MEDINFO 2007
HISA Ltd Australia
© 2007 The authors. All rights reserved.

Piecewise Synonym Construction Using WordNet for UMLS Source Integration

Kuo-Chuan Huang[a], James Geller[a], Michael Halper[b]

[a] *Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102 USA*
[b] *Department of Computer Science, Kean University, Union, NJ 07083 USA*

Abstract

Source vocabulary integration forms the basis of the UMLS's content and its planned continued growth. When integrating a new term, it is essential to recognize situations where the concept expressed by that term already exists in the UMLS, even if it is represented by a different term (i.e., a synonym). A preliminary study has shown that even when using the UMLS's rich set of synonyms, a new term for an existing concept may not match any of them. A process for automatically constructing additional, potential synonyms from multi-word source terms is presented. This process consists of decomposing the source term into words, finding separate synonyms for each in WordNet, and then recombining those synonyms into multi-word terms. We call such synonyms piecewise synonyms, and we present a precise formulation for their construction. The methodology is tested in the context of a (partial) reintegration of the "Minimum Standard Terminology" into the UMLS. The results show that additional "source term/ UMLS term" matches are gained.

Keywords:

UMLS, vocabulary integration, terminology integration, WordNet, synonym

Introduction

The Unified Medical Language System (UMLS) comprises a large terminological database covering the biomedical and health-related fields. This database has been populated via the integration of a variety of source vocabularies, e.g., SNOMED CT , LOINC , and RxNorm . Currently, the number of sources is over 100. Plans call for integration of more terminologies in the future . At present, the Metathesaurus, the UMLS's concept repository, contains over 1,000,000 concepts and over 5,000,000 terms .

The overall process of integrating a new source terminology into the UMLS is defined by the National Library of Medicine to comprise four major phases: (1) analysis and inversion, (2) insertion, (3) human editing, and (4) quality assurance . In general, the process tends to be labor-intensive and error-prone. As such, facilitating source-vocabulary integration is a critical issue facing the UMLS curators. Many algorithmic aides have been developed in this context .

In this paper, we present a methodology for increasing the effectiveness and efficiency of the insertion phase of the integration process. One of that phase's major subtasks is the identification of terms and associated concepts from the source that already exist in the UMLS. Unfortunately, the same concept may be known by several different terms. Thus, it is a distinct possibility that a new source uses a term that differs from the UMLS term for an existing UMLS concept. This problem is exacerbated for terms that consist of several words.

In a preliminary study, we found a surprisingly small number of matches for the terms of a "new" terminology with terms in the UMLS. This is surprising when one considers the large number of terms already in the UMLS. One may conclude that the UMLS is missing valid synonyms for its concepts.

Our methodology automatically generates new synonyms for multi-word terms of a source vocabulary. It does so by first decomposing a term into its constituent words. It then gathers all synonyms for the individual words. These individual synonyms are found in the widely used WordNet , a lexical database for English. Finally, a set of novel, multi-word synonyms, called *piecewise synonyms*, is created for use in a term-matching algorithm. A precise formulation of this synonym-construction process is presented.

To test the effectiveness of our methodology, we utilize it in the process of reintegrating an existing source, called the "Minimum Standard Terminology" into the UMLS. The results show that our algorithmic lexical manipulation offers gains in "source term/UMLS term" matches.

Background

In order to create a baseline for evaluating our methodology, we chose to use a set of terms (a) that had already been included in the UMLS in previous research by other authors, and (b) for which a published record on the integration process exists . The chosen set of gastro-intestinal (GI) terms is known as the Minimum Standard Terminology (MST) . The version of the MST included in the UMLS is designated "MTHMST2001," though we will continue to refer to it simply as "MST."

The MST's designers set out to devise a "minimal" list of terms that could be included within any computer system used to record the results of GI endoscopic examinations. Overall, it comprises 1,944 such terms, which represent 1,636 unique concepts. Of the terms, 289 have explicit synonyms. The concepts also exhibit relationships, e.g., *part_of* (85 concepts), *has_location* (198), *manifestation_of* (235), and *treats (*2).

