Increasing UMLS Coverage and Reducing Ambiguity via Automated Creation of Synonymous Terms: First Steps toward Filling UMLS Synonymy Gaps

François-Michel Lang*	James G. Mork	Dina Demner-Fushman	Alan R. Aronson
flang@mail.nih.gov	mork@mail.nih.gov	ddemner@mail.nih.gov	alan@nlm.nih.gov
301.827.4995	301.827.4996	301.435.5320	301.827.4980

Lister Hill National Center for Biomedical Communications
National Library of Medicine
National Institutes of Health
8600 Rockville Pike
Bethesda, MD 20894 USA
Fax: 301.496.0673

*Corresponding author

Abstract

Background: Although extensive synonymy is one of the greatest strengths of the UMLS Metathesaurus, much research has nonetheless focused on identifying and measuring gaps in UMLS synonymy. This paper proposes a methodology for further extending the UMLS' already rich synonymy by semi-automatically creating new strings not in the UMLS, and including them as additional synonymous strings within existing UMLS concepts.

Results: In this paper we present our methodology for identifying missing UMLS synonymy and semi-automatically creating synonyms to fill these gaps. We created an enhanced Metathesaurus supplemented by these strings, and improved the performance on both biomedical literature and clinical text of two well known named-entity-recognition applications at the US National Library of Medicine, MetaMap and the Medical Text Indexer (MTI).

Conclusions: Our methods propose first steps toward extending the already rich synonymy of the UMLS by filling in some synonymy gaps. We further theorize that some of the newly created strings could also be used to extend the Medical Subject Headings (MeSH) entry terms, and thereby enhance MEDLINE indexing and PubMed queries by better reflecting how authors actually refer to biomedical concepts in the literature.

Keywords

UMLS, Metathesaurus, MeSH, MEDLINE, PubMed MetaMap, Medical Text Indexer, MTI, Synonymy Named-entity recognition, Natural language processing

Introduction

The Unified Medical Language System® (UMLS®[58]) Metathesaurus® [57] provided by the US National Library of Medicine® (NLM®) is a multi-lingual database comprising 12M strings drawn from over 150 biomedical terminologies, and organized into over 3M concepts, each with its own unique Concept Unique Identifier, or CUI. Grouping strings into concepts is one of the most beneficial features of the Metathesaurus, because it provides extensive synonymy that allows named-entity-recognition (NER) applications to map many varied strings to the same UMLS concept. For example, all the strings *Cardiac infarction*, *Myocardial Infarction*, *myocardial necrosis*, *coronary attack*, *heart attack* and sixty-five others (in English alone) share the CUI `C0027051`, which means that in the UMLS, the strings are synonymous.

The biomedical literature contains many articles dealing with synonymy, semantic similarity, and semantic relatedness [20, 29, 43, 44, 45], but the research in this area most often focuses on measuring and testing the semantic similarity and relatedness of existing biomedical strings, rather than creating synonymous strings *de novo*. Hole and Srinivasan [21] report work closer to ours in spirit, because one approach described therein infers term-level synonymy via word-level synonymy, but that paper focused on detecting synonymy of existing UMLS strings that were missed by UMLS editors. The work of Huang et. al. first published in [22] and greatly expanded in [23], builds new synonyms using piecewise synonymy or synonym substitution, which is very similar to our method, albeit on a much smaller scale and for a completely different purpose—source-terminology integration, rather than synonym-set expansion. Their work also uses no corpus validation, which ours does; moreover, [23] constructs its piecewise synonyms using WordNet®[36, 13], which is not a specifically biomedical resource. Indeed, [23] specifically recommends "[e]xperimenting with larger UMLS source terminologies", which is precisely what we have done by using the entire UMLS to generate our synonyms (over 2M strings in 1.5M concepts, rather than the 150K strings in 115K concepts of WordNet).

Our paper focuses on expansion of synonym sets, in particular for NER applications, a topic with a long and distinguished history. Hearst [19] conducted one of the earliest studies in which she developed a rule-based method (consisting of lexico-syntactic patterns) for automatic acquisition of hyponymy relationships from text and a partial implementation of the algorithm to augment a large, manually constructed thesaurus. The method does

not extend well to meronymy but should extend to other relationships. Jacquemin et al. [25] reported on an ambitious study of synonymy in the context of developing a system for automatic generation of French indexing terms using NLP tools for part-of-speech tagging, morphology and parsing. Three types of synonymy are defined, of which subsynonymy is one; but only the other two were actually studied in the paper. Hamon et al. [18] describe a preliminary, corpus-based study of synonymy detection for French technical text. The kind of synonymy studied includes a specific form of subsynonymy limited to two-word terms but also entails a broader notion of synonymy. Jacquemin [24] conducted a much broader study defining a model for describing not only term variation including synonymy, but additional morphological, syntactic and semantic variation as well. His experiments using multiple corpora show high precision in extracting the model variants. Finally, Morin and Jacquemin [37] performed a study, partly inspired by Hearst [19], describing a semi-automated system for discovering single-word relationships from a corpus and then extending them to multi-word relationships. The relationship used in the study is hypernymy.

One of the early papers involving the discovery of biomedical semantic relationships is Verspoor et al. [60], which described a method of extracting hierarchical relationships using the Gene Ontology (GO) for terms found in a corpus of MEDLINE®abstracts (MEDLINE is "the primary component of PubMed" [31], consisting of all and only those documents for which MeSH indexing is available). Using synonymy in GO was left for future work. A biomedical study of extracting synonym sets from MEDLINE/PubMed® [47] was done by McCrae and Collier [28], who devised a method of automatically generating regular expressions by heuristic search, constructing feature vectors and subsequently extracting synonym sets using the vectors. Their results (73.2% precision and 29.7% recall) outperform Wikipedia and MeSH®, NLM's vocabulary for indexing MEDLINE articles, but perform significantly less well than the UMLS. Tsuruoka et al. [56] present an interesting study of biomedical synonymy formulated as a normalization task. They automatically discover normalization rules using an iterative process designed to minimize ambiguity and variability of discovered terms. In a study somewhat related to the current work, Grabar et al. [16] define a method for automatic acquisition of elementary synonyms via analysis of complex, multiword synonyms in GO. And in a subsequent study Grabar et al. [17] define a language-independent method using syntactic analysis and compositionality to infer word-level synonymy from term-level synonymy. They apply this method to the French subset of the Metathesaurus. A more recent study of biomedical synonymy which is more relevant to our current work is the extremely ambitious, brilliantly illustrated, and broad-ranging article by Blair et al. [8], which quantifies the extent of missing Metathesaurus synonymy, and therefore meshes well with our work, which attempts to fill some of these gaps in Metathesaurus synonymy.

