Piecewise Synonyms for Enhanced UMLS Source Terminology Integration

Kuo-chuan Huang, MS[1], James Geller, PhD[1], Michael Halper, PhD[2], James J. Cimino, MD[3]
[1]New Jersey Institute of Technology, Newark, NJ; [2]Kean University, Union, NJ; [3]Columbia University, New York, NY

Abstract

The UMLS contains more than 100 source vocabularies and is growing via the integration of others. When integrating a new source, the source terms already in the UMLS must first be found. The easiest approach to this is simple string matching. However, string matching usually does not find all concepts that should be found. A new methodology, based on the notion of piecewise synonyms, for enhancing the process of concept discovery in the UMLS is presented. This methodology is supported by first creating a general synonym dictionary based on the UMLS. Each multi-word source term is decomposed into its component words, allowing for the generation of separate synonyms for each word from the general synonym dictionary. The recombination of these synonyms into new terms creates an expanded pool of matching candidates for terms from the source. The methodology is demonstrated with respect to an existing UMLS source. It shows a 34% improvement over simple string matching.

Introduction

The Unified Medical Language System (UMLS)[1] is a large terminological database containing medical terms from many sources, e.g., SNOMED CT,[2] LOINC,[3] and NCI.[4] Currently, the UMLS Metathesaurus contains over 100 source vocabularies with more than 1,300,000 concepts and over 6,000,000 terms.[5] The UMLS is continually being extended via integration of new sources.[6] The integration of a new source terminology into the UMLS is labor-intensive and error-prone. The National Library of Medicine has defined four major phases of the integration process to assure the quality of each newly integrated vocabulary.[7] One major task is the identification of terms and associated concepts from the new source that already exist in the UMLS. We present a methodology for increasing the effectiveness of locating such concepts.

The same concept may be expressed in many different ways in different sources. Thus, it is sometimes difficult to match a term from a new source with the correct concept in the UMLS, even with the help of lexical tools provided by the National Library of Medicine, such as MetaMap and norm.[8]

To overcome this difficulty, our methodology takes advantage of preexisting UMLS synonyms in two related but different ways. We first generate additional *general synonyms*. Then we generate *candidate synonyms* from source terms, making use of both preexisting and our new general synonyms. Whenever a candidate synonym matches an existing UMLS term, we can designate its source term as a synonym for a UMLS concept. That source term is referred to as a *piecewise synonym* (*PS*). We motivate and describe the steps of generating general and candidate synonyms. As an experiment, the partial re-integration of an existing source, the Minimum Standard Terminology (MST), is carried out with our methodology. The results are presented.

Background

In order to create a baseline for evaluating our integration methodology, we chose the MST of Gastro Intestinal terms for which a published record of the integration process exists.[9] The MST's designers devised a "minimal" list of terms for recording the results of Gastro Intestinal endoscopic examinations. Overall, it comprises 1,944 terms, which represent 1,636 unique concepts. The concepts also exhibit relationships, e.g., part_of, has_location, treats, etc. The MST was originally integrated[9] into the 2002AA release of the UMLS. Since the MST is not a terminology, but rather a standard (given in tables), the major effort of Tringali et al.[9] focused on creating a terminology from the MST. That terminology then became the source of the integration.

Using the rich data format of the UMLS,[10] we were able to remove the MST from the UMLS 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 preexisting UMLS terms the UMST. In creating the UMLS$^-$ and extracting a version of the MST, we used the 2006AC[11] release of

the UMLS. The UMST has 390 terms with 328 concepts in this release.

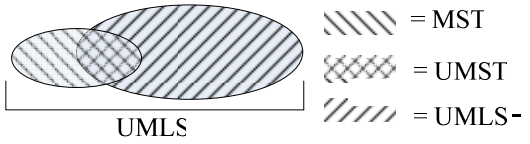

Figure 1. Relationships between UMLS, UMLS⁻, MST, and UMST

A major step in the reintegration of the MST into the UMLS⁻ is the identification of the concepts of the UMST. In a preliminary study, we found a surprisingly low number of string matches between terms from the MST and from the UMLS⁻. Only 208 out of 1,944 terms matched (10.7%). Even syntactic transformations, e.g., removing dashes, did not improve results significantly. Table 1 shows the numbers of UMST terms with their lengths (in words) and for how many of those we found a string match.

