

Research Paper ■

Semantic Classification of Biomedical Concepts Using Distributional Similarity

JUNG-WEI FAN, MS, CAROL FRIEDMAN, PHD

Abstract Objective: To develop an automated, high-throughput, and reproducible method for reclassifying and validating ontological concepts for natural language processing applications.

Design: We developed a distributional similarity approach to classify the Unified Medical Language System (UMLS) concepts. Classification models were built for seven broad biomedically relevant semantic classes created by grouping subsets of the UMLS semantic types. We used contextual features based on syntactic properties obtained from two different large corpora and used α -skew divergence as the similarity measure.

Measurements: The testing sets were automatically generated based on the changes by the National Library of Medicine to the semantic classification of concepts from the UMLS 2005AA to the 2006AA release. Error rates were calculated and a misclassification analysis was performed.

Results: The estimated lowest error rates were 0.198 and 0.116 when considering the correct classification to be covered by our top prediction and top 2 predictions, respectively.

Conclusion: The results demonstrated that the distributional similarity approach can recommend high level semantic classification suitable for use in natural language processing.

■ *J Am Med Inform Assoc*. 2007;14:467-477. DOI 10.1197/jamia.M2314.

Introduction

Enormous amounts of textual biomedical information are becoming available electronically on a daily basis, making it difficult for researchers or indexers to keep current. There is a pressing need to develop automated methods to index, codify, and manage the growing body of information in order to increase its accessibility and use for applications such as information retrieval, knowledge discovery, and database curation. Natural Language Processing (NLP) techniques have been playing an increasing role in extracting and managing biomedical entities and relations from text.¹⁻⁴ However, the development of NLP methods is costly and time-consuming, largely due to the formidably large number of biomedical terms that NLP systems must recognize. Fortunately, much effort has gone into creating biomedical

knowledge resources and tools usable by NLP systems, for instance, the Gene Ontology (GO),⁵ UniProt,⁶ HUGO,⁷ JCBN,⁸ NCBI Taxonomy,⁹ and the Unified Medical Language System (UMLS),^{10,11} thereby substantially reducing the overhead of the development process. These resources serve as huge repositories of biomedical terms. The UMLS also provides associations from the terms to normalized concepts, which are assigned semantic categories.

High-quality semantic classification is generally critical for NLP and for other knowledge-based systems using concepts and/or reasoning. To facilitate the development of ontologies, it is worth exploring an automated method to semantically classify ontological concepts. Ideally the method could also be used to improve the ontology itself and the performance of the systems depending on it. For this study we focus on use of corpus-based distributional properties for such a classification task. According to Harris's sublanguage theory, the syntactic dependence of words on other words in general language and, especially in specialized languages, exhibits different likelihoods.¹² For example, in the biomedical domain the object of the verb "prevent" is more likely to be a disorder than a body part. NLP researchers have explored use of such distributional differences to semantically categorize words.¹³⁻¹⁵ Such an approach differs significantly from the manner by which ontologies are developed, because experts generally use their knowledge of the domain to manually classify ontological concepts. An automatic tool that assists experts in developing and maintaining ontologies for NLP as well as other knowledge-based applications would have substantial benefits of being consistent, reproducible, and high throughput, and when used in conjunction with experts, should help speed the overall process.

Affiliation of the authors: Department of Biomedical Informatics, Columbia University, New York, NY.

The authors thank NLM's Dr. Alan Aronson, Guy Divita, and James Mork for help with the MetaMap program and the MBR database. We also would like to thank Dr. George Hripcsak for performing the expert evaluation to determine the semantic classification for a set of UMLS concepts. We thank Jessica Ancker and Chintan Patel for discussing some questionable UMLS classifications and thank Dr. Peter Hung for validating an example.

This work was supported by Grants R01 LM7659 and R01 LM8635 from the National Library of Medicine.

Correspondence and requests: Carol Friedman, PhD, Department of Biomedical Informatics, Vanderbilt Clinic, 5th Floor, 622 West 168th Street, New York, NY 10032; e-mail: <carol.friedman@dbmi.columbia.edu>.

Received for review: 10/25/2006; accepted for publication: 4/09/2007.

In this paper, we focus on semantic classification within the UMLS, because it is a comprehensive knowledge representation framework with a growing user population and regular maintenance. Many NLP researchers have used UMLS concepts and their associated terms as a source of lexical knowledge. For example, the UMLS has been used in biomedical named entity recognition tasks¹⁶ and in providing mappings to normalized conceptual representations (e.g., for indexing of medical images).¹⁷ The UMLS has also been used by some NLP systems to determine relations among the extracted terms through specified semantic patterns.^{18,19} Systems using the UMLS and linguistic knowledge to determine relations generally depend on recognizing UMLS concepts and associating them with the appropriate semantic categories to function properly. Therefore, semantic classification is critical to relation determination that follows entity recognition. However, the UMLS semantic classifications have been reported to contain inconsistencies,²⁰ with a number of concepts assigned questionable semantic types. For example, concepts are semantically inconsistent under T169 "Functional Concept" (2006AD), which contains procedure-related concepts such as "Intramuscular Injection," function-related concepts such as "Regulatory Pathway," disorder-related concepts such as "Tumor-like lesion," and very high-level concepts such as the verb "Increase."

The corpus-based method we propose is based on distributional properties of terms and is domain-independent. Therefore, it is general and can be used for other ontologies as well, provided that the ontological concepts can be found in text. The requirement of having the concepts occur in text is not a limitation for our work because one of the primary goals of this automated semantic classification method is to increase the utility of an ontology for NLP purposes. Additionally, rather than using the existing UMLS semantic classes, we focus on reclassifying UMLS concepts into broader semantic classes that are more compatible with NLP applications. Our main consideration involves well-defined and clinically relevant classes, and therefore disregards classes such as T066 "Machine Activity," T070 "Natural Phenomenon or Process," and T170 "Intellectual Product." A secondary consideration concerns the granularity of the classification, which is elaborated on in the Background section.

Thus, in this paper we describe the development and evaluation of an automated method for reclassification of UMLS concepts that is based on a distributional similarity approach. We try to address the issues of semantic granularity, objectivity, and consistency. To the best of our knowledge, this is the first time an automated, corpus-based method has been developed for the semantic classification of UMLS concepts.

In the following sections we will first review related research on reorganizing the UMLS semantic classification and associated issues, different approaches for performing semantic classification (with an emphasis on the corpus-based approach), and background about the specific resources we used in building our semantic classifiers. Then we describe the details of the implementation and experiments, followed by the results, our observations, interpretations, and conclusions.

Background

The UMLS Semantic Network

The current release (2006AD) of the UMLS integrates terminologies from 138 source vocabularies into a concept-centered ontology. Each UMLS concept has a corresponding Concept Unique Identifier (CUI), which is assigned to one or more semantic types in the Semantic Network (SN).²¹ The current SN classification consists of 135 semantic types and is organized hierarchically. The semantic categories of the SN provide NLP systems with a framework for detecting relational patterns. For example, the template [(*Pharmacologic Substance*) **lead(s) to** (*Disease or Syndrome*)] can help screen potential adverse drug reactions extracted from text. The ontological relations can further extend the pattern's coverage by inference: a concept of the type *Antibiotic* also satisfies the subject role in the template, because *Antibiotic* IS-A *Pharmacologic Substance* in the SN. The advantage of the UMLS is its comprehensiveness and continual maintenance. On the other hand, using the UMLS is not free of problems. As mentioned earlier, inconsistencies and questionable semantic assignments to the SN have been discovered. Some high-level SN types are known to subsume semantically heterogeneous concepts and tend to be noisier. For example, T058 "Health Care Activity" includes concepts such as "Microdialysis" and "Ergometry," which are procedures applying specific techniques. However, that semantic type also includes many general health care activities such as "Wellness Programs," "Health Policy Monitoring," and "Immunization claim." Moreover, although semantic classification through expert curation is a strength of the UMLS, it involves an inevitable human subjectivity.

