Research Paper ■

Automatic Resolution of Ambiguous Terms Based on Machine Learning and Conceptual Relations in the UMLS

HONGFANG LIU, MS, STEPHEN B. JOHNSON, PHD, CAROL FRIEDMAN, PHD

Abstract **Motivation**. The UMLS has been used in natural language processing applications such as information retrieval and information extraction systems. The mapping of free-text to UMLS concepts is important for these applications. To improve the mapping, we need a method to disambiguate terms that possess multiple UMLS concepts. In the general English domain, machine-learning techniques have been applied to sense-tagged corpora, in which senses (or concepts) of ambiguous terms have been annotated (mostly manually). Sense disambiguation classifiers are then derived to determine senses (or concepts) of those ambiguous terms automatically. However, manual annotation of a corpus is an expensive task. We propose an automatic method that constructs sense-tagged corpora for ambiguous terms in the UMLS using MEDLINE abstracts.

Methods. For a term *W* that represents multiple UMLS concepts, a collection of MEDLINE abstracts that contain *W* is extracted. For each abstract in the collection, occurrences of concepts that have relations with W as defined in the UMLS are automatically identified. A sense-tagged corpus, in which senses of *W* are annotated, is then derived based on those identified concepts. The method was evaluated on a set of 35 frequently occurring ambiguous biomedical abbreviations using a gold standard set that was automatically derived. The quality of the derived sense-tagged corpus was measured using precision and recall.

Results. The derived sense-tagged corpus had an overall precision of 92.9% and an overall recall of 47.4%. After removing rare senses and ignoring abbreviations with closely related senses, the overall precision was 96.8% and the overall recall was 50.6%.

Conclusions. UMLS conceptual relations and MEDLINE abstracts can be used to automatically acquire knowledge needed for resolving ambiguity when mapping free-text to UMLS concepts.

■ **J Am Med Inform Assoc.** 2002;9:621–636. DOI 10.1197/jamia.M1101.

With the widespread use of computers in the biomedical domain, a vast, rich range of biomedical data, including coded data, as well as free-text data has been stored in digital format. Computer applications can interpret coded data automatically while free-text data pose challenges to system developers. To enable access to free text in the biomedical domain, natural language processing (NLP) systems have been developed that facilitate information retrieval (IR), information extraction (IE), and text mining on free text.[1–3] However, all NLP systems require identification of terms (a term can be a single word or a multi-word phrase) in free text with entries in a lexical table.[3,4] Terms in free text can be ambiguous and may have multiple unrelated senses in the lexical table. For example, *capsule* in discharge summaries can mean a unit for a medication, such as in

Affiliations of the authors: City University of New York, New York, New York (HL); Columbia University, New York, New York (SBJ, CF).

This study was supported in part by grants LMO6274 from the NLM and ILS-9817434 from the NSF.

Correspondence and reprints: Hongfang Liu, MS, Department of Medical Informatics, Columbia University, VC-5, 622 W. 168 Street, New York, NY 10032; e-mail: <hongfang.liu@dmi.columbia.edu>.

Received for publication: 2/15/02; accepted for publication: 5/22/02.

"He was put on Dyazide one capsule daily over the past two days" or a body region, such as in "There may be faint lucency in the left internal capsule." It may have related senses, such as the chemical term *potassium*, which can mean a laboratory test item in "Her potassium had been as low as 2.7 on July 27" or a drug item in "Her discharge medications are digoxin five days a week and potassium supplements 10 mEq each week day." It can also be an abbreviation that has multiple full forms, or that has the same spelling as a general English word, such as *HR*, which denotes hour or heart rate, and SOB, which denotes short of breath as well as the general English word *sob*. Resolving the ambiguity of ambiguous terms is a case of word sense disambiguation (WSD). In the general English domain, some WSD systems resolve ambiguity caused by entries in machine-readable dictionaries that have multiple senses. Note that entries in machine-readable dictionaries can be phrases as well as words.

The need for resolving term ambiguity has been realized in information retrieval, information extraction, and text mining applications in the biomedical domain. Aronson[1] found that ambiguity resolution is important for improving the performance of MetaMap, a free text to UMLS concept mapping program. An information extraction system, MedLEE, which was originally developed for radiology reports, encountered ambiguity problems when broadened to a larger domain.[5] Nadkarni et al.[6] found that completely automated concept indexing in medical reports cannot be achieved without resolving ambiguity in free text. In an automatic knowledge discovery system, DAD-system, Weeber et al.[7] found that in order to replicate Swanson's literature-based discovery of the involvement of magnesium deficiency in migraine,[8] it was important to resolve the ambiguity of an ambiguous abbreviation *mg*, which denotes magnesium or milligram.

Several preliminary attempts to resolve term ambiguity in biomedical NLP applications utilized handcrafted rules. Rindflesch and Aronson[9] used a set of handcrafted rules based on semantic types of neighboring words to resolve ambiguity when mapping free text to UMLS concepts. The MedLEE system[10] applies rules based on local contextual information. However, it is expensive and difficult to write a comprehensive set of the necessary disambiguation rules. Furthermore, manual maintenance and further extension of rule sets become increasingly complex.

In the computational linguistics field, the disambiguation knowledge for WSD problem can be acquired automatically through three different sources[11]:

1. Knowledge-bases, usually a machine-readable dictionary, such as WordNet,[12] Longman Dictionary of Contemporary English,[13] Roget's Thesaurus.[14]
2. Sense-tagged corpora, usually manually assembled, such as the Semcor corpus,[12] or the DSO corpus.[15]
3. Raw corpora, a WSD method using raw corpora only, is usually referred as sense discrimination.[16]

Supervised machine learning techniques have been used to acquire disambiguation knowledge from sense-tagged corpora, where senses of ambiguous words have been manually tagged. Performance of the resulting systems was promising,[17–19] but manual sense-annotation of a corpus is an expensive task and inter-agreement among annotators is low.[20] Some researchers attempt to acquire disambiguation knowledge from large resources, such as machine-readable dictionaries with or without combination with large raw corpora.[21–23]

In a previous study,[24] we investigated an unsupervised sense disambiguation approach that consists of two phases: automatic acquisition of sense-tagged corpora from raw corpora using unambiguous synonyms in the UMLS (i.e., terms that are not ambiguous in the UMLS and hold one sense of the corresponding ambiguous term), and automatic construction of WSD classifiers from the acquired sense-tagged corpora using supervised machine learning techniques. The WSD classifiers could accurately determine senses of ambiguous terms in untagged instances. However, there are some limitations of using unambiguous synonyms to derive sense-tagged corpora, as described later. In this paper, we followed the two-phase approach while using conceptual relationships defined in UMLS[25] to acquire sense-tagged corpora. The method was evaluated on a set of ambiguous biomedical abbreviations. The remainder of the paper is organized as follows. The following section provides related work and background knowledge; the Methods section presents our methods; the Experiment section and the Results section describe an experiment and its results; and the last section contains a discussion as well as future directions of the current study.

Related Work and Background

We first introduce previous WSD research that applies conceptual relationships defined in a machine-readable dictionary in the computational linguistics domain, i.e., acquires disambiguation knowledge by using relations defined in a machine-readable dictionary. We then discuss the *one sense per discourse* hypothesis, which is an assumption of our method. Supervised machine-learning techniques are presented next. Finally, the background knowledge about the resources, the summary of our method and differences from related work are provided.

WSD Research Using Relations

Many machine-readable dictionaries contain a rich set of relationships linking senses. For example, all nouns in WordNet,[12] which is a handcrafted online lexicon, are organized into one conceptual network through IS-A relationships. For a term *W*, we define a term that has relations with a sense *S* of *W* in a conceptual network as a **conceptual relative** of *W* via an associated sense *S*. For example, the word *summer* is a conceptual relative of the word *spring* with an associated sense the *springtime sense of spring*, since *summer* and the *springtime sense of spring* are seasons of the year and the relation between them is a sibling relation (i.e., they share a common parent—the season of the year). There are two observations for WSD methods that use conceptual relatives. One is that conceptual relatives with certain relations tend to appear in similar contexts. For example, the *summer* and *springtime sense of spring* can appear in similar contexts, such as "Spring is my favorite season" and "Summer is my favorite season." The other observation is that correct senses for the words in a natural language expression will have closer sense relations (in a conceptual network) than incorrect combinations of senses. For instance, in "Spring is my favorite season," *the springtime sense of spring* has a IS-A relation with *the season of the year*, while other combination of senses (e.g., *spring as a fountain* and *season as sports season*) have weaker relationships.

