Mapping Abbreviations to Full Forms in Biomedical Articles

Hong Yu, MS, MPhil, George Hripcsak, MD, Carol Friedman, PhD

Abstract Objective: To develop methods that automatically map abbreviations to their full forms in biomedical articles.

Methods: The authors developed two methods of mapping defined and undefined abbreviations (defined abbreviations are paired with their full forms in the articles, whereas undefined ones are not). For defined abbreviations, they developed a set of pattern-matching rules to map an abbreviation to its full form and implemented the rules into a software program, AbbRE (for “abbreviation recognition and extraction”). Using the opinions of domain experts as a reference standard, they evaluated the recall and precision of AbbRE for defined abbreviations in ten biomedical articles randomly selected from the ten most frequently cited medical and biological journals. They also measured the percentage of undefined abbreviations in the same set of articles, and they investigated whether they could map undefined abbreviations to any of four public abbreviation databases (GenBank LocusLink, SWISSPROT, LRABR of the UMLS Specialist Lexicon, and BioABACUS).

Results: AbbRE had an average 0.70 recall and 0.95 precision for the defined abbreviations. The authors found that an average of 25 percent of abbreviations were defined in biomedical articles and that of a randomly selected subset of undefined abbreviations, 68 percent could be mapped to any of four abbreviation databases. They also found that many abbreviations are ambiguous (i.e., they map to more than one full form in abbreviation databases).

Conclusion: AbbRE is efficient for mapping defined abbreviations. To couple AbbRE with abbreviation databases for the mapping of undefined abbreviations, not only exhaustive abbreviation databases but also a method to resolve the ambiguity of abbreviations in the databases are needed.

Abbreviations and acronyms are commonly used in biomedical literature. The names of many clinical diseases and procedures, and of common terms in the basic sciences, have widely used abbreviations. Recognizing the full forms associated with abbreviations is important for identifying the meaning of an abbreviation, which in turn facilitates natural language processing of, and information retrieval from, the literature. Two types of abbreviations appear in biomedical articles—common and dynamic abbreviations. Many common abbreviations become accepted as synonyms; for example, CHF (congestive heart failure) and CABG (coronary-artery bypass graft) are listed in standard vocabulary resources, such as the Medical Subject Headings (MeSH) and Unified Medical Language System (UMLS). Obviously, common abbreviations represent terms important in their domains. Federik reported that using common medical abbreviations as search terms for literature citations resulted in more relevant retrievals than did using the full forms as search terms. She found that all 20 common medical abbreviations she chose were recognized by MEDLINE, and all were mapped to the appropriate MeSH headings. In contrast, dynamic abbreviations are defined by an author for convenience in only a particular article; for example, CU might represent Columbia University in one article, computer use in another, and congested udder in a third. Many articles use both common and dynamic abbreviations. Therefore, it is important that automated text processing systems recognize the meanings of both types of abbreviations.

We are investigating two approaches to identify the meanings of abbreviations in electronic articles: 1) detecting abbreviations and mapping them to their full forms solely on the basis of the content of the article, and 2) detecting abbreviations and then mapping them to full forms that we obtain from abbreviation databases. The first approach is limited to those abbreviations that are defined in the article, i.e., their full forms appear in the article. We are exploring using the second approach as an adjunct to the first, to discover the full forms associated with abbreviations not so defined.

The first approach is feasible in part because many scientific journals have rules for the formation and definition of abbreviations; the most common requirement is that an abbreviation be defined on first use in the format <full form> (<abbreviation>). In addition, people apply many common conventions to create an abbreviation. For example, people may form an acronym from the initial letter of the primary words of a phrase (e.g., NLP for natural-language processing), they may create an abbreviation using meaningful portions of the words (e.g., Fig. for figure), or meaningful parts of a neoclassical compound (e.g., APT for aminopropylisothiuronium), or a combination of meaningful units or words and initial letters of component words (e.g., mAb for monoclonal antibody). Therefore, we can use pattern recognition methods to find abbreviations and to map them to their full forms within an article.