The work described in this paper differs from previous research in synonym-set expansion in that it relies on no morphological, lexical, syntactic, or semantic rules. Rather, our new

synonyms are derived from a well-known, richly documented, and universally accessible set of English biomedical terms—the UMLS Metathesaurus, and created through a knowledge-based method supplemented by corpus-based validation.

The initial set of expansion strings is then pruned to the most-likely useful subset using the largest available relevant target corpus (PubMed). Moreover, as we explain in the Methods Section, we test the effectiveness of our method using three biomedical corpora—two drawn from the biomedical literature, and the third from clinical text.

In the remainder of this paper, we describe a semi-automated method of creating new synonyms of existing UMLS strings in order to improve the performance and reduce the ambiguity of results generated by NLM's MetaMap [5] and Medical Text Indexer (MTI) [40, 39, 38], which are well-known NER applications. We now present in some detail the construction of these new synonymous strings, outline our experimental results, and conclude with a discussion of potential next STEPS.

Methods

In this section, we introduce our approach by presenting several examples of missing Metathesaurus synonymy and describe our methods for creating missing synonyms.

Working Examples

We begin with two examples of our technique, which extends Metathesaurus synonymy by creating new strings via an analysis of substring synonymy, or *subsynonymy*. Subsynonymy can best be illustrated using the following two canonical examples:

(1) The two strings *geriatric* and *elderly* are synonyms because they share CUI `C1999167`. Moreover *geriatric patients* is a string in `C0870602`. However, the string resulting from substituting *elderly* for *geriatric*, namely, *elderly patients*, does not occur in the UMLS, even though it occurs nearly 46,000 times in titles or abstracts of PubMed citations (as of this writing; the number will change over time)—nearly ten times as often as the existing UMLS string *geriatric patients*. We therefore propose *elderly patients* as a derived synonym of *geriatric patients* in CUI `C0870602`.

(2) Similarly, although *medications* and *drugs* are synonyms in `C0013227`, and *New medications* is a string in `C1718097`, *New drugs* is not in the UMLS, even though that string occurs over 16,000 times in PubMed—more than fifteen times as often as the existing UMLS string *New medications*. As with the previous example, we propose *New drugs* as a derived synonym of *New medications* in CUI `C1718097`.

We now describe the construction of new, synthetic subsynonymy strings not in the UMLS that are synonyms of existing Metathesaurus strings; Figure 1 illustrates the steps described below.

Insert Figure 1 here

MRCONSO Extract

We first extract from the MRCONSO.RRF UMLS file [34] a unique list of pairs of
CUI | String
in which the strings are

- ASCII
- English (i.e., `LAT == "ENG"`)
- of length between 5 and 50 characters,
- composed of only alphanumerics and whitespace, and
- normalized to lowercase.

This MRCONSO extract contains 2.36M CUI/String pairs, and 2.30M distinct strings, which represents over 30% of all case-normalized English strings in the Metathesaurus. We impose these limitations for several reasons:

- Our downstream applications (MetaMap and MTI) process ASCII English text only, and are not case sensitive,
- strings shorter than 5 characters tend to exhibit excessive ambiguity, and
- strings longer than 50 characters or containing punctuation are less likely to be identified by our applications, or, for that matter, to appear in the biomedical literature.

Synonym Database

Next, we create from the MRCONSO extract a Synonym Database of triples
CUI: $\text{Synonym}_1 = \text{Synonym}_2$
for every CUI and every pair of distinct strings in that CUI. E.g., a few of the synonym triples for `C1999167` are

```
C1999167: elderly    = geriatric
C1999167: elderly    = old age
C1999167: elderly    = senescence
C1999167: elderly    = senium
C1999167: geriatric  = old age
C1999167: geriatric  = senescence
C1999167: geriatric  = senium
C1999167: old age    = senescence
C1999167: old age    = senium
C1999167: senescence = senium
```

We generated 2.38M such triples. The number of triples is lower than one might expect because

- over 1M of the CUIs in the extract (> 43 %) contain only one string, and therefore do not appear in the Synonym Database at all, and
- another 287,000 CUIs ($> 12\%$) contain only two strings, and therefore appear only once.

The Synonym Database will be available at our "Datasets & Test Collections" website [55] under "Subsynonymy Datasets" upon publication.

Superstring Database

We then generate from the MRCONSO extract a Superstring Database of triples
Substring | Superstring | SuperstringCUI
for all substring/superstring pairs of Metathesaurus strings, subject to the constraint that the substring be bounded by the beginning/end of the superstring or a non-alphanumeric character. For example, some Metathesaurus superstrings of *geriatric* are *geriatric patients*, *assessment geriatric*, and *non-urgent geriatric admission*, but not *geriatrician*, *psychogeriatric*, or *psychogeriatrician*. There were 5.8M such superstring triples. The Superstring Database will also be available at our dataset website [55] upon publication.

Retrieve Superstrings

Next, for each entry in the Synonym Database
CUI: $\text{Synonym}_1 = \text{Synonym}_2$
e.g.,

`C1999167: elderly = geriatric`
and its reverse, i.e.,
`C1999167: geriatric = elderly`
we retrieve from the Superstring Database all superstrings of Synonym$_1$ along with those superstrings' CUIs:

```
geriatric patients|C0870602
assessment geriatric|C0017463
non-urgent geriatric admission|C0420534
```

Create New Synthetic Strings

Finally, in each superstring, we replace Synonym$_1$ with Synonym$_2$, as shown in Table 1, thereby creating a collection of new, synthetic strings, which we will call subsynonymy

Table 1: **Creating Subsynonymy Strings Using `geriatric/elderly` Synonymy**

CUI	**Original Metathesaurus String**	**New Subsynonymy String**
`C0870602`	`geriatric patients`	`elderly patients`
`C0017463`	`assessment geriatric`	`assessment elderly`
`C0420534`	`non-urgent geriatric admission`	`non-urgent elderly admission`

strings, and save all such subsynonymy strings that are not already in the Metathesaurus, along with their CUIs.

Need for Filtering

Although the processing described above creates strings that are not in the Metathesaurus, and are therefore potential candidates for filling a synonymy gap in a specific CUI, we could not simply keep all the subsynonymy strings generated because of their enormous volume: We created over 40M distinct triples such as those shown in Table 1—over five times the number of distinct case-normalized English strings in the entire Metathesaurus, and over seventeen times the number of such strings in the MRCONSO extract described in the MRCONSO Extract Section above; extensive filtering was obviously necessary. We next describe briefly several stages of filtering designed to retain only those strings most likely to be useful for biomedical NER; each stage is then explained in more detail.

- Exclude strings not appearing in PubMed: For example, the prospective subsynonymy string *lower abdominalgia*, generated from *lower abdominal pain* in `C0232495` because

of the synonymy of *abdominal pain* and *abdominalgia*, does not appear in PubMed, so it is excluded.

