Published in final edited form as:
Artif Intell Med. 2009 June ; 46(2): 97–109. doi:10.1016/j.artmed.2008.11.008.

Using WordNet Synonym Substitution to Enhance UMLS Source Integration

Kuo-Chuan Huang[a], **James Geller**[a], **Michael Halper**[b], **Yehoshua Perl**[a], and **Junchuan Xu**[a]

[a] Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102−1982 USA

[b] Department of Computer Science, Kean University, Union, NJ 07083−0411 USA

Summary

Objective—Synonym-substitution algorithms have been developed for the purpose of matching source vocabulary terms with existing Unified Medical Language System (UMLS) terms during the integration process. A drawback is the possible explosion in the number of newly generated (potential) synonyms, which can tax computational and expert review resources. Experiments are run using a synonym-substitution approach based on WordNet to see how constraining two methodological parameters, namely, “maximum number of substitutions per term” and “maximum term length,” affects performance. Our hypothesis is that these values can be constrained rather tightly —thus greatly speeding up the methodology—without a marked decline in the additional matches produced. Furthermore, we investigate whether a limitation on only the first of the two parameters is sufficient to achieve the same results.

Methods—A four-stage synonym-substitution methodology using WordNet is presented. A group of experiments is carried out in which the two methodological parameters “maximum number of substitutions per term” and “maximum term length” are varied. The purpose is to examine their effect on the growth in the number of potential synonyms generated and the associated loss of results. The experiments are based on the re-integration of the “Minimal Standard Terminology” (MST) into the UMLS. Synonym-substitution matches found to be inconsistent with the current content of the UMLS and thus deemed to be incorrect are further manually scrutinized as an audit of the original integration of the MST.

Results—An increase of 11% in the number of “MST term/UMLS term” matches was achieved using the synonym-substitution methodology. Importantly, this result prevailed when tight threshold values (such as a maximum of two synonym substitutions per term) were imposed on the parameters. Furthermore, it was found that limiting only the “maximum number of substitutions per term” parameter was sufficient to obtain the performance enhancement. During the additional audit phase, a number of the reported mismatches were actually seen to be correct, representing an additional 10% increase in the number of matches obtained.

Conclusion—A synonym-substitution methodology that utilizes WordNet is a useful automated aide in UMLS source integration. Experiments showed that there was a significant speed-up but no degradation in match results when the methodology's “maximum number of substitutions per term” parameter was relatively tightly constrained. The methodology also helped to discover errors in the MST's original integration, and improve the quality of the UMLS's conceptual content.

a **Address for correspondence** Kuo-Chuan Huang Dept. of Computer Science NJIT Newark, NJ 07102 USA Phone : +1 (973) 596 −3392 Fax: +1 (973) 596−5777 Email: kh8@njit.edu.

Keywords

UMLS; source integration; WordNet; synonym substitution; synonym generation; synonymy; integration audit

1. Introduction

The Unified Medical Language System (UMLS) [1] comprises a large terminological database covering the biomedical and health-related fields. This database has been populated via the integration of a variety of sources, including SNOMED CT [2], LOINC [3], NCI Thesaurus [4], MeSH [5], MedDRA [6], and RxNorm [7]. Currently, the number of sources is over 100, and plans call for the integration of more in the future [8]. The Metathesaurus [9], the UMLS's concept repository, presently contains over 1,500,000 concepts and 3,200,000 English-language terms [10].

The overall process of integrating a new source into the UMLS is defined by the National Library of Medicine to comprise four major phases [8]: (1) analysis and inversion, (2) insertion, (3) human editing, and (4) quality assurance. In general, the integration process tends to be labor-intensive and error-prone. As such, facilitating source terminology integration is a critical issue facing UMLS curators. As noted in [11], "vocabularies are added and updated using sophisticated lexical matching, selective algorithms, and expert review." Many algorithmic aides have been developed in this context. For example, the tools *norm* and MMTX [12], provided by the National Library of Medicine (NLM), and the BLAST (Basic Local Alignment Search Tool) [13], based on the work in [14], have been used to carry out term matching in the process of integrating GO [15] into the UMLS [16].

In [11,17], members of the UMLS editorial team presented a number of techniques used in the process of finding cases of synonymy (which were actually missed by other methods). One of these techniques employs word-level synonym substitution, where known synonyms of individual words, retrieved directly from the UMLS, are substituted into a multiword phrase in an attempt to form new synonyms of the overall phrase. For example, with "renal" being a known synonym of "kidney," the technique infers that "renal failure" is a synonym of "kidney failure" (which, in fact, is true) [11]. As such, these two phrases would be grouped in one concept within the UMLS. A noted drawback to this approach is that it can be very expensive from a computational standpoint.

We have previously formulated and employed a methodology [17] similar to that in [11]. The methodology utilized the UMLS itself as the synonym repository, and also inferred additional synonyms from the UMLS's set of stored synonyms. Re-integration of the Minimal Standard Terminology (MST) [18], a collection of gastro-intestinal (GI) terms, was used as the test-bed. The results showed that the synonym-substitution approach can indeed be helpful in finding more term matches during the insertion phase of the integration.

In this paper, we ran experiments with the synonym-substitution methodology in which two parameters constraining the methodology—namely, maximum number of allowed synonym substitutions per (multi-word) term and maximum term length (in words)—were varied. The goal was to examine how constraining these methodological parameters affects performance. Our hypothesis is that these values can be constrained rather tightly—with an accompanying significant speed-up in the methodology—without a marked decline in the additional matches produced. Moreover, we were interested in determining whether constraining only the "maximum number of substitutions per term" parameter is sufficient to obtain the same performance enhancement.

Unlike in previous work [11,17], our synonym source for the methodology was the widely used WordNet [19]. As our test-bed, we continued to use the MST, which was first removed in its entirety from the UMLS. The experiments were then run in an attempt to re-integrate it.

A domain expert manually scrutinized matches that were found by our experiments but were inconsistent with the content of the UMLS 2008AA and thus deemed incorrect. This constituted an audit of the MST's original integration into the UMLS. As it turned out, a number of these mismatches were actually seen to be correct on further review.

2. Background

2.1. Minimal Standard Terminology (MST)

The MST was originally integrated into the 2002AA release of the UMLS, as described in [20]. 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), *treats* (2), etc.

Since the MST was not created as a terminology *per se* but rather a standard (given in a group of tables) for reports involving GI endoscopy examination results, the major effort in [20] focused on creating a terminology reflecting the MST's content. That terminology then became the actual source of the integration.

