Unsupervised, corpus-based method for extending a biomedical terminology

Olivier Bodenreider **Thomas C. Rindflesch**
Lister Hill National Center
for Biomedical Communications
National Library of Medicine
Bethesda, Maryland, 20894 - USA
{olivier|tcr}@nlm.nih.gov

Anita Burgun
LIM - University of Rennes
Avenue du Pr Léon Bernard
35043 Rennes Cedex - France
Anita.Burgun
@univ-rennes1.fr

Abstract

Objectives: To automatically extend downwards an existing biomedical terminology using a corpus and both lexical and terminological knowledge. **Methods**: Adjectival modifiers are removed from terms extracted from the corpus (three million noun phrases extracted from MEDLINE), and demodified terms are searched for in the terminology (UMLS Metathesaurus, restricted to disorders and procedures). A phrase from MEDLINE becomes a candidate term in the Metathesaurus if the following two requirements are met: 1) a demodified term created from this phrase is found in the terminology and 2) the modifiers removed to create the demodified term also modify existing terms from the terminology, for a given semantic category. A manual review of a sample of candidate terms was performed. **Results**: Out of the 3 million simple phrases randomly extracted from MEDLINE, 125,000 new terms were identified for inclusion in the UMLS. 83% of the 1000 terms reviewed manually were associated with a relevant UMLS concept. **Discussion**: The limitations of this approach are discussed, as well as adaptation and generalization issues.

1 Introduction

Although providing a reasonable coverage of the clinical subdomain, Chute et al. (1996) showed that terminological resources such as the International Classificaton of Diseases, SNOMED International or the UMLS Metathesaurus do not capture all the concepts needed for representing clinical concepts in patient records. In a subsequent study, Chute and Elkin (1997) suggested that qualifiers (including adjectival modifiers) be available as a separate axis, in order to both increase the expressivity of a terminology and reduce its complexity by limiting the number of pre-coordinated terms.

In this study, rather than reducing the complexity, we use modification phenomena in order to investigate a corpus-based methodology for automatically discovering new terms for inclusion in a controlled vocabulary. In other words, our objective is to acquire hyponyms for terms in an original vocabulary that appear in the literature but are not present in the original vocabulary.

2 Background

Terms play a major role in a variety of natural language processing (NLP) applications, including machine translation, text understanding, automatic indexing, and information retrieval. Taking advantage of the availability of large corpora, automatic terminology acquisition methods were developed, for example, by Bourigault and Jacquemin (1999).

Word affinities generally play a central role in these methods. Grefenstette (1994) defines three

Proceedings of the ACL'2002 Workshop "Natural Language Processing in the Biomedical Domain" 2002:53-60.

orders of word affinities. "First order affinities describe collocates of words, second-order affinities show similarly used words, and third-order affinities create semantic groupings among similar words". In term extraction, this analysis is often applied to modifiers in order to establish groups of terms modified by a given modifier or the list of all possible modifiers for a given term.

Hersh et al. (1996) demonstrated the feasibility of applying natural language processing techniques to a corpus of clinical narratives from an electronic medical record (EMR) system. Although the terms extracted were compared to existing terms in the UMLS, the goal of this study was vocabulary discovery, but not the automatic integration of newly discovered terms into the terminology.

The automatic extension of an existing resource based on a corpus has also been studied. For example, Habert et al. (1998) propose a method for extending an existing specialized semantic lexicon.

Although related to these studies, our objective is to automatically extend downwards an existing biomedical terminology using a corpus and a combination of lexical, syntactic, and terminological knowledge.

In this study, the textual source, or corpus, is MEDLINE® [1], the U.S. National Library of Medicine's (NLM) premier bibliographic database. MEDLINE contains over eleven million references to articles from more than 4,600 worldwide journals in life sciences with a concentration on biomedicine.