- Exclude strings exhibiting spurious word duplication: The prospective subsynonymy string *swelling swelling*, generated from *edema swelling* in `C0474434` because of the synonymy of *edema* and *swelling*, contains spurious duplication of *swelling*, so it is excluded.

- Exclude strings exhibiting redundant linguistic variation: The prospective subsynonymy string *tumour marker*, generated from *tumor marker* in `C0041365` because of the synonymy of *tumor* and *tumour*, is simply a linguistic variant of *tumor marker*, so it is excluded. MetaMap's extensive knowledge of linguistic variation, mentioned below in the Linguistic Variation Section makes such variants unnecessary.

- Exclude false positive synonyms discovered via manual inspection: The prospective subsynonymy string *caucasian cells*, generated from *white cells* in `C0023508` because of the synonymy of *white* and *caucasian*, is a manually detected false positive synonym, so it is excluded (the string *caucasian cells* occurs in PubMed (in PMIDs 158862 and 25358733), but obviously does not denote leukocytes.)

All excluded strings, grouped by exclusion criterion, will be available on our dataset website [55] upon publication. We now explain each filtering strategy in more detail, with additional examples.

PubMed Filtering

A simple measure of the potential usefulness of a string for performing NER is its frequency of occurrence in a corpus. We considered measuring the subsynonymy strings' frequency by using either automated queries to an internet search engine (e.g., Bing [7]) or Google's Ngram data [15, 35], but decided against those approaches because we are interested in determining the subsynonymy strings' frequency specifically in biomedical text.

We decided instead to use the NCBI [41] E-Utilities [51] to determine the frequency of each subsynonymy string in the 25M citations in PubMed, because that method would provide text focused on biomedicine. In addition, we were able to generate PubMed citation counts using the E-Utilities for all 40M subsynonymy strings in less than half an hour by running 30 parallel processes, and determined that 235,000 of the subsynonymy strings actually appear in PubMed, as do 364,000 of the corresponding original UMLS strings (i.e., the strings in the second column of Table 1), and 687,000 strings in the 2.36M-string MRCONSO extract described in the MRCONSO Extract Section. The full set of subsynonymy strings and

their counts will be available on our dataset website [55] upon publication. For comparison purposes, Table 2 presents the document (i.e., PubMed citation) frequency of (a) the subsynonymy strings, (b) the original UMLS strings, and (c) the strings in the MRCONSO extract.

Table 2: **PubMed Document Frequency**

Document Frequency	**Subsynonymy Strings**	**Original UMLS Strings**	**Strings in MRCONSO Extract**
100+	30463 (12.94%)	80675 (22.14%)	154310 (22.47%)
500+	8191 (3.48%)	32900 (9.03%)	76811 (11.18%)
1000+	4168 (1.77%)	20402 (5.60%)	55039 (8.01%)
5000+	636 (0.27%)	5099 (1.40%)	23679 (3.45%)
10000+	254 (0.11%)	2377 (0.65%)	16232 (2.36%)

We were pleased to realize that over 4,100 of the 235,000 subsynonymy strings appearing in PubMed were found in at least 1,000 PubMed Citations, and 254 in at least 10,000.

Although the subsynonymy strings exhibit somewhat lower PubMed frequency than do the original UMLS strings and the strings in the MRCONSO extract, their frequency is nonetheless encouraging. Moreover, we know that their distribution does not consist simply of a long tail of low-frequency items, because, as we demonstrate in the Results and Discussion Section, their frequency is sufficient to have a measurable effect on NER.

We deliberately applied this automatic PubMed filtering step first, because it eliminated $> 99\%$ of the initial set of strings, thereby making the later, non-automatic filtering steps described next far more tractable.

Semi-Automatic Filtering

Spot checking of the subsynonymy strings quickly showed that several classes of strings needed to be excluded, as we now explain in greater detail.

Spurious Word Duplication

Manual review of the subsynonymy strings revealed a number of strings exhibiting spurious word duplication: For example, given the string *surgery procedure* in `C0944781` and the synonym pair *surgery/surgical procedure* in `C0543467`, substituting the latter synonym for the former yields the infelicitous string *surgical procedure procedure*. A vast majority of these strings containing repeated words appear in PubMed with punctuation separating the two identical words, e.g., *a standardized 60-minute surgical procedure (procedure 2)* in

PMID 16135206. To perform this filtering, we simply identified all subsynonymy strings containing adjacent identical words, and manually reviewed the 230 strings thus found, discarding 87 in all. The manual review was necessary because we would not want to discard strings containing valid examples of repeated words such as *drug drug interaction*, *infantile beri beri*, or *protein protein domain*.

Redundant Linguistic Variation

Next, we also excluded many subsynonymy strings that differ from existing Metathesaurus strings only via simple linguistic variation, because MetaMap already includes extensive logic based on the NLM's Lexical Variant Generation [42] that enables identification of UMLS concepts via a wide variety of linguistic variants. For example, we discarded the three strings *signaling pathways*, *drug sensitisation*, and *right ventricular* because (1) the Metathesaurus already contains *signaling pathway*, *drug sensitization*, and *right ventricle*, and (2) MetaMap would not need these additional synonyms to identify the original concepts. Excluding new strings exhibiting close linguistic similarity to existing strings removed nearly half our candidate subsynonymy strings, leaving us with about 125,000. Retaining all strings regardless of linguistic variation, however, would probably be advisable for downstream projects that depend on exact string matches or do not benefit from the use of the rich lexical variant generation that MetaMap enjoys.

False Positive Synonyms

Spot-checking the data suggested the need to check for false positive synonyms, so we then undertook a manual review of two samples of the remaining 200,000 strings: (1) The 6,500 strings most frequently occurring in PubMed (about 3% of the total), and (2) A stratified sample of another 6,500 strings (every thirtieth string).

This manual review revealed several classes of false positives, including some amusing ones due to vernacular synonyms of *Methamphetamine* such as *crystal*, *glass* and *speed* in MedlinePlus® [32]:

- *urine crystal* → *urine speed*
- *eye glass* → *eye speed*
- *reading glass* → *reading speed*
- *crystal healing* → *speed healing*

Examination of the original synonym pairs (e.g., *crystal*/*speed*, *glass*/*speed*) of these false positives suggested that most of them tended to be (1) common English words and/or (2)

semantically ambiguous. We accordingly identified all subsynonymy pairs based on original synonyms that are either (1) among the 10,000 most frequent words (excluding stopwords) in MEDLINE [27], or (2) in multiple Semantic Groups [59] (more formally, in CUIs whose associated Semantic Types [52, 10] are in multiple Semantic Groups).

Using the first example above, both *crystal* and *speed* are among the 10,000 most frequent MEDLINE words. Moreover, *crystal* is a string in `C0025611` and `C1704641`, and those CUIs are in multiple Semantic Groups: CHEM (Chemicals & Drugs) and OBJC (Objects). Similarly, *speed* is a string in `C0025611` and `C0678536`, and those CUIs are in multiple Semantic Groups: CHEM and CONC (Concepts & Ideas). *Crystal* and *speed* therefore meet both criteria of MEDLINE frequency and Semantic-Group ambiguity.