The concepts of the MST, as integrated into the UMLS, are assigned a total of 29 semantic types out of the UMLS Semantic Network's 135 semantic types. The MST was originally integrated into the 2002AA release of the UMLS, as described in . Since the MST is not a terminology but rather a standard (given in a group of tables) for reports involving GI endoscopy, the major effort in focused on creating a terminology reflecting the MST's content. That terminology then became the actual source of the integration.

Using the data rich format of the UMLS , we were able to remove the MST to derive what we call the UMLS⁻, which is the UMLS with the MST entirely excluded (Figure 1). Naturally, a number of terms from the MST were also introduced into the UMLS by other terminologies. We call this overlap of MST terms with pre-existing UMLS terms the UMST. In creating the UMLS⁻ and the MST, we used the 2006AC release of the UMLS, installed in our Oracle database server. The UMST has 390 terms with 328 concepts in this release. The goal of our larger research project is to reintegrate the MST (as algorithmically as possible) into the UMLS⁻ and compare the results with those of .

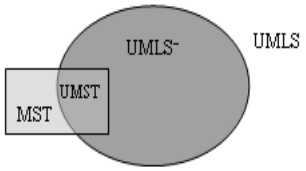

Figure 1 - The relationships between UMLS, UMLS⁻, MST, and UMST

A major step in the reintegration of the MST into the UMLS⁻ is the identification of the concepts of the UMST. Due to the addition of new sources and the update of old ones, there may be more such concepts now than in 2002, when the original integration took place. Thus, as a first step, we concentrated on the problem of finding all concepts residing in the UMST.

In a preliminary study, we found a surprisingly low number of matches between terms from the MTHMST2001 and terms in the UMLS⁻. Only 216 out of 1,944 terms matched (11.11%). Even syntactic transformations, such as removing dashes, did not improve matching results in any significant way.

The low rate of matches between the MST and the UMLS⁻ is surprising because the area of GI diseases is a core medical subject that should be well covered by the UMLS. We hypothesized that many MST terms may exist in the UMLS expressed by their synonyms. Thus, the problem to be solved was to discover new synonyms of terms that already exist in the UMLS. The complex relationships between concepts, terms, and strings in the Metathesaurus are elucidated by Table 1. Each bold number in this table is a unique identifier. As the table shows, one unique concept may correspond to several lexical terms, which are synonyms of each other.

Table 1 - Example of the relationships between Concepts, Terms, Atoms, and Strings in the UMLS

Concept (CUI)	Lexical Terms (LUI)	Strings (SUI)	Atoms (AUI)
C0009375 Colonic tumor Colonic Neoplasm Colonic Neoplasms	**L1834866** Colonic tumor	**S2142448** Colonic tumor	**A0027665** Colonic tumor (MTHMST) **A7591481** Colonic tumor (NCI)
	L0009375 Colonic Neoplasm Colonic Neoplasms	**S0363987** Colonic Neoplasm	**A0401918** Colonic Neoplasm (MSH) **A7576442** Colonic Neoplasm (NCI)
		S0026544 Colonic Neoplasms	**A0041001** Colonic Neoplasms (MSH)

For finding synonyms of individual words, we used WordNet 2.0 and JWNL (Java WordNet Library) 1.3 , a WordNet API. This API (Application Programming Interface) makes it possible to access WordNet from Java programs. Terms in WordNet are grouped into sets of cognitive synonyms, called *synsets*, used to express distinct concepts. Synsets are interlinked by conceptual-semantic and lexical relations such as hyponym, hypernym, etc. The newest version of WordNet is 2.1; however, the JWNL API was so far only implemented for WordNet 2.0. As shown in Table 2, WordNet 2.0 contains 152,059 strings. WordNet 2.1 contains 155,327. This increase is relatively small. Therefore, we found it acceptable to use the data from Version 2.0, allowing us to take full advantage of JWNL.

