BMC Bioinformatics

BioMed Central

Proceedings

Open Access

A comparison study on algorithms of detecting long forms for short forms in biomedical text

Manabu Torii[1], Zhang-zhi Hu[2], Min Song[3], Cathy H Wu[2] and Hongfang Liu*[1]

Address: [1]Department of Biostatistics, Bioinformatics, and Biomathematics, Georgetown University Medical Center, 4000 Resevoir Rd, NW, Washington, DC 20057, USA, [2]Department of Biochemistry and Molecular & Cellular Biology, Georgetown University Medical Center, 3300 Whitehaven St., NW, Washington, DC 20007, USA and [3]Department of Information Systems, New Jersey Institute of Technology, University Heights, Newark, NJ 07102, USA

Email: Manabu Torii - mt352@georgetown.edu; Zhang-zhi Hu - zh9@georgetown.edu; Min Song - min.song@njit.edu; Cathy H Wu - wuc@georgetown.edu; Hongfang Liu* - hl224@georgetown.edu

* Corresponding author

from First International Workshop on Text Mining in Bioinformatics (TMBio) 2006
Arlington, VA, USA. 10 November 2006

Published: 27 November 2007

BMC Bioinformatics 2007, **8**(Suppl 9):S5 doi:10.1186/1471-2105-8-S9-S5

This article is available from: http://www.biomedcentral.com/1471-2105/8/S9/S5

Abstract

Motivation: With more and more research dedicated to literature mining in the biomedical domain, more and more systems are available for people to choose from when building literature mining applications. In this study, we focus on one specific kind of literature mining task, i.e., detecting definitions of acronyms, abbreviations, and symbols in biomedical text. We denote acronyms, abbreviations, and symbols as short forms (SFs) and their corresponding definitions as long forms (LFs). The study was designed to answer the following questions; i) how well a system performs in detecting LFs from novel text, ii) what the coverage is for various terminological knowledge bases in including SFs as synonyms of their LFs, and iii) how to combine results from various SF knowledge bases.

Method: We evaluated the following three publicly available detection systems in detecting LFs for SFs: i) a handcrafted pattern/rule based system by Ao and Takagi, ALICE, ii) a machine learning system by Chang et al., and iii) a simple alignment-based program by Schwartz and Hearst. In addition, we investigated the conceptual coverage of two terminological knowledge bases: i) the UMLS (the Unified Medical Language System), and ii) the BioThesaurus (a thesaurus of names for all UniProt protein records). We also implemented a web interface that provides a virtual integration of various SF knowledge bases.

Results: We found that detection systems agree with each other on most cases, and the existing terminological knowledge bases have a good coverage of synonymous relationship for frequently defined LFs. The web interface allows people to detect SF definitions from text and to search several SF knowledge bases.

Availability: The web site is http://gauss.dbb.georgetown.edu/liblab/SFThesaurus.

Introduction

Much of the new knowledge relevant to biomedical research is recorded as free text in the form of journal articles or annotation fields of databases. The development of reliable natural language processing (NLP) systems, which retrieve relevant documents, extract relevant information, and mine new information from free text, can help biomedical researchers to better handle the overwhelming knowledge recorded in free text [1,2]. One critical component in those systems is the mapping of text strings (i.e., terms) to biomedical concepts. Because of the complexity of the biomedical domain, biomedical terms are often lengthy. They usually contain words that imply their corresponding semantic types, e.g., *virus* in *Epstein-Barr virus* or *protein* in *latent membrane protein*, or words that describe properties of referred entities such as *latent* in *latent membrane protein*. At the same time, for biomedical concepts such as genes or proteins, it may be difficult to come up with short and yet descriptive terms for them. To ease the communication, concise representations of biomedical concepts such as acronyms, abbreviations, and symbols have been used in text for biomedical concepts that either occur frequently or are difficult to describe. However, the use of concise representations has posed great challenges to NLP systems. First, it is difficult to automatically infer their semantic categories from the representation. For example, systems can detect *Epstein-Barr virus* representing a kind of virus but it would be difficult to infer the semantic type *virus* from its acronym, *EBV*. Secondly, concise representations can be highly ambiguous. For example, besides *Epstein-Barr virus*, *EBV* can also represent *estimated blood volume*, among others. In the following, we denote concise representations as short forms (SFs) and their corresponding definitions as long forms (LFs).