Manual review of the 78,000 synonym pairs meeting both criteria revealed 2,500 legitimate false positives, which we removed from the set of subsynonymy strings. Finally, we removed a large number of spurious duplicates; for example, we created the subsynonymy string *benign gastrointestinal tumors* in `C0497538` twenty different ways, but obviously just one was sufficient. After removing duplicates, we were left with 114,000 distinct case-normalized strings, and 142,000 CUI/String pairs in 63,000 CUIs. Those numbers show that several thousand subsynonymy strings appeared in multiple CUIs, leading to increased string ambiguity, but we will show in the Reduced Ambiguity Section that any increase in string ambiguity is outweighed by the increased synonymy, and, perhaps counter to intuition, by significant *reduction* in the ambiguity of our results.

From this set of 114,000 subsynonymy strings, we constructed a synthetic vocabulary called NLMSubSyn, which we merged into a local copy of the UMLS MRCONSO.RRF file, and then created the MetaMap datafiles [33] from this augmented version of MRCONSO.RRF.

Results and Discussion

We next explain our strategy for testing the effect of the NLMSubSyn vocabulary on the performance of MetaMap and MTI on both the biomedical literature and clinical text, present our results, and conclude this section with a discussion of potential benefits of this work to MeSH.

One potential drawback resulting from the addition of 114,000 new strings to our data is a possible increase in runtime for MTI and MetaMap; however, because the 114,000 terms represent less than a 2% increase in the number of English strings used by MetaMap, we are happy to report no observable change in runtime due to the additional strings.

Experiments on Biomedical Literature

Our first experiment used MTI, NLM's application which uses MetaMap and other techniques to map English biomedical text to MeSH, and its test collection [40, 39, 38], which consists of 144,000 randomly selected and recently indexed MEDLINE articles (updated each year) used to train MTI and evaluate changes made to the system throughout the year. We ran MTI twice on the collection: once using the MetaMap datafiles including the NLMSubSyn vocabulary, and a second time using the datafiles *excluding* the NLMSubSyn vocabulary, to measure the effect of the subsynonymy strings on MTI processing. This initial test resulted in an observable and useful, albeit modest, increase in precision (+0.01%), recall (+0.01%), and F_1 (+0.02%) due to the subsynonymy strings, as shown in Table 3. We

Table 3: **MTI Test Collection Results**

SubSyn?	**Precision**	**Recall**	**F_1**
Without	63.53%	62.70%	63.11%
With	63.54%	62.71%	63.13%

were at first disappointed with the results, but further investigations revealed that a number of factors severely constrained any possible improvement in performance: (1) Adding our NLMSubSyn vocabulary to the MRCONSO.RRF file resulted in less than a 2% increase in the total number of English strings; (2) Of the 114,000 distinct subsynonymy strings, fewer than 20,000 appeared in the MTI test collection at all; (3) Of those, fewer than 14,000 are mapped to MeSH by NLM's Restrict-to-MeSH algorithm [9, 50]; (4) Only 11,000 of the remaining strings are not subject to MTI's heavy filtering, which excludes very general MeSH headings (e.g., *Patients*); and finally, (5) MTI's performance results from over a decade of fine-tuning, extensive filtering, and algorithmic enhancements, thereby making it extremely difficult to deliver significant improvements in precision/recall. Moreover, given the large volume of MEDLINE citations processed by MTI each year, *any* improvement in performance, however slight, is nonetheless extremely valuable. In summary, given that less than 10% of the subsynonymy strings are potential candidates for MeSH indexing, and that these strings therefore represent less than a 0.2% increase in the number of strings available for MeSH indexing, the modest improvements are perhaps not surprising.

Modest as these results may be, however, our technique offers two additional benefits: (1) The subsynonymy strings fill some glaring lacunae in Metathesaurus synonymy, such as *elderly patients*, as described in the Working Examples Section above. (2) they are also very useful for assistance with MeSH indexing, which is MTI's purpose: MTI automatically provides NLM's MeSH indexers with suggestions for MeSH headings for approximately 750,000 MEDLINE citations each year, and the improved MTI performance due to subsynonymy

should result in over 1,000 additional correct MeSH-heading recommendations each year.

Our next testing step was designed to verify that the subsynonymy strings could in fact lead to a more significant performance improvement. To test this hypothesis, we again used the NCBI E-Utilities to create a corpus of all PubMed citations (regardless of publication date) containing at least one subsynonymy string. Our results are very encouraging: Over 7M PubMed citations (nearly 30% of all PubMed) contained at least one subsynonymy string, and over 104,000 citations contained at least five. A complete listing of subsynonymy strings with the PMIDs in which they were found will be available at our dataset website [55] upon publication. From this collection of citations guaranteed to contain at least one subsynonymy string, we automatically created a focused sample of the most subsynonymy-rich citations by selecting all those documents (1) for which MeSH indexing was available (i.e., in MEDLINE), and (2) containing at least five subsynonymy strings; applying these criteria yielded a collection of 95,000 MEDLINE documents, which we will henceforth refer to as the **focused collection**. We emphasize that the **focused collection** was constructed completely independently from the MTI test collection; indeed the intersection of the **focused collection** and the 144,000-citation MTI test collection consists of only 940 documents.

As before, we then ran MTI twice on the **focused collection**, and observed a much more dramatic improvement in precision (+0.13%), recall (+0.16%), and F_1 (+0.15%) due to the subsynonymy strings, as shown in Table 4, representing over 2,100 additional MeSH headings correctly identified. In order to estimate the standard error associated with the

Table 4: **Focused Collection Results**

SubSyn?	**Precision**	**Recall**	**F_1**
Without	58.71%	63.44%	60.98%
With	58.84%	63.60%	61.13%

improvements observed in the analysis of the **focused collection**, we then computed precision, recall, and F_1 on 1,000 bootstrap samples with replacement, each consisting of 1,000 citations randomly selected from the **focused collection**. We present the mean and standard deviation (σ) of precision, recall, and F_1, computed as usual both with and without the subsynonymy strings, in Tables 5 (using micro averaging) and 6 (using macro averaging).