Table 2 - Statistics of WordNet 2.0

POS	Unique Strings	Synsets	Total Word-Sense Pairs
Nouns	114648	79689	141690
Verbs	11306	13508	24632
Adjectives	21436	18563	31015
Adverbs	4669	3664	5808
Totals	152059	115424	203145

Methods

Every MST term that is to be integrated into the UMLS⁻ is processed in several steps. Figure 2 contains a flow chart, comprising only the steps discussed in this paper. Step 1: Look for perfect matches and for matches of simple lexical variants, e.g., transformation of all letters into lowercase. Step 2: If the term consists of a single word, use synonyms from WordNet. Step 2 is not a new approach. Such a technique has already been used by Mougin *et al.*, while solving a slightly different problem . Step 3: Employ our novel matching methodology based on "piecewise synonyms."

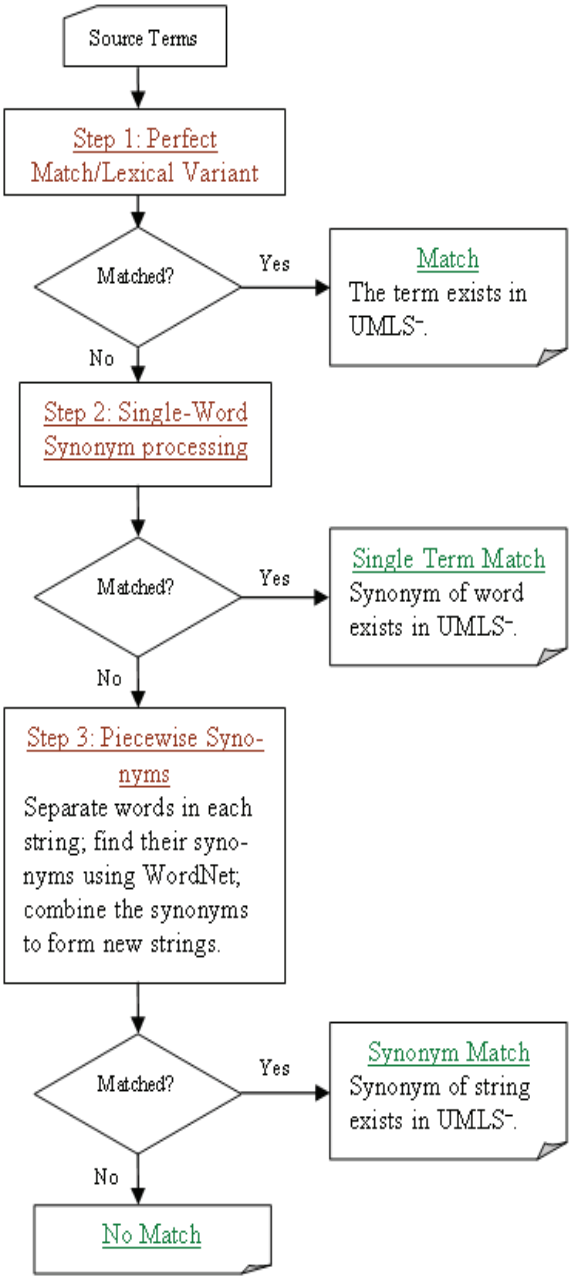

Figure 2 - Partial flowchart for string matching

Our method of constructing *piecewise synonyms* works as follows. If a medical term consists of two (or more) words, our algorithm looks in WordNet for synonyms of each individual word. Then the algorithm reconstructs two-word phrases from these individual synonyms. Then it looks for matches in the UMLS. To provide an intuitive, non-medical example, one might have a term "car insurance" and not find a match for it. Looking for synonyms

of "car" we find "automobile." Let us assume that we do not find any synonyms for "insurance." Now we reconstruct a two word term "automobile insurance" from the synonym of "car" and the original term "insurance" and attempt to find a match for it.

Our example was intentionally the simple case where only one of the two words in a two-word term has a synonym. The situation becomes more complicated if both of the terms have one or more synonyms. In this case, several piecewise synonyms are constructed. If T is a two-word term, we denote its set of piecewise synonyms as Piecewise_Synonyms(T).