The appropriateness of semantic granularity is another open issue and is actually application-dependent. The semantic types of the SN are geared for biomedical knowledge-based applications but sometimes are too fine-grained for a general-purpose biomedical NLP system because the semantic patterns occurring in biomedical text are coarser than the types of semantic classes needed for knowledge-based applications. For example, text may have the same semantic patterns involving *Fungus*, *Virus*, *Rickettsia* or *Chlamydia*, *Bacterium*, and *Archaeon* as those involving microorganisms in general, and therefore these classes would not be important to differentiate for NLP purposes. However, such differentiation could be important for other biomedical applications. Considerable research has been performed on simplifying the UMLS semantic classification as well as auditing the errors in it. The motivation for reducing the complexity of the SN is to make it easier for human comprehension and for system integration (including NLP systems). Several approaches have been proposed for aggregating the SN types into fewer semantic groups, such as rule-based regrouping,²² regrouping based on SN relations,²³ and regrouping by string matching of SN type definitions.²⁴ Some SN auditing approaches were derived directly from a coarser-grained classification (i.e., meta-schema).^{25,26} Others performed auditing by checking concepts assigned to mutually exclusive SN types,²⁷ by checking redundant SN assignments,²⁸ or by applying principles to enforce ontological soundness.²⁹

Basically the above approaches can be considered “ontological” and thus are still susceptible to the limitations of the existing SN structure. For example, we consider the concept “Intramuscular Injection” should be more appropriately assigned a procedure-related class instead of the broad “Functional Concept.” However, this can not be achieved by a re-grouping approach that always aggregates “Functional Concept” with other broad types such as “Qualitative concept.” Also, none of the above auditing methods involve automating the process of determining the most appropriate classification(s) for each concept. Our work differs from the above related work in that we reclassify the concepts directly and automatically by statistical models.

Ontology-based Semantic Similarity

To semantically classify ontological concepts, a key step is to define a semantic similarity measure. A widely applied measure is to use the hierarchy of the ontology to calculate the distance between two concepts. The basic idea is that two concepts with smaller distance are considered to be more similar than that with larger distance, and the distance is generally defined using either a node-based³⁰ or an edge-based approach.³¹ The ontological approaches are suitable for implementing unsupervised clustering of the concepts, and the node-based approach can also take into account the probability estimation of the concepts in real language usage. However, the dependence on the ontological structure makes it susceptible to potential biases inherent in the hierarchical arrangement, and erroneous ontological relations could have a deterministic effect and propagate over the clustering. Additionally, Caviedes and Cimino³² showed that the clustering could become intractable when applied to a huge and complex ontology like the entire UMLS, due to the extreme density and sparseness of edge distribution present in different parts of the ontology.

Corpus-based Semantic Similarity

A different paradigm of semantic similarity originates from Firth’s epigram “You shall know a word by the company it keeps.”³³ A similar theory was also elaborated by Harris with his distributional hypothesis,³⁴ which emphasizes semantic and syntactic constraints based on contexts of words when applied to a specialized domain. That is, words belonging to the same semantic class co-occur more frequently in specific syntactic relations with a restricted set of words. Therefore, to measure the similarity between “cryotherapy” and “cryosurgery,” we would have to measure and compare the similarity of between the distributions of their contexts in a corpus. The contexts can be simply co-occurring words, or, following Harris’ hypothesis, they can be words associated with specific syntactic relations, such as being an adjective modifier of the head noun or the direct object of a verb. Karlgren and Sahlgren³⁵ claimed that the distributional hypothesis ideally accounts for Wittgenstein’s renowned semantics argument “meaning is use,”³⁶ and can represent a semantic space with features of empirical language usage, which thus tends to be free from the bias of any predefined ontological meanings. The distributional approach has been adopted in a range of NLP research, such as word classification¹³⁻¹⁵ and automatic thesaurus construction.^{37,38}

Although the implementation varies with different similarity formulae, the basic idea is that the similarity of two words w_1, w_2 can be calculated through the two distributions

$P(C|w_1)$ and $P(C|w_2)$, where C is the union of the w_1, w_2 context features, which in the simplest form would be words that co-occur in a corpus with w_1 and w_2 . That is, the more overlap there is between the contexts that co-occur with both w_1 and w_2 , the more similar w_1 and w_2 are semantically. A practical issue of implementing distributional similarity is the probability estimation for unseen (c_i, w_i) pairs, where c_i is an element of C that may co-occur with w_1 but not w_2 or vice versa. Recently, a widely cited solution has been Lee’s α -skew divergence,³⁹ which was derived from the Kullback-Leibler (KL) divergence (or relative entropy):⁴⁰

$$D(P \parallel Q) = \sum_x P(x) \cdot \log \frac{P(x)}{Q(x)} \quad (1)$$

The KL divergence quantifies the coding inefficiency of assuming the distribution is Q while the true distribution is P .⁴¹ It is preferably called “divergence” because it does not satisfy the symmetric property (i.e., $D(P \parallel Q) \neq D(Q \parallel P)$) or the triangle inequality (i.e., $D(P \parallel Q) \leq D(P \parallel R) + D(R \parallel Q)$) of Euclidean distance measures. The α -skew divergence is an asymmetric generalization of the KL divergence (here we substitute the above example into the following general formula):

$$S\alpha(P(C|w_1), P(C|w_2)) = D(P(C|w_2) \parallel \alpha \cdot P(C|w_1) + (1 - \alpha) \cdot P(C|w_2)) \quad (2)$$

where $0 \leq \alpha \leq 1$, and D is the KL divergence. The linear combination in $S\alpha$ helps estimate the conditional probabilities of unseen $\{c_i, w_i\}$ pairs in $P(C|w_1)$ by a magnitude of $(1 - \alpha) \cdot P(c_i|w_2)$ and thus smoothes the model.