We use the Unified Medical Language System® (UMLS®) Metathesaurus® [2] as the terminology to be extended. The Metathesaurus, also developed by NLM, is organized by concept or meaning. A concept is defined as a cluster of terms representing the same meaning (synonyms, lexical variants, acronyms, translations). For example, names for the disease multiple sclerosis include *multiple sclerosis*, *MS*, '*multiple sclerosis, NOS*', *disseminated sclerosis*, and *sclérose en plaques*. The 13th edition (2002) of the UMLS Metathesaurus contains over 1.5 million unique English terms drawn from more than sixty families of medical vocabularies, and organized in some 775,000 concepts.

[1] www.nlm.nih.gov/pubs/factsheets/medline.html
[2] umlsks.nlm.nih.gov

In order to address the large size of the Metathesaurus, we limited our study to a significant subdomain of clinical medicine: disorders and procedures (currently about 615,000 unique terms, corresponding to some 157,000 disorder concepts and 95,000 medical procedure concepts).

In the UMLS, each concept is categorized by semantic types (ST) from the semantic network. McCray et al. (2001) designed groupings of STs that provide a partition the Metathesaurus and, therefore, can be used to extract consistent sets of concepts corresponding to a subdomain, such as disorders or procedures.

Disorder and procedure terms were restricted to terms suitable for natural language processing, excluding, for example, such terms as *abdominal injury, NOS*. The notation "NOS", meaning "not otherwise specified", is a marker for underspecification often found in terminological resources. When identified in the Metathesaurus, obsolete and truncated terms were also excluded. 477,491 unique terms were selected for further processing.

3 Methods

The approach we propose for discovering candidates for Metathesaurus concepts is to compare phrases extracted from MEDLINE to current UMLS phrases. We capitalize on differences in modification structure between the MEDLINE phrase and the UMLS phrase to determine candidates for inclusion in the Metathesaurus. The crucial difference is between a phrase containing adjectival modification and a similar phrase "demodified" by removing its adjectives.

A phrase from MEDLINE becomes a candidate term in the Metathesaurus if the following two requirements are met: 1) a demodified term created from this phrase is found in the terminology and 2) similarly modified terms exist in the terminology, for a given semantic category. For example, the phrase *pancreatic bronchogenic cyst* is a candidate term for a disorder in the Metathesaurus because *bronchogenic cyst* exists in the Metathesaurus (concept: C0006281) and other Metathesaurus disorder terms are modified by the same adjective *pancreatic* (e.g., *pancreatic hemorrhage*).

3.1 Processing phrases from MEDLINE

Recently, Srinivasan et al. (2002) performed a shallow syntactic analysis on the entire MEDLINE

collection, using only titles and abstracts in English. Although their goal was to find Metathesaurus concepts in MEDLINE citations, an interesting side-effect of their analysis was the production of some 175 million noun phrase types that are available for further research.

From these phrases, we selected the subset of "simple" phrases, i.e., noun phrases excluding prepositional modification or any other complex feature. Examples of simple MEDLINE noun phrases include *abdominal aneurysmal aortitis* and *radical aggressive tumor resection*. Out of some forty million simple noun phrases, we randomly selected a subset of three million phrases to be used as our corpus, representative of the noun phrases found in MEDLINE.

The phrases in our sample were then submitted to an underspecified syntactic analysis described by Rindflesch et al. (2000) that draws on a stochastic tagger (see Cutting et al. (1992) for details) as well as the SPECIALIST Lexicon, a large syntactic lexicon of both general and medical English that is distributed with the UMLS. Although not perfect, this combination of resources effectively addresses the phenomenon of part-of-speech ambiguity in English.

The resulting syntactic structure identifies the head and modifiers for the noun phrase analyzed. Each modifier is also labeled as being adjectival, adverbial, or nominal. Although all types of modification in the simple English noun phrase were labeled, only adjectives and nouns were selected for further analysis in this study. For example, the term *catastrophic cervical spinal cord injuries* was analyzed as:

```
[[mod([catastrophic,adj]),
  mod([cervical,adj]),
  mod([spinal,adj]),
  mod([cord,noun]),
  head([injuries,noun])]]
```

A similar analysis was performed on UMLS terms for disorders and procedures.