Our experiments were conducted in the process of re-integrating the MST after its removal from the UMLS. Since the MST's original integration has been well documented [20], it serves as a good baseline with which to compare the results of our experiments.

2.2. WordNet

WordNet 2.0 [21] is a large lexical database of the English language. Terms in WordNet are grouped into sets of cognitive synonyms, called *synsets*. Each synset is used to express a distinct concept. Synsets are interlinked by conceptual-semantic and lexical relations such as hyponym (“subclass”), hypernym (“superclass”), synonym (“also see”), antonym, cause, coordinate term (“sibling”), entailment (“follows from”), holonym (“whole of”), meronym (“part of”), and attribute. WordNet 2.0 contains 152,059 strings with 115,424 synsets. Table 1 shows the distribution of words across the parts of speech.

2.3. The *norm* Tool

The NLM provides tools for lexical processing of UMLS terms [12], dealing, for example, with capitalization, syntactic variants, etc. One of them is *norm*, which takes one term and creates other “normalized” terms (e.g., in word form and capitalization) that have the same meaning. It transforms the original string into a lowercase version, without punctuation (such as a hyphen ‘-’), genitive markers (such as an apostrophe expressing possession), stop-words (e.g., ‘a,’ ‘the,’ ‘of,’ etc.), diacritics (i.e., symbols such as accent marks as in ‘Protégé’), and ligatures (two letters bound into one). It also transforms verbs into infinitive form and nouns into singular. In the case of a multi-word term, the constituent words are sorted in alphabetical order. In some situations, there are two different ways to normalize the same term, i.e., normalization is not unique. For example, “scleroses” could be the plural of the noun “sclerosis” or the third person singular of the verb “sclerose.” Thus, as a result of the normalization of a

list of terms, the length of the list often increases, providing additional terms that can be used for matching. We use *norm* to supplement two stages of the methodology employed in our experiments.

3. Methods

3.1. Experimental Design

The goal of any source-integration methodology is the identification of an existing UMLS concept that represents and can thus "house" a given term residing in a new integration source. In the absence of such a concept, the methodology would conclude that a new UMLS concept needs to be created. In our integration experiments, we have broken this process down into a sequence of string-matching stages, as illustrated in Figure 1. Let us emphasize that we are only processing multi-word source terms in the experiments. The reason for this is the fact that no combinatorial explosion is possible with single-word terms and our experiments aim at limiting such problems.

The first two stages serve to filter out source terms that can be found in the UMLS with the use of conventional techniques, namely, exact string matching (Stage 1) and normalized matching (Stage 2). The second stage uses *norm* as its normalization mechanism. If a match is found for a source term at Stage 1, then it is assumed to be valid, and no further processing is carried out. Of course, it is possible that the exact match is, in fact, invalid. (See, e.g. [22],.) However, such an assessment can only be made by a domain expert, and our concern here is primarily with the results of the synonym-substitution stages: Stages 3 and 4.

The same stance concerning successful matches is taken at Stage 2. If a normalized version of the source term—derived using *norm*—matches an existing concept, then processing of that term is halted and the match is deemed to be valid. Terms that match in Stage 1 or 2 are not passed on to Stage 3.

Stage 3 is the first of the synonym-substitution phases. At this stage, we attempt to algorithmically construct new synonyms of a given multi-word source term in order to find a match for it with an existing UMLS concept. These new synonyms, called *candsyns* (short for "*candidate synonyms*") [17], are derived with the use of the respective WordNet synsets of the words in the term. In the following, we formalize this synonym-substitution procedure.

Let T be a source term consisting of n (≥ 2) words. Its entire set of candsyns can be expressed using the Cartesian product. Recall that the Cartesian product ($\times$) of two sets, say, $\{a, b\}$ and $\{c, d\}$ is defined as:

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

In the following, we use "word(T, k)" to be the k^{th} word of the term T. The WordNet synset of a word w will be denoted synset(w). Note that synset(w) always includes w itself. The set of candsyns of T is then:

$$\{\text{``}w_1\ w_2\ \cdots\ w_n\text{''} \mid (w_1, w_2, \ldots, w_n) \in \text{synset}(\text{word}(T, 1)) \times \text{synset}(\text{word}(T, 2)) \times \cdots \times \text{synset}(\text{word}(T, n))\} - \{T\} \tag{2}$$

In (2), "$w_1\ w_2$ ... w_n" stands for the string consisting of w_1 as its first word, followed by w_2 as its second word, all the way through w_n as its last word, such that (w_1, w_2, ..., w_n) is an element of the n-way Cartesian product of the synsets of the words in T. Stated differently, the candsyns are those terms derived from T by replacing one or more of its words with their synonyms from WordNet. The words of T themselves are utilized in the construction of the candsyns. For

example, one candsyn of *T* comprises a synonym for its first word followed by *T*'s remaining original words (2 through *n*). Note that we explicitly exclude *T* itself from the set of candsyns because we have by this point already failed to find a match for it in Stages 1 and 2. Formally, this exclusion is written with the use of the "set difference" operator (denoted "−") applied to the singleton set {*T*} at the end of (2). The set defined in (2) will be denoted CandSyns(*T*).

As an example, consider the two-word term "biliary tumor" in the MST. To construct CandSyns ("biliary tumor"), the synsets of the individual words "biliary" and "tumor" are retrieved from WordNet. The synset of "biliary" is {"biliary", "bilious"}, and the synset of "tumor" is {"tumor", "neoplasm", "tumour"}. Then we have:

$$\begin{aligned}
&\text{CandSyns}(\text{``biliary tumor''})\\
&\quad = \{\text{``biliary''}, \text{``bilious''}\} \times \{\text{``tumor''}, \text{``neoplasm''}, \text{``tumour''}\} - \{\text{``biliary tumor''}\}\\
&\quad = \{\text{``biliary tumor''}, \text{``biliary neoplasm''}, \text{``biliary tumour''}, \text{``bilious tumor''},\\
&\qquad\quad \text{``bilious neoplasm''}, \text{``bilious tumour''}\} - \{\text{``biliary tumor''}\}\\
&\quad = \{\text{``biliary neoplasm''}, \text{``biliary tumour''}, \text{``bilious tumor''}, \text{``bilious neoplasm''},\\
&\qquad\quad \text{``bilious tumour''}\}
\end{aligned}$$

Note that there are a total of $2 \cdot 3 - 1 = 5$ candsyns for "biliary tumor." As it happens, this MST term is not found in the version of the UMLS with the MST excluded. Thus, an attempted re-integration of the MST would conclude that a new concept is needed for "biliary tumor." However, one of its candsyns, "biliary neoplasm," can indeed be found there. Thus, the synonym-substitution process would yield the result that "biliary tumor" should be denoted as a synonym of the existing UMLS concept "biliary neoplasm."