Review of Previous Research

Other researchers have developed automatic methods for identifying abbreviations and pairing those abbreviations with a definition. Hisamitsu and Niwa identified technical terms—including company names, organization names, law names, and theory—names from Japanese newspaper articles. They first, through bi-gram statistics, selected phrases associated with parentheses (the parenthetical phrase and the outer phrase co-occur more frequently than random); they then applied a set of simple rules to identify whether the parenthetical phrase was an abbreviation of the outer phrase. For example, a rule indicated that a phrase was an abbreviation of a full form if the letters of the phrase appeared in order in the full form. Their evaluation of this approach demonstrated 97 percent precision.

KEP (for knowledge extraction program) is another system that identifies paired abbreviations and full forms. The system first detects a word as an abbreviation when all the letters of the word are uppercase. It then fragments the sentence that contains the abbreviation into a set of t-word strings, where t ranges from 1 to n+3 (n is the total number of letters in the abbreviation). For each string, KEP takes the initial letter of each word and forms a shortened string. KEP considers the string as a full form of the abbreviation if the letters of the shortened string match over 70 percent of the letters of the abbreviation. KEP has been shown to have 73 percent recall and 84 percent precision.

PNAD-CSS (for Protein Name Abbreviation Dictionary Construction Support System) extracts paired a protein name (e.g., eukaryotic initiation factor 2) and its abbreviation (e.g., elf2) from biological abstracts. The program was built on top of PROPER,
a program that uses morphologic features (e.g., uppercase letters combined with numbers) to recognize proper nouns as protein terms in biological abstracts. For example, PROPER recognizes “elf2” as a protein term because it contains a numeric value (in this case, “2”).

PNAD-CSS also uses TEX82, a program that breaks up words in a phrase into several components. PNAD-CSS first finds the parentheses associated with the protein terms recognized by PROPER; it then determines whether the parenthetical phrase is an abbreviation of the outer phrase. PNAD-CSS uses TEX82 to break up words of the preceding phrase and determines whether the parenthetical abbreviation candidate maps to the initial letters of the broken-up phrase.

Consider the phrase megestrol acetate (megace), for example. TEX82 parses “megestrol acetate” as “meges trol ac etate,” which PNAD-CSS then matches with “megace” because it matches the initial letters of the components (e.g., “meg,” “ac,” and “e” in “megace” match the initial letter(s) of “meges,” “ac,” and “etate,” respectively). PNAD-CSS had 95.56 percent recall and 97.58 percent precision.

All three systems have limitations that may affect their use in the biomedical domain. Hisamitsu and Niwa’s approaches rely on statistical significance of the two terms that are associated with parentheses; the approach might miss abbreviations and full forms that are newly introduced into the literature. KEPS considers as abbreviations only words in which all letters are uppercase, and matches only letters (not other symbols, such as numbers). These restrictions do not apply to many biomedical abbreviations, which often consist of both upper- and lowercase letters (e.g., Ab for Antibody) and include numbers (e.g., Ig1 for lateral gastrocnemius 1). PNAD-CSS was built on top of PROPER and may miss paired abbreviations and full forms that were not recognized by PROPER.

Hisamitsu and Niwa’s approaches and KEPS have not been evaluated in the biomedical domain. PNAD-CSS was developed to extract protein names and their abbreviations; no one has yet evaluated whether it can be generalized to recognize other full forms and associated abbreviations in other settings or in whole articles rather than abstracts. Mapping abbreviations in whole articles may be more challenging since the linguistics of an article body may be more sophisticated than its abstract.

Hisamitsu and Niwa’s approaches, KEPS, and PNAD-CSS all apply sets of pattern-matching rules for mapping an abbreviation to its full form. However, Hisamitsu and Niwa’s pattern-matching rules are preliminary and can introduce false matches. For example, column would be falsely recognized as an abbreviation of Columbia University, because the letters of column appear in order in Columbia University.

KEPS applies the n-gram approach to identify full forms and therefore may have difficulty in identifying a full form boundary. For example, KEPS may mistake the full form of BPI as bactericidal permeability increasing instead of bactericidal permeability increasing protein, since the initial letter of protein is not in the abbreviation. In addition, KEPS’s pattern-matching rules consider only the initial letters of words in a phrase; they may miss those abbreviations that represent the middle letters of words (e.g., APT for aminopropylisothiuronium).

KEPS does apply approximate matching (i.e., if the string formed from initial letters of a sequence of words matches over 70 percent of the abbreviation, KEPS considers the sequence of words as its full form), and the approximation may indirectly include some matches from the middle letters. It is not clear how suitable the approximation is in the biomedical domain, however.