Features for Corpus-based Semantic Similarity

The simplest features used for distributional similarity are words within a fixed window of contexts. For example, using a window of preceding/following three words without crossing a sentence boundary, the features of “tumor” within the sentence “The cryotherapy prevented local tumor recurrence” would be “cryotherapy,” “prevented,” “local,” and “recurrence.” Another type of feature is based on shallow (or partial) parsing, in which pattern matching is applied over shallow-parsed sentences (i.e., sentences where only simple phrases are identified instead of the complete sentence structures) to extract surface dependencies. Cimiano and Völker⁴² used such an approach to extract contextual features, which they called “pseudo-syntactic dependencies” in their named entity classification task. For example, applying their pattern matching method over the example sentence, the direct preceding noun “cryotherapy” would be extracted as the subject of the verb “prevented.” Although many true syntactic dependencies can still be captured using this method, it has limited ability in extracting dependencies of longer distances. For example, it would be more difficult to identify “cryotherapy” as the subject of “prevented” in the sentence “The cryotherapy which was performed by his urologist 6 months ago prevented local tumor recurrence” because of the relative clause “which was performed by his urologist 6 months ago” that modifies “cryotherapy.” However, the shallow parsing is easier to implement and more efficient than using a full parsing approach. Cimiano and Völker showed that the pseudo-syntactic dependencies outperformed adjacent words within a fixed window. In a task that automatically clustered adjectives, Hatzivassiloglou also re-

ported syntactic features from shallow parsing outperformed window-based co-occurrences.⁴³

Another important issue about the syntactic dependencies is feature-weighting. Assigning a weight to each syntactic dependency with respect to a particular word estimates the importance of that dependency in characterizing the word. For example, if the adjective “malignant” is frequently used to modify disorder terms such as “tumor,” then “malignant” should be assigned a relatively high weight with respect to “tumor.” In contrast, “malignant” rarely modifies a procedure type of term such as “cryotherapy” and thus should be weighed less correspondingly. In Cimiano and Völker’s work mentioned above, they reported that α -skew divergence performed best with conditional probability as the feature-weighting function, and therefore we adopted it for the α -skew divergence implemented in this paper (as described in detail in the Materials and Methods section).

Corpus-based Semantic Similarity of Concepts

Although earlier corpus-based semantic classification focused on measuring the similarity of words, the same approach can readily be applied to measuring the similarity of concepts. For example, the contextual features of two synonymous terms (according to 2006AD UMLS) “joint inflammation” and “inflammatory arthritis” can be pooled because both terms correspond to the same concept C0003864. The following related researches suggested that computing distributional similarity at the concept or semantic class level is not only feasible but desirable. In some tasks semantic class-based statistics were shown to be especially useful for addressing the data sparseness problem^{44,45} because more contexts are acquired when using synonyms. Geffet and Dagan indicated that distributional similarity principles should hold at the concept (word sense) level rather than the word level.⁴⁶ Interestingly Mohammad and Hirst modified Firth’s words into “You shall know a sense by the company it keeps”⁴⁷ and reported that concept-based distributional measures correlated more closely with human judgment than word-based measures in a word-pair ranking task. In the biomedical domain, Pedersen et al.⁴⁸ recently used the Mayo Clinic Thesaurus to aggregate synonymous occurrences of individual SNOMED-CT concepts in the Mayo Clinic Corpus of Clinical Notes. They created a window-based context vector for each concept, and calculated the cosine values to measure the semantic relatedness between the concepts. Their results showed that the distributional approach achieved the highest correlations with both physicians and medical coders than the other ontology-based approaches.

The methods proposed in this paper differ from the above related work because to the best of our knowledge, this is the first time a distribution-based method has been proposed for automatically classifying UMLS concepts. The proposed distributional approach uses syntactic dependencies from shallow parsing to classify UMLS concepts. In addition, we used existing knowledge resources provided by the National Library of Medicine (NLM) to build our models for reclassifying and validating UMLS concepts.

Resources for Automatic Semantic Classification

1) The UMLS Semantic Network

As discussed above, some high-level SN types (e.g., “Health Care Activity”) are known to subsume semanti-

cally heterogeneous concepts. These types are not suitable to be used in training distributional models. However, there are also well-defined SN types (e.g., “Diagnostic Procedure”) that generally contain semantically homogenous concepts. These well-defined SN types can be grouped to form broader semantic classes for which distributional classifiers can be built, as will be described in Material & Methods. The SN type assignments of the CUIs are also subject to manual curation and changes over different UMLS releases. These updates were made to improve the UMLS and thus should be relatively accurate. Therefore, it should be possible to use the updates as an economical way to obtain a test set (and gold standard) of a moderate sample size.

2) MetaMap and the MEDLINE/PubMed Baseline Repository

Associated with the UMLS, the MetaMap⁴⁹ program was developed by the NLM to map terms in free text to UMLS concepts. The NLM also maintains the MEDLINE/PubMed Baseline Repository (MBR)⁵⁰ annual databases that represent a static view of all completed citations in MEDLINE up to a particular year, with 14,792,864 citations of the whole 2005 database processed by MetaMap into human- and machine-readable formats. The MetaMap machine-readable outputs can serve an important role for implementing concept-based distributional similarity in two ways: 1) the machine-readable outputs contain detailed information about the shallow parsing and concept mapping (see Supplement 1, available as an online data supplement at www.jamia.org), which is valuable for obtaining the syntactic dependencies; 2) MetaMap facilitates the aggregation of synonymous terms along with their syntactic dependencies into concepts (e.g., both “cryotherapy” and “cold therapy” are mapped to C0010412).

Materials and Methods

The two major components of the material were: 1) the UMLS 2006AA and 2005AA MRSTY.RRF files, from which we obtained an initial test set containing 5,996 CUIs that had been assigned different SN type(s); 2) two disjoint training corpora requested from the 2005 MBR database in MetaMap machine-readable format: one consisted of 100,000 title/abstracts and the other consisted of 99,313 (we will refer to them as the 100K and the 99K set hereafter).

Figure 1 provides an overview of different steps of the classification method, which can be summarized as: 1a) the MBR corpus was used to determine syntactic dependencies of the CUIs; 1b) seven training classes were determined, the CUIs belonging to each class were identified, their syntactic dependencies collected (however, test CUIs were excluded), weights were calculated and normalized, and a distribution of the dependencies was formed for each class; 1c) for each test CUI, the syntactic dependencies of the CUI were collected, weights of the corresponding features were calculated and normalized, and a distribution of the dependencies was formed for the CUI. Finally, as shown on the right-hand side between Figure 1b and 1c, the α -skew divergence was used to compute the distributional similarities, and then the class with the closest similarity score was proposed as the appropriate class. In the following subsections we describe each step of the methods in more detail.