There are two different WSD methods using conceptual relatives based on the above two observations. One method uses unambiguous conceptual relatives with certain relations, such as IS-A or synonymy, in a concept-oriented dictionary to derive sense-tagged corpora automatically for use with supervised WSD classifiers. Our previous work,[24] as described earlier, belongs to this category. For example, we can build a sense-tagged corpus for *CSF* by extracting all instances that contain the full forms *cerebral spinal fluid* or *colony-stimulating factor* and then replacing the full forms by the ambiguous term *CSF* along with the corresponding senses. A WSD classifier is then constructed automatically from the sense-tagged corpus by applying supervised machine learning techniques (the detailed information about supervised machine learning will be described later). An experiment demonstrated that classifiers trained on the derived sense-tagged corpora achieved an overall precision of about 97% for terms that had unambiguous synonyms found in the UMLS, with greater than 90% precision for each individual ambiguous term.

The other method consists of looking up the conceptual relatives of an ambiguous term *W* in the context. The method takes a number of conceptual relatives via a relation and a formula to measure the relatedness of conceptual relatives with each of the senses of *W* in a concept-oriented dictionary, and then uses them to determine the sense of *W* in the context. In the general English domain, researchers usually choose WordNet as the concept-oriented dictionary. Sussna[26] used several relation types (such as IS-A relation, synonymy) in WordNet and chose a measure that takes account of the shortest path, the number of edges associated with the same type leaving a node, the depth of a given edge in the overall tree, and a weight assignment for each relation type. Sussna evaluated his method on five documents from *Time* magazine and compared it to human experts using the same evidence, with a precision measure of 52.3%. Agirre and Rigau[23] proposed a method that used conceptual relatives via the relation IS-A and chose a measure which is sensitive to the following parameters: the length of the shortest path that connects the concepts involved, the depth in the hierarchy and the density of concepts in the hierarchy. Agirre and Rigau[27] evaluated their method on the noun portion of a document that contained 2,079 words. The overall performance was measured in terms of precision and recall, with 66.4% for precision and 58.8% for recall.

One Sense per Discourse

In their experiments with WSD, Gale, Church, and Yarwosky[28] observed a strong relationship between discourse and sense. They proposed a hypothesis: **one sense per discourse**. When a word occurs more than once in a discourse, all occurrences of that word will share the same meaning. They conducted an experiment using 9 ambiguous words and a total of 82 pairs of concordance lines for those words and showed that 94% occurrences of ambiguous words from the same discourse have the same meaning. However, Krovetz[29] reported that 94% is only true for coarse-grained distinctions (i.e., to distinguish between *bank* as a *bank of a river* or as a *financial bank*) among senses; it is only 67% using fine-grained distinction (i.e., *a financial bank* sense of *bank* will split into several senses such as a *depository financial institution, savings bank,* or *the funds held by a gambling house*) among senses when evaluated on two manually tagged corpora, Semcor[12] and DSO,[30] using WordNet as the sense inventory. However, Krovetz's[29] result was based on two manually sense-tagged corpora in which the agreement was 57%[20] for the overlapping part. The hypothesis has been used by Yarowsky[22] for WSD in the general English domain.

Supervised Machine Learning Techniques on Sense-tagged Corpora

Generally, supervised machine learning techniques[31] use annotated data to make a decision for un-annotated data. It is widely used to learn classifiers for decision support systems in the biomedical domain. For example, given a set of medical reports each describing a pregnancy and a birth using 200 features (e.g., patient's *weight, height*), machine-learning techniques can be used to learn a classifier to categorize patients with high risk of emergency cesarean section. In order to use supervised machine-learning technique, the first step requires transforming each annotated instance to a feature representation, usually a feature set $fv = \{(f_1, v_1), (f_2, v_2), \ldots (f_n, v_n)\}$, where f_i is a feature and v_i is its corresponding value. For example, each medical report is transformed to a set with 200 elements in the above example. After transforming each instance to a feature representation, a supervised learning algorithm can be used to build classifiers, i.e., to categorize an un-annotated instance to a fixed number of categories. A large number of learning algorithms have been proposed in the literature including Bayesian probabilistic algorithms, artificial neural networks, decision trees, support vector machines, and instance-based algorithms.[31]

When supervised machine-learning techniques are used to resolve the ambiguity of a term *W*, we first need to transform each sense-annotated instance of *W* to a feature vector based on the context of *W*, which is formed by terms surrounding *W*. An important issue in using supervised learning for generating a WSD classifier is the appropriate choice of features. Intuitively, a good feature should capture sources of knowledge critical for determining the sense of *W*. Various kinds of features representing different knowledge sources have been presented in WSD research.[17] Features can be the surrounding terms of W in a fixed window size and their values can be Boolean values to indicate their existence in a given instance. For instance in **Sentence I** below, the features with nonzero values in the corresponding feature vector when considering the ambiguous word *CSF* and a window size of 3, are *the, origin, of, and, the, processes*. Features can also be derived from the surrounding terms of *W* in a fixed window size in the universe through different mappings, such as the orientation and/or distance of the surrounding terms (e.g., the features for **Sentence I** when considering the orientation and distance in a window size 3 are *the/L3, origin/L2, of/L1, and/R1, the/R2, processes/R3*, where *L* is for left, *R* is for right and the number is for the distance), local collocations (i.e., a short sequence of words near *W* taking the word order into account), utilization of further linguistic knowledge such as part of speech (POS) tags (i.e., verb or noun) or stemming techniques (*discharge, discharged, discharging*, and *discharges* are treated as the same feature *discharg*), or domain knowledge, such as the semantic categories in a semantic lexicon.

Sentence 1. *After a brief summary of current views on the origin of CSF and the processes underlying its elaboration, the author discusses studies of isolated chorid plexus in extracorporeal perfusion.*

Several supervised learning algorithms have been implemented in the general English domain including Naïve Bayes algorithm, decision lists, neural network, and instance-based learning. We used Naive Bayes algorithm. For other supervised machine learning algorithm, readers can refer to a survey paper of Marquez.[32] A Bayes classifier applies the Bayes decision rule when choosing a sense *S* from a set of senses $\{S1, S2, \ldots, Sm\}$ given a feature set *fv*, the rule that minimizes the probability of error[33]:

Equation 1. Decide S if $P(S \mid fv) > P(Sk \mid fv)$ for all $k = 1, \ldots, m$ that $Sk \neq S$,

The Bayes decision rule is optimal because it minimizes the probability of error. For each individual case, it chooses the sense with the highest conditional probability and hence the smallest error rate. The conditional probability $P(SK \mid fv)$ is computed using Bayes' Theorem:

Equation 2. $$P(Sk \mid fv) = \frac{P(fv \mid Sk)\, P(Sk)}{P(fv)}$$

$P(Sk)$ is the prior probability of sense *Sk*: the probability that we have an instance of *Sk* if we do not have any knowledge about the context. $P(fv \mid Sk)$ is the likelihood probability of fv given Sk, and $P(fv)$ is the prior probability of the feature set fv; fv usually can be eliminated (since it is a constant for all senses and hence does not influence what the maximum is). By applying the logarithm on the probabilities, **Equation 1** is equivalent to the following equation:

Equation 3. $S = \underset{Sk}{\text{argmax}}(\log P(fv \mid Sk) + \log P(Sk))$

$P(fv \mid Sk)$ is usually estimated using the Naïve Bayes assumption, i.e., features are conditionally independent of each other:

Equation 4. $P(fv \mid Sk) = P(\{(f_j, v_j) \mid j = 1, \ldots, n\} \mid s_k)$

$$= \prod_{j=1}^{n} P((f_j, v_j) \mid s_k)$$

Using the Naive Bayes assumption, we get the following modified decision rule:

Equation 5.

$$S = \underset{Sk}{\text{argmax}}[\log P(Sk) + \sum_{(f_j, v_j) \in fv} \log P((f_j, v_j) \mid Sk)]$$

$P((f_j, v_j), Sk)$ and $P(Sk)$ are computed via maximum likelihood estimation from the sense-tagged corpus:

Equation 6.

$$P((f_j, v_j) \mid Sk) = \frac{\text{occu}((f_j, v_j), Sk)}{\text{occu}(Sk)},$$

where $\text{occu}((f_j, v_j), Sk)$ is the number of (f_j, v_j) co-occurring with sense Sk in the sense-tagged corpus, $\text{occu}(Sk)$ is the number of occurrences of Sk in the sense-tagged corpus, and $\text{occu}(W)$ is the total number of occurrence of W.

The power of Naïve Bayes learning is due to its efficiency and its ability to combine evidence from a large number of features that are derived from a large number of instances.

Resources

Below we present information about several resources on which our method and evaluation are based.