Usually, authors provide the corresponding LF of an SF in their writing using patterns such as parentheses or phrases such as *stands for*. For example, readers can know what *EBV* stands for in a document when introduced first as following in the document, *Epstein-Barr virus (EBV) is a member of*. However, parentheses can also be used for other purposes besides defining SFs. Several systems have been developed to detect parentheses used for defining SFs and extract the corresponding LFs [3-6] with F-measures reported as over 90%. However, authors may not always define SFs especially for well-known biomedical concepts in the domain. In this situation, automated systems or readers unfamiliar with the domain would need an SF knowledge base that lists all LFs associated with a given SF, and a method to associate the SF with the correct LF in a document. Some terminological knowledge bases such as the UMLS [7] have included SFs as synonyms of their LFs. For example, *EBV* and *Epstein-Barr virus* have been assigned to the same UMLS conceptual identifier *C0014644*. Terminological knowledge bases specifically listing the definitions of SFs also exist. For example, the file LRABR in the SPECIALIST lexicon provides definitions for SFs that are present in the lexicon.

This study was designed to answer several questions. First, how do various systems perform in detecting LFs for SFs from parenthetical expressions given a large collection of novel text? To avoid evaluating systems on their development data set, abstracts of recently published articles were used in the study. Secondly, what is the coverage for various terminological knowledge bases to record SFs as synonyms of their LFs? Since those terminological knowledge bases often contain rich semantic information about terms, it will be beneficial to map SFs and LFs to them. Additionally, how can we combine the results from various systems and SF knowledge bases? To answer those questions, we evaluated several LF detection systems that are publicly accessible using a corpus consisting of MEDLINE abstracts published between January 2006 and May 2006. We used two terminological knowledge bases, the UMLS and BioThesaurus, to see the coverage of LFs and the coverage of including SFs as synonyms of their LFs. In addition, we implemented a web interface that utilizes several SF knowledge bases for detecting or searching LFs.

Background

In the following, we provide background information about SFs in the biomedical domain, review studies published relevant to detecting LFs for SFs in text, and summarize two terminological knowledge bases in biomedicine.

SFs of biomedical concepts

SFs are universal phenomena, occurring in all languages and writings and they can be formed in several ways [8,9]:

Truncating the end, e.g. *adm* for *administration* (or *administrator*),

First letter initialization, e.g. *AAA* for *abdominal aortic aneurysm*,

Syllabic initialization, e.g. *BZD* for *benzodiazepine*,

Combination initialization, e.g. *ad lib* for *ad libitum*, and

Symbols/synonyms substitution or initialization e.g. *ASD I* for *Primum atrial septal defect*; *Fe* for *iron*.

In the clinical domain, writing favors brevity because time pressures often prevent medical specialists from describing clinical findings fully. Many clinical words and phrases are long, and SFs are a way to ease the communication [10,11]. In the biomedical domain, biological enti-

ties such as genes or proteins are usually represented as symbols in text which can be derived by initializing their descriptive names or assigned by nomenclature committees. Note that some of the symbols may never be defined in text [1,12].

Methods for detecting LFs and/or assembling SF knowledge bases

Existing methods for detecting LFs in text or assembling SF knowledge bases can be categorized into one or combination of the following four types: (i) alignment-based approach, (ii) machine learning approach, (iii) template/rule-based approach, and (iv) collocation-based approach. We summarize a few systems for each type in the following.

Alignment-based approach

The basic assumption of the alignment-based approach is that LFs can be found in neighboring phrases that subsume all or almost all the letters of the corresponding SF (in the same order). For example, Taghva et al. [13] developed a detection system based on the longest common subsequence (LCS) algorithm. Their system assumes that an SF consists of the initial letters of the words contained in the LF (the common letters should appear in the SF and the LF in the same order), and the system seeks candidate LFs of an SF accordingly. Yu et al. [14], in their study of abbreviations in biology and medical papers, used several patterns to detect LFs, which also reflects the alignment idea. Similarly, the method by Yoshida and colleagues [15] detects LFs based on the assumption that the first several letters of each syllable in the words of LFs constitute the corresponding SFs.

Another alignment-based algorithm is proposed by Schwartz and Hearst [4]. Given an SF candidate in parentheses, their algorithm seeks the shortest phrase that immediately precedes the parentheses and subsumes all the letters in the SF candidate in the same order, while the leftmost letter of the phrase and that of the SF should be the same. Despite its simplicity, the performance of their algorithm is highly competitive [3,5,6].

Machine learning approach

Machine learning has also been explored for LF detection. For example, Chang et al. [5] proposed a supervised machine learning approach to extract (SF, LF) pairs from MEDLINE abstracts. The system employs the LCS algorithm to search for different alignments between a candidate SF in parentheses and the text string preceding the parentheses. Alignments detected are then evaluated using a machine learning method (logistic regression), and the one yielding the highest score is considered as the LF. The features considered in their approach for machine learning include the ratio of SF letters that are aligned with the initial letters of the words in a candidate LF, and the ratio of SF letters that are aligned with the initial letters of the syllables in the words of a candidate LF, and among others. Nadeau and Turney [16] also used machine learning to select the best definition phrases among the set of candidate phrases that are assembled using heuristics based on previous studies [6,17].