Macro averaging, which gives equal weight to each MeSH heading identified, shows greater improvement, which is probably understandable because in micro-averaging, which gives equal weight to each per-document decision, the contribution of the relatively few subsynonymy strings added is less visible among the average of approximately 10 MeSH main headings assigned to each MEDLINE document (see the third column ("Min/Avg/Max Occurrences") of the "MeshHeading" and "DescriptorName" rows of [30], which show that an average of 10.16 MeSH headings are assigned to documents in the 2015 MEDLINE Baseline

Table 5: **Micro-Averaged P/R/F_1 Statistics for Focused Collection**

SubSyn?	Precision		Recall		F_1
	Mean	σ	Mean	σ	
Without	58.72%	0.46	63.45%	0.52	60.99%
With	58.84%	0.46	63.61%	0.51	61.13%

Table 6: **MACRO-Averaged P/R/F_1 Statistics for Focused Collection**

SubSyn?	Precision		Recall		F_1
	Mean	σ	Mean	σ	
Without	58.93%	0.47	65.32%	0.52	61.96%
With	59.07%	0.47	65.49%	0.52	62.11%

[26]).

Although the subsynonymy strings contributed to performance improvement for both the MTI test collection and the **focused collection**, as shown in Tables 3 and 4, respectively, it is perhaps surprising that MTI performed better on the test collection (Table 3) than the **focused collection** (Table 4)—independently of any improvements due to subsynonymy. The lower performance of the **focused collection**, however, has a sound explanation that is due entirely to precision: Although the recall observed for the **focused collection** is somewhat greater than that for the MTI test collection (because of the intentionally greater frequency of subsynonymy strings), the **focused collection's** precision is significantly lower than the MTI test collection's—for two reasons: (1) As noted above, the MTI test collection consists of MEDLINE citations that are no more than a year old, and therefore recently indexed; furthermore, the collection is used throughout the year to train MTI and improve its results (mainly precision), and the results presented in this paper were generated late in the year, after nearly a full year of continuous training and improvement, when MTI's performance was at its peak. (2) The **focused collection**, by contrast, is drawn from all of MEDLINE, regardless of publication date; consequently, the average indexing year of the citations in the **focused collection** is 2001 (the oldest dates back to 1965!), so its MeSH indexing is far from current. Moreover, as NLM indexing policy changes over time, MTI processing is modified accordingly, but previously indexed citations are never re-indexed, which necessarily adversely affects MTI's performance on older documents.

Experiments on Clinical Text

To test the effect of the subsynonymy strings on clinical text, we used the 298 clinical notes in the ShARe corpus and their annotations from Task 7 of SemEval-2014 [54, 46, 53]. The annotations provide 7,776 disorder mentions along with the disorders' CUIs and the text spans in the note pinpointing the location in the text of the disorder mention, which provided the gold standard for the clinical experiments. A simplified form of the annotations for document ID `00098-016139` is shown below:

```
00098-016139|C0149651|1218|1226
00098-016139|C0010520|1228|1236
00098-016139|C0013604|1241|1246
00098-016139|C0917996|1327|1344
00098-016139|C1290398|1389|1392|1412|1420
```

The first annotation line above shows that in document `00098-016139`, human annotation identified the UMLS Metathesaurus concept whose CUI is C0149651 in text spanning character positions `1218` and `1226`. Our initial processing of the clinical notes was similar to that of the MEDLINE citations described above, but used MetaMap instead of MTI, because our gold standard for this collection consists of UMLS concepts, and not MeSH headings. We ran MetaMap twice on each note (as before, once with and once without the subsynonymy strings), but this time conforming to the SemEval-2014 task 7 guidelines [54], which restricted annotations to UMLS concepts in the Disorder Semantic Group [59]. We used MetaMap to generate Fielded MetaMap Indexing (MMI) Output [14], and retained only those concepts whose lexical category was `noun`.

We then calculated MetaMap's performance against the ShARe corpus gold-standard annotations, counting as true positives those MetaMap results that matched the gold-standard annotations as follows: (1) exact CUI match, and (2) at least a partial text-span match. The results of this initial experiment are shown in Table 7. To measure the influence on our

Table 7: **Clinical Results (strict)**

SubSyn?	**Precision**	**Recall**	**F_1**
Without	47.90%	64.55%	54.99%
With	47.43%	63.96%	54.47%

results of common, frequently occurring terms, we also calculated the same performance measures restricting the results to unique occurrences of terms in a given document; these results are shown in Table 8, and, as expected, show somewhat lower performance than

the results in Table 7, but the difference in performance is not so great as to suggest that commonly occurring terms exert an undue influence on the results reported above. We

Table 8: **Clinical Results (strict; unique occurrences)**

SubSyn?	Precision	Recall	F_1
Without	46.70%	61.93%	53.25%
With	45.94%	61.39%	52.55%

were obviously disappointed that the subsynonymy strings did not lead to improved performance, but a detailed review of MetaMap's false positives revealed that many of them identified a concept similar to the gold standard's—albeit in a different CUI. For example, from the text `upper extremity deep vein thrombosis` in document `05163-019624`, MetaMap did not identify the gold standard's *deep vein thrombosis*, finding instead the more specific *upper extremity deep vein thrombosis*. More examples of this phenomenon are presented in Table 9. In order to analyze the reduced performance due specifically to these partial matches, which our scoring methodology counted as false positives and negatives, we reviewed all 261 false positives occurring only in the analysis with the subsynonymy strings by examining their appearance in the original clinical documents, and determined that 187 of them are actually true positives that were either (1) not annotated at all, or (2) annotated, but identified as one of the 3,391 "CUI-less" annotations, which we ignored in our original calculations, because they could not provide a CUI match. In these cases, the subsynonymy strings provided a synonym that enabled MetaMap to map text to a specific disorder and CUI, which the human annotators were not able to do, presumably because the original text was not sufficiently close to any existing Metathesaurus string to be recognized as a Metathesaurus concept. Similarly, we reviewed the text spans of 64 false negatives occurring only in the analysis with the subsynonymy strings and found 62 cases in which MetaMap had in fact identified a concept with an overlapping span, albeit not the one in the gold standard. Relaxing our criteria to count such partial matches as true positives, the results using the subsynonymy strings are significantly improved, as shown in Table 10.

We also noted that when the concept identified by MetaMap provided more specific information than the gold standard's, such as the *deep vein thrombosis* example presented above and the other cases presented in Table 9, MetaMap did find the gold standard's concept as well, but discarded it in favor of the longer concept with greater phrase coverage, and therefore a higher score, as described in [4].