We take a more cautious approach in Stage 3 regarding the presumed validity of matches than in Stages 1 and 2. If a match is found between a candsyn *C* (of a multi-word term *T*) and a UMLS term associated with a concept having unique concept identifier (CUI) *U*, then a triple (*T*, *C*, *U*) is inserted into a table called the *potential-match table* (*PMT*). Processing then continues on with any remaining candsyns of *T*. That is, if a match is found, it does not imply a cessation of processing at this stage. The overall output of this stage is the PMT, which is supplied to a domain expert for review. The reason for operating in this manner is that it is very possible that multiple candsyns match UMLS terms and that some of the matches are incorrect. Thus, stopping the processing early with a single match could mean that an incorrect match precludes the discovery of a correct one. We are, of course, interested in knowing the number of correct candsyn matches that Stage 3 produces, even if these are accompanied by some extraneous invalid matches. Note that *T* might be matched against a CUI *U* via more than one candsyn. That would provide further evidence to support *T*'s merger into *U*. *T* might also be matched against multiple CUIs. In that case, some of those matches will certainly be incorrect, assuming as a first step that the UMLS itself is error-free.

One may be concerned that some candsyns will be nonsensical. For example, one candsyn of "false diverticulum"—a special diverticulum of the intestine—is "untrue diverticulum," derived from the synonymy of "untrue" and "false" in WordNet. This is clearly an absurd construction. However, this is not a problem because there is no way that this candsyn will result in a match. Thus, the UMLS itself serves as a filter to exclude nonsensical combinations. Any candsyn that is found in the UMLS is by definition meaningful.

If Stage 3 produces an empty PMT (i.e., no matches are found), then processing continues on at Stage 4. This final stage operates similarly to Stage 3. However, instead of trying to match *T*'s candsyns themselves (which was done at Stage 3), it attempts to match normalized versions of those candsyns. Specifically, the normalization of the candsyns is carried out using the

norm tool. For the sake of efficiency, the entire set of candsyns, CandSyns(*T*), already generated at Stage 3 is passed along to Stage 4. Formally, the attempted matches are with respect to the following set of terms denoted NormCandSyns(*T*):

$$\text{NormCandSyns}(T) = \bigcup_{C \in \text{Candsyns}(T)} norm(C),$$

where *norm*(*C*) is the set of terms generated by the *norm* tool with *C* as its input. Note that the length of a term in NormCandSyns(*T*) is not necessarily *n* words.

The output of Stage 4 is another table, called the *potential normalized match table* (*PNMT*), comprising triples (*T*, *O*, *U*), where *T* is the original source term, *O* is a normalized version of one of its candsyns, and *U* is the CUI of the UMLS concept containing a term that matched *O*. The entire PNMT is delivered to the domain expert for analysis. Therefore, as a result of reaching Stage 3 or Stage 4, the expert is presented with a table, either the PMT or the PNMT, for review. If, however, both stages produce empty tables, then, in all likelihood, the source term expresses a concept that does not currently exist in the UMLS and needs to be created.

3.2. Experiments with Different Parameter Threshold Values

There is the possibility of a combinatorial explosion when generating the set CandSyns(*T*), particularly when *T* consists of many words [17]. The UMLS does contain quite a few long terms, such as “absence of bleeding of edematous duodenal mucosa.” Our test-bed, the MST, has terms comprising 11 words! Consider the term “Ischemic colitis as reason for lower g. i. examination.” It alone would produce a set of more than 500,000 candsyns (2·2·17·11·1·38·2·8 = 508,288 combinations).

Generating all the candsyns of many such long terms would result in excessive computational runtimes, and hence hinder the usefulness and effectiveness of the synonym-substitution approach. This is especially true if the new source terminology contains tens of thousands or even hundreds of thousands of terms.

We are therefore interested in running experiments in which limits are imposed on the following parameters of the synonym-substitution methodology:

1. The maximum number of words per term that are allowed to be replaced by their WordNet synomyns.
2. The maximum length of a term (in words) to be processed.

We have performed a number of experiments adjusting these parameters to see how the results compare to the unrestricted methodology. We use the notation S_xL_y to denote such an experiment, where *x* is the maximum number of substitutions allowed per source term at a time. The length of a term to be processed is between 2 and *y*. $S_\infty L_\infty$ denotes the unrestricted methodology described in the previous section. In the context of the current work, $S_\infty L_\infty$ was done as a basis for performance comparisons.

To get an idea of the effect of these constraints on the behavior of the algorithm, consider the term “bleeding gastric tumor.” In WordNet, synset(“bleeding”) = {“bleeding”, “hemorrhage”, “haemorrhage”}, synset(“gastric”) = {“gastric”, “stomachic”, “stomachal”}, and synset (“tumor”) = {“tumor”, “tumour”, “neoplasm”}. Now, let us consider experiments $S_\infty L_\infty$ and $S_1 L_\infty$. Experiment $S_\infty L_\infty$ will generate $3 \cdot 3 \cdot 3 - 1 = 26$ candsyns at Stage 3 (see Table 2). $S_1 L_\infty$ which allows only one word to be replaced by its synonyms, will generate only $(3 - 1)$

+ (3 − 1) + (3 − 1) = 6 candsyns at that stage (Table 2). The number of candsyns generated by S_1L_∞ is lower than that of $S_\infty L_\infty$ by a significant factor.

In particular, we have performed the four experiments S_2L_5, S_2L_9, S_4L_5, and $S_\infty L_\infty$. Note that the candsyns generated by an experiment with a longer maximum term length and more synonym substitutions per term will include the candsyns generated by an experiment with a tighter maximum term-length constraint and fewer allowed substitutions. This necessarily implies fewer matches at Stage 3 (and Stage 4). For example, formally speaking, $\|S_2L_9\| \geq \|S_2L_5\|$ where "‖ ‖" means the total number of candsyns that match existing UMLS terms.

A natural question is whether we need to limit both parameters—the length of the terms and the maximum number of words per term that can replaced—in order to keep the number of generated candsyns manageable while keeping the matches at an acceptable level. Maybe it is sufficient to limit only one of them. If so, which one should it be? In order to check this possibility, we performed additional experiments. Specifically, we tried all combinations of S_xL_y where y was either 5, 9, or ∞ (no limitation on term length) and x was either 2, 4, or ∞ (no limitation on the number of synonym substitutions with respect to a given term).