PNAD-CSS relies on TEX82 to break up words into components; therefore, TEX82 needs to be evaluated to determine how well it breaks words in biomedical fields other than protein science.

To date, Hisamitsu and Niwa’s approaches and KEPS have been evaluated by the developers, but not by independent researchers. PNAD-CSS was evaluated by a person who was not a biomedical specialist. The evaluation of PNAD-CSS also assumed that PROPER had 100 percent recall and 100 percent precision in identifying protein terms and that PNAD-CSS recognized a correct abbreviation as an abbreviation of a protein name even if the abbreviation was not. Therefore, PNAD-CSS’s recall and precision may be lower than reported.

General Approach

The program we developed, AbbRE, differs from the three approaches just described. AbbRE was developed to handle full biomedical articles. AbbRE searches for parenthetical expressions for paired abbreviations and full forms. AbbRE does not break up words into components; it relies only on a set of pattern-matching rules for mapping an abbreviation.
to its full form. The pattern-matching rules were generalized from the common conventions by which people create an abbreviation. As described in this paper, AbbRE has been evaluated by domain experts.

Any method that attempts to define abbreviations solely on the basis of information in the articles in which they appear obviously cannot interpret abbreviations that are undefined in those articles. We therefore attempted to augment AbbRE by mapping undefined abbreviations to externally developed abbreviation databases.

Because people recognize that understanding abbreviations is important for information retrieval, there are many such databases. They include databases containing protein- and gene-name abbreviations (e.g., GenBank LocusLink, SWISSPROT, Yeast Genome Database, and Genome Database Bank), common-abbreviation databases such as those used for the natural language processing lexicon (e.g., LRABR), and those created for computer linkages between abbreviations among different disciplines (e.g., BioABACUS). We chose to use Genbank LocusLink, SWISSPROT, LRABR from the UMLS Specialist Lexicon, and BioABACUS because they are maintained by domain experts and many of them are supported by government organizations; they also have a good coverage.

- Genbank LocusLink is a Web source developed recently by the National Center for Biotechnology Information (NCBI) to facilitate retrieval of gene-based information and to provide a reference sequence standard. LocusLink contains a database (stored in the file LL_out) of 54,719 genes; it lists both their abbreviations and their full forms.

- SWISSPROT is an annotated protein-sequence database established in 1986 and maintained collaboratively by the Swiss Institute for Bioinformatics (SIB) and the European Bioinformatics Institute (EBI). SWISSPROT currently has 88,800 protein abbreviations and their full forms.

- The LRABR file of more than 10,000 abbreviations is part of the UMLS SPECIALIST lexicon. The National Library of Medicine (NLM) built the UMLS Knowledge Sources to improve the ability of computer programs to “understand” the biomedical meaning of user inquiries and to use this understanding to retrieve and integrate relevant machine-readable information for users. The UMLS SPECIALIST lexicon is an English-language lexicon of biomedical terms from a variety of sources, including MEDLINE citation records and the UMLS Metathesaurus.

- BioABACUS is a public database of common abbreviations that creates computer linkages between abbreviations and their meanings. The database was generated manually from literature and from other databases; it covers only biotechnology and computer science. It contains more than 6,000 abbreviations and their full forms.

**Methods**

Our study had three components—development of AbbRE, evaluation of AbbRE, and determination of the percentage of undefined abbreviations that could be mapped to entries in each of four abbreviation databases (GenBank LocusLink, SWISSPROT, LRABR, and BioABACUS).

**Development of AbbRE**

We have developed a set of rules that define a well-formed abbreviation. The rules were generalized from review of all the abbreviations and their full forms in 200 Science articles, a randomly selected subset of articles related to signal-transduction pathways. Table 1 summarizes these rules.

By implementing these rules in a computer code (Perl), we developed AbbRE (abbreviation recognition and extraction program), which maps abbreviations and full forms from computer-readable versions of scientific articles and produces as output paired abbreviations and full forms. AbbRE performs its work in four steps.

**Step 1: Parenthesis Detection**

AbbRE preprocesses the article to remove html tags and certain parentheses that are not associated with abbreviations, such as parentheses containing only numbers, numbers with percentage symbol (%), and certain keywords—fig, table, lane, pH, page, inside, inset, and column. After preprocessing, AbbRE parses the article into sentences and selects for further analysis the remaining sentences that contain parentheses.