Template/rule-based approach

Most studies using handcrafted templates/rules are for constructing SF knowledge bases from MEDLINE abstracts such as AcroMed [6], ARGH [18], and SaRAD [19]. Another example of using templates/rules is ALICE [3] which includes templates/rules for 320 different patterns. ALICE assembled several sets of stop words to avoid proposing SFs and LFs containing inappropriate words. For example, one of the sets contains stop words (e.g., *of*) for the leftmost word of LFs. Note that some studies such as the work by Yu et al. [14] in alignment-based approach can also be considered as the template/rule-based approach.

Collocation-based approach

Motivated by the fact that the majority of LFs for SFs have been defined using parenthetical expressions many times in a large corpus and the parenthetical expressions can be considered as collocations, we extracted an SF knowledge base from parenthetical expressions in MEDLINE abstracts using a collocation-based approach [20]. Okazaki and Ananiadou [21] and Zhou et al. [22] also used a collocation-based approach to build SF knowledge bases. One advantage of collocation-based approaches is the correct detection of LFs for SFs that are created through symbols/synonyms substitution/initialization. For example, collocation-based methods can successfully detect the definition for *1H-MRS* is *proton magnetic resonance spectroscopy* where 1H is a symbol for proton.

Knowledge bases

As we have discussed, several knowledge bases for SFs have been constructed automatically using MEDLINE abstracts. Other resources to obtain SF knowledge are biomedical terminology sources. Those sources contain synonym relationship between terms, and SFs can be considered as synonyms of corresponding LFs. Since not all SFs are defined in text, it is important to have such knowledge bases. In this study, we used two terminological knowledge bases in the biomedical domain: the Unified Medical Language System (UMLS) and BioThesaurus.

The UMLS [7] contains terms from a set of large scale terminological knowledge sources in biomedicine. Among many components in the UMLS, we used MetaThesaurus, which associates synonyms with unique concept identifi-

ers, and the file LRABR in the SPECIALIST lexicon that associates SFs with the corresponding LFs.

BioThesaurus is a knowledge source providing mapping between gene/protein names and protein entries in UniProtKB, the most comprehensive protein knowledge base [23]. Through unique accession numbers assigned to each protein entry in UniProtKB, BioThesaurus groups terms referring to the same gene/protein entities.

In the general English domain, one popular SF site is AcronymAttic (Table 1), which consists of 2,982,000 human-edited entries and it was claimed to be the world's largest and most comprehensive dictionary of acronyms. Other online resources include Special Dictionary, http://abbreviations.com, and http://acronyma.com (see Table 1).

Methods

As we have discussed, there are four types of approaches for detecting LFs for SFs. However, it is not clear how well they perform given novel text. We conducted a comparison study to evaluate their performance in detecting LFs. The method involves several steps. The first step is to identify systems that are publicly accessible. The second step is to define common criteria to select candidate sentences for detecting LFs. Because different systems may use different criteria for selecting candidate sentences, it is important to include only sentences that all systems consider them as candidate sentences. The third step is to obtain a list of candidate sentences from MEDLINE abstracts published between January 2006 and May 2006. We then ran the systems on these sentences, followed by a detail assessment of the results.

Table 1 provides a summary of systems and resources used in the study. We used three systems in our comparison study in detecting LFs given novel text (indicated using "*"): i) a handcrafted pattern/rule based system by Ao and Takagi, ALICE, ii) a machine learning system by Chang et al., and iii) a simple alignment-based program by Schwartz and Hearst. Note that we did not include the collocation-based approach in the comparison study since it is not suitable for detecting LFs in text but for assembling SF knowledge bases from a large corpus. As introduced in the previous section, the CSA system proposes an LF for a given SF with a score (between 0 and 1). Higher scores indicate more confidence in detecting LFs. After reviewing the systems' outputs as well as reported performance [2,6], we chose ≥0.03 as the threshold for the system to propose SFs and LFs.

Candidate SF detection

To compare how well each system performs in detecting LFs, we focused on sentences containing parentheses. As we have indicated, different systems have different criteria in considering a text string as a candidate SF. For instance, CSA proposes a phrase containing a comma-space sequence as an SF, but the other two systems discard tokens after a comma or semicolon, e.g., *BMI, in kg/m2* vs. *BMI* given *body mass index* (*BMI, in kg/m2*) in the example below. CSA also recognizes an SF with only one letter, while S&H requires SF candidates containing at least two letters. S&H does not recognize SF candidates in sentences containing nested parentheses.