Table 1. UMST with Perfect Match

Term length (words)	# in UMST	# Perfectly Matched
1	75	66
2	159	100
3	104	32
4	23	5
5	21	3
6	3	2
> 6	5	0
Total:	390	208

The low rate of matches between the MST and the UMLS⁻ is surprising because the area of Gastro Intestinal diagnoses should be well covered by the UMLS. We hypothesized that many MST terms may exist in the UMLS, expressed by their synonyms in the Metathesaurus. Thus, the problem is to discover new synonyms of terms that already exist in the UMLS.

Methods

Starting with a term from the new source (Figure 2), we first generate an entire set of candidate synonyms ("candsyns," for short) that are used in an attempt to find a corresponding (existing) concept in the UMLS Metathesaurus. The candsyns are generated in two ways. One approach uses only pairs of preexisting synonyms of the UMLS in a process of word substitution in the source term. The second approach first generates additional synonyms from the UMLS synonyms, based on a subsequence analysis. Then the same method as in the first approach is applied. If a candsyn is found matching a UMLS term, then we refer to its *source term* as a piecewise synonym (due to the manner in which the discovery was made) and postulate that the respective UMLS concept should have the source term as a new synonym. That is, there is no need to create a new concept in the UMLS. Of course, this must eventually be confirmed by expert review. Our methodology and the processing of one term are shown in Figure 2. In the following, we describe the details of each major step.

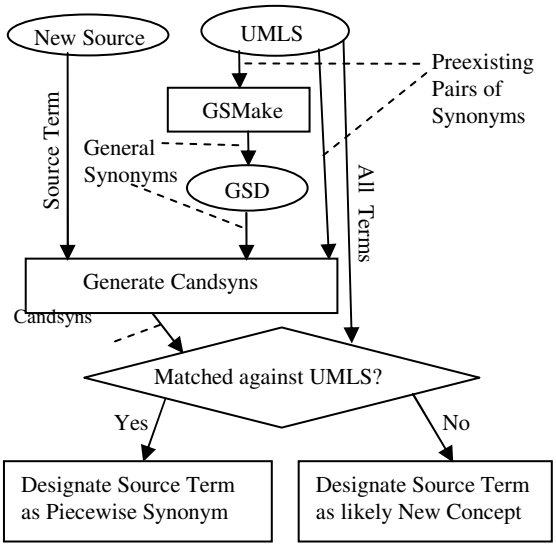

Figure 2. Overall Flow of Synonym Generation

Let us first describe the process of generating candsyns based on preexisting UMLS synonyms. A source term is decomposed into its constituent words and the UMLS synonyms are retrieved for each word individually. The candsyns are then created by combining single-word synonyms with each other. For example, the term "colon anastomosis" is broken down into the words "colon" and "anastomosis" and the synonyms of each are retrieved. There are eight synonyms for "colon," two of which are "colonic" and "large intestine," and two synonyms for "anastomosis," one of which is "anatomical anastomosis." The candsyns created according to the above examples are shown in Table 2. Showing all candsyns for all synonyms would take too much space. When comparing the candsyn list to terms in the UMLS⁻ we find that one candsyn, "large intestine anastomosis" is actually a concept in the UMLS⁻. An expert review confirms that these two terms are synonyms and "colon anastomosis" is in the UMST.

Table 2. Some candsyns for "colon anastomosis"

colon anatomical anastomosis
colonic anastomosis
colonic anatomical anastomosis
large intestine anastomosis
large intestine anatomical anastomosis

The process of expanding the pool of synonyms afforded by the UMLS is carried out with an algorithm called GSMake (Figure 2). Its underlying assumption is that if two multi-word synonyms share common subsequences of words, then their non-common parts may well be synonyms that have not been explicitly recorded in the UMLS. For example, if "large bowel anastomosis" is a synonym of "colonic anastomosis," then there is a good chance that "colonic" is a synonym of "large bowel."