**Step 2: Parenthesis Separation**

Using the selected sentences from step 1, AbbRE first parses a sentence into components by the right parenthesis; for each component it then pairs the phrase preceding the left parenthesis (the outer phrase) with the phrase after the left parenthesis (the inner phrase). For example, in the sentence Transmembrane domain (TM), DD (death domain), and the negative regulatory domain (NR) are labeled, the
three paired outer and inner phrases for further analysis are Transmembrane domain (TM), DD (death domain), and And the negative regulatory domain (NR).

Step 3: Abbreviation Detection

Using the selected paired phrases from step 2, AbbRE partitions any inner phrase that contains certain punctuation marks, such as a semicolon or comma, and extracts the part of the inner phrase to the left of the punctuation mark for further analysis. For example, with TNFRI-associated death domain protein (TRADD; Hsu et al., 1995, 1996a), AbbRE parses the inner phrase, TRADD; Hsu et al., 1995, 1996a, and extracts TRADD as a new inner phrase for further analysis.

AbbRE assumes that an abbreviation consists of only one word and recognizes that an abbreviation is shorter than its full form. Either an outer phrase or an inner phrase may contain an abbreviation or a full form. If the inner phrase contains more than one word, then AbbRE assumes that the inner phrase contains a potential full form and the word right before the left parenthesis is the potential abbreviation. For example, in DD (death domain), AbbRE recognizes the inner phrase death domain as containing a potential full form, and the word right before the left parenthesis, DD, as a potential abbreviation.

If an inner phrase contains only one word, then the inner phrase is judged to be an abbreviation and the outer phrase is judged to contain the full form. It is possible, however, that a full form consists of only one word. For example, the full form of the abbreviation T is temperature. To recognize this type of abbreviation, AbbRE applies the following strategies.

When an inner phrase contains only one word and the number of letters in the inner phrase is more than the number of letters in the word right before the left parenthesis, AbbRE not only considers the inner phrase as a potential abbreviation and the outer phrase as a potential full form, but also considers the inner phrase as a potential full form of the word right before the parenthesis. In the amount of Ab (antibody), AbbRE not only considers the inner phrase, antibody, as a potential abbreviation, with its full form contained in the outer phrase, the amount of Ab, but also considers antibody as a potential full form of Ab.

Step 4: Full Form Detection

Next, AbbRE applies the pattern-matching rules that we developed (Table 1) to map an abbreviation to its full form. Since the first letter of the abbreviation always corresponds to the first letter of the first meaningful word of the full form, AbbRE selects the words in a potential full form when these words begin with the first letter of the potential abbreviation. Then AbbRE extracts a list of strings of words starting from the selected word to the end of the phrase, and recognizes each string as a potential full form.

In death domain (DD), for example, both death and domain are marked up (because both words begin with a letter d, which is the first letter of the potential abbreviation D); AbbRE recognizes two strings—domain and death domain—as potential full forms.

<table>
<thead>
<tr>
<th>Rule</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. The first letter of an abbreviation matches the first letter of the meaningful word of the full form.</td>
<td>The Unified Medical Language System (UMLS)</td>
</tr>
<tr>
<td>2. The abbreviation matches the first letter of each word in the full form.</td>
<td>tumor necrosis factor (TNF)</td>
</tr>
<tr>
<td>3. A word in the full form can be skipped if the abbreviation letter matches the first letter of the following word.</td>
<td>extracellular signal-regulated protein kinase 1 (ERK1)</td>
</tr>
<tr>
<td>4. The abbreviation letter matches consecutive letters of a word in the full form.</td>
<td>insulin receptor (InR)</td>
</tr>
<tr>
<td>5. The abbreviation letter matches the last letter of a word in the full form if the letter is an s and if the first letter of the word matches the abbreviation.</td>
<td>cysteine-rich domains (CRDs)</td>
</tr>
<tr>
<td>6. The abbreviation letter matches a middle letter of a word in the full form if the first letter of the word matches the abbreviation.</td>
<td>immunoglobulin G1 (IgG1)</td>
</tr>
</tbody>
</table>

