and is shown in **Table 1**.

Source	Total Lexemes	Total Meaning Clusters	Clusters Used in Index
Mayo Synonym Set	2856	1508	1196
SNOMED-RT Single Words	4538	2001	1964
Combination	7103	3280	2931
Overlap	291	229	229
Overlap %	4.1%	7.0%	7.8%

Table 1 – Overlap with target terminology

Only 4% of the single words representing SNOMED concepts appeared within the Mayo synonym set. Even when we ignored words in the Mayo synonym set that didn't appear anywhere in the SNOMED-RT beta, less than 8% of the meaning clusters appeared to overlap with SNOMED-RT concepts.

Cross Referencing the Terminology

We programmatically scanned all of the active SNOMED-RT beta terms and created a permutated cross-reference with the synonym meaning clusters. Almost 50% of the 155,241 terms in the SNOMED-RT contained one or more words that appeared in the Mayo synonym set. Concepts for procedures, findings, diseases and body sites had the most complete coverage. This was as expected, as the SNOMED-RT beta was most complete in these areas, and the synonym list itself focused on clinical findings.

Analyzing the Coverage

We needed to know what portion of the terms in the SNOMED-RT beta were already lexical or synonym variants of each other, as we wouldn't expect the addition of indexing to have an impact in those cases. **Table 2** shows, less than 1% of the terms turned out to be simple lexical variants of each other[1] while close to 7% of all of the terms already represented full synonym permutations. As would be expected, the axes with the greatest level of indexing also showed the highest degree of redundancy.

One potentially undesirable side effect of mapping terms containing words with similar meaning into the same underlying set of meaning clusters was that terms that actually represented *different* concepts might map to the same underlying representation as well – an effect that we called "confounding".

[1] Only the words within a phrase were normalized. The application of additional techniques, such as phrase inversion, etc. would probably yield a considerably higher figure.

Axis	Lexical Variants	Synonym Variants
F-01001 (Finding, conclusion AND/OR assessment)	1.7%	12.9%
DF-00000 (Disease)	2.6%	15.3%
P0-00000 (Procedure)	0.1%	5.6%
L-00000 (Living Organism)	0.0%	0.2%
F-61002 (Substance)	0.0%	0.6%
C-00000 (Chemical, drug AND/OR biological product)	0.0%	0.6%
T-D0004 (Topographic Region)	0.1%	1.6%
M-00000 (Morphology)	1.0%	9.9%
T-D0010 (Body as a whole	0.2%	1.5%
J-00000 (Occupation)	0.0%	0.0%
A-000F1 (Physical agent, activity AND/OR force)	0.1%	0.3%
F-00000 (Biological function)	0.2%	5.1%
G-00F1 (Modifier, linkage term AND/OR qualifier)	1.1%	7.1%
S-00000 (Social Context)	0.0%	0.9%
Total for Entire Terminology	0.7%	7.0%

Table 2 - Redundant Terms by Axis

	Terms Representing More than one Concept	Percent of Total
Baseline	32	0.0%
Lexical Normalization	385	0.3%
Synonym cross referencing	2272	1.5%

Table 3 - Confounded Terms

As **Table 3** above shows, the process of lexical normalization did confound about 1% of the terms. Analysis showed that the effect appeared to be at least partially due to the fact that our normalization process ignored word order, and terms such as *...characterized by cell separation, contact minimal or absent* and *...characterized by cell contact, separation minimal or absent* were treated as being identical.

The Impact on Concept Recall and Precision

We used two different sets of terms to test the impact of synonyms, with and without clustering, on the recall and precision of text based concept retrieval. The first set consisted of 49 common diagnostic terms that had been independently graded and coded for the purposes of evaluating programmatic term composition. It should be noted that this set had previously used as a test case for the design of synonym algorithms, so bias could be present.

The second test set was based on the 200 most frequently used entries in the master sheet at the Mayo Clinic from Jan 1, 1994 through April 1, 1995. After removing obvious Mayo specific colloquialisms we ended up with 173 phrases that

accounted for approximately 19% of the master sheet entries. Each of these phrases was then hand-coded into what was deemed the most appropriate code or codes available in the February 1999 SNOMED-RT beta.