Example 1.*The objective was to describe the association of waist circumference (WC) and body mass index (BMI; in kg/m2) with plasma circulating oxidized LDL (ox-LDL) and C-reactive protein (CRP) [PMID:16400046].*

In order to investigate how well each system associates LFs with SFs without being confused with different schemes to identify SF candidates, we only consider sentences where all three systems attempt to detect LFs. Specifically, we consider an occurrence of parentheses for detection when the text string inside parentheses consists only of alphabetic letters, numbers or hyphen, and contains at least one upper case letters with a total length between two and ten inclusive.

Table 1: Systems and SF search engines considered in the study. The sign * indicates the system was used for the comparison study. The sign + indicates the system was included by the web interface.

System/Resource	Reference	Method	Website
ALICE*	Ao and Takagi	Templates/rules	http://uvdb3.hgc.jp/ALICE/program_download.html
ARGH+	Wren & Garner	Templates/rule	http://invention.swmed.edu/argh
CSA* (BAS+)	Chang	Machine Learning	http://abbreviation.stanford.edu/ (BAS)
S&H*	Schwartz & Hearst	Alignment	http://biotext.berkeley.edu/software.html
ADAM+	Zhou	Collocation	http://128.248.65.210/arrowsmith_uic/adam.html
Acromine+	Okazaki & Ananiadou	Collocation	http://www.nactem.ac.uk/software/acromine/
AcronymAttic+	NA	NA	http://www.acronymattic.com
Special Dictionary (Acronym)+	NA	NA	http://www.special-dictionary.com/acronyms/
Abbreviation+	NA	NA	http://www.abbreviations.com/
ACRONYMA+	NA	NA	http://www.acronyma.com

Evaluation data

The MEDLINE records published between January 2006 and May 2006, which are not part of the development/training corpora for any of the three systems, were used for evaluation.

For each MEDLINE record (identified using PubMed identifier PMID), we used a Perl script to extract all sentences containing parentheses with candidate SFs. Each occurrence of parentheses with candidate SFs can be uniquely identified as (PMID, SF) where PMID is the PubMed unique identifier and SF is the text string inside parentheses found in the corresponding MEDLINE record. For example, there are four parentheses associated with the sentence shown in Example 1, where three of them are considered as candidate definition occurrences (identified as (16400046, *WC*), (16400046, *ox-LDL*), and (16400046, *CRP*)), and one, i.e., *(BMI; in kg/m2)* is not considered since it contains characters other than letters, numbers, and hyphen.

Assessment of LF detection

After we obtained a list of candidate sentences containing at least one candidate SF occurrence, we then obtained LF detection results. For ALICE and S&H, we downloaded the programs available in their project web pages (see Table 1), and executed them locally [3,4]. For CSA, we submitted the sentences to the system running on their project web site (see Table 1).

For each system, we obtained a collection of tuples (PMID, SF, LF), where the pair (PMID, SF) indicates the candidate definition occurrence and LF denotes the long form proposed by the system. We derived a Venn diagram to show the overlapping information about the three collections. For each area in the Venn diagram, we sampled 100 instances and manually judged the detection accuracy. Note that one candidate occurrence (PMID, SF) may correspond to multiple tuples if different systems extract different LFs. We also provided an analysis on those occurrences.

We then predicted the recall of the systems using the UMLS and BioThesaurus as knowledge sources. For pairs (PMID, SF) where none of the systems propose any LF, we used the UMLS MetaThesaurus as a knowledge source of synonyms to find out missing LFs using the following steps:

Look up a given SF (e.g., *APAP*) candidate in the UMLS MetaThesaurus. If found, record the corresponding concept IDs (CIDs) (e.g., *APAP* → CID: C0000970.)

Gather all the phrase strings associated with the recorded CIDs, e.g., CID: C0000970 → {"*APAP*", "*Acetaminophen*, *N-(4-Hydroxyphenyl)acetanilide*",...}

Look up any of the gathered strings in the text prior to the corresponding parenthetical expression. If found and it is longer than the SF candidate, propose it as the LF for the given SF.

When looking up SFs or LFs in text or knowledge bases, we first tokenize phrases, where tokens can be words, numbers, or special tokens such as Roman numerals, Greek letters, or digits. Tokens are then normalized by converted into base forms using the UMLS SPECIALIST lexicon. We also ignored case difference during the lookup.