Our experiments with the clinical text showed appreciably greater improvements than those in the biomedical literature for two reasons: (1) Processing the biomedical literature using MTI involves extensive filtering of results, which the clinical experiments do not, because

Table 9: **ShARe Corpus: Gold Standard vs. MetaMap**

Document	Gold Standard Concept	MetaMap Concept
07429-001857	pain	pain in extremities
17467-010718	fracture	fracture of hip
09339-028983	cataract	bilateral cataract
07352-013977	dysplasia	high grade dysplasia
25003-338492	pneumonia	right lung pneumonia
24786-014472	dizziness	dizziness of unknown cause
19138-025729	atrial fibrillation	Intermittent atrial fibrillation
02115-010823	strokes	multiple strokes
11439-014138	anemia	severe anemia
00381-006281	wound	abdominal wound
17467-010718	fracture	fracture of hip
21413-012450	Hemangioma	Hemangioma of the liver
02136-017465	stenosis	left anterior descending artery stenosis
20701-013632	abdominal pain	left upper quadrant abdominal pain
09339-028983	cataract	bilateral cataract
19138-025729	tenderness	calf tenderness
00534-017453	bicuspid aortic valve	congenital bicuspid aortic valve
05163-019624	deep vein thrombosis	upper extremity deep vein thrombosis
17652-018982	upper extremity weakness	right upper extremity weakness
01314-028800	displaced fracture	displaced fracture of proximal phalanx left thumb

Table 10: **Clinical Results With Relaxed Matching Criteria**

SubSyn?	Precision	Recall	F_1
Without	47.90%	64.55%	54.99%
With	49.51%	65.58%	56.42%

they were carried out simply using MetaMap. (2) The clinical experiments were restricted to results in the Disorder Semantic Group, which allowed much more targeted processing.

Reduced Ambiguity

Ambiguity is arguably one of the leading *bêtes noires* of NER systems. Although the improved accuracy resulting from the subsynonymy strings is certainly helpful, we expect

that a greater benefit will result from the reduced ambiguity of results generated using the subsynonymy strings. We next explain why the addition of the subsynonymy strings leads to reduced ambiguity and provide a number of examples, and then present statistics showing the effect of the reduced ambiguity on the analysis of the **focused collection** described earlier.

To demonstrate the reduction in ambiguity due to the subsynonymy strings, we performed two experiments: First, we ran MetaMap twice on each of the subsynonymy strings themselves using the same strategy as described in the previous section: once using our now-enhanced data including the NLMSubSyn vocabulary, and a second time excluding the NLMSubSyn vocabulary. The run without the new NLMSubSyn vocabulary showed over 300% more ambiguity on average than the run including the NLMSubSyn vocabulary. The explanation is clear: The subsynonymy strings exhibit far less ambiguity than their component substrings. For example, using the subsynonymy-enriched data, MetaMap correctly maps the subsynonymy string *life threatening ventricular tachycardia* to a single UMLS concept (`C1556245`). Without the subsynonymy data, however, the entire string is not mapped to a single concept; the two substrings that *are* mapped, however, *life threatening* and *ventricular tachycardia*, exhibit three- and fourfold ambiguity, respectively, and are therefore each mapped to multiple concepts [3], as shown in Table 11. MetaMap's mapping

Table 11: **Reduced Ambiguity**

SubSyn?	**CUI**	**String**
With	C1556245	life threatening ventricular tachycardia
Without	C2826244	Life Threatening
	C1517874	LIFE THREATENING
	C1546953	Life threatening
	C3537125	LIFE THREATENING
	C0042514	TACHYCARDIA, VENTRICULAR
	C0344428	VENTRICULAR TACHYCARDIA
	C1963247	Ventricular tachycardia

algorithm [2] then computes the Cartesian product of the sets of concepts identified in the two substrings, and produces the highly undesirable result of twelve final mappings (each final mapping represents MetaMap's best interpretation of the text analyzed, and consists of a subset of the set of concepts identified). Subsynonymy in this case clearly produces a vastly improved, more specific, and more compact final result: one concept covering the entire phrase, rather than twelve distinct combinations of smaller concepts. Such ambiguity is hardly atypical: When we ran MetaMap on the subsynonymy strings themselves as input, but excluding the NLMSubSyn vocabulary from our data, over 4% of the subsynonymy

strings exhibited at least tenfold ambiguity; the worst offenders are the two strings *Radiation treatment groups* and *Surfactant combinations*, which each generate over 200 final mappings when run without the NLMSubSyn strings, instead of just one each, because of the extreme ambiguity of their component substrings. The cause of the reduced ambiguity may be evident, but the extent of the reduction is nonetheless remarkable.

Our second, and more significant, ambiguity experiment involved analyzing the results generated from the **focused collection**. As described earlier, we ran MetaMap twice on the entire collection (once with the subsynonymy strings and again without them), and then computed (1) the number of final mappings generated, and (2) the number of concepts appearing in them. The results presented in Table 12 show a 7.65% reduction in the number of final mappings, and a 13.09% reduction in the number of concepts participating in the final mappings.

Table 12: **Focused Collection Ambiguity Results**

SubSyn?	**# of Final Mappings**	**# of Concepts in Mappings**
Without	25.74M	79.68M
With	23.77M (-7.65%)	69.25M (-13.09%)

Next Steps

Some future plans for extending UMLS subsynonymy research include the following:

1. Improving false-positive detection, perhaps via crowdsourcing, e.g., via Amazon Mechanical Turks [1].

2. Expanding subsynonymy generation to include British/American variants of prospective subsynonymy strings.

3. Expanding the MRCONSO extract beyond alphanumeric/whitespace strings.

4. Exploring an idea proposed in [21] and similar to a strategy presented in [6] for discovering missed synonymy by eliminating shared words in synonym pairs. E.g., *family members died* and *relatives died* are synonyms in `C0425043`; eliminating the shared word *died* suggests the synonymy of *family members* and *relatives*; however, although those two strings both appear in the Metathesaurus (in `C0086282` and `C0080103`, respectively), no CUI contains them both, so they are not currently synonyms in the UMLS. In fact, this is an excellent example of undocumented UMLS synonymy as

discussed in [8]. Adding each string to the other's concept before generating the synonym list described at the beginning of the Methods Section might further reduce ambiguity and increase recall.

5. Bootstrapping our process by recursively applying all the logic described above: Instead of starting from the out-of-the-box MRCONSO.RRF file, we would instead use the MRCONSO file augmented with the subsynonymy strings.

6. PubMed searches are carried out using an indexed phrase list created and maintained by NCBI, as explained in the "Default Boolean Combinations and Phrase Searching" section of [11]. Constructing our own phrase list, or, more realistically, supplementing NCBI's list with any of our candidate subsynonymy strings not already included, could identify additional subsynonymy strings that are not in PubMed, because PubMed returns no results for strings that are not in the indexed phrase list—even if they do appear in PubMed.