Table 1: Pattern-matching Rules for Mapping an Abbreviation to Its Full Form
AbbRE starts with the string with fewer words (e.g., *domain*) and maps the string to its potential abbreviation by applying rules 2 through 7. If the abbreviation does not match the full form (e.g., rules 2 through 7 do not apply), the next larger string, which is *death domain*, is processed. Once the abbreviation matches its full form, AbbRE outputs the paired abbreviation and its full form and moves on to the next paired phrases. The output is in the form “abbreviation | full form | article-identification number:"

**DD | death domain | 1067.html |**

**Evaluation of AbbRE**

We evaluated AbbRE by recall (the number of correct abbreviations present in the reference standard and found by AbbRE, divided by the total number of abbreviations in the reference standard) and by precision (the number of correct abbreviations in the AbbRE output divided by the total number of abbreviations in that output), typically applied to evaluate the outcomes of information retrieval. The reference standard was determined by a majority vote of biomedical experts.

We evaluated AbbRE in the biomedical domain, which we further divided into medical domain and biological domain. Accordingly, we selected three medical experts (experts 1 to 3), all of whom hold MD degrees, as well as three biological experts (experts 4 to 6), all of whom hold doctoral degrees in biological science. Medical experts were asked to evaluate the application of AbbRE in medical articles, and biological experts were asked to evaluate the application of AbbRE in biological articles.

We selected five journals from each domain. The selection of the journals was determined by whether the journal articles are available in electronic form and whether they were among the most frequently cited in biomedical articles. The five selected biological journals were *Cell, Science, Trends in Neurosciences (TNS), Proceedings of the National Academy of Sciences (PNAS)*, and *Journal of Biological Chemistry (JBC)*; the five selected medical journals were *New England Journal of Medicine, CA: A Cancer Journal for Clinicians, Journal of the National Cancer Institute (JNCI), Journal of the American Medical Association (JAMA)*, and *Lancet*. All ten selected journals were among the 100 most frequently cited biomedical journals in 1999.

For each journal, we randomly downloaded (we randomized the time of publication) five electronic articles; we therefore obtained a total of 50 articles from the ten selected journals. We assigned each article a unique article-identification number.

We divided the evaluation process into part A and part B. For part A, experts selected abbreviations and full forms from the evaluation articles; for part B, experts judged the correctness of the abbreviations and full forms that were selected by AbbRE from the evaluation articles.

Ten articles were used in part A of the evaluation; each article was randomly selected from the five downloaded articles of any of the ten selected journals. We gave five articles to each expert in his domain (medical or biological). We e-mailed all the experts their evaluation articles (in HTML format) with their article-identification numbers. We also e-mailed the experts instructions for identifying abbreviations and entering them into two output files. One output file consisted of defined abbreviations, their corresponding full forms, and their unique article-identification numbers. The other output file consisted of undefined abbreviations and their unique article-identification numbers. The experts then e-mailed us their output files.

The same ten articles were analyzed by AbbRE. The AbbRE output consisted of the defined abbreviations, their full forms, and their article-identification numbers. All the defined abbreviations selected by either the experts or AbbRE were pooled to create a list of abbreviations with the full forms and the unique article-identification numbers. All the undefined abbreviations selected by the experts were pooled to create a list of abbreviations and unique article-identification numbers. The two pooled lists of defined and undefined abbreviations were re-evaluated by the experts. Experts selected abbreviations in the pooled lists. The reference standard consisted of those abbreviations that were selected by two or three experts (i.e., a majority for each domain).

We calculated the recall and precision of AbbRE in relation to the reference standard. In addition, we obtained the recall and precision of each expert by comparing his original selection of abbreviations with the reference standard. In this analysis, the total number of abbreviations was the sum of unique pairs of abbreviations and full forms within each article, not the pairs that were unique across all articles. We also obtained overall agreements as an index of strength of expert agreement, when experts selected abbreviations as well as when experts agreed and disagreed on the pooled abbreviations. We also man-
usually mapped all the abbreviations that were selected by the experts but not detected by AbbRE to their original articles and reviewed the causes of the AbbRE failures. We also obtained the percentage of abbreviations that were defined in the articles.