We will formalize the Piecewise_Synonyms function using the Cartesian product. In simple terms, a Cartesian product creates every possible combination of members from one set with members of a second set, while maintaining the order between members. Thus, the Cartesian product ($\times$) of the set $\{a, b\}$ and the set $\{c, d\}$ would be:

$$\{a, b\} \times \{c, d\} = \{(a, c), (a, d), (b, c), (b, d)\} \quad (1)$$

In the following, we use the notation "word(T, n)" to be the nth word of the term T. The WordNet synset of a word w will be denoted synset(w). The operator $\cup$ is set union.

$$\text{Piecewise_Synonyms}(T) = \{\text{“}w_1\ w_2\text{”} \mid (w_1, w_2) \in (\text{synset}(\text{word}(T, 1)) \cup \{\text{word}(T, 1)\}) \times (\text{synset}(\text{word}(T, 2)) \cup \{\text{word}(T, 2)\})\} - \{T\} \quad (2)$$

In (2), "$w_1\ w_2$" stands for the string (i.e., potential synonym) consisting of w_1 as its first word and w_2 as its second word, such that (w_1, w_2) is an element of the Cartesian product of the synonyms of the two words in T. Note that the words of T are also utilized in the construction of the piecewise synonyms, as indicated by the two set union operations. For example, a potential synonym for T consists of its second word preceded by a synonym for its first word. This case was demonstrated above with "automobile insurance." Therefore, the complete term T must be excluded as a potential synonym of itself. This is done with the explicit use of the "set difference" operator (denoted "–") applied to the singleton set $\{T\}$ at the end of (2).

If any of the piecewise synonyms of T match an existing term in the UMLS$^-$, then we say that T is a newly discovered synonym of an existing UMLS$^-$ term. Our above approach may be easily extended to terms consisting of n (≥ 3) words. In that case, (2) would comprise an n-way Cartesian product. Its result would consist of n-tuples of words, instead of pairs of words.

One may be concerned that computing piecewise synonyms will result in nonsensical terms. For example, if "New York" is passed to the Piecewise_Synonyms() function, it might find that "Novel" is a synonym for "New" and return "Novel York." However, this is not a problem, for the following reason. We do not use a resulting piecewise synonym unless it is found as a matching term in the UMLS. Thus, we are using the UMLS itself as a filter to exclude nonsensical combinations. Any piecewise synonym that is found in the UMLS is by definition meaningful. Of course, it is recommended that a human expert review all results.

We now present a medical example. Table 3 shows the results of computing piecewise synonyms for "biliary tumor." This MST term was not found in the UMLS$^-$. However, one of the piecewise synonyms, "biliary neoplasm," was indeed found in the UMLS$^-$. Thus, we have now identified biliary tumor as a new, valid (after human review) synonym of biliary neoplasm.

Table 3 - Piecewise Synonyms of "Biliary tumor"

Original term	Biliary tumor	
Synonyms of separate words, including the words themselves	Biliary	tumor
	Bilious	neoplasm
		tumour
Piecewise synonyms, excluding the original term	Bilious tumor	
	Biliary neoplasm	
	Bilious neoplasm	
	Biliary tumour	
	Bilious tumour	

The UMLS contains many long terms. For example, "absence of bleeding of edematous duodenal mucosa" consists of seven words. Terms of more than eight words and terms with more than 100 characters occur. This would result in excessive computational runtimes. For example, the term "Ischemic colitis as reason for lower g.i. examination" would produce more than half a million potential piecewise synonyms ($2 \cdot 2 \cdot 20 \cdot 11 \cdot 1 \cdot 37 \cdot 2 \cdot 8$ = 520,960 combinations). To avoid this explosion of synonyms, our program only processes terms with fewer than six words. Even with this limitation, it takes hours to compute all potential piecewise synonyms and to check each of the generated piecewise synonyms against the whole UMLS list of terms, to verify a match.

Results

We have generated 785,155 synonyms from the MST with word length less than six. Table 4 shows the results of

matching piecewise synonyms against the UMLS. The "Matched Synonyms" column indicates how many synonyms were found. The "Matched Terms" column expresses the number of MST terms that generated these synonyms. As seen in the table, there are 132 synonyms, which came from 91 different MST terms, matched in the UMLS⁻. Of these 132 synonyms, 78 are piecewise synonyms from 65 different MST terms. The other 54 synonyms consist of one word.