We estimated the coverage of two existing terminological knowledge bases, the UMLS and BioThesaurus, regarding to LFs and synonymous relationships between SFs and the corresponding LFs. For pairs (SF, LF) agreed by the three systems, we grouped them according to their frequencies in the result collection (i.e., the number of occurrences of the corresponding tuples (PMID, SF, LF)). For each group, we measured (i) coverage of LFs, and (ii) coverage of pairs (SF, LF) as synonyms.

Results and discussion

Statistics and performance comparison

The MEDLINE evaluation dataset contains about 210 thousands records with a set of candidate sentences containing about 258 thousands candidate SF occurrences. Figure 1 shows the Venn diagram of the results where each area is labeled with Roman numerals (i.e., I, II,...,VII). For example, there are totally four areas associated with ALICE (i.e., I, II, III, V) with a total number of tuples as 226,684 (the summation of 214,886, 896, 3,978, and 6,924).

From Figure 1, we can see that the three systems agreed with each other for a large portion of the tuples (over 94%). S&H detected as many tuples as the other two elaborated systems did, though the algorithm of S&H is very simple. This significant overlap is mainly because most SFs were obtained through various kinds of initialization as discussed in the background session. For example, there are 87,057 unique pairs (SF, LF) corresponding to Area I (i.e., detected by all three systems), 61% of the SFs were formed through First Letter Initialization from their corresponding LFs, e.g., *AAA* for *abdominal aortic aneurysm*.

The estimation of the precision for each area is also shown in Figure 1. For example, the precision in Area I is 100% when assessed using 100 randomly sampled tuples. We found that generally the more systems detect the same LF, the more accurate the detection is. For example, the detection in areas III and IV tends to be reasonably accurate

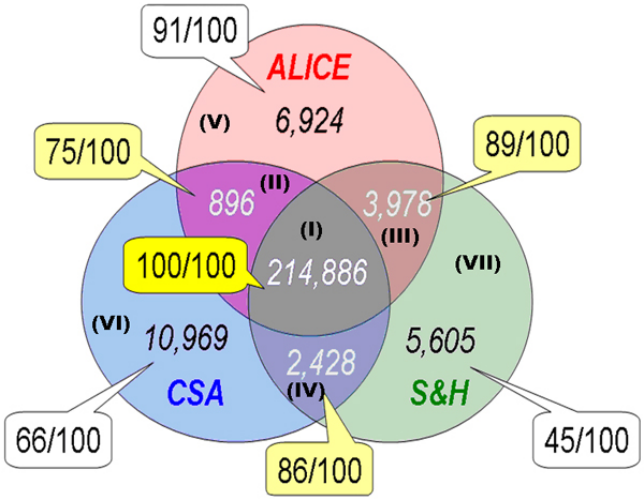

Figure 1
Venn diagram of the results obtained from three systems: ALICE, CSA, and S&H. Each area is labeled with Roman numerals (i.e., I, II,...,VII). Statistics for each area includes the number of tuples (PMID, SF, LF) and the estimation of the precision for each area.

(i.e., 89% and 86%) when S&H and one other system proposed the same LF for a given candidate SF occurrence.

Table 2 shows the statistics of SF candidate occurrences associated with multiple LFs for different combinations and the number of correct detection associated with each system when assessed using 53 pairs randomly sampled. For example, the third column of the third row indicates that there are totally 2,572 SF candidate occurrences where ALICE and CSA proposed the same LF and S&H proposed a different one. When inspecting 53 pairs randomly selected from 2,572 occurrences, ALICE and CSA proposed the correct LFs for 51 pairs, S&H proposed the correct LF for one pair, and none of the systems proposed the correct LF for one pair.

Table 2: Statistics of SF candidate occurrences when multiple LFs proposed by ALICE, CSA, and S&H. The same superscript indicates the two systems proposed the same LFs. The superscript 0 indicates the corresponding system did not propose an LF.

Cases	# Correct (of 53)	# SF candidate occurrence
ALICE^1, CSA^2, S&H^3	17, 20, 13	53
ALICE^1, CSA^1, S&H^3	51, 51, 1	2,572
ALICE^1, CSA^2, S&H^1	36, 11, 36	765
ALICE^1, CSA^2, S&H^2	33, 18, 18	901
ALICE^1, CSA^2, S&H^0	27, 12, ---	167
ALICE^1, CSA^0, S&H^2	37, ---, 6	160
ALICE^0, CSA^2, S&H^3	---, 13, 28	325