7. Potential Benefit to MeSH: After analyzing these subsynonymy strings and their distribution in Metathesaurus source vocabularies, we also realized this work could lead to potential long-term benefits for MeSH, and, in particular, MeSH indexing. Our processing added nearly 800 subsynonymy strings, each appearing at least 500 times in PubMed, to CUIs containing a MeSH main heading; because of the rich overlap of our subsynonymy strings with MeSH, we are pursuing two aspects of this work that could benefit MeSH:

   (a) Some of these new subsynonymy strings might serve as potential indicator phrases for our indexing-assistance tool MTI. For example, consider the subsynonymy string *preterm babies*, which was created from the original UMLS string *preterm infants* in `C0021294` because of the synonymy of *infants* and *babies* in `C0021270`. The subsynonymy string occurs in the title or abstract of a citation 1,346 times in PubMed[1] and 1,207 times in MEDLINE[2]; however, 995 of those 1,346 PubMed citations lack the original UMLS string *preterm infants*[3] as do 899 of those in MEDLINE.[4] In analyzing those citations that contain *preterm babies* but not *preterm infants*, the subsynonymy string *preterm babies* could prove to be a useful indicator for MTI: If MTI finds the phrase *preterm babies* in the text, it is very likely a good indication that MTI should recommend the MeSH main heading *Infant, Premature*. We can validate this assumption by querying PubMed to determine how often indexers recommended *Infant, Premature* when (a) *preterm babies* occurs in the title or abstract, but (b) *preterm infants* does not: We find that indexers chose the MeSH heading *Infant, Premature*[5] in 509 of those 899 MEDLINE citations (56.62%), thereby showing that *preterm babies* would indeed be a good indicator phrase for MTI to use in recommending *Infant, Premature*.

(b) MeSH defines MeSH entry terms as "synonyms, alternate forms, and other closely related terms in a given MeSH record that are generally used interchangeably with the preferred term for the purposes of indexing and retrieval, thus increasing the access points to MeSH-indexed data" [12]. A large number of our new subsynonymy strings represent how authors denote real-world biomedical entities in PubMed, so including some of them as MeSH entry terms might prove useful. For example, the subsynonymy string *robotic surgery*, found 2,449 times in PubMed and 1,785 times in MEDLINE, could be a good candidate for inclusion as a MeSH entry term for the MeSH main heading *Robotic Surgical Procedures*. Expanding the list of MeSH entry terms for a given MeSH main heading could help in two ways: (1) Provide more indexing consistency by the human indexers, because there would now be additional defined MeSH terms exactly matching the actual phrases found in the text, and (2) Improve PubMed Automatic Term Mapping [48, 49] by using these new subsynonymy strings to map additional user query terms to specific MeSH terms, which should improve search results by providing more relevant articles to users executing PubMed queries.

Conclusions

This paper has described a UMLS synonymy-based technique of creating strings that appear in PubMed, but are not in the UMLS, and are therefore potentially useful for increasing coverage and recall, and reducing ambiguity of biomedical NER applications such as MetaMap and MTI. Even with a relatively small addition of only 142,000 CUI/string pairs (less than a 2% augmentation in the number of English CUI/strings pairs in the UMLS MRCONSO.RRF file), we were able to report improved performance and reduced ambiguity for both literature and clinical data.

Availability of Supporting Data

The data sets supporting the results of this article will be available at our "Datasets & Test Collections" website [55] under "Subsynonymy Datasets" upon publication.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This work was supported by the Intramural Research Program of the National Institutes of Health and the National Library of Medicine. All authors read, revised, and approved the final manuscript.

Notes

[1]PubMed query: `"preterm babies" [tiab]`

[2]`"preterm babies" [tiab] AND MEDLINE [sb]`

[3]`"preterm babies" [tiab] NOT "preterm infants" [tiab]`

[4]`"preterm babies" [tiab] NOT "preterm infants" [tiab] AND MEDLINE [sb]`

[5]`"preterm babies" [tiab] NOT "preterm infants" [tiab] AND MEDLINE [sb] AND "Infant, Premature" [mh:noexp]`

References

[1] Amazon. Amazon Mechanical Turk - Welcome. `https://www.mturk.com/mturk/welcome`, 2016.

[2] A. Aronson. The MetaMap Mapping Algorithm. `https://ii.nlm.nih.gov/Publications/Papers/mm.mapping.pdf`, 2000.

[3] A. Aronson. MetaMap Candidate Retrieval. `https://ii.nlm.nih.gov/Publications/Papers/mm.candidat` 2001.

[4] A. Aronson. MetaMap Evaluation. `https://ii.nlm.nih.gov/Publications/Papers/mm.evaluation.pdf`, 2001.

[5] A. Aronson and F.-M. Lang. An Overview of MetaMap: Historical Perspective and Recent Advances. *JAMIA*, 17(4):229–236, 2010.

[6] R. Baud, C. Lovis, A.-M. Rassinoux, P.-A. Michel, and J.-R. Scherrer. Automatic Extraction of Linguistic Knowledge from an International Classification. In *Proc MEDINFO*, pages 581–585, 1998.

[7] Bing. Bing. `https://www.bing.com`, 2016.

[8] D. Blair, K. Wang, S. Nestorov, J. Evans, and A. Rzhetsky. Quantifying the Impact and Extent of Undocumented Biomedical Synonymy. *PLoS Comput Biol*, 10(9), 2014.

[9] O. Bodenreider, S. Nelson, W. Hole, and H. Chang. Beyond Synonymy: Exploiting the UMLS Semantics in Mapping Vocabularies. In *Proc AMIA*, pages 815–819, 1998.

[10] Current Semantic Types. Current Semantic Types. https://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html, 2016.

[11] Entrez Help. Entrez Help - Entrez Help - NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK3837, 2016.

[12] Entry Terms. Entry Terms and Other Cross-References. https://www.nlm.nih.gov/mesh/intro_entry.html, 2016.

[13] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[14] Fielded MMI. Fielded MetaMap Indexing (MMI) Output Explained (Updated for MetaMap 2016 Output). https://metamap.nlm.nih.gov/Docs/MMI_Output_2016.pdf, 2016.

[15] Google Ngram Viewer. Google Ngram Viewer. https://books.google.com/ngrams, 2016.

[16] N. Grabar, M. Jaulent, and T. Hamon. Combination of endogenous clues for profiling inferred semantic relations: experiments with gene ontology. In *Proc AMIA*, 2008.

[17] N. Grabar, P. Varoutas, P. Rizand, A. Livartowski, and T. Hamon. Automatic acquisition of synonym ressources and assessment of their impact on the enhanced search in ehrs. volume 48(2), pages 149–154, 2009. PMID 19283312.

[18] T. Hamon, A. Nazarenko, and C. Gros. A step towards the detection of semantic variants of terms in technical documents. In *Proc ACL*, pages 498–504, 1998.

[19] M. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, COLING '92, pages 539–545, 1992.

[20] A. Henriksson, M. Conway, M. Duneld, and W. Chapman. Identifying Synonymy Between Snomed Clinical Terms of Varying Length Using Distributional Analysis of Electronic Health Records. In *Proc AMIA*, pages 600–609, 2013.

[21] W. Hole and S. Srinivasan. Discovering Missed Synonymy in a Large Concept-Oriented Metathesaurus. In *Proc AMIA*, pages 354–358, 2000.

[22] K. Huang, J. Geller, M. Halper, and J. Cimino. Piecewise synonyms for enhanced umls source terminology integration. In *Proc AMIA*, 2007.