In part B of the evaluation, we ran AbbRE using the remaining 40 articles (20 medical articles from five medical journals and 20 biological articles from five biological journals). The output of AbbRE consisted of defined abbreviations, their full forms, and their unique article-identification numbers as well as the sentences that contained the abbreviations and full forms. We asked the experts to judge the correctness of each abbreviation and its full form listed in the AbbRE outputs. The reference standard consisted of those abbreviations that were agreed on by two or three experts. We obtained the precision of AbbRE for medical and biological journals separately as well as for the aggregate.

**Determination of the Percentage of Undefined Abbreviations That Could Be Mapped to Abbreviation Databases**

We randomly selected a subset of the undefined abbreviations (30 from medical articles and 30 from biological articles) from the reference standard determined in part A of the evaluation and judged the existence of those abbreviations in any of four abbreviation databases (Genbank LocusLink, SWISSPROT, LRABR, and BioABACUS). We further calculated the percentages of those abbreviations that could be identified in the four abbreviation databases, individually and in combination.

**Results**

**Evaluation of AbbRE**

In part A of the evaluation, a total of 46 defined abbreviations were pooled from three medical experts (experts 1 to 3) and the AbbRE, of which 45 were selected as the reference standard on the basis of agreement by two or three of the experts. A total of 51 defined abbreviations were pooled from three biological experts (experts 4 to 6) and the AbbRE, of which 44 were selected as the reference standard. Table 2 lists the results of part A of the evaluation for those defined abbreviations.

For defined abbreviations, as shown in Table 2, the average recall and precision of the three medical experts were 0.8 and 1.0, respectively; the recall and precision of AbbRE for medical articles were 0.78 and 0.97, respectively. Among the three medical experts, the overall agreement before and after pooled abbreviations was 0.70 and 1.00, respectively. The average recall and precision of the three biological experts were 0.79 and 0.96, respectively; the recall and precision of AbbRE for biological articles were 0.61 and 0.93, respectively. Among the three biological experts, the overall agreement before and after pooled abbreviations was 0.75 and 0.80, respectively. The recall and precision of AbbRE for both medical and biological articles were 0.70 and 0.95, respectively.

A total of 132 and 250 undefined abbreviations were selected by the experts from five medical articles and five biological articles, respectively, of which 132 and 137 were chosen as the reference standard. Therefore,
the percentages of abbreviations that were defined in five medical articles, five biological articles, and both medical and biological articles were 25 percent, 24 percent, and 25 percent, respectively. The overall agreements among medical experts before and after the pooled abbreviations were 0.42 and 1.00, respectively. The overall agreements among biological experts before and after the pooled abbreviations were 0.40 and 0.66, respectively.

In part B of the evaluation, AbbRE extracted 160 and 157 defined abbreviations and full forms from 20 medical articles and 20 biological articles, respectively, of which two or three experts agreed with 144 and 135 medical and biological abbreviations and full forms, respectively. Abbreviations selected by AbbRE on which the experts disagreed included of alternative medicine (oam) and gst fusion vector, cydr was first expressed as a gstfusion protein (gst-cydr).

We noticed that 3 medical abbreviations and full forms and 14 biological abbreviations and full forms were given question marks by experts because the full forms were attached to an HTML tag (e.g., presenilin 1nbsp was a full form of ps1). After we removed the HTML tag, all experts agreed with those abbreviations and full forms. We therefore added those abbreviations to the reference standard. Thus, the reference standard consisted of 147 and 149 medical and biological abbreviations and full forms, respectively.

The precision of AbbRE was 0.92 (95% CI, 0.90–0.94) and 0.95 (95% CI, 0.93–0.97) for medical and biological articles, respectively. The precision of AbbRE for both domains was 0.93 (95% CI, 0.92–0.94). Among the experts, the overall agreement for medical articles was 0.88; the overall agreement for biological articles was 0.94.

AbbRE failed to recognize some abbreviations and full forms selected by experts; we therefore manually mapped all the abbreviations selected by the experts and those included in the AbbRE output to their original articles and identified the causes of the failure.

We found that most abbreviations that failed to be recognized by AbbRE were not associated with their full forms through parentheses. Many abbreviations were defined not in the article body but in a special section of the articles. For example, the Journal of Biological Chemistry has a special abbreviation section that includes some chemical abbreviations and full forms (e.g., Cbz, benzylxoycarbonyl) that are not defined in the articles. Some abbreviations were defined in different parts of the articles. For example, AJT, which was used in the article body of a Lancet article, are the initials of the author, Andrew J. Thompson, which appeared in the author section of the article. Other abbreviations and full forms were not suitable to be mapped by the pattern-matching rules. An example was 100 mL 0.01 M phosphate buffer and 0.9% sodium chloride [PH 7.4], with 1.0 g bovine serum albumin and 0.1 mL Tween 20 (PBA).