When predicting the recall of the systems using the UMLS and BioThesaurus, we obtained 21,657 SF candidate occurrences where none of the systems proposed LFs. After mapping, we found 1,029 pairs with a total of 396 unique pairs failed to be detected by all three systems according to the synonymous relationship in the knowledge bases. The most frequently observed pairs were *5-HT* for *serotonin* (111 times), *Pb* for *lead* (44 times), *CsA* for *cyclosporine* (33 times). Additionally, several chemical names such as *RDX* for *hexahydro-1,3,5-trinitro-1,3,5-triazine* were observed. A few pairs were observed where the derivation of SFs from the LFs involves substitution of synonymous tokens, e.g., *NIS* for *sodium iodide symporter* (*sodium* → *natrium*), or reordering of words, e.g., *PLAU*for *urokinase plasminogen activator*. We found that over 70% of the detected unique pairs were ones where the SFs could not be implied from LFs through initialization, e.g., *5-HT* and *serotonin*. To identify these pairs and to properly handle them in various applications, it is necessary to incorporate terminology knowledge bases such as the UMLS, or to explore the collocation-based approach [20,22].

Finally, given a precision over 100 instances for different partitions in the Venn diagram (Figure 1), we may speculate the precision of the three systems for the entire data set, e.g., ALICE proposed 214886 (I) + 896 (II) + 3978 (III) + 6924 (V) pairs, and considering the corresponding precisions in Figure 1, 214886*1.0 + 896*0.75 + 3978*0.89 + 6924*0.91 may be correct pairs. Similarly, incorporating the recall study using the UMLS above, ALICE, CSA, and S&H would achieve recalls of 97%, 96%, and 96%. These performance measures are much higher than the precision and recall values reported before on these systems. One reason for the better performance may be our highly selective choices of sentences that were passed to the systems for evaluation.

As we have shown, most problematic cases (i.e., inconsistent or failed detection of LFs among the systems) are chemical/protein/gene symbols. It may be caused by the following reasons: the symbols may be assigned by nomenclature committee or be created through symbols/synonyms initialization/substitution.

Coverage of terminological knowledge bases

Figure 2 shows the coverage results where X-axis is the frequency bin with the $[2^n, 2^{n+1})$ where n from 0 to 8, the first Y-axis (left side) is the coverage and the second Y-axis (right side) is the number of unique pairs. From Figure 2, we can see that there are 66 thousands of unique pairs (SF, LF) defined only once in our data with the corresponding coverage of LFs as 31% and the corresponding coverage of pairs (SF, LF) around 11%. The coverage increases when the frequency of a pair being defined increases. For example, for pairs being defined [16, 32) times, the coverage of