The terms in the test sets were normalized and automatically mapped to the corresponding SNOMED-RT concept. There were four different mappings:

1) **No Synonyms** – A SNOMED-RT term matched a test term if they had one or more normalized words in common.
2) **Mayo Synonyms without Clustering** – A SNOMED-RT term matched a test term if they had *any* word in the transitive closure of the synonyms for the words in the term in common.
3) **Mayo Synonyms with Clustering** – A SNOMED-RT term matched a test term if they had any word in common that belonged to the same synonym cluster.
4) **Mayo Synonyms + SNOMED Synonyms** – Identical to 3) above except that the synonym clusters consisted of both the Mayo synonyms and the set of SNOMED-RT beta single word synonyms.

The results of each mapping were then compared to the set of predetermined values for the test set.

	Recall	Precision
No Synonyms	0.54	0.45
Mayo Synonyms without Clustering	0.81	0.40
Mayo Synonyms with Clustering	0.81	0.43
Mayo Synonyms + SNOMED Synonyms	0.81	0.42

Table 4 - 49 Diagnostic Terms

	Recall	Precision
No Synonyms	0.78	0.49
Mayo Synonyms without Clustering	0.86	0.40
Mayo Synonyms with Clustering	0.86	0.43
Mayo Synonyms + SNOMED Synonyms	0.87	0.39

Table 5 – 173 Master Sheet Terms

In both cases, the use of synonyms significantly increased the recall. In the first case, the recall went from 54 to 81% ($p < .0001$, Pearson method) and in the second from 78 to 86% ($p < .0197$). The only statistically significant change in precision was the drop that was observed in the Master Sheet Terms, where it dropped from 49 to 40% ($p < .0094$) when synonyms were introduced. None of the other results were statistically significant.

Conclusions

A small, focused set of synonyms can significantly improve the recall of terms in a reference terminology. The use of synonyms, however, contains within it an inherent lack of precision – a lack of precision that is made manifested during the text-based retrieval of concept codes. Synonym clustering was thought to be one technique that could be utilized to reduce the precision loss, but the results of this study did not bear that out. Given the trend toward improvement in precision with synonym clustering, it is possible that this technique may prove useful with a larger more complete set of synonyms.

Disclaimer

The terminology used as the basis for this study was an early beta release (Feb. '99) of SNOMED-RT, which was utilized because we expect that it will soon become an important and useful tool for clinical coding. We recognize that content of this release was not yet finalized and it is not our intent to make any statements about its coverage or completeness. It was used solely for the purpose of comparative evaluation.

References

1. Spackman KA, Campbell KE. SNOMED RT: A Reference Terminology for Health Care. In: AMIA Annual Fall Symposium, 615-619, 1998.
2. Spackman KA, Campbell KE. Compositional Concept Representation Using SNOMED: Towards Further Convergence of Clinical Terminologies. In: AMIA Annual Fall Symposium, 740-748, 1998
3. Brown PJ, O'Niel M, Price C. Semantic Representation of Disorders in Version 3 of the Read Codes. *Methods Inf Med* 1998. Nov;37:415-19.
4. Rogers JE, Rector AL. Terminological Systems: Bridging the Generation Gap. In: AMIA Annual Fall Symposium, 610-614, 1997.
5 Rossi Mori A, Consorti F, Galaezzi E. Standards to Support Development of Terminological Systems for Healthcare Telematics. *Methods Inform Med* 1998;37(4-5):551-63.
6. Chute CG, Elkin PL, Sheretz DD, Tuttle MS. Desiderata for a Clinical Terminology Server. In: AMIA Annual Fall Symposium, 42-46, 1999
7. McCray AT, Srinivasan S, Brosn AC. Lexical Methods for Managing Variation in Biomedical Terminologies. *Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care,* 1994: 235-239.
8. SNOMED®RT DOCUMENTATION. (RT beta 0.5) College of American Pathologists, 1999.
9. Gondy L, Tolle KM, Hsinchun C. Customizable and Ontology-Enhanced Medical Information Retrieval Interfaces. *Proceedings of IMIA Working Group 6 – Medical Concept Representation and Natural Language Processing.* In Press 1999
10. Miller GA, Beckwith R, FellBaum C, Gross D, Miller K. Introduction to WordNet: An On-line Lexical Database. August 1993. ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.ps
11. Microsoft Office97/Visual Basic Language Reference. 1997 Microsoft Press, Redmund WA.
12. UMLS Knowledge Sources. National Library of Medicine, 1998.
13. Chute CG, Solbrig HR, Elkin PL. Terminology Services as Software Components: An Architecture and Preliminary Efforts. *Proceedings of IMIA Working Group 6 – Medical Concept Representation and Natural Language Processing.* In Press 1999