3.2 Comparing MEDLINE phrases to UMLS phrases

The method we use can be summarized as follows. Starting with a random subset of three million simple noun phrases from MEDLINE, we excluded those that were already present in the UMLS by mapping them to the Metathesaurus. We then performed a shallow syntactic analysis of the phrases in order to select those consisting of one or more modifiers followed by a head noun.

Demodified terms were created by removing every possible combinations of modifiers in the terms. The same process was applied to disorder and procedure terms in the Metathesaurus in order to obtain a list of allowable adjectival modifiers for these two categories. Such modifiers in the Metathesaurus serve as a filter for MEDLINE phrases, since finding a similarly modified term in the UMLS is one of the two requirements for candidate terms. Demodified terms created from *accidental arterial perforations* include *arterial perforations*, *accidental perforations*, and *perforations*.

Demodified terms derived from MEDLINE phrases whose modifiers are all allowable are then mapped to the Metathesaurus. In this example, both *accidental* and *arterial* are adjectives found in the Metathesaurus in disorder or procedure terms. The second requirement for candidate terms is that at least one associated demodified term be mapped to a concept in the Metathesaurus. Two terms from our example map to Metathesaurus concepts: *arterial perforations* and *perforations*. The term *accidental perforations* does not map to any concept and is therefore eliminated from further processing. The last step ensures that, in case of multiple demodified terms, the finest-grained is selected. *Arterial perforations* is selected over *perforations* for this reason.

Figure 1 illustrates the sequence of methods used in the study and the interactions between processing MEDLINE phrases and Metathesaurus terms. It also presents the number of MEDLINE phrases and Metathesaurus terms present before and after each of the six steps detailed below.

Step1. Mapping phrases to the UMLS

In order to identify MEDLINE phrases that already exist in the Metathesaurus, all MEDLINE phrases in our sample were mapped to the UMLS by first attempting an exact match between input term and Metathesaurus concept. If an exact match failed, normalization was then attempted. This process makes the input and target terms potentially compatible by eliminating such inessential differences as inflection, case and hyphen variation, as well as word order variation. Duplicate names were removed from each set prior to mapping to the UMLS.

Step 2. Identifying (adj+, noun*, head) phrases

Since this method is based on adjectival modification, the syntactic analysis was used to restrict the original sets of MEDLINE phrases and Metathesaurus terms to phrases and terms having the following structure: `(adj+, noun*, head)`.

The phrase is required to start with an adjectival modifier, possibly followed by other adjectives and end with a head noun, possibly preceded by other nouns. This specification excludes both simple terms (e.g., one isolated noun) and complex terms, not suitable for our analysis.

Step 3. Creating demodified terms

When adjectival modifiers are identified in a term O, a set of demodified terms $\{T_1, T_2,\ldots,T_n\}$ is created by removing from term O any combinations of adjectival modifiers found in it. While the structure of the demodified terms remains syntactically correct, the semantics of some terms may be anomalous, especially when adjectives other than the leftmost are removed. Since most of them are semantically valid, we found it convenient to keep all demodified terms for further analysis. Demodified terms with incorrect semantics will be filtered out later in the experiment, since they will not map to an existing concept.

The number of demodified terms is $2^m - 1$, m being the number of adjectival modifiers. For example, the term *chronic sciatic constriction injury* starts with the two adjectival modifiers *chronic* and *sciatic*, so that the following three demodified terms are generated *sciatic constriction injury*, *chronic constriction injury*, and *constriction injury*.

Although there is no need to demodify UMLS terms in this study, the removal of adjectival modifiers was used to establish a list of adjectives occurring in disorder and procedure terms. These adjectives constitute the list of allowable modifiers for the two categories of terms studied.

Step 4. Searching for similarly modified terms in the Metathesaurus

In this study, one requirement for candidate terms is that a similarly modified term be present in the terminology. The list of allowable modifiers computed from Metathesaurus terms at the previous step provides a simple way to implement this constraint. For a given category, an allowable modifier indicates that some terms from this category are modified by this modifier, i.e., that a similarly modified term exists in the Metathesaurus.