Table 4 - MST matches based on synonyms

	Matched Synonyms	Matched Terms
Single-word Synonyms	54	26
Piecewise Synonyms	78	65
Total:	132	91

Table 5 shows the results of applying the matching process of Figure 1. The first and second rows (with two and three sub-rows, respectively) give the results of Step 1, i.e., perfect matches and lexical-variant matches, as defined by a transformation to lowercase letters only. For example, there were 142 perfect matches involving multi-word terms. (This number represents 3.81% of the total number of terms in the MST.) This number increased to 164 with lowercase letters only. Overall, the lexical-variant matching had an improvement of 10.19% over perfect matching. The third row in the table presents the results of Steps 2 and 3 in its first two respective sub-rows. For single-word terms, 26 matches (1.34%) were found. Finally, the main results show that the piecewise-synonyms match yielded 65 matches (3.34%). Of those 65, six are brand-new discoveries. That represents a 3.66% improvement over the results of the multi-word lowercase match.

Table 5 – Matching results

		# Matched Terms	% Matched
Perfect Match	**Single-Word**	74	3.81
	Multi-Word	142	7.30
Lower-case Match	**Single-Word**	74	3.81
	Multi-Word	164	8.43
	Improvement over Perfect	22 (10.19%)	
Synonym Match	**Single-Word**	26	1.34
	Piecewise	65	3.34
	Piecewise improvement over Lowercase	6 (3.66%)	

Table 6 shows the number of matched terms with respect to the number of words per term. Because the UMLS is concept-oriented as opposed to term-oriented, we also present the corresponding numbers for concepts. For example, two-word terms in the MST matched a total of 38 terms in the UMLS⁻. (They also matched 38 concepts.) In total, two-word terms matched 50 synonyms.

Table 6 - Matched terms with respect to number of words

	Matched Terms	Matched Concepts	Matched Synonyms
Single-Word	26	25	54
Two-Word	38	38	50
Three-Word	20	20	20
Four-Word	5	4	6
Five-Word	2	2	2
Total:	91	89	132

Table 7 shows all new matches for MST terms achieved by the piecewise synonyms methodology (Step 3). For example, the piecewise synonym "Biliary neoplasm" was used to find a match for the term "Biliary tumor" in the UMLS⁻.

Table 7 - New terms found by generating piecewise synonyms

MST Term	Piecewise Synonym
Biliary tumor	Biliary neoplasm
Colo-rectal cancer	Colorectal cancer
2nd part of the duodenum	Second portion of the duodenum
Bile ducts stone	Bile duct stone
Modification of bowel habits	Change of bowel habit
Extravasation of contrast medium	Extravasation of contrast media

Discussion

As in , we are using WordNet to improve matching of terms with the UMLS. The approach in used "Data Elements" as input and WordNet to either validate a match or select between several matches. In our approach, WordNet matching is used as one step in a larger terminology integration project, with the focus of this paper on the generation of piecewise synonyms. The key positive result of this research was the ability to generate 78 *new, valid, multi-word terms* (new from the point of view of the MST). "Validity" means that meaningful multi-word terms, which exist in the UMLS, were found, as opposed to terms such as "Novel York."

We generated these 78 new terms from 65 MST source terms; 59 of these source terms already existed in the UMLS; however, six terms did not exist. Thus, six genuinely new mappings were found. While this may appear to

be a small number, one needs to keep in mind three points: (i) The MST is fairly small, as medical terminologies go, and the 78 new terms have to be considered in relation to that fact; (ii) the UMLS is a huge repository of terms, and finding *any* new terms in a core area such as GI is important; (iii) only 216 out of 1,944 MST terms were found in the case of perfect matching, which was unexpectedly low. Thus, even a small absolute improvement by other methods of matching is meaningful.