Determination of the Percentage of Undefined Abbreviations That Could Be Mapped to Entries in Each of Four Abbreviation Databases

We randomly selected 30 undefined medical abbreviations and 30 undefined biological abbreviations from the reference standard described above, and manually identified the existence of these abbreviations in the four abbreviation databases—GenBank LocusLink, SWISSPROT, LRABR, and BioABACUS. Table 3 lists the numbers and percentages of these abbreviations that can be mapped to each database and to any of the four combined databases.

We observed that many abbreviations were covered by more than one database. For example, EDTA (ethylenediaminetetraacetic acid) was found in both LRABR and BioABACUS, and TRADD (TNFRSF1A-associated via death domain) was found in GenBank LocusLink, SWISSPROT, and BioABACUS. FELIX, SPSS, and U-test
are examples of abbreviations that could not be mapped to any of the four databases.

We also observed that many abbreviations were ambiguous. Different full forms of an abbreviation could be found within a database or across databases. For example, Ltd mapped to laron-type dwarfism, leukotriene d, and long-term disability in LRABR, lightoid in GenBank LocusLink, and Long-term Depression in BioABACUS.

Discussion

AbbRE achieved reasonable overall performance (recall 0.70, precision 0.95). The results indicate that AbbRE may be a useful tool for mapping defined abbreviations. However, the overall percentage of defined abbreviations may be small (average, 25 percent). Thus, it is unlikely that we will capture all the abbreviations in literature articles by applying AbbRE alone; other approaches need to be integrated.

We explored mapping undefined abbreviations to four abbreviation databases—GenBank LocusLink, SWISSPROT, LRABR, and BioABACUS. However, an average of only 68 percent of the undefined abbreviations could be mapped to any of four databases. Our results suggest that the four databases we tested do not provide exhaustive coverage and that we would need a more comprehensive abbreviation database to map undefined abbreviations effectively.

Our future research plans include the use of AbbRE itself to create a more comprehensive abbreviation database, either by applying it to a large body of electronic articles or to all the MEDLINE abstracts in PubMed, under the assumption that abbreviations are usually defined in the abstracts when they are first introduced into the literature. Another assumption is that even though not all the abbreviations in an article are defined in the abstract, they might be defined in the abstracts of other articles.

Our results indicate another obstacle to mapping undefined abbreviations to an abbreviation database: Some abbreviations have more than one full form. Abbreviations that have many forms are common. Abbreviations are not well standardized in medical, biological, or pharmaceutical science; each scientist uses his or her own judgment in choosing abbreviations. For example, in medicine, PID stands for both pelvic inflammatory disease and prolapsed intravertebral disc.1

Although researchers are working to standardize medical and biological abbreviations,1-4 the standardization is limited to specific domains, such as cardiology or vertebrate virus species. Therefore, the same abbreviation may become ambiguous when we search across several domains. For example, in molecular biology, CAT means chloramphenicol acetyl transferase; in computer science, it means computer-aided testing; in cell biology, it means computer-automated tomography; and in medicine, it means computed axial tomography.25

Disambiguating an abbreviation is a case of word sense disambiguation, the problem of resolving semantic ambiguity. There are many computational linguistic approaches, including lexicon and corpus-based approaches, to disambiguating the meaning of words.29-30 Most approaches, however, target the general English word, such as bank. Hatzivassiloglou et al.31 applied machine-learning techniques to disambiguating symbols to determine whether they represent proteins, genes, or RNA. However, the approach does not identify the meanings (or the full forms) of gene or protein symbols.

We propose identifying the knowledge domain to which an abbreviation belongs. The rational is that there are fewer ambiguous abbreviations within a knowledge domain than across knowledge domains. Thus, identifying the knowledge domain to which an abbreviation belongs may disambiguate the abbreviation. This approach requires a database that contains not only the abbreviation and its concept but also the knowledge domain.