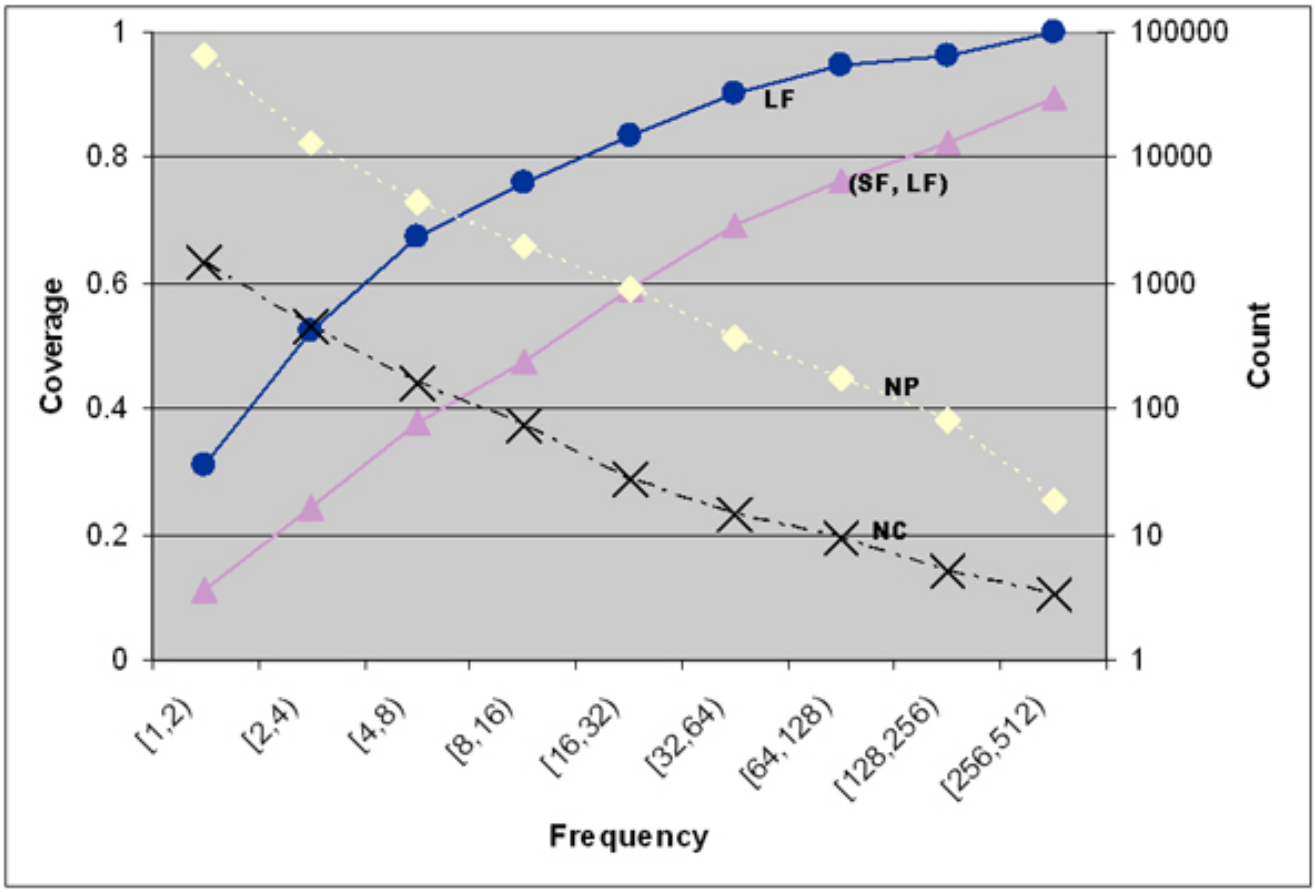

Figure 2
Results of the coverage study. X-axis is the frequency bin $[2^n, 2^{n+1})$ where n from 0 to 8, the first Y-axis (left side) is the coverage and the second Y-axis (right side) is the number of unique pairs. Four lines mean: Line NP – the total number of unique pairs for each bin. Line LF – the percentage of unique pairs (SF, LF) where LF can be mapped to the knowledge base. Line (SF, LF) – the percentage of pairs where the synonymous relationship between SF and LF can be inferred. Line NC – the percentage of pairs where LF cannot be mapped to knowledge bases.

LFs increased to 83% and the coverage of (SF, LF) pairs increased to 59%. For 19 pairs defined [256, 512) times, all LFs were mapped to the UMLS (i.e., 100% coverage) and 17 (SF, LF) pairs were mapped to the UMLS (i.e., 90% coverage).

Although the coverage of existing terminological knowledge bases was found to be low for less frequently defined LFs, they have an advantage of associating LFs with SFs that are not derived through simple initialization (e.g., *1H MRS* for *Proton Magnetic Resonance Spectroscopy*). Also, the mapping of LFs to existing terminological knowledge bases provides useful additional information, such as semantic type information (e.g., UMLS Semantic Types) or links to biological sequence databases (e.g., BioThesaurus).

A web interface for LF detection and search

Observing that i) different LF detection methods may propose different LFs, and ii) the more systems proposed the same LF, the more accurate the detection is, we have implemented a web interface so that users can search for LFs associated with a given SF from different SF knowledge bases. It is a virtual integration of various SF knowledge bases including ones assembled from MEDLINE abstracts and ones available in the general English domain (indicated using "+" in Table 1).

The web interface provides two functions. One is to use various SF knowledge bases to detect definitions in text for SFs in parenthetical expressions. Given a document, the system searches each SF candidate and retrieves corresponding LFs from those knowledge bases. For each

knowledge base, it compares retrieved LFs to the text prior to the corresponding parenthetical expression and associates the one that appears in the text. When multiple overlapping LFs found, the system returns ones with the highest precedence score in those SF knowledge bases applicable (e.g., the frequency of SF being defined as LF in MEDLINE). For example, when searching ARGH, we retrieved 229 LFs for CRP including *C-reactive protein*, *cAMP receptor protein, cyclic AMP receptor protein*. Comparing to the text prior to the parenthetical expression in Example 1, we consider *C-reactive protein* as the associated LF in the given text.

Another function is to search various SF knowledge bases. For a given SF, we send queries to all search engines and provide a table summarizing the retrieved results where syntactic variants have been grouped and LFs were ranked according to the number of SF knowledge bases containing them. Figure 3 shows the screenshots of the web interface. Note that the last column of the result tables indicates the existence of synonymous relationship between SF and LF in BioThesaurus (i.e., if SF and LF can be mapped to the same protein entity in BioTheasurus, we consider BioThesaurus captures the synonymous relationship). We do not include the UMLS in the result tables of our interface due to issues relevant to the UMLS license agreement.