Practically, MEDLINE phrases whose adjectival modifiers do not all belong to the list of allowable modifiers are excluded from further analysis, because, by definition, there will be no similarly modified term in the Metathesaurus.

Step 5. Searching for demodified terms in the Metathesaurus

The second requirement for a MEDLINE phrase to become a candidate term is that a demodified term created from this phrase be found in the terminology. Using only MEDLINE phrases whose adjectival modifiers all belong to the list of allowable modifiers, the demodified terms created from these phrases are mapped to the UMLS using the procedure previously described. MEDLINE phrases with no demodified term mapped to a UMLS concept are definitely excluded. Demodified terms mapping to concepts in categories other than disorders or procedures are also eliminated.

As explained earlier, the compatibility of the modifiers of the candidate terms is checked against the list of allowable modifiers for the category of the Metathesaurus concept(s) to which a demodified term mapped. In some cases, a cadidate term is eliminated because the modified term maps to a disorder concept, while its modifiers are compatible with procedures (or the other way around).

Step 6. Hooking candidate terms to the terminology

The remaining step consists of finding the appropriate hook in the terminology for the candidate term. Based on the fact that modification is normally associated with a hyponymic relation, tentative parents for the candidate term will be those that map to the demodified terms generated from this term.

When only one demodified term maps to a Metathesaurus concept, this concept is selected as the tentative parent for the candidate term. When several demodified terms map to Metathesaurus concepts, the preference is given to the concept that is likely to be closest to the term. As a surrogate for closeness, we use the following heuristics: 1) the fewer modifiers removed, the closer the terms, and 2) the candidate term and the demodified term are closer if the modifier removed is the

leftmost modifier. In the rare cases where several demodified terms are deemed equally close to the candidate term, they are all selected as tentative parents.

4 Evaluation

A subset of 1000 candidate terms was randomly selected to evaluate this method. The existence of a hyponymic relationship between the candidate term and the Metathesaurus concept(s) selected as valid mappings for the demodified terms created from the candidate term was evaluated by a manual review performed by the authors. A secondary objective of this evaluation was to gain insights about how these methods could be tuned in order to prevent inaccurate mappings and select the most useful candidate terms.

The following classification was used to describe the quality of the hyponymic relationship between the candidate term and the Metathesaurus concept(s) selected: "relevant" means that the hooking of the candidate term to the terminology was relevant, even if a more specific concept was available; "non relevant" means that none of the Metathesaurus concepts selected was a correct hook for the candidate term; "more or less relevant" means that the Metathesaurus concepts selected were not irrelevant as hooks, but were distant ancestors, i. e., too general for the relationship to be fully informative. Finally, for polysemous candidate terms, it was not possible to evaluate the quality of the relationship with certainty.

5 Results

Out of the 3 million randomly selected simple MEDLINE phrases, 125,464 phrases were selected as candidate terms with (at least) one Metathesaurus concept to hook them to. Details about the number of phrases selected at each step of the processing are given in Figure 1.

The total number of adjectival modifiers found in a MEDLINE phrase ranged from 1 to 7. Phrases with one (42% of the phrases) or two (46% of the phrases) modifiers predominated. The candidate terms resulted from removing one modifier from the original phrase in 66% of the cases, and two modifiers in 30% of the cases. The modifier(s) removed included the leftmost modifier in 95% of the cases. The list of the most frequent modifiers in existing terms and candidate terms for disorders and procedures is given in Table 1.

In 78% of the cases, only one demodified term was generated from the original phrase. Two demodified terms were generated in 17% of the cases. In 61% of the cases, only the leftmost adjective was removed. The first two adjectives in the phrase were removed in 29% of the cases.

Out of the 1000 candidate terms reviewed as hyponyms of some Metathesaurus concept, 834 were considered relevant, 28 more or less relevant, and 138 not relevant.