Each new synonym resulting from GSMake is called a *general synonym* and is placed into the *general synonym dictionary* (*GSD*), used eventually in the generation of candsyns. GSMake processes each pair of multi-word terms t_1 and t_2 having the same concept ID and compares their component words looking for common word(s) and/or phrase(s). If t_1 and t_2 contain such common components, those pieces are discarded. The remaining words (phrases) in t_1 and t_2 are possible general synonyms. GSMake tries to find the longest common phrase. For example, assume t_1 = "A B C G" and t_2 = "D F G A B," where each letter represents one word. GSMake will find that "A B" is the longest common subsequence in both terms and that "G" in both t_1 and t_2 is a common word. The remaining parts, "C" from t_1 and "D F" from t_2, are potential general synonyms.

The remaining parts are not considered to be general synonyms in the following two cases: (1) there are non-consecutive words in any one of them; (2) the common phrase or word has a local overlap. For example, t_3 = "A C D E" and t_4 = "C D F" do not give rise to a general synonym because the remaining parts of t_3 are "A" and "E," which are not consecutive words. The two terms t_5 = "A B C D A" and t_6= "A B E" do not have general synonyms either, because the common word "A," from the end of t_5 and the beginning of t_6, and the common phrase "A B" have an overlap, namely, "A." Further details of GSMake are omitted for the sake of brevity.

As an example, in the UMLS$^-$, the terms "Abdominal neoplasm" and "Abdominal tumor" have the same concept ID and are thus synonyms. GSMake eliminates the common word "Abdominal" and postulates that *tumor* and *neoplasm* are synonyms. An entry (neoplasm, tumor) is made in the GSD. Once the GSD has been fully populated, the creation of candsyns proceeds nearly exactly as before. However, instead of accessing the UMLS directly to obtain the synonyms, the GSD is utilized for this purpose. Continuing the example, the GSD entry (neoplasm, tumor) and the MST term "Biliary tumor," which does not exist in the UMLS$^-$, yield a candsyn "Biliary neoplasm" that in fact exists there. Thus, the source term "Biliary tumor" is now recognized as a piecewise synonym of "Biliary neoplasm" and assigned to an existing concept.

One may be concerned that some general synonyms will be nonsensical. However, this is not a problem because the goal is to match the source term with a UMLS concept. Obviously, candsyns derived from nonsensical general synonyms will fail in this regard, and will not lead to wrong piecewise synonyms. In addition, semantic rules based on semantic types could be used to eliminate undesirable combinations.

As noted, we are using the extracted MST as a test-bed for our approach. In this context, the UMLS$^-$ (the version of the UMLS prior to MST integration) played the role of the UMLS in Figure 2. To have a proper basis for comparison, we performed three separate experiments. First, we used only the preexisting UMLS synonyms to create candsyns. Then we used only the GSD as the synonym source. Finally, the two groups of synonyms were combined.

Let us note that many entries in the GSD are expected to have dozens or even hundreds of synonyms, which would result in excessive computational runtimes. For example, the term "Benign intrinsic colonic stenosis" would produce more than two million candsyns (79*16*37*44 = 2,057,792 combinations). However, in a preliminary study, most of the piecewise synonyms were discovered by substituting only one or two words with their synonyms no matter what the length (in words) of the term was. To avoid an explosion of the number of candsyns, our program replaces at most two words in each source term by their synonyms. Stop words, such as "the," "of," "a," etc., are not processed in this study.