Limiting ourselves to terms with at most five words appears acceptable, as our results show that the longer a term is, the less likely it is that its synonym has a match. We found only two matched synonyms for five-word terms, but 50 synonyms matched two-word terms. Among the 65 matched source terms, we found piecewise synonyms with matches in the UMLS^{-}, which were semantically wrong, because they belonged to concepts outside the GI domain. This reinforces the need for human review of all results.

Another unexpected result was that "subdivision of pancreas" and its piecewise synonym "branch of pancreas" belong to two different UMLS concepts, which should not be the case. This suggests a new method for auditing UMLS terms.

Conclusions

In this paper, a new methodology for enhancing automated term matching between a source vocabulary and the UMLS was presented. The methodology avails itself of the WordNet lexical database in the construction of new, potential synonyms for multi-word terms. These piecewise synonyms were shown to be another useful tool in the toolbox of medical terminology integrators and curators. Additionally, we found that our approach can have an impact on auditing the UMLS. The work described herein suggests a number of directions for future research: (1) increased use of syntactic categories (adjectives versus nouns); (2) use of one-word synonyms from the UMLS itself for constructing piecewise synonyms; (3) use of WordNet subclasses (hyponyms) and superclasses (hypernyms) of given terms; (4) use of piecewise mapping with synonyms of partial multi-word terms; and (5) use of piecewise-synonym generation as part of a complete algorithm for integrating a terminology into the UMLS.

Acknowledgments

This work was partially supported by the United States National Library of Medicine under grant R 01 LM008445-01A2.

References

[1] Humphreys BL, Lindberg DAB, Schoolman HM, Barnett GO. The Unified Medical Language System: An Informatics Research Collaboration. JAMIA 1998;5(1):1–11.

[2] SNOMED CT - Systematized Nomenclature of Medicine-Clinical Terms. URL: http://www.nlm.nih.gov/ research/ umls/Snomed/snomed_main.html. **This and all Web sites were accessed on Nov. 30, 2006.**

[3] LOINC - Logical Observations Identifiers, Names, Codes. URL: http://www.nlm.nih.gov/research/umls/loinc_main.

[4] RxNorm - a standardized nomenclature for clinical drugs. URL: http://www.nlm.nih.gov/research/umls/rxnorm/.

[5] NLM - National Library of Medicine. URL: http:// www.nlm.nih.gov/.

[6] UMLS Release Notes and Problems. URL: http:// www.nlm.nih.gov/research/umls/release_notes.html.

[7] Vocabularies in the UMLS Metathesaurus. URL: http:// www.nlm.nih.gov/research/umls/source_faq.html.

[8] Sarkar IN, Cantor MN, Gelman R, Hartel F, Lussier YA. Linking Biomedical Language Information and Knowledge Resources in the 21st Century: GO and UMLS. Pacific Symposium on Biocomputing 2003; p. 439–450.

[9] Miller GA. WordNet: A Lexical Database for English. Communications of the ACM 1995;38(11):39–41.

[10] Tringali M, Hole WT, Srinivasan S. Integration of a standard gastrointestinal endoscopy. In: Kohane IS, editor. Proc. 2002 AMIA Annual Symposium. San Antonio, TX; 2002. p. 801–805.

[11] OMED - World Organisation of Digestive Endoscopy. URL: http://www.omed.org/index.php/resources/.

[12] UMLS Metathesaurus. URL: http://www.nlm.nih.gov/ research/umls/meta2.html.

[13] Schuyler PL, Hole WT, Tuttle MS, Sheretz DD. The UMLS Metathesaurus: Representing different views of biomedical concepts. Bull Med Libr. Assoc 1993;81(2):217–222.

[14] Old versions of WordNet. URL (download): http:// wordnet.princeton.edu/oldversions.

[15] JWNL - Java WordNet Library. URL (download): http:// sourceforge.net/projects/jwordnet.

[16] Mougin F, Burgun A, Bodenreider O. Using WordNet to improve the mapping of data elements to UMLS for data sources integration. Proc. 2006 AMIA Annual Symposium. Washington, DC; 2006. p. 574–578.

Address for correspondence

Kuo-Chuan Huang, Dept. of Computer Science, NJIT, Newark, NJ 07102, USA.
Email: kh8@njit.edu