6 Discussion

This study confirms the observations made in two previous studies taking advantage of adjectival modification phenomena in various tasks related to terminologies, in particular to suggest hyponymic relations among medical terms [Bodenreider et al. (2001)] and to assess the consistency of a biomedical terminology [Bodenreider et al. (2002)].

Although a larger-scale evaluation would be required to fully assess the results, the major finding is that the method is effective at automatically identifying many new terms for inclusion into an extended terminological resource. However, the evaluation revealed some limitations which are analyzed below. Adaptation and generalization issues will be addressed as well.

Limitations

The errors discovered during the manual review illustrate some of the limitations of this method. More exactly, these limitations are common to many NLP applications. Although acronyms were sometimes associated with their correct meaning in the Metathesaurus, in the set of terms reviewed manually, the presence of acronyms was responsible for 22% of the non-relevant associations. For example, the MEDLINE term *individual black rats*, whose two adjectival modifiers are allowable disease modifiers, is wrongly identified as a hyponym of *recurrent acute tonsillitis* because the acronym *RAT* is associated (as a synonym) with the disease *recurrent acute tonsillitis* in the Metathesaurus. In some cases, failure to identify the correct part of speech also resulted in inaccurate associations (e.g., *controlling stress* to *stress* where *controlling* was actually not an adjective). Not all

truncated terms present in the Metathesaurus synonyms of some concepts are identified as such. When not identified, truncated terms are used for the mapping, sometimes resulting in inaccurate associations. For example, the candidate term *urinary protein* is wrongly associated with the concept *protein measurement* because *protein* is considered a synonym for the procedure *protein measurement* in the Metathesaurus.

Sometimes, the association is not inaccurate, but the concept associated with the candidate term is very general, and the relationship weakly informative. For example, once demodified, *aplastic syndrome* is associated with *syndrome*, a concept close to the top of the hierarchy. Although *aplastic syndrome* is a valid hyponym of *syndrome*, it would be more accurately categorized as a kind of hematologic syndrome, which requires domain knowledge unavailable here.

Finally, in some cases, because hyponymy is the only relation considered, the association of a candidate term with a Metathesaurus concept, although relevant, is not necessarily the closest possible. For example, the term *colonic vaginal fistula* is correctly associated with its hypernym *vaginal fistula*, but fails to be identified as a synonym of the concept *fistula of vagina to large intestine*. Practically, in a completely automatic setting, the use of this algorithm could result in creating several concepts for the same meaning.

Tuning

This algorithm can be tuned from a strict mode, allowing fewer phrases to automatically become candidate terms, but with greater precision, to a relaxed mode, selecting a larger number of candidate terms when recall is the priority. The latter would require some supervision prior to integrating the candidate terms into the terminology.

Almost all the limitations mentioned above can be addressed. Terms containing acronyms could be identified and eliminated before mapping to the Metathesaurus. Part of speech taggers trained on a terminology would more accurately identify the part of speech of words that can be both adjectives and nouns. Truncated Metathesaurus terms should be systematically excluded from the index used for mapping. Methods for identifying synonymy based on derivational variation or other techniques could also be investigated.

Moreover, additional refinement could be brought to this method. For example, when demodified terms are created, the removal of adjectives could be restricted to the leftmost, thus maximally preserving the structure of the remaining noun phrase, and therefore limiting the risks of association with a semantically distant concept.

Finally, using statistical information about the distribution of adjectival modifiers could provide a surrogate for the strength of the association. For example, knowing that many diseases can be acute, if this adjective is found in the corpus as the modifier of a disease concept, this association could be accepted with a confidence proportional to the relative frequency of this modifier for all diseases, in the case of acute for a disease, a high confidence.

Generalization

The method presented was voluntarily restricted to the domain of disorders and procedures, to adjectival modification, and to the biomedical literature.

Generalizing to other domains would pose no problems as long as terms of their terminology is amenable to natural language processing techniques and modification phenomena. This would include domains such as anatomy or physiology. However, domains such as molecular biology, with many gene and gene product names, and chemistry, with many chemical names would probably yield fewer candidate terms.