Results

Side-by-side comparisons of the results of the three experiments are shown in Table 3. The GSD column represents the results of using the general synonyms dictionary only. The Preexisting column lists the results of using preexisting UMLS synonyms. The "Both" column contains the results when using general and preexisting synonyms together. For example, the candsyns generated using only the GSD resulted in 251 PSs (see "Matched Terms" row in Table 3). Of these, 139 (in the "Correct Mappings" row) were correct, as defined by the original integration of the MST[9], for a rate of 55.4% (=139/251). The other 112 (44.6%) represented wrong mappings. With the GSD, 20.3% of the terms in the UMST were not matched when they should have been. The candsyns actually allowed us to discover 48 matched terms beyond those found

strictly with perfect matching (see last row of Table 3). As several candsyns are generated for many terms, there are cases when one candsyn matches a correct UMLS⁻ term, while another candsyn matches a wrong term. Those cases are called "Hybrid Mappings" in Table 3. What stands out in Table 3 is that most of the correct PSs were discovered by replacing only one word with its synonyms. (Those are in the row labeled "1-word synonymy.") For example, 101 such correct PSs were found using preexisting synonyms.

Table 3. Matching results

	GSD	Preexisting	Both
Matched Terms	251	109	259
Correct Mappings	139	101	148
Wrong Mappings	112	8	111
Hybrid Mappings	106	3	111
1-Word Synonymy	134	101	143
Additional Terms Found w/ candsyns	48	24	49

As shown in Table 1, the UMST contains 315 multi-word terms, about half of which are two-word terms. Our PS approach, using both the GSD and preexisting synonyms, matched a total of 148 of these. Most of the matches (109/148 = 74%) occurred with respect to two-word terms (see second column in Table 4). In comparison, the perfect match approach had a total of 142 matches, with 100 of these being two-word. The combination of perfect matching and our PS approach found a total of 191 matches, with 129 two-word matches. Thus, there was an improvement of 49 total (and 29 two-word) additional matches over the perfect match approach. This allowed for the discovery of a total of 191 matches for the 315 multi-word UMST terms (a 60% rate) using the GSD and preexisting synonyms with the combined "PS/Perfect Matching" approach.

Table 4. Correct matches with all approaches

Term length (words)	PS	Perfect	PS + Perfect	Improvement of PS + Perfect over Perfect
2	109	100	129	29
3	27	32	45	13
4	6	5	8	3
5	4	3	6	3
6	2	2	3	1
> 6	0	0	0	0
Total:	148	142	191	49

As seen in Table 4, our methodology discovered 49 new term matches between the MST and UMLS⁻, as compared to simple Perfect Matching, an improvement of 34.55. This evaluation was based on a comparison with an available gold standard, namely the previous integration of the MST into the UMLS.[9]

For example, the MST term "bile duct fistula" has a candsyn "biliary tract fistula" found in the UMLS⁻. Additional examples are in Table 5. The first column contains the source terms (now recognized as piecewise synonyms) of the MST that do not have string matches in the UMLS⁻. The second column represents the candsyns that created the matches. Note that one PS may be mapped to the UMLS⁻ through different candsyns. For example, "colon anastomosis" generated three candsyns which were all matched to the same concept in the UMLS⁻.

The GSD contains about one million entries, and it took about an hour of computer time to create it. The sizes of the candsyn tables for all three experiments (preexisting synonyms only, GSD only, combined) were about forty millions rows each, and they took about six hours each to generate. It then took about two hours of computing time, in each case, to determine all matching candsyns and identify all PSs.

Table 5. Sample of terms found with PS approach

	MST PS	UMLS⁻ Terms
1	bile duct fistula	biliary tract fistula
2	main bile duct tumor	common bile duct neoplasms
3	surgical gastrostomy	creation of gastrostomy
4	2nd part of the duodenum	second portion of the duodenum
5	thermal therapy	thermal techniques
6	colon anastomosis	1. colonic anastomosis 2. large bowel anastomosis 3. large intestine anastomosis
7	liver ducts	1. hepatic duct 2. structure of hepatic duct
8	ampullary tumor	1. ampulla of vater neoplasm 2. ampulla of vater tumor
9	peg (procedure)	percutaneous endoscopic gastrostomy (procedure)

Discussion

We found (Table 5) that the GSD offers the following advantages: (a) syntactic category replacement, e.g., of noun and adjective, such as for "bile" and "biliary" and "liver" and "hepatic"; (b) synonym discovery, such as "duct" and "tract"; (c) discovery of non-synonyms that are used by different sources in synonymous ways, like "main" and "common"; (d) normalization of expressions, like the match of "2nd" and "second"; (e) discovery of abbreviations, e.g., "peg" for "percutaneous endoscopic gastrostomy."