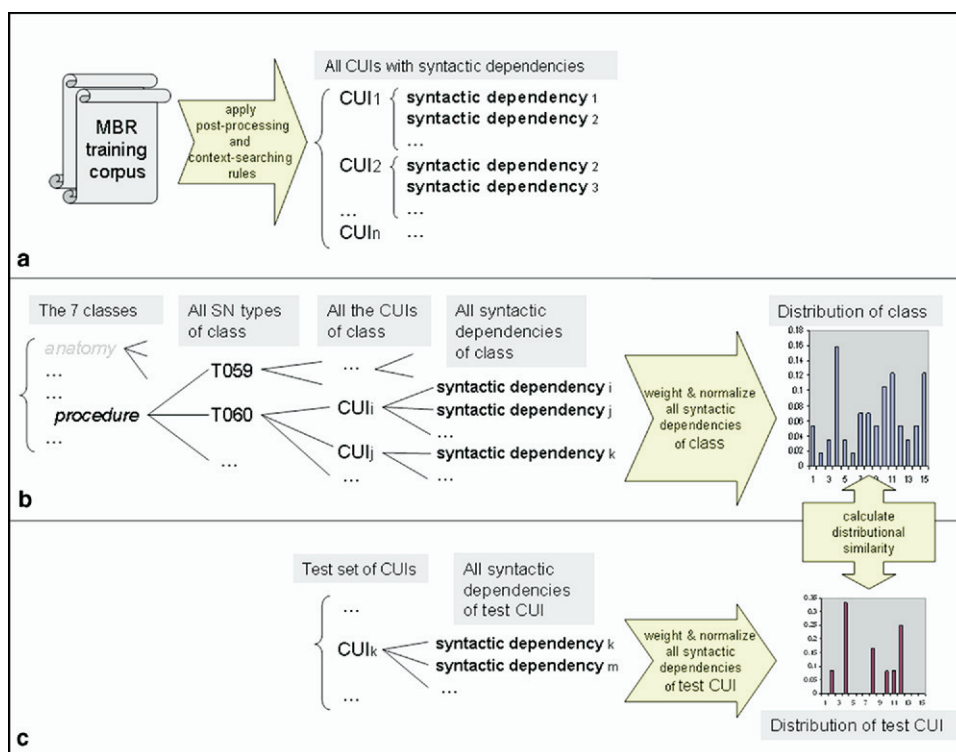


Figure 1. (a) Preparing the syntactic dependencies. We post-processed the MBR corpora and applied the context-searching rules to extract syntactic dependencies of the CUIs in the corpora. The extractions then provided the syntactic dependencies of the CUIs required for training and testing. (b) Constructing the training classes and their distributions. The *procedure* class is used as an example to illustrate the hierarchy of SN types and CUIs of the SN types included for training. All syntactic dependencies of each class were assigned weights and normalized to form the distribution profile. (c) Constructing the testing distributions. The syntactic dependencies of each test CUI were assigned weights and normalized to form a distribution for that CUI. The α -skew divergence was used to calculate the distributional similarities between each test CUI and the seven classes.

Preparing the Syntactic Dependencies

To prepare syntactic dependence features from the MetaMap-processed corpora, we created Perl scripts to post-process the MetaMap output. The post-processing involved extracting the part of speech (POS) tags, identifying the phrase types (e.g., noun phrase or prepositional phrase), and tracing the mapped concepts to their original extractions in text. For example, the machine-readable format (see Supplement 1, available as an online data supplement at www.jamia.org, for the complete output) was reconstructed as shown in Figure 2 for the sample sentence “The cryother-

apy prevented local tumor recurrence.” Then we used a set of context-searching rules to collect the syntactic dependencies for each concept so that for each concept all the rules were applied to the entire sentence containing that concept. Basically the rules were applied under disjoint conditions (i.e., the type of phrase the concept occurs in and the syntactic role the concept plays in the phrase). We extracted syntactic dependencies of those terms that were unambiguously mapped to only single CUIs, assuming they were more reliable than ambiguous mappings. For example, the syntactic dependencies in Figure 3b were extracted by the rules

S: The cryotherapy prevented local tumor recurrence.			
P: The cryotherapy NP			
W: The	determiner		
W: *cryotherapy	noun		
E: C0010412	Cold Therapy		Therapeutic or Preventive Procedure
P: prevented VP			
W: prevented	verb		
E: C0309873	PREVENT		Pharmacological Substance
E: C1292733	Prevents		Functional Concept
P: local tumor recurrence NP			
W: local	adjective		
W: tumor	noun		
W: *recurrence	noun		
E: C0205276	Local		Spatial Concept
E: C0027651	Neoplasms		Neoplastic Process
E: C0034897	Recurrence		Phenomenon or Process

Figure 2. A sentence with POS tags, phrase types, and CUI mappings identified from the MetaMap machine-readable format. Notations: S: sentence, P: phrase and phrase type, W: word and POS, asterisk marks the head noun, E: CUI, mapped concept, SN type.

1) If a term belongs to a noun phrase and it is the head noun (e.g. <i>cryotherapy</i>), then
a) consider the nouns within the noun phrase as adjunct nouns
b) consider the closest preceding verb as a passive verb, but stop when encountering any conjunction, noun phrase, complement, auxiliary, or any prepositional phrase
c) consider the closest subsequent verb (e.g. <i>prevented</i>) as an active verb, but stop when encountering any conjunction, noun phrase, complement, or auxiliary
2) If the term is not the head noun (e.g. <i>tumor</i>), consider the head noun within the noun phrase as the host noun (e.g. <i>recurrence</i>).
a
C0010412 “Cold Therapy”: active_verb(prevented)
C0027651 “Neoplasms”: host_noun(recurrence)
b

Figure 3. (a) The rules applied to extract the syntactic dependencies from Figure 2. (b) The syntactic dependencies extracted from Figure 2.

in Figure 3a based on the information shown in Figure 2. Because the MetaMap output provided richer shallow-parsed information than just POS tags, such as the phrase types obtained by the above process, we applied comprehensive syntactically based context-searching rules to the output. Readers can refer to Supplement 3 (available as an online data supplement at www.jamia.org) for the full set of our context-searching rules. An overview of the process is shown in Figure 1a.

Determining the Training Classes and Forming Their Distributions

Based on our experience with clinical applications and knowledge of the UMLS, we selected a clinically relevant subset (totaling 64) of the SN semantic types to form the seven broader semantic classes: *biologic function*, *anatomy* (above the molecular level), *disorder*, *gene protein*, *microorganism*, *procedure*, and *substance*. In addition to biomedical relevance, the reliability of the SN types and their homogeneity of granularity were both considered. For example, although many concepts of T033 “Finding” belong to the *disorder* class (e.g., “abnormal fluid balance” and “xanthopsia”), it also contains different types of concepts such as “postoperative state,” and therefore was not used for training. Each of the seven classes corresponded to multiple SN types (see Supplement 2, available as an online data supplement at www.jamia.org, for the composition of the classes), and we automatically identified the CUIs belonging to each particular class. As illustrated by the *procedure* class in Figure 1b, for each class we obtained a list of CUIs associated with that class, and then by looking up the CUI-syntactic dependencies obtained in Figure 1a, we retained for training only those CUIs having syntactic dependencies.

The syntactic dependencies associated with each broad class were pooled to form a distribution profile of the class, so that there were seven such training distributions (in Figure 1b we only show the *procedure* distribution but the process for the others was the same). We adopted the conditional probability function $P(\text{class}|\text{syntactic dependency})$ to compute the weights, and each conditional probability was estimated by $\text{frequency}(\text{class} \cap \text{syntactic dependency})/\text{frequency}(\text{syntactic dependency})$, according to the frequencies in the corpus. For each class, the weighted syntactic dependencies were then normalized to form a probability distribution, e.g., in Figure 1b, the X-axis of the *procedure* distribution represents the union of all the syntactic dependencies extracted from the corpus, with their probabilities summing to 1.

Determining the Test CUIs and Forming Their Distributions

We determined the CUIs in the initial test set that occurred in the two MetaMap-processed corpora and obtained 182 and 193 test CUIs with respect to the 100K and the 99K corpus. The 182 and 193 CUIs formed the test sets of this study and therefore were excluded from the training data. In addition, there were 157 overlapping CUIs between the 182 and 193 test sets, and we used the overlap to perform a controlled evaluation on the effects that were due to the differences of the training corpora. The gold standard of the classification was based on the 2006AA SN assignments, assuming that this was the correct version. For example, the SN type of C1442395 “Small T cell lymphoma” had been

changed from T031 “Body Substance” to T191 “Neoplastic Process” between the releases, thus in the gold standard the concept belonged to the *disorder* class (see Supplement 2, available as an online data supplement at www.jamia.org, for the SN type to broad class mapping).