MEDLINE [34] is the NLM bibliographic database that contains over 11 million references to journal articles in life sciences with a concentration on biomedicine. Each entry contains the citation information for the corresponding journal article, and also often contains an abstract.

The UMLS integrates various terminologies pertaining to biomedicine.[35] The Metathesaurus (META) is one component of the UMLS that contains information about biomedical concepts and terms from many controlled terminologies. The META is organized by concept, where each distinct concept has been assigned a unique concept identifier (CUI). All concept names corresponding to the same concept are assigned the same CUI. For instance, *congestive heart failure* and *biventricular heart failure* are two different concept names with the same CUI (C0018802). Each concept name has a term status to indicate whether it is the preferred concept name of the corresponding concept, or if it is suppressed (i.e., abbreviated or problematic). In the 2001 version of the UMLS, the table MRCON lists all concept names with their corresponding CUIs. It has 797,359 English concepts and 1,462,202 different English concept names (in the following, the 2001 version of the UMLS is assumed).

Another table called MRREL lists relationships between UMLS concepts. There are 9,524,132 entries in MRREL. Among them, 9,518,798 were derived directly from the source vocabularies. The remaining 5,334 entries are relationships between different sources that were created during the construction of the UMLS. There are 9 different relationship types, including broader (RB), narrower (RN), other-related (RO), parent (PAR), child (CHD), sibling (SIB), similar (RL), qualifier (AQ), and be-qualified (QB). However, since relations in the UMLS were mostly derived from different source vocabularies, the definition of relationship types may not be consistent. Two concepts may have multiple relationship types defined in the MRREL table. For example, the concepts C0004015 (i.e., *aspartic acid*) and C0085845 (i.e., *aspartate*) have a parent relation and a broader relation from source vocabulary AOD99; they have a narrower relation from source vocabulary MSH2001; in source vocabulary LNC10o, they have an other-related relation. A concept may have a relation with itself. For example, the concept C0022709 *angiotensin = converting enzyme* has RO relation with itself in source vocabularies CSP2000 and LNC10o.

The SPECIALIST Lexicon, an English language lexicon, is another component of the UMLS. The lexicon consists of a set of lexical entries with one entry for each spelling or set of spelling variants with a particular part of speech. The table LRAGR lists all variant forms for each entry in the lexicon. The UMLS also contains a Semantic Network where each CUI in META has been assigned to one or multiple semantic categories.

There are two different kinds of ambiguities presenting in the UMLS: conceptual and semantic. Conceptual ambiguity refers to the ambiguity caused by terms with multiple concepts, whereas semantic ambiguity refers to the ambiguity caused by terms that have multiple semantic categories. For example, concepts that are *organic chemicals* most likely are also *pharmacologic substances*.

There are 187,943 (of 797,359) concepts possessing multiple semantic categories in the UMLS. There were 4,547 conceptually ambiguous terms that represented 11,178 concepts in the UMLS ambiguous term table AMBIG.SUI, with an average ambiguity of 2.46. Johnson[4] investigated the semantic ambiguity of a semantic lexicon that was based on the UMLS and discharge summaries and proposed a set of preference rules to reduce the semantic ambiguity. For example, in the discharge summary domain, chemical concepts occur only under the semantic category *chemicals viewed functionally* instead of under *chemical viewed structurally*. After applying his preference rules to the derived semantic lexicon, occurrences of entries with multiple semantic types were reduced from 9.41% to 1.46% in discharge summaries. Rindflesch and Aronson[9] considered the conceptual ambiguity of the UMLS and proposed to use semantic categories of neighboring concepts to resolve the ambiguity. They conducted a preliminary study and found that a manually crafted set of rules based on semantic categories of neighboring concepts successfully resolved conceptual ambiguity around 80% of the time. Aronson and colleagues[37] proposed that machine-learning techniques could be used to derive rules instead of the manual crafting process. However, there is no published study of this approach according to our knowledge.

Below we introduce two programs that are discussed in the Methods section and the Experiment section: MetaMap[1] and the UMLS abbreviation extraction program.[38]

MetaMap[1] is a highly configurable program that maps biomedical text to concepts in the META. Options control MetaMap's output as well as its internal behavior, such as how aggressive to be in generation of word variants, whether or not to ignore META strings containing very common words, and whether to consider or to ignore word order. The initial purpose of the MetaMap program was to improve retrieval of bibliographic material such as MEDLINE citations. It has also been applied to several data mining efforts such as to detect clinical findings,[39] molecular binding expressions,[40] or novel relationships between drugs and diseases.[7]

The **UMLS abbreviation extraction program**[38] was developed for extracting (abbreviation, full form) pairs from the UMLS. It utilizes several fixed patterns associated with abbreviations in the META as well as the fact that abbreviations are considered synonyms in the META. The program was executed using the 2000 version of the UMLS. It extracted 163,666 different (abbreviation, full form) pairs with a recall of 96% and precision of 97.5%. The UMLS abbreviations were highly ambiguous: 54% of abbreviations with three characters had multiple full forms; the number of different full forms for all abbreviations with three characters was 3.05, while it was 10.9 for abbreviations with two characters. The UMLS abbreviations covered 66% of unique full forms found in the medical reports, and for frequently occurring abbreviations, the coverage was around 80%.

Our Method and Its Difference from Related Work

The method proposed in this article considers conceptual ambiguity in the UMLS. It utilizes conceptual relations defined in the UMLS to automatically derive sense-tagged corpora for ambiguous terms. WSD classifiers can then be automatically constructed using the derived sense-tagged corpora.

Our methods differ from related work in several ways. The previous investigations for resolving UMLS ambiguity were based on manually crafted rules. Johnson's method concentrated on reducing semantic ambiguity in a semantic lexicon; it was based on a particular sub-domain and manually crafted rules, but provided no solution for resolving ambiguity in the context. The method suggested by Aronson and colleagues,[37] which proposes to apply machine learning techniques to derive a set of WSD rules based on semantic categories of neighboring concepts, requires an annotated corpus, where semantic categories of neighboring concepts are annotated and senses of the corresponding ambiguous terms are also annotated.

The previous WSD work[23,26,27] that utilized conceptual relatives disambiguated all nouns in general English text; the conceptually oriented dictionary used was WordNet. Our method concentrates on automatic construction of a sense-tagged corpus for each individual ambiguous biomedical term in the biomedical domain. Instead of using WordNet, we use the UMLS as our conceptually oriented knowledge source. WordNet and the UMLS are different in the following ways. The goal of WordNet is to provide a database of lexical relations for computational linguistic research, and the power of WordNet lies in a manual handcrafting process and domain-independence. The goal of the UMLS is to overcome retrieval problems caused by differences in terminology by integrating different electronic biomedical terminologies into one concept-oriented knowledge base. Almost all relationships are directly mapped from original source terminologies. WordNet has a strict definition about its relationship types such as IS-A etc. However, in the UMLS, the definition of relationship types in MRREL is vague, and different terminologies have their own definition for the same relationship type. For example, a parent relation may be considered as a child relation in two different sources, causing circular hierarchical relationships.[41]

Our previous work,[24] which used unambiguous synonyms to derive a sense-tagged corpus, is limited because it can only be used to build a WSD classifier for an ambiguous term *W* under the following two assumptions: each sense of *W* has unambiguous synonyms; and there are enough sense-tagged instances for each sense. However, some senses may not have unambiguous synonyms, or may not occur in enough instances. The current method does not require that each sense have unambiguous synonyms. It utilizes conceptual relatives that occur in the context to determine the sense and then to derive a sense-tagged corpus. Our previous method and the current method can be combined to increase our ability to acquire sense-tagged corpora automatically, and then used to train supervised WSD classifiers.

Methods

Let *W* be an ambiguous word and let the set SEN(*W*) = {S_1, S_2, ..., S_m} be its m senses. The method is based on the observation that terms with related senses of *Si* tend to co-occur with the sense *Si* rather than other senses of *W*. We assume that multiple occurrences of *W* hold the same sense in the MEDLINE abstract, i.e., **one sense per abstract**. The context for acquiring disambiguation knowledge in this paper is the whole abstract.

Let CUI_{Si} be the concept identifier that represents the sense *Si*. We denote the concept identifier sense representation set {CUI_{S1}, CUI_{S2}, ..., CUI_{Sm}} as SCUI(*W*). For example, the SCUI(*W*) for the following four senses of the abbreviation *CSF* is {C0007806, C0009392, C0072454, C0893357}:

- CSF_1: cerebrospinal fluid (C0007806),
- CSF_2: colony stimulating factor (C0009392),
- CSF_3: cytostats factor (C0072454),
- CSF_4: competence and sporulation factor (C0893357).