One way to obtain the knowledge domain is to assign MeSH concepts to paired abbreviations and full forms. Each MEDLINE article has manually indexed MeSH concepts. The assigned MeSH concepts usually define the knowledge domain of its article. Therefore, the abbreviations used in the article are within the scope of the list of MeSH concepts. We may use AbbRE to extract defined abbreviations in abstracts, as well as the list of MeSH concepts indexed to the articles. (Assigned MeSH concepts are available in electronic format along with the abstracts.)

When a particular abbreviation is not defined in an article, we may map this abbreviation, as well as the list of MeSH concepts indexed to the article, to the abbreviation database we developed, by using AbbRE to determine the actual meaning of the abbreviation. In addition, context-based disambiguation may also be a way to disambiguate abbreviations.29-31

Another approach to identifying the full forms of undefined abbreviations is to link the abbreviations to citations to the articles in which they appear, to refer-
ferences in the articles in which they appear, and to related articles; all functions are provided by PubMed. The assumption is that all the abbreviations must be defined in the articles when the abbreviations are first introduced in literature, and those articles may be listed in the citations. Both citation and related-articles approaches were applied and evaluated to sufficiently improve information retrieval in other systems.32,33

Our results indicate that AbbRE may enhance information retrieval by two means. First, AbbRE may be used to recognize the full forms of defined abbreviations; full-form recognition may increase term frequency, a measurement widely used in information retrieval, when the full form is used as the search term. The rationale is that we expect less occurrence of a full form in the article when its abbreviation is used in the article. Second, AbbRE may be used indirectly to recognize the full forms of undefined abbreviations, in that AbbRE may be applied to create an exhaustive abbreviation database, which may be used to map undefined abbreviations. The abbreviation database created by AbbRE may further facilitate abbreviation disambiguation.

In our study, we used the opinions of domain experts to evaluate the performance of AbbRE. Developing analyzers that yield a conceptual representation of biomedical narratives has long been a research topic in biomedical informatics.34–38 In order to validate the usage of the program, evaluation is a necessary step and a reference standard is needed for an evaluation. Usually, domain experts are chosen for that purpose.35–38 However, domain experts are human and therefore may be error prone themselves. In order to be fair to the computer program, we determined the reference standard by having experts re-evaluate pooled selections from both the experts and the AbbRE output.

We measured overall agreement to indicate the experts’ agreement. Our results showed that the overall agreements were different for defined abbreviations and undefined ones. For example, the overall agreements in the selection of defined abbreviations in both part A and part B evaluations were all above 0.70, and the overall agreements in the part B evaluation reached 0.88 and 0.94 for medical and biological articles, respectively. However, the overall agreements of both medical and biological experts in selecting undefined abbreviations were lower (0.42 and 0.40, respectively). The results indicated that experts are more likely to agree on defined abbreviations than on undefined abbreviations.

The results are consistent with the frustration many experts expressed in identifying whether a term was an abbreviation or a symbol. For example, experts disputed “pi,” “NiCl2512,” and “stage III” as abbreviations. Our results also indicate that the overall agreements among both medical and biological experts after pooled abbreviations were higher than before pooled abbreviations, and that the overall agreements in validating an abbreviation in part B of the evaluation were higher than the overall agreements in selecting an abbreviation in part A of the evaluation; the results suggest that experts agreed more in validating an abbreviation than in finding an abbreviation.39

Conclusion

Mapping an abbreviation to its full form in an electronic article is a nontrivial but important task. The mapping not only facilitates the natural language processing but also is important for information retrieval. In this study, we showed that a software program AbbRE was efficient for pairing abbreviations with full forms when the abbreviations were defined in the articles (0.70 recall and 0.95 precision). We determined, however, that only 25 percent of the abbreviations were defined in the articles and only 68 percent of undefined abbreviations could be mapped to any of the four databases (GenBank LocusLink, SWISSPROT, LRARBR, and BioABACUS). In addition, an abbreviation could be mapped to more than one full form in the databases. Future work will concentrate on creating an exhaustive abbreviation database and an algorithm for disambiguation. We also plan to apply AbbRE to all MEDLINE abstracts to create such an exhaustive abbreviation database, and we may link paired abbreviations and full forms to their knowledge domain to solve the problem of ambiguous abbreviations.

The authors thank Alexa McCray, Marie-Claude Blatter, Andrey Rzhetsky, and Vasileios Hatzivassiloglou for their contributions to the project.

References