The syntactic dependencies associated with each test CUI (e.g., that of CUI_k in Figure 1c) were pooled to form the testing distribution of that CUI, i.e., each test CUI was associated with its own distribution profile (in Figure 1c we only show the distribution for one test CUI). For the test CUIs, the weights of the syntactic dependencies were calculated using $P(\text{CUI}|\text{syntactic dependencies}) = \text{frequency}(\text{CUI} \cap \text{syntactic dependency})/\text{frequency}(\text{syntactic dependency})$, and they were normalized in the same way as the class distributions.

Computing the Distributional Similarities

Lee’s α -skew divergence was used to compute the distributional similarities. For example, we substitute the distributions of Figure 1b and 1c into formula 2 as: $S_\alpha(\text{procedure distribution}, \text{CUI}_k \text{ distribution})$. We set $\alpha = 0.99$ in the skew divergence formula, as suggested by Lee.⁵¹ To classify a test CUI, for its distribution the skew divergence was calculated against each distribution of the seven classes, and we assigned the CUI the class with the closest similarity score (lowest divergence value).

Evaluation Studies

1) Evaluation of the gold standard

In order to verify the quality of the gold standard, 50 CUIs in the gold standard set were randomly selected and shown to an expert with a biomedical background. The CUIs along with all their associated strings (with parenthesized annotations removed) from the MRCONSO.RRF table were provided to the expert. He was asked to select one or more of the seven classes that he considered to be most appropriate for each CUI. A disagreement was counted when the expert’s classification disagreed with the gold standard; an agreement was counted when the expert’s classification overlapped totally with the gold standard, and a partial agreement was counted when the two partially overlapped. For the C1442395 “Small T cell lymphoma” example above, if the expert considered it to be both an *anatomy* and a *disorder* class, then it was counted as a partial agreement with the gold standard, but if the expert considered it to be *disorder* only, it was considered as a complete agreement. After the first evaluation round, the expert was shown the partial agreements and disagreements, and was asked if he would change his mind on some of his previous decisions. Then a second round evaluation was made accordingly.

2) Evaluation of the classification method

The main outcome measure was error rate (see formula 3 below), which was calculated by comparing differences in classification between the automated method and the gold standard. We experimented with three main factors in the evaluation: 1) the number of syntactic dependencies was varied, 2) two different training corpora (i.e., the 100K and 99K) were used, and 3) two sets of error rates were calculated so that one was based on the best prediction only and the other was based on consideration of the top two predictions. When counting the correct/

Table 1 ■ Error Rates for the Distributional Similarity Method on Testing Sets with at Least 1, 5, and 10 Syntactic Dependencies

# of syntactic dependencies	By Top Prediction		Within Top 2 Predictions	
	100K	99K	100K	99K
≥ 1	0.398 (N = 182)	0.306 (N = 193)	0.212 (N = 182)	0.218 (N = 193)
≥ 5	0.370 (N = 119)	0.246 (N = 126)	0.176 (N = 119)	0.151 (N = 126)
≥ 10	0.337 (N = 86)	0.198 (N = 91)	0.116 (N = 86)	0.121 (N = 91)

incorrect classifications, the count for a single test CUI was divided by the number of classes it was associated with in the gold standard. For example, if in the gold standard a CUI belonged to both classes *gene protein* and *substance*, and in the top 2 predictions of the classifier only *substance* was obtained, it was counted as a 0.5 correct classification and a 0.5 misclassification. A tie was defined when the correct/incorrect predictions received equal similarity scores from the classifier. The error rates were calculated by the formula:

$$\frac{M + 0.5 \cdot T}{N} \quad (3)$$

where M is the number of misclassifications, T is the number of ties, and N is the total number of CUIs tested. The error rate can be understood as complementary to classification accuracy. We chose the misclassifications made by our optimal model (trained on the 99K corpus and having at least 10 syntactic dependencies) and summarized those for the misclassified classes. In addition, for each misclassified CUI we examined the ranking that was determined by our method to see where in the ranking the correct class occurred.

3) Estimation of the classification coverage

We estimated the potential number of CUIs that can be classified by our method. First we estimated the recall of MetaMap by counting the number of unique CUIs identified in the whole 2005 MBR corpus, as well as how many belonged to the seven classes (according to Supplement 2, available as an online data supplement at www.jamia.org). Then within our 199K training corpus (i.e., obtained by combining the 100k and 99k corpora), we calculated the number of unique CUIs identified by MetaMap and the number of these unique CUIs that have at least one syntactic dependency extracted by our context-searching rules, which provided estimation of the recall of our context-searching rules. Within the CUIs having at least one syntactic dependency, we created a histogram of the number of syntactic dependencies and normalized the frequencies, which provided estimation of the proportion of the CUIs with varying abundance of features, assuming the proportions remain consistent with the expansion of the training corpus. Through the cascade described above, we could approximate the number of CUIs that would be classified under varying accuracies.

Results

Expert Agreement with the Gold Standard

After the first round, the expert agreed completely with the gold standard classification for 40 of the 50 randomly sampled CUIs (80%). There were five disagreements and

five partial agreements (see Supplement 4, available as an online data supplement at www.jamia.org). After seeing the disagreements and partial agreements, the expert thought he would change three of his previous disagreements to agreements, resulting in 86% (43/50) complete agreement with the gold standard.

Error Rates of the Distributional Classification

The error rates calculated by formula 3 are presented in Table 1, where N stands for the size of the testing set. Table 1 shows that the error rates decrease as the number of syntactic dependencies increase. The lowest error rate for the top prediction was 0.198, achieved by the model trained on the 99K corpus and tested on CUIs with at least 10 syntactic dependencies. When the top two predictions were considered, the error rates were much lower. The lowest error rate for the top two predictions was 0.116, achieved by the model trained on the 100K and tested on CUIs with at least 10 syntactic dependencies. The model trained on the 99K corpus performed generally better than the 100K corpus, and the error rate for the 157 overlapping CUIs (between the 182 and 193 data sets shown in Table 1) was 0.392 and 0.296 for the 100K and 99K corpus, respectively. Note that this intersection set consisted of test CUIs with at least one syntactic dependency and was used to compare performance when using the different training corpora.

Summary of Misclassifications

We analyzed the 18 misclassifications made by the best model associated with the top prediction (error rate of 0.198 in Table 1). The misclassifications belonged to four classes of the gold standard: *anatomy* (4), *biologic function* (5), *disorder* (6), and *substance* (3). For example, the *anatomy* "Fibril" and "Desmosomes" were misclassified as *gene protein*. The *disorder* "Immune tolerance" and "Frameshift mutation function" were misclassified as *biologic function*. The *substance* "Products used to treat thrombocytopenic purpura" was misclassified as *disorder*. Please refer to Supplement 5 (available as an online data supplement at www.jamia.org) for detailed information about the misclassifications.