Figure 1 illustrates the construction process of a sense-tagged corpus of W as well as the process of constructing a WSD classifier. The conceptual relatives of *W* are acquired through the MRREL table. A collection of MEDLINE abstracts that contain *W*, denoted as CMA(*W*), is extracted. For each abstract in the collection, occurrences of conceptual relatives of *W* from the collection are automatically identified. The sense-tagged corpus of *W* is derived using conceptual relatives identified in the context. A supervised WSD classifier is then constructed using the derived sense-tagged corpus. In the following, we first introduce the method for establishing conceptual relatives for *W* from the UMLS. We then discuss the method, CRMap (for Conceptual Relatives Mapping program), which maps conceptual relatives in abstracts. The construction of the sense-tagged corpus is presented next.

Establishing Conceptual Relatives for Each Sense

For each sense *Si* of *W*, we use concepts that have a direct relation with *Si* (i.e., concepts with CUIs that co-occur with CUI_{Si} in the MRREL table) to derive conceptual relatives. We consider concepts with relation types that are conceptual such as “Broader”, “Narrower,” and “Parent.” We exclude concepts with qualifier relation types (i.e., “Qualifier” or “Be-Qualified”) since they have high frequency, and provide little sense disambiguation information. Each *CUI* is added to the relative CUI set of *W* (RCUI(*W*)) with its associated sense *Si* and the relations. We consider that each concept has a synonymy relation with itself, and add each CUI_{Si} of W in the relative CUI set of W with its associated sense *Si* and a relationship type synonymy, but disregard relations among different senses of *W* in the MRREL table. For example, C0020255 (i.e., hydrocephalus) and C0007806 (i.e. CUI_{S1} of CSF) have an “Other” relation in MRREL; therefore C0020255 is added to the relative CUI set of *CSF* with its associated sense CSF_1.

For each CUI in the relative CUI set of *W*, we gather all unambiguous English concept names. Because concept names with a term status “suppressed” are problematic, we exclude them; in addition, because abbreviations are highly ambiguous, we exclude those

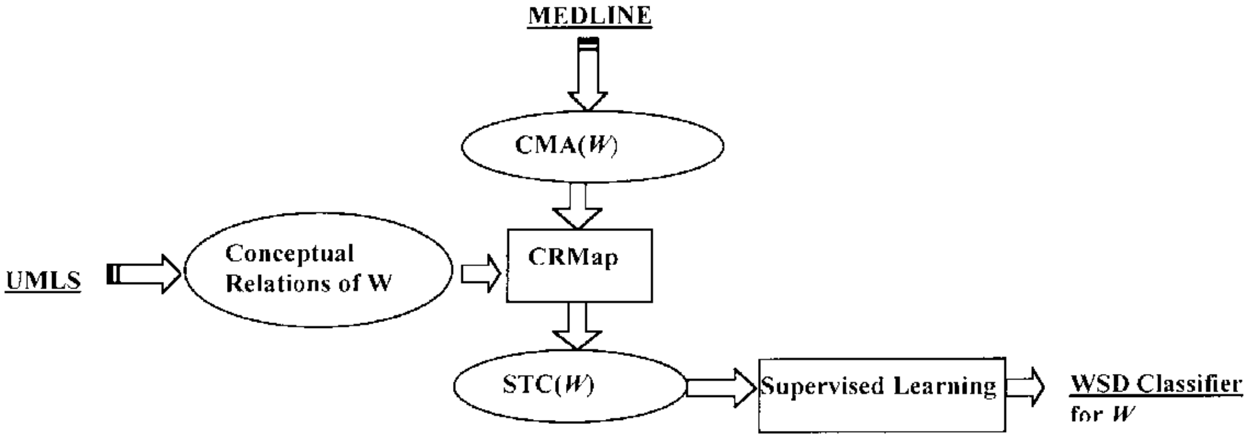

Figure 1 The process of constructing a sense-tagged corpus (STC(*W*)) for a specified word W from MEDLINE abstracts based on UMLS conceptual relations. CMA(*W*) consists of MEDLINE abstracts that contain W. The conceptual relations of *W* are acquired through the UMLS. Conceptual relatives of *W* (i.e., terms have conceptual relations with *W*) in each abstract are automatically identified using a program we developed called CRMap. The sense-tagged corpus STC(*W*) is derived using the majority vote of associated senses of identified conceptual relatives. Supervised learning is then used to automatically construct a WSD classifier for *W*.

identified as abbreviations by the UMLS abbreviation extraction program. Each concept name is normalized by changing it to lower-case, removing symbols such as *nos* in *cerebrospinal fluid, nos* or *unspecified* in *hydrocephalus, unspecified*, removing some patterns such as parenthetical expressions (*ck*) in *creatine kinase* (*ck*), or *ck* – in *ck* – *creatine kinase*, and substituting punctuations by blanks. The resulting strings with length greater than 4 are considered as conceptual relatives of W with corresponding associated senses. Note that strings with length less than or equal to 4 are excluded because they usually have a high level of ambiguity. In addition, they may be problematic because many of them are abbreviations in the context but are not identified as abbreviations by the UMLS abbreviation extraction program. For example, *FOX* is not an abbreviation in the UMLS; but in the text *Lipid hydroperoxides were measured by the FOX assay*, it abbreviates *ferrous oxidation with xylenol orange*.

Identifying Conceptual Relatives in Each Abstract

CRMap identifies conceptual relatives in an abstract using the following phases: preprocessing, exact-string matching, UMLS-SPECIALIST normalization matching, and stem normalization matching.

In the preprocessing phase, we remove parenthetical expressions that contain a capitalized term with fewer than six characters. This is based on the observation that parenthetical expressions containing a short capitalized term inside are usually abbreviation-type parenthetical expressions. The punctuation is replaced by blank and the text is changed to lower case. As an example, the text *The influence of prednisone on S-angiotensin-converting enzyme (S-ACE) activity was examined* is changed to *the influence of prednisone on s angiotensin converting enzyme activity was examined*.

The three matching phases are processed subsequently. All three matching phases match conceptual relatives of the longest possible length. The matching phases differ in whether they require normalization or not, and if so, the normalization method used. In the exact-matching phase, conceptual relatives are used without normalization, while in the UMLS-SPECIALIST normalization matching phase, CRMap normalizes each word in the conceptual relative set and abstracts and maps it to its base-form in accordance with the SPECIALIST Lexicon LRAGR table, if applicable. In the stem-normalization matching phase, CRMap uses the Porter-stemmer[42] to normalize each word to its stem.

For example, in **Abstract 1**, which contains the abbreviation *CSF*, three conceptual relatives are identified, each associated with the sense *CSF*$_1$ (i.e., *cerebrospinal fluid*): *hydrocephalus*, *spinal cord*, and *brain*).

> **Abstract 1**. *The* ***brain*** *(CSF$_1$_SIB) from an infant with a cystic occipital mass present at birth is examined in serial section. The occipital mass proved to be a rhombic roof ventriculocele. Within the posterior fossa, it was bound to an occipital lobe encephalocele which issued as a diverticulum of the left lateral ventricle through a microgyric cortical defect in the territory*

> *of the left posterior cerebral artery. The posterior medial aspects of both cerebral hemispheres were herniated downward into the widened tentorial gap. Craniolacunae were prominent on the inner aspect of the skull. The aqueduct and central canal of the* ***spinal cord*** *(CSF_1_SIB) were widely dilated, although the lateral ventricles were collapsed. It is suggested that* ***hydrocephalus*** *(CSF_1_RO) secondary to obstruction to flow of CSF through the rhombic roof entrained a sequence of events giving rise to the rhombic roof ventriculocele and causing occlusion of the posterior cerebral artery and subsequent diverticulation of the lateral ventricle through an infarcted region of the posterior-medial hemisphere.*

Besides CRMap, the MetaMap program can also be used to identify concepts that have relations with senses of *W*.

Deriving a Sense-Tagged Corpus

For an abstract that contains *W*, since all occurrences of *W* in the abstract hold the same sense *S* based on the one sense per abstract assumption, we call *S* the sense of *W* in that abstract. Note that not every abstract contains conceptual relatives that can be identified using CRMap. For an abstract that has conceptual relatives identified by CRMap, if all conceptual relatives are associated with one sense, the sense of W in the abstract is assigned with that sense without any doubt. For example, the sense of *CSF* in **Abstract 1** is assigned to CSF_1 (i.e., cerebrospinal fluid). For abstracts that contain conceptual relatives with multiple associated senses, a sense assignment scheme is needed in order to assign senses to those abstracts. For simplicity, we use the following sense assignment scheme while leaving other possible sense assignment schemes to discussion: assign the majority vote of the associated senses of the identified conceptual relatives; if there is a tie, randomly choose one of the tied senses.