[23] K. Huang, J. Geller, M. Halper, Y. Perl, and J. Xu. Using wordnet synonym substitution to enhance umls source integration. *Artificial Intelligence in Medicine*, 46(2):97–109, 2009.

[24] C. Jacquemin. Syntagmatic and paradigmatic representations of term variation. In *Proc ACL*, pages 341–348, College Park, Maryland, USA, June 1999. Association for Computational Linguistics.

[25] C. Jacquemin, J. Klavans, and E. Tzoukermann. Expansion of multi- word terms for indexing and retrieval using morphology and syntax. In *Proceedings ACL*, pages 24–31, 1997.

[26] MBR. MEDLINE/PubMed Baseline Repository (MBR). `https://mbr.nlm.nih.gov`, 2016.

[27] MBR. MEDLINE/PubMed Baseline Repository (MBR) Single Words. `https://mbr.nlm.nih.gov/Download/2013/WordCounts/singleWords.gz`, 2016.

[28] J. McCrae and N. Collier. Synonym set extraction from the biomedical literature by lexical pattern discovery. *BMC Bioinformatics*, 9, 2008.

[29] B. McInnes, T. Pedersen, and S. Pakhomov. UMLS-Interface and UMLS-Similarity: Open Source Software for Measuring Paths and Semantic Similarity. In *Proc AMIA*, pages 431–435, 2009.

[30] MEDLINE. 2015 MEDLINE®/PubMed®Baseline: 23,343,329 Citations Found. `https://www.nlm.nih.gov/bsd/licensee/2015_stats/2015_LO.html`, 2015.

[31] MEDLINE. MEDLINE Fact Sheet. `https://www.nlm.nih.gov/pubs/factsheets/medline.html`, 2016.

[32] MedlinePlus. Methamphetamine: MedlinePlus. `https://www.nlm.nih.gov/medlineplus/methamphetamine` 2016.

[33] MetaMap Data Versions. MetaMap Data Versions. `https://metamap.nlm.nih.gov/DescriptionOfDataVersions.shtml`, 2016.

[34] Metathesaurus RRF. Metathesaurus - Rich Release Format (RRF) - UMLS Reference Manual - NCBI Bookshelf. `https://www.ncbi.nlm.nih.gov/books/NBK9685` (Section 3.3.4), 2016.

[35] J.-B. Michel, Y. Shen, A. Aiden, A. Veres, M. Gray, W. Brockman, and et. al. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331(6014):176–182, 2011.

[36] G. Miller. Wordnet: A lexical database for english. *Comm ACM*, 30(11):39–41, 1998.

[37] E. Morin and C. Jacquemin. Automatic acquisition and expansion of hypernym links. *Computers and the humanities*, 38:363–396, 2003.

[38] J. Mork, D. Demner-Fushman, S. Schmidt, and A. Aronson. Recent Enhancements to the NLM Medical Text Indexer. *BioASQ*, 2014.

[39] J. Mork, A. J. Yepes, and A. Aronson. The NLM Medical Text Indexer System for Indexing Biomedical Literature. *BioASQ*, 2013.

[40] MTI. Indexing Initiative: Medical Text Indexer (MTI). `https://ii.nlm.nih.gov/MTI/index.shtml`, 2016.

[41] NCBI. National Center for Biotechnology Information. `https://www.ncbi.nlm.nih.gov`, 2016.

[42] NLM Lexical Tools. NLM Lexical Tools. `https://lexsrv2.nlm.nih.gov/LexSysGroup/Projects/lvg/2014/docs/userDoc/tools/lvg.html`, 2016.

[43] S. Pakhomov, B. McInnes, T. Adam, Y. Liu, T. Pedersen, and G. Melton. Semantic Similarity and Relatedness between Clinical Terms: An Experimental Study. In *Proc AMIA*, pages 572–576, 2010.

[44] S. Pakhomov, T. Pedersen, B. McInnes, G. Melton, A. Reggieri, and C. Chute. Towards a Framework for Developing Semantic Relatedness Reference Standards. *J Biomed Inform*, 44(2):261–265, 2011.

[45] T. Pedersen, S. Pakhomov, S. Patwardhan, and C. Chute. Measures of Semantic Similarity and Relatedness in the Biomedical Domain. *J Biomed Inform*, 40(3):288–289, 2007.

[46] S. Pradhan, N. Elhadad, B. South, D. Martinez, L. Christensen, A. Vogel, and et. al. Task 1: ShARe/CLEF eHealth Evaluation Lab 2013. *Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop*, 2013.

[47] PubMed. Home - PubMed - NCBI. `https://ncbi.nlm.nih.gov/pubmed`, 2016.

[48] PubMed ATM. How PubMed works: automatic term mapping. `https://www.ncbi.nlm.nih.gov/books/NBK3827/#pubmedhelp.How_PubMed_works_automatic_te`, 2016.

[49] PubMed ATM. PubMed's Automatic Term Mapping Enhanced. NLM Technical Bulletin. 2004 Nov-Dec. `https://www.nlm.nih.gov/pubs/techbull/nd04/nd04_atm.html`, 2016.

[50] RTM. Restrict to MeSH Algorithm. https://ii.nlm.nih.gov/MTI/Details/RTM.shtml, 2016.

[51] E. Sayers. A General Introduction to the E-utilities. https://www.ncbi.nlm.nih.gov/books/NBK25497, 2016.

[52] Semantic Network. Semantic Network - UMLS®Reference Manual - NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK9679, 2016.

[53] SemEval-2014. Relevant Publications: SemEval-2014 Task 7. http://alt.qcri.org/semeval2014/task7/index.php?id=relevant-publications, 2014.

[54] ShARe/CLEF. ShARe/CLEF eHealth 2013 Shared Task: Guidelines for the Annotation of Disorders in Clinical Notes. https://drive.google.com/file/d/0B7oJZ-fwZvH5VmhyY3lHRFJhWkk/edit, 2013.

[55] Test Collections. Indexing Initiative: Datasets & Test Collections. https://ii.nlm.nih.gov/DataSets/index.shtml, 2016.

[56] Y. Tsuruoka, J. McNaught, and S. Ananiadou. Normalizing biomedical terms by minimizing ambiguity and variability. *BMC Bioinformatics*, 9(Suppl 3):S2, 2008.

[57] UMLS. UMLS Reference Manual, Bethesda (MD): National Library of Medicine (US). https://www.ncbi.nlm.nih.gov/books/NBK9684, 2016.

[58] UMLS. Unified Medical Language System (UMLS) - Home. https://www.nlm.nih.gov/research/umls, 2016.

[59] UMLS Semantic Groups. The UMLS Semantic Groups. https://semanticnetwork.nlm.nih.gov, 2016.

[60] C. Verspoor, C. Joslyn, and G. Papcun. The gene ontology as a source of lexical semantic knowledge for a biological natural language processing application. pages 51–56, 2003.