Estimation of the Classification Coverage

There were 300,431 unique CUIs unambiguously identified by MetaMap in the entire 2005 MBR corpus, which was about 22% of the total number of CUIs in the current UMLS, and 211,683 of the 300,431 ($211,683/300,431 = 70\%$) belonged to the well-defined SN types covered by our seven classes. In the 199K corpus, there were 42,121 CUIs identified by MetaMap, of which 36,961 (88%) had at least one syntactic dependency. The ratio of 88% was used as an estimate for CUIs with ≥ 1 syntactic dependencies, i.e., the recall of our context-searching rule. The reverse cumulative distribution of CUIs with up to ≥ 20 syntactic dependencies is provided in Figure 4, which shows that about 41% of the

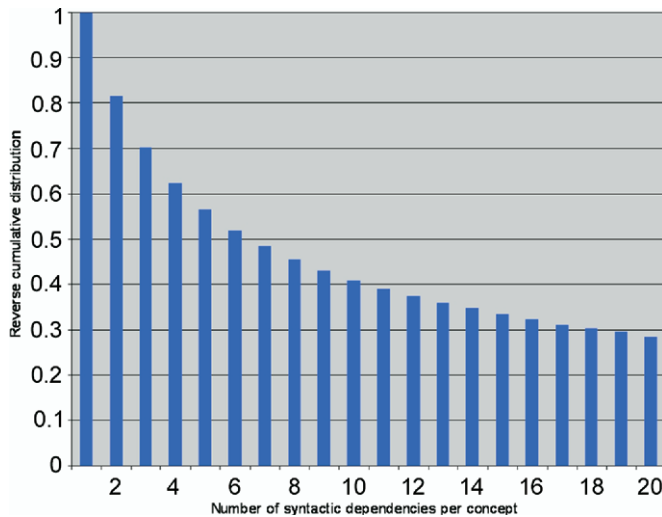


Figure 4. (Reverse cumulative) Distribution of CUIs with up to ≥ 20 syntactic dependencies per concept, estimated from the union of the 100K and 99K training corpus. For example, all CUIs extracted with ≥ 1 syntactic dependency have at least one syntactic dependency (the first bar), but only about 81% of these CUIs have at least two syntactic dependencies (the second bar).

CUIs with ≥ 1 syntactic dependencies had at least 10 syntactic dependencies. Therefore, we estimate that there would be about 76,375 CUIs (i.e., $211,683 \cdot 0.88 \cdot 0.41$) that belong to the seven classes and can be classified with the lowest error rate estimated in Table 1. However, the preceding estimation should be between 76,375 and 108,395 (by $300,431 \cdot 0.88 \cdot 0.41$), because our method can also be used to classify concepts in other less well-defined SN types such as “Findings” and “Functional Concept.”

Discussion

Our goal was to reclassify UMLS concepts into broader classes that would be more suitable for the needs of NLP applications, and therefore we did not intend to apply the method to classifying the concepts into fine-grained classes (e.g., the original 135 SN types). However, the differentiation by our method of *gene protein* from *substance* revealed a potential for coping with a finer granularity. The feasibility study based on the seven broad classes showed promising results, and more importantly, demonstrated how NLP and a knowledge base may reciprocally support each other. In the following subsections we discuss various issues related to our results and methods.

Expert Agreement with the Gold Standard

Taking into account the remaining five (10%) partial agreements with the 86% complete agreements obtained in the second round evaluation, the automatically generated gold standard could be considered dependable. Many disagreements showed that semantic classification for an expert is not trivial and is dependent on the viewpoint of the expert at the time of classification. In particular, it was more difficult to achieve agreement for concepts that could semantically belong to multiple or overlapping classes. For example, C0392525 “Nephrolithiasis” (i.e., kidney stone) was assigned *disorder*, *anatomy*, and *substance* by the expert, but in the gold standard it only belonged to *disorder*.

Similarly, C0012817 “Diverticulum” was an anatomical abnormality (*disorder*), but the expert considered it to be an *anatomy*. The above nontrivial disagreements suggested that the test sets of recent SN updates requiring corrections may have contained more ambiguous or difficult concepts to classify, which therefore could have resulted in an underestimation of the performance of our methods in general.

Another issue involves the information that was made available to the expert for the classification task. The expert was shown only concept strings in order to avoid having him see existing ontological knowledge encoded in the UMLS, since this could influence his judgment. This was also the reason why the parenthesized annotations were removed from the strings because they contained classifications. Although in this approach, the information available to the expert was more limited, the expert made his own judgments under a more objective condition. For example, the expert’s diverse assignments for “Nephrolithiasis” above implied his domain knowledge sufficed in spite of the limited information. In addition, had the expert been given more information such as the contextual relations with other concepts, his agreement with the gold standard would likely have been even higher. For example, the parent concept “Kidney Diseases” could have influenced him to classify “Nephrolithiasis” only as *disorder*, which would then result in complete agreement with the gold standard.

Issues about the Misclassifications by the Automated Method

Two *anatomy* concepts (“Fibril” and “Desmosomes”) that are cell components were misclassified as *gene protein*. We hypothesize that it was because the components are made of structural proteins, so that their frequent occurrences in contexts similar to other proteins confused our classifier. The two *disorder* concepts misclassified as *biologic function* were also not trivial for our classifier. According to the gold standard, “Immune tolerance” is a type of pathologic function, but our classifier failed to differentiate it from a normal biologic function. Possibly normal and pathologic functions have similar distributional properties. Similarly, as a type of molecular dysfunction, “Frameshift mutation function” was misclassified as a biologic function. An interesting case was the concept C0034153 “Products used to treat thrombocytopenic purpura,” misclassified as a *disorder*. In the UMLS release used by the 2005 MBR MetaMap, the CUI represented “Purpura, Thrombocytopenic” and was assigned T047 “Disease or Syndrome,” and therefore the syntactic dependencies extracted for the CUI also corresponded to the disorder sense. However, since 2005AC the same CUI has been changed to stand for the pharmacologic substance (assigned T121) used to treat the disorder. This example not only reveals some potential problem in the UMLS curation process but justifies the benefit of using an up-to-date MBR corpus in the future.

Issues about MetaMap and the MBR database

Although MetaMap had filtered out some strings (e.g., suppressible synonyms) that might introduce noise to the mapping process, and we reduced noise by using only single CUI mappings (i.e., non-ambiguous mappings) when searching the syntactic dependencies, the precision of MetaMap still could not be guaranteed. MetaMap does not support word sense disambiguation, and this could intro-

duce some noise in training and could decrease the recall (by discarding many useful contexts due to unresolved ambiguities). Aside from the ambiguity issue, MetaMap had other limitations associated with complexities related to concept mapping. For example, implicit references to previous information (e.g., in “family history of *the* disease,” *the* is referring to a previous disease mention) are not resolved and coordinating conjunctions (e.g., in “hereditary or sporadic retinoblastoma,” “hereditary retinoblastoma” will be missed) are not expanded,⁵² which can affect the recall of syntactic dependencies.

There were also issues in using the MBR 2005 database as the training data. First, the behavior of different MetaMap versions reflects the changing configurations of each UMLS release. For example, in one of the 2005 MBR training abstracts “posterior cricoarytenoideus” was correctly identified as C0448337, but all later occurrences of its abbreviation “PCA” were uniquely mapped to C0030625 “passive cutaneous anaphylaxis,” which contaminated the features of the *procedure* class with that of the *anatomy* class. By examining the error on the Web-based MetaMap version, we found it was due to the 2005 Knowledge Sources. Second, the old SN classification used in the previous MetaMap version was more error-prone (e.g., C0333343 “Body cavities” was incorrectly assigned to T190 “Anatomical Abnormality”), but this could be solved by synchronizing it with the up-to-date SN classification. Third, we found that in the MBR 2005 database the POS tagger used by the MetaMap at that time had a bug in processing all the “do”- and “have”-derived auxiliary verbs. We dealt with the known errors by discarding all sentences with the affected auxiliary verbs from our training corpora. Estimated from the 199K corpora, the discarded sentences constituted 10.5% (100,855/960,491) of all the sentences which were associated with CUIs. This also resulted in underestimation of the recall of our context-searching rules, which should be higher than the 88% reported in Results. We believe that the performance of our method will gain improvement from the future improved releases of the MBR databases and that MetaMap will continue to be a valuable resource. Fourth, using directly the pre-processed MBR corpus might not be optimal for building the distributional models because some potentially better parameters of MetaMap were not explored.