3.3. Test-Bed: MST Re-integration into the UMLS

As a test-bed source for our experiments, we use the MST, which has been previously integrated into the UMLS. We started off by completely removing the MST from the UMLS. Our experiments deal with re-integrating the MST. The version of the UMLS entirely excluding the MST will serve as the target of the re-integration process. We refer to it as the "UMLS$^-$" (see Figure 2). Naturally, a number of concepts arising from the MST were also introduced into the UMLS via other terminologies. We call this overlap of MST-introduced concepts with pre-existing UMLS concepts the "UMST" (Figure 2). The UMLS concepts introduced exclusively by MST terms were removed along with those terms to make the intersection of the MST and the UMLS$^-$ meaningful.

In deriving the UMLS$^-$ and the UMST, we used the 2008AA release of the UMLS. Therefore, "UMLS08AA" denotes the UMLS with the MST included. In that release, the UMST has 331 concepts and 391 terms. Among its terms are 75 one-word terms and 316 multi-word terms. Ideally, our experiments should match all 391 MST terms originally residing in the UMST with their original UMLS concepts and fail to match all the remaining 1,553 terms residing exclusively in the MST. These latter terms would require the creation of new UMLS concepts.

The rationale behind experimenting with the re-integration of the MST rather than the integration of a brand new source is two-fold. First, the original integration of the MST is well documented [20]. Second, and more importantly, there is no need to involve a domain expert to determine the accuracy of the results. They can be checked automatically by simply consulting the original version of the UMLS prior to the MST's removal.

At this point, let us define the notions of *valid match* and *mismatch* with respect to the results of Stages 3 and 4. (As we remarked earlier, we are not so concerned about the accuracy of the matches obtained at Stages 1 and 2, which serve more as filters for the input to the synonym-substitution stages.) The entry (*T*, *C*, *U*) in the PMT produced by Stage 3 is a *valid match* if the CUI of *T* is *U* in the UMLS08AA. Otherwise, (*T*, *C*, *U*) is a *mismatch*. The two notions are defined analogously with respect to the PNMT at Stage 4. Again, let us emphasize that valid matches and mismatches can be determined automatically due to the fact that we are using terms whose concepts in the UMLS08AA are already known.

In a preliminary study, we found a surprisingly low number of exact matches between terms from the MST and terms in the UMLS$^-$. Only 217 out of 1,944 terms matched (11.16%). Even

syntactic transformations, such as removing dashes, did not improve the 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 even prior to the introduction of the MST. We assumed that many MST terms in fact exist as concepts in the UMLS but are denoted synonymously. Thus, the MST makes a good test-bed for our experiments.

Table 3 shows the distribution of terms in the UML−, the MST, and the UMST based on the length of each term (in words). For example, the UMLS^{-} contains 740,148 two-word terms; the MST has 264; and the UMST has 147. For the sake of completeness, the number of one-word terms is also shown, even though they are not processed in our experiments, rect the original MST integration effort [20]. In fact, we did discover some *incorrect mismatches*. That is, in some cases, our methodology reported a potential match between a MST term and a concept in the UMLS^{-}, but the MST term's original CUI was not the same as the matched concept's, indicating a mismatch. However, a human review contradicted this finding and showed that the mismatch was, in actuality, a correct match. Let us present two examples here. A complete list will be given below in the Results section.

In the first example, we find that the MST term *Gastric mass* (UMLS08AA CUI: C0038356) is returned as a match for the term *gastric mass* (C0577018) in the UMLS^{-}. When originally integrating the MST into the UMLS, *gastric mass* was made a synonym of *Stomach Neoplasms* (C0038356). However, it should have been made a synonym of *Mass of stomach* (C0577018), which already had a synonym *Gastric Mass*. Thus, the perceived mismatch of the methodology is really a mistake of the original integration of the MST into the UMLS. This example is of special interest because we can establish a relationship between the incorrect concept and the correct one. The concept *Stomach Neoplasms* should be in a narrower relationship to *Mass of stomach.*

In the second example, a new concept *Prosthesis result* (C0941227) was created specifically for the MST term *Prosthesis result*. A review of the UMLS^{-} shows that *Effect prosthetic device* (C0497149) was introduced as a concept into the UMLS by ICPC [23]. These two concepts basically carry the same meaning. Therefore, *Prosthesis result* should have been mapped to *Effect prosthetic device* (C0497149).

4. Results

4.1. Experiments S_2L_5 S_2L_9 S_4L_5 $S_\infty L_\infty$

The unconstrained experiment $S_\infty L_\infty$ was first run as a basis for comparing the effects of the threshold values. At Stage 1, 1,868 MST terms were processed, yielding 143 matches. Of these, 141 were valid and two were mismatches (see Table 4). With 143 terms eliminated from consideration at Stage 1, only 1,725 terms were processed at Stage 2. These produced 66 matches: 58 valid matches and eight mismatches. These results are identical for the other experiments, S_2L_5, S_2L_9, S_4L_5, because the parameters do not take effect until Stage 3.

During Stage 3 of $S_\infty L_\infty$, a total of 1,659 terms were processed. These yielded 41,731,186 candsyns. Among these, 11 matched concepts appearing in the UMLS^{-}, for a “hit rate” of 0.66% (= 11 / 1,659). As an example, the MST term “Biliary tumor” correctly matched the UMLS^{-} concept with CUI C0005426 via the candsyn “Biliary neoplasm.” The number of valid matches, as determined by inspecting the UMLS08AA, was three. There were eight mismatches, for an error rate of 72.7%. For example, a mismatch occurred for the term “Branches of Pancreas”, whose candsyn “subdivision of pancreas” incorrectly matched the UMLS^{-} concept with CUI C0733964. The overall processing time for this stage was recorded at 85 minutes. See the first row of Table 5.

Experiment $S_\infty L_\infty$ failed to find any matches for 1,624 terms in Stage 3, so these were subsequently processed in Stage 4. The total of 41,730,445 candsyns generated for these terms at Stage 3 were passed along for normalization, producing 80,716,976 normalized candsyns (see Table 6). Out of these, 12 yielded matches with the UMLS$^-$, for a rate of 0.74%. An example is the MST term "Bile leak" that correctly matched the UMLS$^-$ concept "Leakage of bile" (CUI: C0400997) via the normalized candsyn "bile leakage." It turned out that that was the only valid match. The 11 others were mismatches. Thus, the error rate was 91.7%. An example mismatch occurred between the normalized candsyn "device effect prosthetic" derived from the term "Prosthesis result" and the UMLS$^-$ concept with CUI "C0497149." This stage took a staggering 1.32 days to complete.