The collection of abstracts in which the sense of *W* can be determined using conceptual relatives is the resulting sense-tagged corpus for *W* (STC(*W*)). More sense-tagged abstracts can be derived by constructing a supervised WSD classifier of *W* that is trained on the derived sense-tagged corpus and then by performing the disambiguation task on the whole abstract collection of *W*. A large sense-tagged corpus for *W* is formed by combining instances in STC(*W*) and instances that are sense-tagged by the constructed WSD classifier of *W*.

Experiment

The proposed method can be applied to almost all ambiguous terms in the UMLS. There are 1,262,668 (CUI_1, CUI_2) unique relation pairs defined in MRREL, where CUI_1 is one of the 11,178 concepts that contain an ambiguous concept name. An average of 113 concepts have relations with each CUI_1. There are only 6 (out of 4,547) ambiguous terms, where one concept has no relations defined in MRREL.

We performed a study to evaluate the above method using a set of ambiguous abbreviations. There were several reasons for using ambiguous abbreviations to evaluate the method:

- From the study of UMLS abbreviations, we have frequency information for abbreviations in a collection of one-year' worth of medical reports in the following domains: discharge summary, radiology, neurophysiology, pathology, GI endoscopy, ob/gyn, cardiology, and surgery.
- Abbreviations paired with their full forms in a parenthetical expression provide an automatic way to annotate the senses of abbreviations, and therefore provide a gold standard for the study.
- The same abbreviation throughout one abstract corresponds to the same full form (i.e., one sense per abstract for abbreviations, even though one sense per abstract may need to be measured for other terms).
- Abbreviations are highly ambiguous compared with other terms.

In the following, we discuss the derivation of ambiguous abbreviations, and the derivation of the gold standard. The evaluation method is presented next.

Derivation of Ambiguous Abbreviations

The UMLS (2001 version) was processed using the UMLS extraction program to obtain a list of (abbreviation, full form) pairs. We chose three-letter abbreviations since they were moderately ambiguous in the UMLS, and appeared frequently in writing. We kept only those pairs where the abbreviation was listed as an ambiguous term in AMBIG.SUI, had multiple full forms, appeared more than 100 times in the collection of medical reports, and full form was a UMLS concept name.

Derivation of the Gold Standard

For each abbreviation *A*, conceptual relatives for each sense of *A* were then established using MRREL. We extracted a collection of MEDLINE abstracts that contained *A* in upper case. The collection was divided into two sets I and II : Set I consisted of abstracts with an occurrence of *A* inside a parenthetical expression, and Set II consisted of all others. The gold standard sense of *A* for each abstract in Set I was then automatically derived using synonyms of *A* (i.e., conceptual relatives with concept identifiers from the sense representation set of *A*).

For each abstract in Set I, if a synonym (SYN) of *A* occurred in the abstract with a pattern SYN (*A*), we considered the gold standard for the correct sense of *A* in the abstract to be the associated sense of SYN. After determining the correct sense of A, the abstract was automatically modified by replacing the pattern SYN (*A*) with *A*, and then added to the gold standard set of *A* (GSS(*A*)). For instance, **Abstract 2** was modified to **Abstract 2′** with the gold standard sense attached at the beginning (separated using the sign "|"). Only modified abstracts were used for further processing. The sense at the beginning of each abstract was used for evaluation purposes to determine correctness, but not used by the disambiguation method itself.

Abstract 2. *After a brief summary of current views on the origin of cerebrospinal fluid (CSF) and the processes underlying its elaboration, the author discusses studies of isolated chorid plexus in extracorporeal perfusion. . . .*

Abstract 2′. *CSF*$_1$ | *After a brief summary of current views on the origin of CSF and the processes underlying its elaboration, the author discusses studies of isolated chorid plexus in extracorporeal perfusion. . . .*

Evaluation

We identified conceptual relatives in each abstract from the MEDLINE abstracts collection of *A* using CRMap and subsequently derived a sense-tagged corpus for *A* (STC(*A*)). Additional sense-tagged abstracts were derived by constructing a supervised WSD classifier that was trained on STC(*A*) and then performing the disambiguation task on the entire abstract collection of *A*. A large sense-tagged corpus for *A* was formed by combining instances in STC(*A*) and instances that were sense-tagged by the WSD classifier of *A*.

We evaluated STC(*A*) using the gold standard set of *A* (i.e., GSS(*A*)) with two measures: recall, i.e. the ratio of the number of abstracts with correctly identified sense to the total number of abstracts in GSS(*A*), and precision, the ratio of the number of abstracts with correctly identified sense to the number of abstracts that were sense-tagged using conceptual relatives.

We constructed a WSD classifier for *A* by applying the Naive Bayes algorithm on the derived sense-tagged corpus for *A*. We transformed each abstract in the corpus to a feature set where features were words after stemming (*A* was excluded for consideration as a feature, and the Porter Stemmer[42] was used) and values were Boolean values that indicate their existence in a context. For example, features with the value 1 in the feature set for **Abstract 2′** are *after, a, brief, summari ,..., in, extracorpor*, and *perfus*. Note that in the Naive Bayes algorithm, the inclusion of stop words (i.e., words that occur frequently disregarding different instances and terms) as features does not affect the decision process, because they contribute almost the same conditional probability to all senses.

The constructed WSD classifier of *A* was executed to assign senses for the MEDLINE abstracts that contained A but were not in STC(*A*). A larger sense-tagged corpus for *A* was then derived by combining instances in STC(*A*) and instances that were sense-assigned by the WSD classifier of A and consisted of all abstracts that contained A. The quality of the corpus was related to the quality of STC(*A*) and the performance of the WSD classifier. The performance of the WSD classifier was evaluated for precision on abstracts that were in the gold standard set of *A* but were not in STC(*A*). Note that there is no recall measure since a Naive Bayes WSD classifier determines the sense for each instance. We also compared the quality of sense-tagged corpora using two different mapping programs (i.e., CRMap and MetaMap, for several abbreviations).

Results

Thirty-five abbreviations from the UMLS abbreviation extraction program met the criteria for the experiment. The average ambiguity for the set (i.e., the average number of senses), was 3.77, with a standard deviation of 1.91. The ambiguity here was the potential ambiguity captured by the UMLS. The detailed information for a few representative abbreviations is provided in Table 1. For example, the two full forms of *BSA* are *body surface area* and *bovine surface area*, which have been assigned sense identifiers *BSA*$_1$ with a CUI C0005902 and *BSA*$_2$ with a CUI C0036774.

We extracted 155,723 abstracts from MEDLINE; 80,681 of them had an occurrence of the corresponding abbreviation inside a parenthetical expression;

70,764 had gold standard senses identified for the corresponding abbreviations and consisted of the gold standard set. The average ambiguity in the gold standard set was 3. The number of abstracts in the derived sense-tagged corpus was 85,554. Detailed information about the collection of MEDLINE abstracts, the derived sense-tagged corpus, and the gold standard set for each abbreviation is listed in columns 2, 3, and 4 of Table 2. For example, the number of abstracts in the collection of MEDLINE abstracts for *CSF* was 34,483, and the number of abstracts in the gold standard set of *CSF* was 10,771.

The average recall of the derived sense-tagged corpus was 48.0% when evaluated on the gold standard set, and the average precision of the derived sense-tagged corpus was 92.5% when evaluated on the joint set of the derived sense-tagged corpus and the gold standard set. The average precision of the WSD classifier was 87.4% when evaluated on instances that were in the gold standard set but not in the sense-tagged corpus. The average precision of the large size sense-tagged corpus was 90.0% when evaluated on the gold standard set. The detailed information about the performance for each abbreviation is listed in Columns 5, 6, 7, and 8 of Table 2. In Table 2, we see that performance differed widely among the abbreviations. For example, the sense-tagged corpora for 27 out of 35 abbreviations had a precision of over 94% (e.g., *ACE, CAD*), while there were four abbreviations (i.e., *ASP, DVT, EMG*, and *MAC*) for which the sense-tagged corpus had a precision that was lower than 80%. Figure 2 shows the performance of the WSD classifier in relation to the precision of the sense-tagged corpus for different abbreviations. The X-axis represents abbreviations ordered by ascending order of the precision of the sense-tagged corpus. The Y-axis represents the precision. Based on Figure 2, the performance of the WSD classifier was generally related to the precision of the sense-tagged corpus: WSD classifiers trained on sense-tagged corpora with high precision tended to perform better than those on sense-tagged corpora with low precision. However, there were some exceptions, for example, the sense-tagged corpus for *CPI* had a precision of 100%, but the WSD classifier trained on the sense-tagged corpus for *CPI* had a precision of 10.7%.