We estimated in Results that only 22% of UMLS concepts were identified in the entire 2005 MBR database. However, a recall of 22% underestimates the coverage of usable concepts with respect to text processing tasks. Many UMLS concepts never appear in text because they are too specific or artificial to occur in natural language (e.g., C0232298 “Left axis deviation greater than -90 degrees by EKG”). This phenomenon correlates well with the estimation by McCray et al.⁵³ that only about 10% of the UMLS strings could be found in MEDLINE. Therefore, many UMLS concepts are not applicable for text processing applications. In addition, many multi-term concepts are identified by MetaMap as separate constituent concepts. For example, although the term “Hypertensive spasm of cardiac sphincter” corresponding to C0232593 cannot be mapped fully, the component concepts C0857121 “Hypertensive,” C0037763 “Spasm,” and C0227192 “Cardiac sphincter” can be mapped appropriately.

Issues about the Distributional Similarity Method

In building the seven broad classes, we selected the SN types that we considered to be semantically representative of each class and to contain semantically homogeneous constituent CUIs. However, the classes may not have been optimally chosen. For example, C0852201 “Saliva” was included in our *anatomy* class through the SN type “Body Substance” (see Supplement 2, available as an online data supplement at www.jamia.org), but it semantically straddles the *anatomy* and *substance* classes, and might have introduced noise into our training. Such impurity in the classes posed a main source of training noise. Another layer of noise came from incorrect MetaMap mappings, which have affected the quality of syntactic dependencies, as discussed earlier.

We approximated syntactic dependencies from the shallow-parsed sentences by manually created rules, since it was an economical way to use the MBR corpus. Our approach was similar to Cimiano and Völker’s, but we used richer parse information obtained from MetaMap and applied more comprehensive context-searching rules. However, there were some errors caused by using an incomplete parse. For example, the rule for searching the active verb of a head noun, i.e., rule 1c of Figure 3a, would fail to differentiate verbs as the past tense or as the perfect participle. In the sentence “normal peripheral blood B cells **expressed** high levels of CDw75...” from a training abstract, the rule captured “expressed” correctly as the active verb of “B cells,” but incorrectly chose “expressed” as the active verb of “cell-surface antigen” in the sentence “. . . CDw75 is a sialylated cell-surface antigen **expressed** in a number of tissue-specific isoforms.” The wrong syntactic dependencies could have introduced some noise into the features of the class distributions as well as into some test CUI distributions. In this paper we have not explored variations of our rules, but in the future we will experiment with stricter rules and use of a larger corpus to address the potential sparseness issue.

The results related to the number of syntactic dependencies agreed with our general intuition that having more features increases the classification accuracy. In practice we will request a larger training corpus to achieve higher accuracy and to cover more concepts. However, this should not adversely affect the use of an automated method because although the corpus-based approach will require large training data, the classifiers are not subject to frequent changes once they are built. Moreover, by studying the number of syntactic dependencies (per concept) extracted from the training corpora (both had median of 13), the result that the 99K outperformed the 100K on the same 157 test CUIs implied that the quality of the corpus is more important than the quantity. Therefore, finding methods to obtain a more compact but qualified training corpus is another topic worth further study.

The error rates when considering the top two predictions were much lower than when considering the top prediction, which implied the ranking mechanism of the distributional classification worked reasonably. In addition, we observed that many of the CUIs that did not belong to the seven classes and were excluded from training could be classified properly. We also foresee the method can offer the flexibility to create different semantic perspectives by grouping the semantic classes differently, which enables generating dif-

ferent semantically-oriented versions of the ontology for different applications. These derived topics are left for our future work.

Limitations

The seven classes were manually determined and thus were not completely objective. The evaluation of the gold standard was based on the judgment of only one expert, using a moderate sized sample. The test sets from the UMLS updates might be biased and may be more difficult to classify than a random sample, but it is also possible that they could be easier to classify because they were selected from the more well-defined SN types. Due to the computationally demanding process of shuffling large training corpora, we have not performed multifold cross validation. The recall of our classification method is bound by the concepts available in the training corpus, and in this work by the concepts identified in the MBR database. We directly used the MetaMap-processed outputs of the MBR database, but those default parameters of MetaMap might not be optimal for building the distributional models. Our methods may be inadvertently taking advantage of some properties and associated resources peculiar to the UMLS infrastructure, and this may not be easily generalized to especially other ontologies beyond the UMLS coverage. For example, not all ontologies are associated with a concept-mapping and annotation program like MetaMap, or a huge corpus that is already processed like the MBR database.

Future Work

Future work will involve: 1) using multiple experts in evaluating the gold standard, 2) evaluating the method using random sets other than the UMLS updates, 3) using an up-to-date release of the MBR databases, 4) exploring different MetaMap parameters in preparing the training corpus because they may be better, 5) improving the classifier by using a larger training corpus, 6) refining the context-searching rules, 7) exploring other similarity measures, 8) classifying concepts that did not specifically belong to the seven classes but could be of potential use for NLP purposes (e.g., CUIs in T033 "Finding" and T169 "Functional Concept"), and 9) adding other semantic classes that are of clinical interest.

Conclusion

In this paper we propose the use of a distributional similarity approach to classify and/or validate the semantic classification of UMLS concepts for NLP applications. This was achieved by utilizing a corpus processed by MetaMap, which proved to be a valuable resource for this work. The initial results of classifying CUIs according to seven broad semantic classes showed that the method is indeed feasible, with an estimated lowest error rate of 0.198. The performance was even more promising when considering the correct classification to be covered by the top two predictions, where the estimated lowest error rate was 0.116. We believe that the method can be further improved by implementation refinements and that the performance currently would be effective as an advising system.