The experiment S_2L_5 processed only 1,211 terms at Stage 3 due to the "L_5" (i.e., five-word) restriction on the term length (see Table 5). These terms yielded a total of 113,459 candsyns, which represents a 99.7% reduction with respect to $S_\infty L_\infty$. The processing time for S_2L_5 was reduced accordingly to just half a minute (Table 5), 0.59% of the time required for $S_\infty L_\infty$. The experiment S_2L_9 produced 222,626 candsyns from 1,641 terms. (Only 18 MST terms that got past Stages 1 and 2 were greater than nine words in length.) Again, this represents a sharp reduction compared to $S_\infty L_\infty$, and we see a corresponding reduction in the processing time. These results confirm our intuition that the "S_2" restriction has a tremendous impact on the computational time required for the experiments.

The experiment S_4L_5 also processed 1,211 terms at Stage 3, the same number as S_2L_5, and produced 559,419 candsyns (Table 5). This is about five times as many as S_2L_5, but only 1.3% of the number produced by $S_\infty L_\infty$. The processing time for S_4L_5 was a little less than three times more than S_2L_5's, and, again, far below $S_\infty L_\infty$'s.

Interestingly, the numbers of matches (11), valid matches (3), and mismatches (8) were exactly the same for all four experiments at Stage 3. In fact, all the matches were exactly the same! It turned out that all the candsyns responsible for matches required no more than two synonym substitutions for their creation. Their numbers are shown in Table 7 based on their term's length. For example, of the candsyns derived from terms of length two, ten matched UMLS$^-$ concepts. Five of them were produced with just one synonym substitution; the remainder, with two synonym substitutions. Note that the reported numbers of matches (Column 2) include multiple matches for individual source terms. For example, the MST term "Gastrointestinal bleeding" has matched candsyns from both one synonym substitutions, such as "Gastrointestinal bleed," and two synonym substitutions, such as "GI bleed." Also, let us note that no candsyn derived from a term of length six or more resulted in a match. The hit rate of 0.9% for S_2L_5 and S_4L_5 was about 36% higher than for $S_\infty L_\infty$.

The Stage 4 results of S_2L_5, S_2L_9, and S_4L_5 mirror those of Stage 3, except for the orders of magnitude in time reduction compared to $S_\infty L_\infty$. Experiment S_2L_5 only required 3.7 minutes to carry out this stage (see Table 6). S_2L_9 took about twice as long, and S_4L_5 used 16 minutes. Again, we find equal numbers of matches (12), valid matches (1), and mismatches (11) for all four experiments S_2L_5, S_2L_9, S_4L_5, and $S_\infty L_\infty$. As noted above, the only valid match occurred for the MST term "Bile leak" using the normalized candsyn "bile leakage."

In aggregate, the four stages of our experiments found 203 correct matches in the UMLS$^-$. The rate of correct matches over all multi-word terms of the MST was 64% (= 203/316), while the rate of correct matches over all matched terms was 88% (= 203/232).

To conclude, Table 8 lists all valid matches for MST terms achieved by Stages 3 and 4 of the synonym-substitution methodology. For example, as noted above, the MST term "Biliary tumor" is matched using its candsyn "Biliary neoplasm" to a concept in the UMLS$^-$ at Stage 3.

4.2. Additional Experiments: Term Length vs. Number of Synonym Substitutions

In order to determine whether it was sufficient to limit only one of the two parameters, instead of restricting both, we looked at the results of nine experiments S_xL_y, with y having a value of either 5, 9, or ∞, and x having a value of either 2, 4, or ∞. Since the four experiments S_2L_5, S_2L_9, S_4L_5, and $S_\infty L_\infty$ reported on in the previous section all yielded the same number of matches, we know that our five additional experiments will do the same. Therefore, it is sufficient to examine the number of candsyns, in total, that each yielded. The runtime is proportional. These numbers appear in Table 9. For example, S_2L_5 generated 113,459 candsyns, while S_4L_9 generated 3,007,601.

As seen in the first column of Table 9, for a replacement of two words, the change in length did not change the magnitude of the number of generated candsyns. On the other hand, for each length, the change from replacement of two words to the replacement of four words caused an increase in magnitude. (For a term length of five, it is actually about a fivefold increase, as seen in the first row.) Our conclusion is that it is sufficient to limit the number of words that are replaced to two per term, but not to limit the term length. Hence, the methodology will be able to discover matches between very long source terms and UMLS terms that differ (synonymously) in up to two words.

We note that for the replacement of four words, the increase in length caused an increase in magnitude. The reason is that for length nine, for example, there are 126 (mathematically: choose 4 out of 9 without consideration of order) ways of choosing which four words to replace. Nevertheless, the magnitude is such that it is not prohibitively expensive to generate candsyns of all lengths.

4.3. Correction of Mismatches

As we discussed above, a number of candsyn matches were rejected as mismatches because their CUIs were different from those defined in the original state of the UMLS. On review by one of the authors (JX) who holds an MD, six of these seeming mismatches were deemed to be actually correct matches, thus exposing problems introduced during the original integration of the MST into the UMLS [20]. All of these are listed in Table 10.

As noted above, the MST term *Prosthesis result* is currently associated with the UMLS concept having CUI C0941227, but our experiments matched it to the UMLS$^-$ concept *Effect prosthetic device* (C0497149). In actuality, these two concepts should be consolidated into one.

In each of the other five cases listed in the table, the meaning of the MST term was broader than the meaning of the UMLS concept with which it is currently associated. The UMLS$^-$ concept that it matched via a candsyn turned out to have the same broader meaning and was thus deemed to be more suitable. For example, a gastric mass is not necessarily a neoplasm, but the UMLS08AA has the MST term gastric mass associated with the concept *Stomach Neoplasms* (C0038356). The concept matched using a candsyn, *Mass of stomach* (C0577018), is a better fit.