Nominal modification is common in English and in principle can be addressed with a methodology similar to the one discussed here. Nominal modifiers often express a quality more closely related semantically to the head than do adjectives. Details in the methodology would be adjusted to accommodate this characteristic.

Generalization to other corpora such as patient records and electronic textbooks of medicine would likely yield additional terms.

Finally, although this method relies on features of the UMLS such as the semantic categorization of the concepts, it could also be applied to other terminologies that do not provide this feature, such as the Medical Subject Headings (MeSH). In this case, the concept hierarchy itself could be used as a surrogate for the categorization. For example, if the candidate term *chronic rheumatic fever* is associated with the MeSH term *rheumatic fever*, its

category is disease because the polyhierarchical structure in which *rheumatic fever* is involved ultimately converges to the top of the C hierarchy, i.e., the term *diseases*.

References

Bodenreider, O., Burgun, A., and Rindflesch, T. C. (2001). *Lexically-suggested hyponymic relations among medical terms and their representation in the UMLS*. Proceedings of TIA'2001 "Terminology and Artificial Intelligence", 11-21.

Bodenreider, O., Burgun, A., and Rindflesch, T. C. (2002). *Assessing the consistency of a biomedical terminology through lexical knowledge*. Proceedings of the Workshop on Natural Language Processing in Biomedical Applications (NLPBA'2002), 77-83.

Bourigault, D., and Jacquemin, C. (1999). *Term Extraction + Term Clustering: An Integrated Platform for Computer-Aided Terminology*. Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL'99).

Chute, C. G., Cohn, S. P., Campbell, K. E., Oliver, D. E., and Campbell, J. R. (1996). *The content coverage of clinical classifications. For The Computer-Based Patient Record Institute's Work Group on Codes & Structures*. J Am Med Inform Assoc *3*, 224-233.

Chute, C. G., and Elkin, P. L. (1997). *A clinically derived terminology: qualification to reduction*. Proc AMIA Annu Fall Symp, 570-574.

Cutting, D. R., Kupiec, J., Pedersen, J. O., and Sibun, P. (1992). *A practical part-of-speech tagger*. Proceedings of the Third Conference on Applied Natural Language Processing, 133-140.

Grefenstette, G. (1994). Corpus-derived first, second and third-order word affinities. Paper presented at: EURALEX (Amsterdam).

Habert, B., Nazarenko, A., Zweigenbaum, P., and Bouaud, J. (1998). *Extending an existing specialized semantic lexicon*. Proceedings of the First International Conference on Language Resources and Evaluation, 663-668.

Hersh, W. R., Campbell, E. H., Evans, D. A., and Brownlow, N. D. (1996). *Empirical, automated vocabulary discovery using large text corpora and advanced natural language processing tools*. Proc AMIA Annu Fall Symp, 159-163.

McCray, A. T., Burgun, A., and Bodenreider, O. (2001). *Aggregating UMLS semantic types for reducing conceptual complexity*. Medinfo *10*, 216-220.

Rindflesch, T. C., Rajan, J. V., and Hunter, L. (2000). Extracting molecular binding relationships from biomedical text. In "Proceedings of the 6th Applied Natural Language Processing Conference" (San Francisco, Morgan Kaufmann Publishers), pp. 188-195.

Srinivasan, S., Rindflesch, T. C., Hole, W. T., and Aronson, A. R. (2002). *Finding UMLS Metathesaurus concepts in MEDLINE*. Proc AMIA Annu Fall Symp, (submitted).

Table 1. Most frequent modifiers identified in existing terms (UMLS) and candidate terms (MEDLINE) for disorders and procedures.