Table 1 ■

Full Forms, Assigned Sense Identifiers (SID), and Corresponding Concept Identifiers (CUI) for Some Abbreviations (AW)

AW	SID	CUI	Full Form
ACE	ACE$_1$	C0001044	acetylcholinesterase
	ACE$_2$	C0022709	angiotensin converting enzyme
	ACE$_3$	C0050385	doxorubicin cyclophosphamide
	ACE$_4$	C0108844	doxorubicin cyclophosphamide etoposide
	ACE$_5$	C0286421	amsacrine cytarabine etoposide
	ACE$_6$	C0304721	adrenocortical extract
	ACE$_7$	C0473028	antegrade colonic enema
APC	APC$_1$	C0003315	antigen-presenting cells
	APC$_2$	C0032580	adenomatous polyposis coli
	APC$_3$	C0033036	atrial premature complexes
	APC$_4$	C0085171	aphidicholin
	APC$_5$	C0809732	activated protein c
ASP	ASP$_1$	C0038013	ankylosing spondylitis
	ASP$_2$	C0003431	antisocial personality
	ASP$_3$	C0003993	asparaginase
	ASP$_4$	C0004015	aspartic acid
	ASP$_5$	C0052546	aspartylglycine
	ASP$_6$	C0085845	aspartate
BSA	BSA$_1$	C0005902	body surface area
	BSA$_2$	C0036774	bovine serum albumin
CSF	CSF$_1$	C0007806	cerebrospinal fluid
	CSF$_2$	C0009392	colony stimulating factors
	CSF$_3$	C0072454	cytostatic factor
	CSF$_4$	C0893357	competence and sporulation factor
EMG	EMG$_1$	C0004903	exomphalos macroglossia gigantism
	EMG$_2$	C0013839	electromyography
	EMG$_3$	C0180677	electromyographs
	EMG$_4$	C0393125	electromyogram
IBD	IBD$_1$	C0021390	inflammatory bowel diseases
	IBD$_2$	C0022104	irritable bowel syndrome
MAS	MAS$_1$	C0016065	mccune albright syndrome
	MAS$_2$	C0025048	meconium aspiration syndrome
	MAS$_3$	C0451273	macandrew alcoholism scale
PVC	PVC$_1$	C0032624	polymer vinyl chloride
	PVC$_2$	C0151636	premature premature complex
	PVC$_3$	C0280556	cisplatin cyclophosphamide etoposide
RSV	RSV$_1$	C0035236	respiratory syncytial virus
	RSV$_2$	C0086943	rous sarcoma virus
VCR	VCR$_1$	C0042679	vincristine
	VCR$_2$	C0182936	videocassette recorder
	VCR$_3$	C0526312	vanadyl ribonucleoside complex

The result of the comparison between CRMap and MetaMap is summarized in Table 3. CRMap was significantly better than MetaMap with respect to the quality of the derived sense-tagged corpora, except for *APC* and *BSA*. CRMap was superior to MetaMap with respect to the performance of the WSD classifier for all abbreviations except for *VCR* and *APC*. The large sense-tagged corpus derived using CRMap had a better precision that that derived using MetaMap for all abbreviations except *VCR*.

Table 2 ■

Statistical Information Associated with the Collection of MEDLINE Abstracts (CMA), Derived Sense-tagged Corpus (STC), Gold Standard Set (GSS), and Performance Measures*

	Number of Abstracts			Performance (%)			
ABBR	CMA	STC	GSS	STC-R	STC-P	WSD-P	LSTC-P
ACE	9,387	6,750 (6,714)	5,856	76.7 (76.8)	97.9 (98.2)	79.6 (92.8)	93.9 (97.0)
ANA	1,898	1,206 (1,199)	896	73.3	100	98.3 (99.6)	99.6 (99.9)
APC	5,079	4,099	2,310	68.8 (68.6)	84.3	84.9 (83.5)	84.4 (84.2)
ASP	643	300 (299)	141	63.8 (63.8)	74.4	55.0	71.6
BPD	1,351	494 (482)	906	39.5 (39.7)	97.5 (99.2)	93.7 (97.1)	95.3 (97.9)
BSA	6,794	1,913	3,162	9.0	89.9	93.9 (94.3)	93.5 (93.9)
CAD	4,762	3,480 (3,478)	3,325	85.5 (85.1)	99.9	99.6 (94.5)	99.8 (99.1)
CAT	3,421	1,534 (880)	36	41.7	100	90.5 (95.2)	94.4 (97.2)
CML	5,075	3,236	3,350	61.5	99.0	96.2 (95.9)	97.9 (97.8)
CMV	7,841	4,344 (4,276)	4,944	63.1	99.4 (100)	88.2 (99.6)	95.3 (99.8)
CPI	430	48	72	22.2	100	10.7 (12.5)	30.6 (31.9)
CSF	34,483	23,469	10,771	38.4	88.6	86.7 (88.6)	87.5 (88.6)
CVA	584	408 (407)	226	76.1	100	77.8 (100)	94.7 (100)
CVP	1,094	172	587	11.4	100	98.8 (99.8)	99.0 (99.8)
DIP	649	87	112	28.6	94.1	91.0	92.0
DOB	194	25 (11)	2	100	100	NA	100
DVT	1,891	1,607	1,598	26.3	33.0	15.4 (14.5)	29.4 (29.2)
EMG	10,317	2,186 (2,149)	3,770	8.8	38.7 (39.0)	49.9 (49.2)	47.4 (46.9)
FDP	1,280	765	431	55.0	100	95.9 (96.4)	98.1 (98.4)
HSV	9,195	5,890	3,479	38.9	99.9	99.5 (99.6)	99.7 (99.7)
IBD	1,634	1,201	1,149	80.7 (81.7)	96.2 (100)	94.1 (100)	95.8 (100)
LAM	445	127	183	30.6	87.5	80.7 (82.4)	83.1 (84.2)
LDH	8,140	4,451 (4,450)	3,390	48.3	100	99.9	100
MAC	3,873	1,340 (1,308)	862	58.5	78.3 (78.5)	63.8 (64.5)	74.6 (74.9)
MAS	900	114	112	60.7	98.6	100	99.1
MCP	2,670	1,715 (1,712)	461	72.2	98.2	84.4 (85.2)	94.6 (94.8)
PCA	3,788	847	1,553	22.7 (23.3)	94.4 (98.6)	96.2 (96.7)	95.8 (97.2)
PCP	3,534	2,000 (1,991)	2,225	50.6 (50.0)	94.5 (96.3)	76.9 (99.4)	86.3 (96.9)
PEG	2,233	1,190	70	34.3	100	100	100
PSA	5,179	1,528 (1,512)	3,227	28.0 (27.6)	98.5 (100)	97.7 (99.5)	97.9 (99.6)
PVC	1,483	475 (463)	571	25.4	94.2 (98.6)	59.0 (92.0)	68.5 (93.7)
RSV	2,933	739	1,954	17.6	99.7	93.7 (94.0)	94.7 (95.0)
SLE	9,300	6,094	6,772	59.8	99.5	99.2	99.4
TPN	2,200	1,170 (1,145)	1,623	47.4 (48.3)	96.6 (100)	98.3 (99.8)	97.5 (99.9)
VCR	1,043	550	638	65.7	100	63.5 (100)	87.5 (100)
Total	155,723	85,554 (84,565)	70,764	48.0 (47.4)	92.5 (92.9)	87.4 (90.4)	90.0 (91.1)

*STC-R is the recall measure of the sense-tagged corpus when evaluated on the gold standard set. STC-P is the precision measure of the sense-tagged corpus when evaluated on instances that were in both the sense-tagged corpus and the gold standard set. WSD-P is the precision measure of the WSD classifier that used Naïve Bayes algorithm as the machine learning algorithm and bag of stemmed word as features through a measure when evaluated on instances that were in the gold standard set but not in the sense-tagged corpus. LSTC-P is the precision measure of the large sense-tagged corpus when evaluated on the gold standard set (note the number of instances in the large sense-tagged corpus is the same as the number of instances in the collection of MEDLINE abstracts). The number inside the parentheses is the corresponding number after removing rare sense for each abbreviation.

Discussion

Our method can function for terms that are not abbreviations without any change. However, in this study we used abbreviations only because we could automatically obtain the gold standard, and then use it to evaluate the quality of the derived sense-tagged corpus automatically. Therefore, we avoided the expense and effort that is associated with obtaining a gold standard set using experts.