5. Discussion

Overall, the contributions of this paper include a formal treatment of a synonym-substitution methodology for term matching in UMLS source integration, as well as experiments varying two parameters that constrain the methodology in order to examine its efficiency. We also used WordNet as a synonym resource rather than the UMLS itself. A domain expert manually examined reported mismatches produced during an attempted re-integration of the MST source vocabulary into the UMLS. This effectively allowed us to audit aspects of a completed integration effort with algorithmic assistance. The examination of the mismatches revealed

that mistakes (e.g., incorrect term/concept associations) had been introduced into the UMLS during the original integration of the MST.

Like the technique in [24], we used WordNet to improve matching of terms with the UMLS. In [24], WordNet synsets were used to either validate/disambiguate a "data element" (DE) of a source if the DE had direct matches to the UMLS, or indirectly matched the DE to the UMLS via WordNet if the DE did not have such a direct match. In resolving unmatched concepts, the approach in [24] took the longest spanning syntagms for multi-word DEs, found their synsets in WordNet, and then found the synonyms or parents of the synsets to match against UMLS terms. Our methodology differs in that we first decompose the multi-word terms into individual words, find their synsets in WordNet, re-compose these synsets into candsyns, and finally match the candsyns against existing UMLS terms. We use WordNet exclusively for the generation of the candsyns.

While the synonym substitution methodology generated a lot of extraneous candsyns, it did manage to generate and match 23 multi-word terms (11 from Stage 3 and 12 from Stage 4) existing in the $UMLS^{-}$, as seen in Table 5 and Table 6. This process was completely automated, so the unmatched candsyns were no burden on the human editor.

In comparison, the rectification of the incorrect mismatches contributes 10% (= 6/58) more correct matches. When combining the contribution of the synonym-substitution methodology with the needed corrections uncovered by human review of the incorrect mismatches, a 17% (= 10/58) increase over the normalization process (Stage 2) is obtained. We note that the six matches that were originally judged to be mismatches appeared in various stages of the process. One arose in Stage 1, four in Stage 2, and one more in Stage 4.

The most significant finding of our work was that limiting the length (number of words) of terms, and the maximum number of words that may be replaced by their WordNet synonyms, dramatically reduced the total number of generated candsyns without affecting the quantity of the results. When integrating a large terminology source into the UMLS, this is critical, because the computing resources may be taxed by generating all candsyns (i.e., running the $S_{\infty}L_{\infty}$ experiment) of the whole terminology. While there is no guarantee that the results will always be optimal, as in our experiments, one may assume as a first approximation that the loss incurred by replacing only two words per term will be minimal.

A second significant finding was that there was *no* need to worry about the lengths of the terms being processed. The more significant parameter was the number of synonym replacements per term. Our experiments showed that it is sufficient to limit that parameter to a value of two. Removing the restriction on the term length did not entail any real penalty in regard to the number of candsyns generated.

One limitation of this study is the fact that WordNet does not contain a complete set of medical terms. A preliminary study using the UMLS itself to provide synonyms can be found in [17]. MST is a relatively small source terminology. Experimenting with larger UMLS source terminologies is needed to further assess the results of this study.

The work described herein suggests a number of directions for future research: (1) use of WordNet subclasses (hyponyms) and superclasses (hypernyms) of given terms; (2) use of multi-word phrase substitution instead of single-word synonym substitution; and (3) use of candsyn generation as part of a complete algorithm for integrating a terminology into the UMLS.

6. Conclusions

Algorithmic aides to the process of source term integration are necessary for the continued expansion of the UMLS. In this paper, we experimented with a methodology that employs WordNet synonym substitution as a means for producing matches between source terms and existing UMLS concepts that would not otherwise be found using simple string comparison. We were particularly interested in seeing what effect varying two methodological parameters, namely, “maximum number of substitutions allowed per term” and “maximum term length,” had on the performance of the methodology. Using the Minimal Standard Terminology (MST) as our test source—in a re-integration effort—the results showed that the methodology was effective in finding additional matches, and there was no degradation in its performance when the parameters were relatively tightly constrained. Thus, the methodology was seen to be able to perform very well in a reasonable amount of time. It is not necessarily subject to an overwhelming explosion of generated terms often accompanying synonym-substitution approaches. In fact, it is unnecessary to limit the lengths of the source terms being processed in order to avoid such an explosion.

An additional benefit of our experiments was an audit of the MST's original integration into the UMLS. The methodology found some “MST term/UMLS concept” matches that were inconsistent with the current content of the UMLS 2008AA and therefore deemed to be incorrect. However, on further inspection, some of these matches were found to actually be correct and should supplant the originals. Overall, this enhanced the performance of the methodology with 10% more matches.

Acknowledgments

This work was partially supported by the NLM 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. Journal of the American Medical Informatics Association 1998;5(1):1–11. [PubMed: 9452981]
2. SNOMED CT - Systematized Nomenclature of Medicine-Clinical Terms. [14 August 2008]. Available from: http://www.ihtsdo.org/our-standards/snomed-ct/
3. LOINC - Logical Observations Identifiers, Names, Codes. [14 August 2008]. Available from: http://www.nlm.nih.gov/research/umls/loinc_main.html
4. NCI - National Cancer Institute. [Accessed: 14 August 2008]. Available from: http://nci.nih.gov
5. MeSH - Medical Subject Headings. [14 August 2008]. Available from: http://www.nlm.nih.gov/mesh/meshhome.html
6. MedDRA - Medical Dictionary for Regulatory Activities Terminology. [14 August 2008]. Available from: http://meddramsso.com/
7. RxNorm - a standardized nomenclature for clinical drugs. [14 August 2008]. Available from: http://www.nlm.nih.gov/research/umls/rxnorm/
8. Vocabularies in the UMLS Metathesaurus. [14 August 2008]. Available from: http://www.nlm.nih.gov/research/umls/source_faq.html
9. Schuyler PL, Hole WT, Tuttle MS, Sherertz DD. The UMLS Metathesaurus: representing different views of biomedical concepts. Bulletin of the Medical Library Association 1993;81(2):217–222. [PubMed: 8472007]
10. UMLS Release Notes and Problems. [14 August 2008]. Available from: http://www.nlm.nih.gov/research/umls/release_notes.html
11. Hole, W.; Srinivasan, S. Discovering Missed Synonymy in a Large Concept-Oriented Metathesaurus.. In: Overhage, JM., editor. AMIA Annual Symp. Los Angeles, CA: 2000. p. 354-358.