We used one pass to create the GSD. We can apply our algorithm repeatedly to get more GSD entries. For example, if we have two entries in the GSD (A B, C D) and (A, C) and reapply our algorithm, we would get (B, D). However, this would lead to a further

explosion of the use of computing resources, with no guarantee of many new valid synonyms.

While creating the GSD, we found cases of obvious errors, e.g, "Barrett's esophagus" and "Barrett's oesophagitis" are two terms assigned to the same concept. Based on this, our algorithm proposed a GSD entry (oesophagitis, esophagus), which is erroneous. In general, PS matching produces fewer wrong mappings (7.34%) with preexisting synonyms than with our GSD (44.62%), because there are fewer incorrect preexisting synonyms. To overcome this limitation, we will have to use methods to validate the correctness of matched terms, like Mougin et al.[12]

We note that the 49 new matches did come at a price. There were in fact 111 (42.9%=111/259) wrong mappings (or mismatches). The percentage of wrong mappings generated is indeed high; however, this is outweighed by the importance of the newly discovered matches. Also, millions of candsyns had to be generated. Many were not utilized, and this required significant computational resources. Additional efficient filtering techniques will be needed to control this combinatorial explosion.

Our methodology also finds some ambiguous concept assignments in the UMLS. For example, the MST term "gastric tumor" has one candsyn "stomach mass" mapped to the UMLS⁻. It is hard to think of a stomach mass that is *not* a tumor (unless it is an intracavitary mass, like a hairball—which is really a "mass in the stomach"). But, according to the MST, the concept "gastric tumor" is different from "stomach mass," which makes the term ambiguous.

Our approach is not intended to replace previously developed syntactic methods for term matching. Rather it is complementary to them. In future work, combinations of this new method with syntactic methods should be investigated. Our approach also does not solve the homonym problem. If one term expresses two concepts, then generating a piecewise synonym does not determine which of the two original meanings should be associated with the candsyn.

Conclusions

A new methodology for enhancing automated term matching between a new source terminology and existing UMLS concepts was presented. At the foundation of this methodology is the construction of previously unknown, unrecorded synonyms from those already appearing in the UMLS. The actual synonym construction is carried out algorithmically using a word-sequence analysis and word substitution. As an experiment, the MST, a current UMLS source, was removed and then partially re-integrated. The experiments showed an improvement in matching between source terms and UMLS concepts of 34%. As a side effect, our newly created synonyms also revealed ambiguous UMLS terms.

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. Available at http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html. **This and all other Web sites were accessed on Feb. 26, 2007.**
3. Logical Observations Identifiers, Names, Codes (LOINC). Available at http://www.nlm.nih.gov/research/umls/loinc_main.html.
4. NCI - National Cancer Institute. Available at http://nci.nih.gov.
5. UMLS Release Notes and Problems. Available at http://www.nlm.nih.gov/research/umls/release_notes.html.
6. NLM - National Library of Medicine. Available at http://www.nlm.nih.gov.
7. Vocabularies in the UMLS Metathesaurus. Available at http://www.nlm.nih.gov/research/umls/source_faq.html.
8. Cantor MN, Sarkar IN, Gelman R, Hartel F, Bodenreider O, Lussier YA. An evaluation of hybrid methods for matching biomedical terminologies: mapping the gene ontology to the UMLS. Studies in Health Technology and Informatics. 2003;62–7
9. 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.
10. UMLS Metathesaurus. Available at http://www.nlm.nih.gov/research/umls/meta2.html.
11. UMLS July Release 2006AC. Available at http://www.nlm.nih.gov/research/umls/archive/2006AC/umlsdoc.html.
12. 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.