We analyzed the causes of low precision for the derived sense-tagged corpora, and there were two causes: relatedness among different senses and the existence of poor conceptual relatives. The cause of low precision of the WSD classifier trained on the

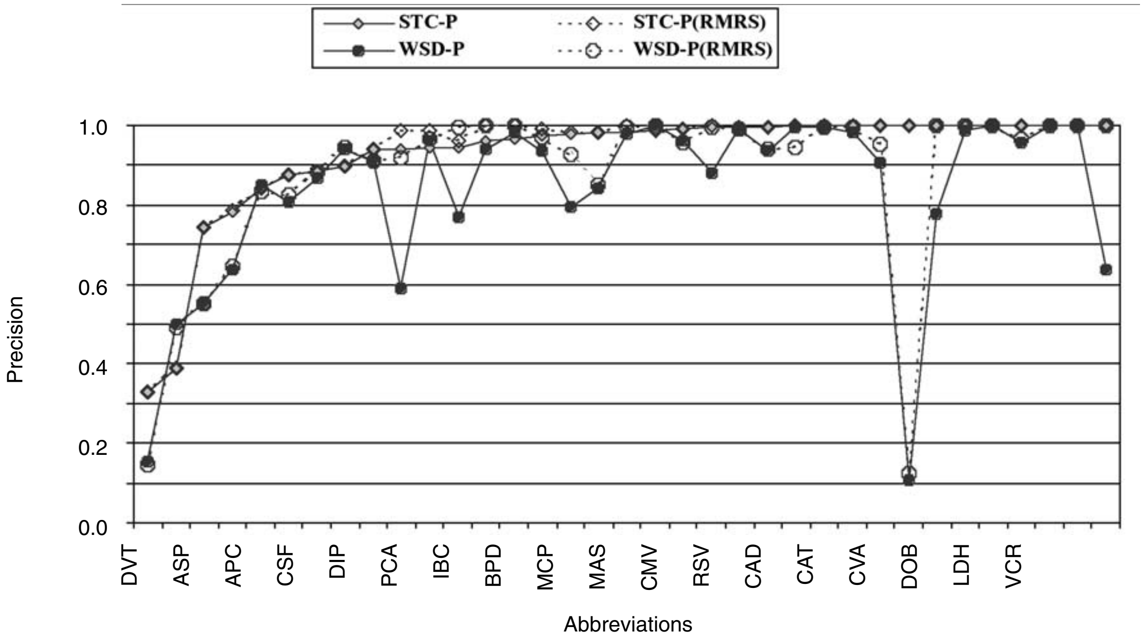

Figure 2 The precision of the WSD classifier in relation to the precision of the sense-tagged corpus. The x-axis represents abbreviations ordered by the ascending order of the precision of STC (STC-P). Lines with diamond points denote the precision of STC (STC-P) and the precision of STC after remove rare senses (STC-P (RMRS)); and lines with circular points denote the precision of the WSD classifier (WSD-P) and the precision of the WSD classifier after removing rare senses (WSD-P) (RMRS).

derived sense-tagged corpora with high precision was the lack of enough training instances.

The low precision of the derived sense-tagged corpora for some abbreviations was caused by the existence of closely related senses. For example, there were four senses of *EMG*: *EMG*$_1$ (exomphalos macroglossia gigantism), *EMG*$_2$ (electromyography), *EMG*$_3$ (electromyographs), and *EMG*$_4$ (electromyogram). Three of them (i.e, *EMG*$_2$, *EMG*$_3$ and *EMG*$_4$), were closely related (every pair had a relation defined in MRREL). The precision of STC(*EMG*) was 37%. *ASP* had two closely related senses: *ASP*$_4$ (aspartic acid) and *ASP*$_6$ (aspartate). They had relations defined in MRREL, and they also related to 21 concepts in common in MRREL. The precision of the sense-tagged corpus for *ASP* was 74.4%. All four abbreviations with sense-tagged corpora with a precision lower than 80% had closely related senses. After ignoring these four abbreviations, the average

Table 3 ■

Comparing Results of Two Mapping Programs: CRMap and Metamap*

	Performance (%) Original (Remove-Rare-Sense)							
	STC-R		STC-P		WSD-P		LSTC-P	
ABBR	CRMap	Metamap	CRMap	Metamap	CRMap	Metamap	CRMap	Metamap
APC	68.8 (68.6)	60.6	84.3	86.3	84.9 (83.5)	82.7	84.4 (84.2)	85.2
BSA	9.0	29.9	89.9	89.8	93.9 (94.3)	90.7	93.5 (93.9)	90.4
LAM	30.6	16.4	87.5	63.8	80.7 (82.4)	59.6	83.1 (84.2)	60.7
MAS	60.7	8.0 (53.6)	98.6	13.6 (98.4)	100.0	30.4 (100)	100.0	20.5 (99.1)
PVC	25.4	17.5	94.2 (98.6)	56.8	59.0(92.0)	39.7 (49.6)	68.5 (93.7)	45.0 (51.8)
VCR	65.7	57.5 (58.0)	100.0	99.2 (100)	63.5 (100)	97.0 (99.6)	87.5 (100)	98.3 (99.8)

*Refer to Table 2 for notation of STC-R, STC-P, WSD-P, and LSTC-P.

recall of the sense-tagged corpora, the average precision of the sense-tagged corpora, the precision of the WSD classifiers, and the precision of the large sense-tagged corpora were 50.6%, 96.4%, 91.9%, and 94.3% respectively.

The quality of the sense-tagged corpus was also related to the quality of conceptual relatives for each sense. For example, a conceptual relative of APC_1 (antigen presenting cell), was cells, and textual variants of cells occurred in many abstracts; therefore our method favored APC_1.

The low precision of the WSD classifiers was due to the existence of rare senses (i.e., those occur less than 0.5% of the total occurrences). We excluded those rare senses from the sense definition and re-evaluated performance. For example, we excluded VCR_3 (vanadyl ribonucleoside complex) from the sense definition of *VCR* since it occurred only once. The change of measures is presented in Table 2 using parenthetical expressions. The change of measures is presented in Figure 2 using dotted lines. Generally, the precision of sense-tagged corpora improved only slightly, but there was a dramatic improvement in WSD classifiers. For example, the precision of the WSD classifier for *CMV* increased from 88.2% to 99.6%; and the precision of the WSD classifier for *VCR* increased from 63.5% to 100%. There were a few abbreviations with senses that had less than 10 instances in sense-tagged corpora, but had more than 0.5% of the total number of occurrences (for example, CPI_1 occurred once and CPI_3 occurred once). Since those senses were not considered to be rare, the corresponding WSD classifier performed poorly. For example, the precision of the WSD classifier for *CPI* was only 12.5%. After ignoring four abbreviations with closely related senses and removing rare senses, the average recall of the sense-tagged corpora, the average precision of the sense-tagged corpora, the precision of the WSD classifiers, and the precision of the large sense-tagged corpora were 50.6%, 96.8%, 95.3%, and 96.0% respectively.

The difference in performance using CRMap and MetaMap indicated the different goals of the two programs. The goal of CRMap is to match only conceptual relatives, while the goal of MetaMap is to map every noun phrase in the context to UMLS concepts. MetaMap fails to find conceptual relatives that contain prepositional noun phrases whereas CRMap does not have such limitation. For example, MetaMap failed to identify *persistent pulmonary hypertension of the newborn*, which is a sibling of MAS_2 (*meconium aspiration syndrome*), as a relative of MAS_2 in abstracts that contain it (e.g., *MAS* can easily develop *persistent pulmonary hypertension* of *the new born*). The running time of CRMap is much faster than MetaMap since CRMap considers only conceptual relatives of a specific term.

Sense-tagged corpora derived using the proposed method can be combined with sense-tagged corpora derived using our previous method, i.e., using unambiguous synonyms, to construct WSD classifiers for NLP systems that use the UMLS as a sense inventory. The applied domain of WSD classifiers can be MEDLINE abstracts as well as other text in the specialized domain, such as medical reports. Since MEDLINE abstracts consist of free-text from almost all subdomains in biomedicine, such as the biological domain, clinical domain, and bioinformatics.

We assumed one sense per discourse in our method. It is almost certainly valid in MEDLINE abstracts for domain-specific terms. However, for general English terms, such as *cold* or *discharge*, the validity of this assumption remains to be tested.