12. NLM Lexical Tools. [14 August 2008]. Available from: http://lexsrv3.nlm.nih.gov/SPECIALIST/Projects/lvg/current/index.html
13. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic Local Alignment Search Tool. Journal of Molecular Biology 1990;215(3):403–410. [PubMed: 2231712]
14. Krauthammer M, Rzhetsky A, Morosov P, Friedman C. Using BLAST for Identifying Gene and Protein Names in Journal Articles. Journal of Functional and Evolutionary Genomics 2000;259(1 −2):245–252.
15. Gene Ontology Consortium - Gene Ontology. [14 August 2008]. Available from: http://www.geneontology.org/
16. Lomax J, McCray AT. Mapping the Gene Ontology into the Unified Medical Language System. Comparative and Functional Genomics 2004;5(5):354–361. [PubMed: 18629164]
17. Huang, K.-c.; Geller, J.; Halper, M.; Cimino, JJ. Piecewise Synonyms for Enhanced UMLS Source Terminology Integration.. In: Teich, JM.; Suermondt, J.; Hripcsak, G., editors. Proc. AMIA Annual Symp. Chicago, IL: 2007. p. 339-343.
18. OMED - World Organisation of Digestive Endoscopy. [14 August 2008]. Available from: http://www.omed.org/index.php/resources/
19. Miller GA. WordNet: A Lexical Database for English. Communications of the Association for Computing Machinery 1995;38(11):39–41.
20. Tringali, M.; Hole, WT.; Srinivasan, S. Integration of a standard gastrointestinal endoscopy.. In: Kohane, IS., editor. Proc. AMIA Annual Symp. San Antonio, TX: 2002. p. 801-805.
21. WordNet 2.0. [14 August 2008]. Available from: http://wordnet.princeton.edu/oldversions
22. Johnson, HL.; Cohen, KB.; Hunter, L. A fault model for ontology mapping, alignment, and linking systems.. In: Altman, RB., et al., editors. Pacific Symposium on Biocomputing. Maui, HI: 2007. p. 233-244.
23. ICPC - The International Classification of Primary Care. [14 August 2008]. Available from: http://www.who.int/classifications/icd/adaptations/icpc2/en/index.html
24. Mougin, F.; Burgun, A.; Bodenreider, O. Using WordNet to improve the mapping of data elements to UMLS for data sources integration.. In: Bates, D., editor. Proc. AMIA Annual Symp. Washington, DC: 2006. p. 574-578.