Disorder terms		Procedure terms	
MEDLINE	**UMLS**	**MEDLINE**	**UMLS**
severe	congenital	using	surgical
chronic	acute	clinical	serum
acute	accidental	two	diagnostic
primary	intentional	routine	dental
human	chronic	conventional	local
recurrent	pulmonary	surgical	therapeutic
pulmonary	malignant	initial	total
multiple	cerebral	human	patient
two	renal	total	percutaneous
malignant	benign	standard	cardiac

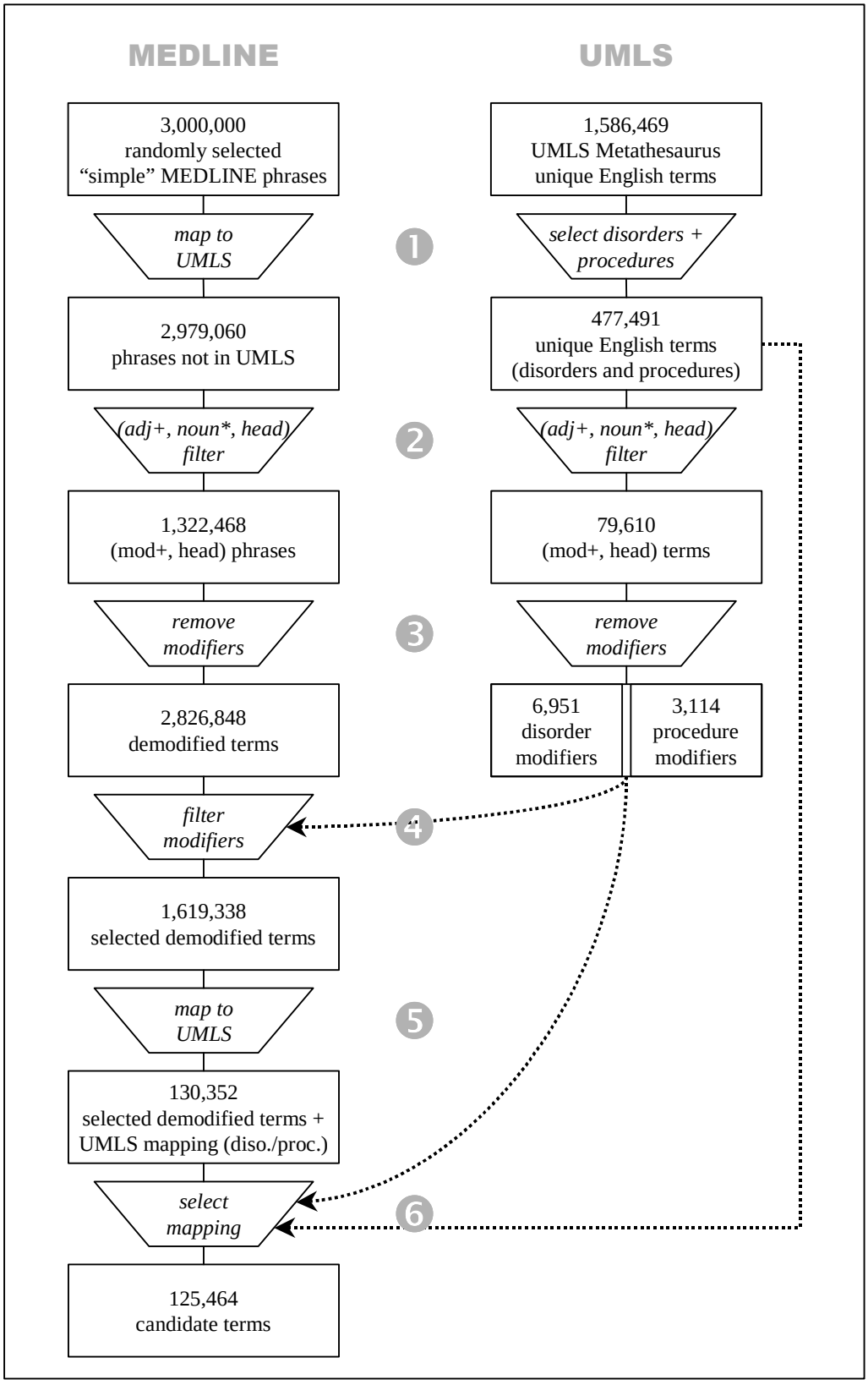

Figure 1. Summary of the methods for comparing MEDLINE phrases to UMLS terms.