References ■

1. Craven M, Kumlien J. Constructing biological knowledge base by extracting information from text sources. *Proc Int Conf Intell Syst Mol Biol* 1999;77-86.
2. Friedman C, Kra P, Yu H, Krauthammer M, Rzhetsky A. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinform* 2001;17:S74-82.
3. Cohen AM, Hersh WR. A survey of current work in biomedical text mining. *Brief Bioinform* 2005;6(1):57-71.
4. Natarajan J, Berrar D, Hack CJ, Dubitzky W. Knowledge discovery in biology and biotechnology texts: a review of techniques, evaluation strategies, and applications. *Crit Rev Biotechnol* 2005;25(1/2):31-52.
5. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nature Genet* 2000;25:25-9.
6. Bairoch A, Apweiler R, Wu CH, et al. The Universal Protein Resource (UniProt). *Nucleic Acids Res* 2005;33:D154-9.
7. Eyer TA, Ducluzeau F, Sneddon TP, Povey S, Bruford EA, Lush MJ. The HUGO Gene Nomenclature Database, 2006 updates. *Nucleic Acids Res* 2006;34:D319-21.
8. IUPAC-IUBMB Joint Commission on Biochemical Nomenclature (JCBN). Available at: <http://www.chem.qmul.ac.uk/iupac/jcbn/>; accessed October 27, 2006.
9. Wheeler DL, Barrett T, Benson DA, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2000;28(1):10-4.
10. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. *Methods Inf Med* 1993;32(4):281-91.
11. Campbell KE, Oliver DE, Shortliffe EH. The Unified Medical Language System: toward a collaborative approach for solving terminologic problems. *J Am Med Inform Assoc* 1998;5(1):12-6.
12. Harris ZS. *A Theory of Language and Information: A Mathematical Approach*. New York: Oxford University Press; 1991.
13. Hirschman L, Grishman R, Sager N. Grammatically-based automatic word class formation. *Inf Process Manage* 1975;11:39-57.
14. Hindle D. Noun classification from predicate-argument structures. *Proc Annu Meet Assoc Comput Linguist* 1990;268-75.
15. Pereira F, Tishby N, Lee L. Distributional clustering of English words. *Proc Annu Meet Assoc Comput Linguist* 1993;183-90.
16. Lee H, Kim T. Language resource and rule construction for biological named entity system using UMLS. *Genome Inform* 2003;14:691-2.
17. Woods JW, Sneiderman CA, Hameed K, Ackerman MJ, Hatton C. Using UMLS metathesaurus concepts to describe medical images: dermatology vocabulary. *Comput Biol Med* 2006;36(1):89-100.
18. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *J Biomed Inform* 2003;36(6):462-77.
19. Rindflesch TC, Libbus B, Hristovski D, Aronson AR, Kilicoglu H. Semantic relations asserting the etiology of genetic diseases. *AMIA Annu Symp Proc* 2003;554-8.
20. Cimino JJ, Min H, Perl Y. Consistency across the hierarchies of the UMLS Semantic Network and Metathesaurus. *J Biomed Inform* 2003;36(6):450-61.
21. McCray AT. An upper level ontology for the biomedical domain. *Comp Funct Genom* 2003;4:80-4.
22. McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. *Medinfo* 2001;10(Pt 1):216-20.
23. Chen Z, Perl Y, Halper M, Geller J, Gu H. Partitioning the UMLS semantic network. *IEEE Trans Inf Technol Biomed* 2002;6(2):102-8.
24. Zhang L, Perl Y, Halper M, Geller J, Hripcsak G. A lexical metaschema for the UMLS semantic network. *Artif Intell Med* 2005;33(1):41-59.
25. Gu HH, Min H, Perl Y, Zhang L, Perl Y. Using the metaschema to audit UMLS classification errors. *Proc AMIA Symp* 2002;310-4.

26. Gu HH, Perl Y, Elhanan G, Min H, Zhang L, Perg Y. Auditing concept categorizations in the UMLS. *Artif Intell Med* 2004;31: 29–44.
27. Cimino JJ. Auditing the Unified Medical Language System with semantic methods. *J Am Med Inform Assoc* 1998;5(1):41–51.
28. Peng Y, Halper MH, Perl Y, Geller J. Auditing the UMLS for redundant classifications. *Proc AMIA Symp* 2002;612–6.
29. Schulze-Kremer S, Smith B, Kumar A. Revising the UMLS semantic network. *Medinfo* 2004;1700–4.
30. Resnik P. Using information content to evaluate semantic similarity in a taxonomy. *Proc Intl Joint Conf Artif Intell* 1995;v1: 448–53.
31. Leacock C, Chodorow M. Combining local context and WordNet similarity for word sense identification. In: Fellbaum C (ed). *WordNet: An Electronic Lexical Database*, pp. 265–83. Cambridge (MA): MIT Press; 1998.
32. Caviedes JE, Cimino JJ. Towards the development of a conceptual distance metric for the UMLS. *J Biomed Inform* 2004;37(2): 77–85.
33. Firth JR. A synopsis of linguistic theory, 1930–1955. In: *Studies in Linguistic Analysis, Special Volume of the Philological Society*, pp. 1–32. Oxford: Blackwell; 1957.
34. Harris ZS. *Mathematical Structures of Language*. New York: Wiley; 1968.
35. Karlgren J, Sahlgren M. From words to understanding. In: Uesaka Y, Kanerva P, Asoh H (eds). *Foundations of Real-World Intelligence*, pp. 294–308. Standord: CSLI Publications; 2001.
36. Wittgenstein L. *Philosophical investigations*. Translated by Anscombe GEM. Oxford: Blackwell; 1953.
37. Grefenstette G. *Explorations in Automatic Thesaurus Discovery*. Boston: Kluwer Academic; 1994.
38. Tokunaga T, Iwayama M, Tanaka H. Automatic thesaurus construction based-on grammatical relations. *Proc Intl Joint Conf Artif Intell* 1995;1308–13.
39. Lee L. Measures of distributional similarity. *Proc Annu Meet Assoc Comput Linguist* 1999;25–32.
40. Kullback S, Leibler RA. On information and sufficiency. *Ann Math Stat* 1951;22:79–86.
41. Cover TM, Thomas JA. *Elements of Information Theory*. New York: John Wiley & Sons; 1991.
42. Cimiano P, Völker J. Towards large-scale, open-domain and ontology-based named entity classification. *Proc Intl Conf Recent Adv Nat Lang Process* 2005;166–72.
43. Hatzivassiloglou V. Do we need linguistics when we have statistics? A comparative analysis of the contributions of linguistic cues to a statistical word grouping system. In: Klavans JL, Resnik P (edotors). *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, pp. 67–94, Cambridge(MA): MIT Press; 1996.
44. Resnik P. Semantic classes and syntactic ambiguity. *Proc ARPA Workshop Human Lang Technol* 1993;278–83.
45. Clark S, Weir D. A class-based probabilistic approach to structural disambiguation. *Proc Intl Conf Comput Linguist* 2000; 194–200.
46. Geffert M, Dagan I. The distributional inclusion hypotheses and lexical entailment. *Proc Annu Meet Assoc Comput Linguist* 2005;107–14.
47. Mohammad S, Hirst G. Distributional measures of concept-distance: a task-oriented evaluation. *Proc Conf Empir Methods Nat Lang Process* 2006;35–43.
48. Pedersen T, Pakhomov SV, Patwardhan S, Chute CG. Measures of semantic similarity and relatedness in the biomedical domain. *J Biomed Inform* (e-pub ahead of print). Jun 10, 2006. DOI 10.1016/j.jbi.2006.06.004. Available at: <http://www.sciencedirect.com/science/journal/15320464>; accessed September 16, 2006.
49. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp* 2001: 17–21.
50. MEDLINE/PubMed Baseline Repository (MBR). Available at: <http://mbr.nlm.nih.gov/>; accessed May 25, 2006.
51. Lee L. On the effectiveness of the skew divergence of statistical language analysis. *Proc 8th Int Workshop Artif Intell Stat* 2001;65–72.
52. Divita G, Tse T, Roth L. Failure analysis of MetaMap Transfer (MMTx). *Proc MedInfo* 2004;763–7.
53. McCray AT, Bondenreider O, Malley JD, Browne AC. Evaluating UMLS strings for natural language processing. *Proc AMIA Symp* 2001;448–52.