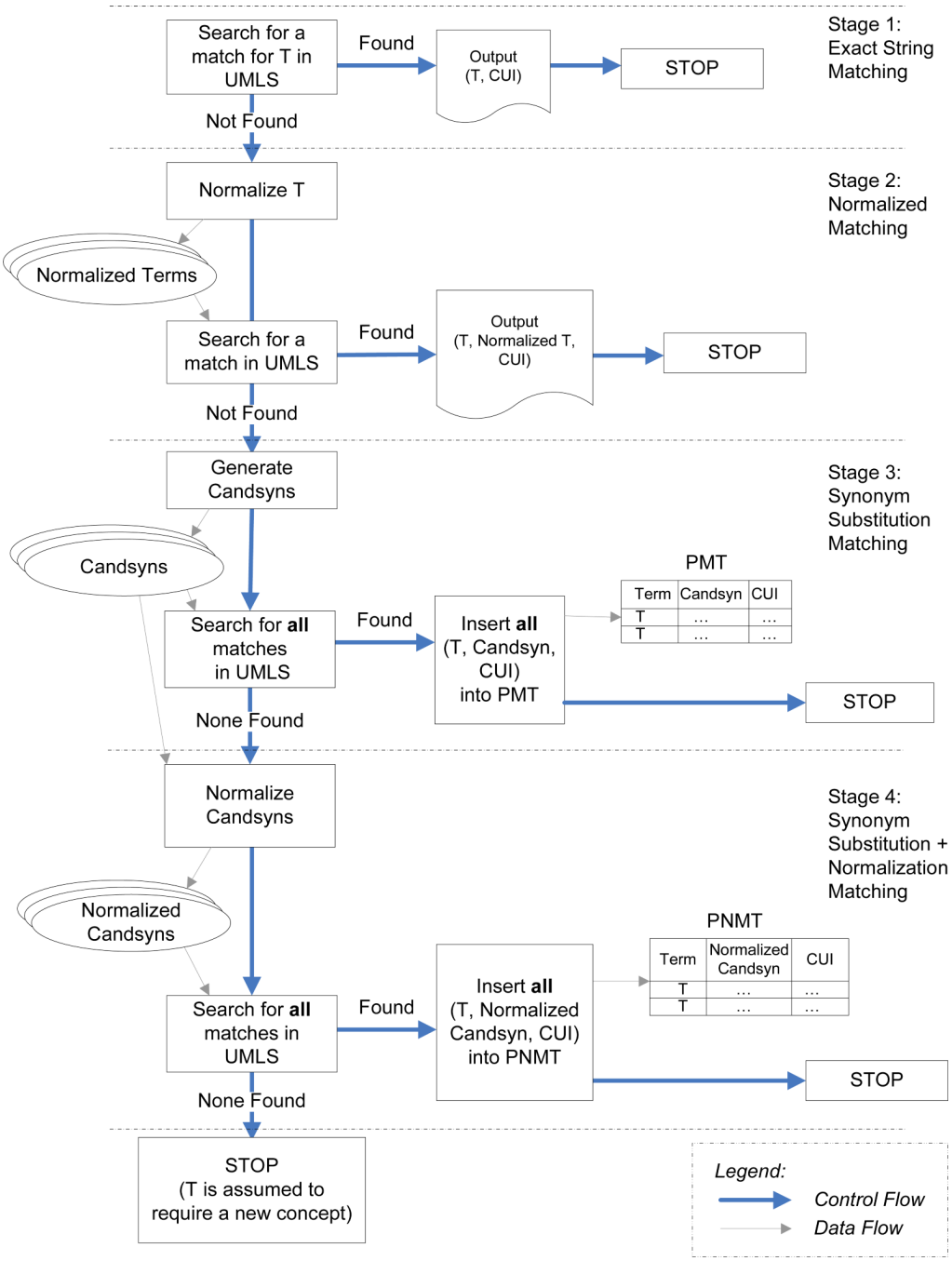

Figure 1.
Overall flow for processing a source term *T*

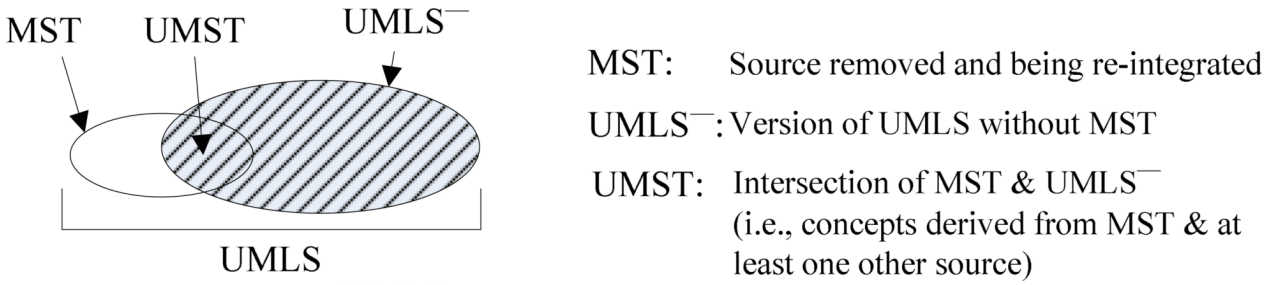

Figure 2.
Relationships between the UMLS, UMLS⁻, MST, and UMST

Table 1

Word distribution in WordNet 2.0

Part of speech	# Unique strings	# Synsets
Noun	114,648	79,689
Verb	11,306	13,508
Adjective	21,436	18,563
Adverb	4,669	3,664
Total:	**152,059**	**115,424**

Table 2

CandSyns("bleeding gastric tumor") in experiments $S_{\infty}L_{\infty}$ and S_1L_{∞}

Candsyns generated in experiment $S_{\infty}L_{\infty}$	**1 word substituted:** *hemorrhage* gastric tumor, *haemorrhage* gastric tumor, bleeding *stomachic* tumor, bleeding *stomachal* tumor, bleeding gastric *tumour*, bleeding gastric *neoplasm* **2 words substituted:** *hemorrhage stomachic* tumor, *haemorrhage stomachic* tumor, *hemorrhage stomachal* tumor, *haemorrhage stomachal* tumor, *hemorrhage* gastric *tumour*, *hemorrhage* gastric *neoplasm*, *haemorrhage* gastric *tumour*, *haemorrhage* gastric *neoplasm*, bleeding *stomachic tumour*, bleeding *stomachic neoplasm*, bleeding *stomachal tumour*, bleeding *stomachal neoplasm*, **3 words substituted:** *hemorrhage stomachic tumour*, *haemorrhage stomachic neoplasm*, *hemorrhage stomachal tumour*, *hemorrhage stomachal neoplasm*, *haemorrhage stomachic tumour*, *hemorrhage stomachic neoplasm*, *haemorrhage stomachal tumour*, *haemorrhage stomachal neoplasm*
Candsyns generated in S_1L_{∞}	*hemorrhage* gastric tumor, *haemorrhage* gastric tumor, bleeding *stomachic* tumor, bleeding *gastric* tumour, bleeding *stomachal* tumor, bleeding gastric *neoplasm*

Table 3

Term distribution by length in the UMLS⁻, MST, and UMST

Term length (# words)	# in UMLS⁻	# in MST	# in UMST
1	332,282	77	75
2	740,148	264	147
3	697,913	512	112
4	547,317	412	28
5	439,517	230	21
6	309,609	184	2
7	205,280	98	1
8	146,411	105	5
9	103,072	44	–
10	75,042	13	–
≥ 11	193,886	5	–
Total:	**3,790,477**	**1,944**	**391**

Table 4

Results of Stages 1 and 2 for $S_{\infty}L_{\infty}$

Stage	# Terms processed	# Matches	# Valid matches	# Mismatches
1	1,868	143	141	2
2	1,725	66	58	8
Total:	N/A	209	199	10

Table 5

Results of Stage 3 for the different experiments

Experiment	# Terms processed	# Candsyns generated	# Matches	# Valid matches	# Mismatches	Processing time (in minutes)
$S_{\infty}L_{\infty}$	1,659	41,731,186	11	3	8	85.0
S_2L_5	1,211	113,459	11	3	8	0.5
S_2L_9	1,641	222,628	11	3	8	0.8
S_4L_5	1,211	559,419	11	3	8	1.4

Table 6

Results of Stage 4 for the different experiments

Experiment	# Terms processed	# Normalized candsyns generated	# Matches	# Valid matches	# Mismatches	Processing time (in minutes)
$\mathbf{S_{\infty}L_{\infty}}$	1,624	80,716,976	12	1	11	1,895.0
$\mathbf{S_2L_5}$	1,176	136,902	12	1	11	3.7
$\mathbf{S_2L_9}$	1,606	291,393	12	1	11	7.3
$\mathbf{S_4L_5}$	1,176	677,856	12	1	11	16.4

Table 7

Matched candsyns for different term lengths

Length of term (# words)	# Matched candsyns	# Produced by a single synonym substitution	# Produced by exactly two synonym substitutions
2	10	5	5
3	2	2	–
4	1	–	1
5	1	–	1
≥ 6	–	–	–
Total:	**14**	**7**	**7**

Table 8

Valid matches obtained using candsyn or normalized candsyns (* = normalized)

MST term	Matched candsyn
Biliary tumor	Biliary neoplasm
2nd part of the duodenum	Second portion of the duodenum
Bile leak	bile leakage*
Modification of bowel habits	Change of bowel habit

Table 9

Numbers of candsyns as a function of term length and number of synonym substitutions

Maximum number of synonym substitutions allowed / **Maximum term length (# terms processed)**	**2**	**4**	∞
5 (1,211)	113,459	559,419	560,707
9 (1,641)	222,628	3,007,601	27,573,352
∞ (1,659)	238,811	4,091,355	41,731,186

Table 10

Incorrect mismatches resolved by domain expert analysis

MST term	Current incorrect UMLS concept (CUI)	UMLS⁻ concept matched and proposed as correct (CUI)
Prosthesis result	Prosthesis result (C0941227)	Effect prosthetic device (C0497149)
cannulation duct pancreatic	Endoscopic insertion of stent into pancreatic duct (C0400522)	Cannulation of pancreatic duct (C0176945)
Papillary stenosis	Papillary stenosis as diagnosis for pancreas (C0700377)	PAPILLARY STENOSIS (C0238340)
Bleeding of duodenal mass	Bleeding of duodenal tumor (C0947627)	DUODENAL MASS BLEEDING (C0743305)
Gastric mass	Stomach Neoplasms (C0038356)	Mass of stomach (C0577018)
Esophageal mass	Esophageal Neoplasms (C0014859)	Esophageal mass (C0577008)