One limitation of the proposed method is that there may not be enough instances for each sense. We think that WSD classifiers derived using Naive Bayes algorithm are not appropriate for terms that have senses with less than 10 instances in the training set. Without enough instances, the evidence for assigning rare senses is insufficient, and the existence of them in the sense-tagged corpus will affect the overall performance. For example, there was only one abstract with sense VCR_3 in the sense-tagged corpus for *VCR* and no abstract with sense VCR_3 in the gold standard set for *VCR*; this caused the WSD classifier to incorrectly assign 80 (out of 219) abstracts to VCR_3, although here 77 abstracts had the gold standard sense VCR_1, and 3 abstracts had the gold standard sense VCR_2.

We believe that rare-senses should be separated from frequent senses when constructing a WSD system. A hybrid WSD system seems to be unavoidable. One part of the system should handle rare senses by using handcrafted WSD rules or WSD knowledge learned using machine learning techniques that do not depend on statistical information, such as instance-based methods,[43] where the sense assignment depends on the similarity measure of two instances. The other part of the system should handle frequent senses using WSD knowledge learned from many instances.

In this study, each conceptual relative appeared to contribute equally to the sense assignment. However, different relations and different sources may have different levels of contribution to the sense assignment. We plan to further investigate relations defined in different sources and to formulate a new sense assignment scheme. For abstracts with conceptual relatives from multiple senses, a weight-sense assignment scheme should be possible to formulate. In addition, we plan to use clustering techniques to find instances that are associated with rare senses or unknown senses. Finally, we plan to evaluate our method on a manually sense-tagged WSD test collection, the National Library of Medicine WSD test collection,[36] which consists of 50 ambiguous terms from the UMLS and 100 sense-tagged instances for each term.

Conclusion

The mapping of free-text to UMLS concepts is an important task for NLP applications. To improve the performance of the mapping, a method to resolve terms that possess multiple concepts is necessary. Several preliminary attempts were based on manual handcrafted rules, which were often incomplete and unscalable. Supervised machine-learning techniques have been used to construct WSD classifiers automatically from sense-tagged corpora. However, manual sense-annotation of a corpus is also a manual task. In this article, we acquired sense-tagged corpora automatically by utilizing conceptual relations in the UMLS and abstracts in MEDLINE so that WSD classifiers can be constructed automatically. The method can be used on almost all ambiguous terms in the UMLS. The derived sense-tagged corpora had a precision of 96.8% and a recall of 50.6% when evaluated on majority senses of a set of abbreviations after ignoring abbreviations with closely related senses. The large size sense-tagged corpora that contained all abstracts in MEDLINE with an occurrence of the corresponding abbreviation had a precision of 96.0% when evaluated on the majority senses of a set of abbreviations after ignoring abbreviations with closely related senses. The gold standard set used for the evaluation was derived automatically. This work demonstrated that sense-tagged corpora can be used to construct WSD classifiers using Naive Bayes learning for NLP systems, provided that there are enough instances for each sense.

The authors thank Dr. Alan Aronson at National Library of Medicine for sharing the MetaMap.

References ■

1. Aronson A. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. Proc AMIA Symp. 2001; 17–21.
2. Friedman C, Hripcsak G, DuMouchel W, et al. Natural language processing in an operational clinical information system. Nat Lang Eng. 1995; 1(1): 83–108.
3. Spyns P. Natural Language Processing in Medicine: An Overview. Meth Inform Med. 1996; 35:285–301.
4. Johnson SB. A semantic lexicon for medical language processing. J Am Med Inf Assoc. 1999; 6(3): 205–18.
5. Friedman C. A Broad Coverage Natural Language Processing System. Proc AMIA Symp. 2000; 270–4.
6. Nadkarni P, Chen R, Brandt C. UMLS Concept Indexing for Production Databases. J Am Med Inf Assoc. 2001; 8:80–91.
7. Weeber M, Klein H, Aronson A, Mork J. Text-based discovery in biomedicine: The architecture of the DAD-system. Proc AMIA Symp. 2000; 903–7.
8. Swanson DR. Migraine and magnesium: Eleven neglected connections. Perspect Biol Med. 1988; 31(4): 526–57.
9. Rindflesch TC, Aronson AR. Ambiguity Resolution while Mapping Free Text to the UMLS Metathesaurus. Proceedings of the Eighteenth Annual Symposium on Computer Applications in Medical Care, pp 240–4.
10. Friedman C. A Broad Coverage Natural Language Processing System. Proc AMIA Symp. 2000; 270–4.
11. Jurafsky D, Martin J. Word Sense Disambiguation and Information Retrieval. Speech and Language Processing. New York, Prentice Hall, 2000: 631–66.
12. Fellbaum C. WordNet: An Electronic Lexical Database. 1998.
13. Longman Dictionary. <http://www.longman-elt.com/dictionaries/research/resnlapp.html>, 2001.
14. Chapman R. Roget's International Thesaurus. New York, Harper & Row, 1977.
15. Ng H, Lee H. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. Proc ACL. 1996; 40–7.
16. Schütze H. Automatic Word Sense Discrimination. Artif Intell. 1998; 24[1]: 97–123.
17. Ng HT, Zelle J. Corpus-based approaches to smenatic interpretation in natural language processing. AI Magazine. 1997; Winter: 45–64.
18. Ide N, Veronis J. Introduction to the special issue on word sense disambiguation: The state of the art. Comput Ling. 1998; 24[1]: 1–40.
19. Ng HT. Getting serious about word-sense disambiguation. Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What and How? 1997; 1–7.
20. Kilgarriff A. Gold standard datasets for evaluating word sense disambiguation programs. Comput Speech Lang. 1998; 12(3).
21. Gale WA, Church KW, Yarowsky D. Using bilingual materials to develop word sense disambiguation methods.Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation. 1992, pp 101–12.
22. Yarowsky D. Unsupervised word sense disambiguation rivaling supervised methods. Proc ACL 33; 189–96.
23. Agirre E, Rigau G. Word Sense Disambiguation using Conceptual Density. Proceedings of COLING, 1996.
24. Liu H, Lussier Y, Friedman C. Disambiguating ambiguous Biomedical terms in biomedical narrative text: An unsupervised method. J Biomed Informat. 2001; 34 (4): 249–62.

25. Unified Medical Language System. NIH. 2000.
26. Sussna M. Word sense disambiguation for free-text indexing using a massive semantic network. Proceedings of the International Conference on Information and Knowledge Management. 1993; 67–74.
27. Agirre E, Rigau G. A proposal for word sense disambiguation using conceptual distance. Amsterdam Studies in the Theory and History of Linguistic Science: Current Issues in Linguistic Theory. 1997, pp 161–73.
28. Gale WA, Church KW, Yarowsky D. One Sense Per Discourse. Proceedings of the ARPA Workshop on Speech and Natural Language 1992, pp 233-237.
29. Krovetz R. More than One Sense Per Discourse. Proceedings of the ACL-SIGLEX Workshop, 1998.
30. Ng HT. Examplar-Based Word Sense Disambiguation: Some Recent Improvements. Proc EMNLP-2, 1997.
31. Mitchell T. Machine Learning. New York, McGraw Hill, 1997.
32. Marquez L. Machine Learning and Natural Language Processing. 2000.
33. Duda R, Hart P. Pattern Classification and Scene Analysis. New York, John Wiley & Sons, 1973.
34. <medline. http://www.nlm.nih.gov>, 2001.
35. UMLS Knowledge Sources, 2000 Edition. Washington, DC, U.S. Dept of Health and Human Services, National Institutes of Health, National Library of Medicine, 2000.
36. Weeber M, Mork J, Aronson A. Developing a Test Collection for Biomedical Word Sense Disambiguation. Proc AMIA Symp. 2001; 746–50.
37. Aronson AR, Rindflesch TC, Browne A. Exploiting a large thesaurus for information retrieval. Proc RIAO. 94; 197–216.
38. Liu H, Lussier Y, Friedman C. A study of the UMLS abbreviations. Proc AMIA Symp. 2001; 393–7.
39. Sneiderman CA, Rindflesch TC, Aronson AR. Finding the findings: Identification of findings in medical literature using restricted natural language processing. Proc AMIA Symp. 1996; 239–43.
40. Rindflesch TC, Hunter L, Aronson AR. Mining molecular binding terminology from biomedical text. Proc AMIA Symp. 1999, pp 127–31.
41. Bodenreider O. Circular Hierarchical Relationships in the UMLS: Etiology, Diagnosis, Treatment, Complications and Preventions. Proc AMIA Symp. 2001; 57–61.
42. Porter MF. An algorithm for suffix stripping. Program. 1980; 14(3): 130–7.
43. Cardie C. A case-based approach to knowledge acquisition for domain-specific sentence analysis. Proceedings of the National Conference of AI. 1993; 798–803.