Conclusion

In this work, we conducted a comparison study of three LF detection systems, ALICE, CSA, and S&H, which reflect three different approaches in LF detection for SFs. We observed that the majority of (SF, LF) pairs in MEDLINE abstracts were formed in a relatively simple way (i.e., ini-

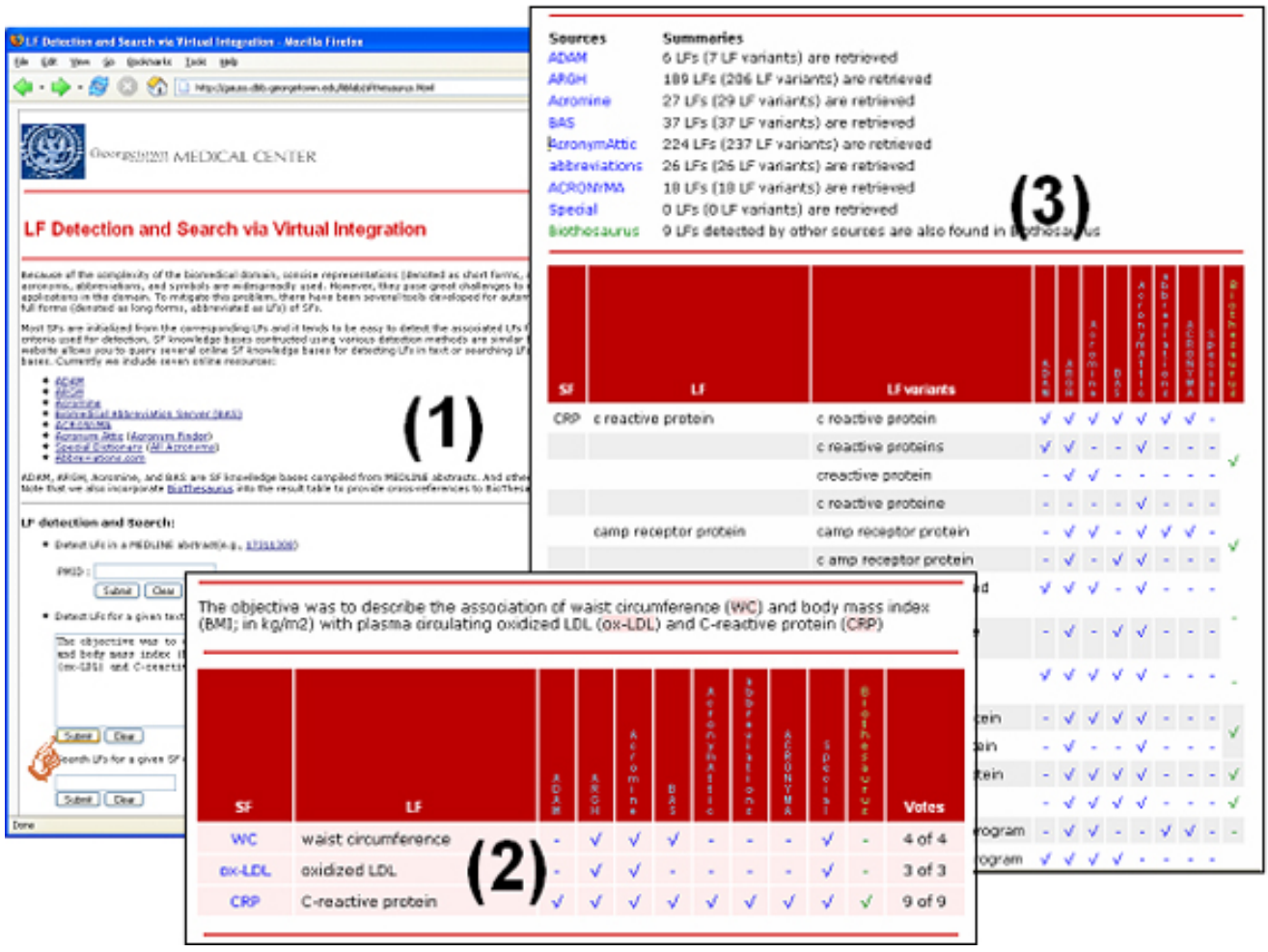

Figure 3
Screen shots of the web interface for virtual integration of various SF knowledge bases. 1. The main page of the web interface which provides two functions: i) LF detection in text, and ii) LF search from various knowledge bases. 2. Results for LF detection in text for Example 1. 3. Results for LF search from various knowledge bases.

tialization) and can be detected by almost all three systems. We also investigated the coverage of existing terminology knowledge sources, namely the UMLS and BioThesaurus, and the results showed that they have better coverage for pairs that are frequently defined. We implemented a web interface to provide virtual integration of SF knowledge bases derived using various detection methods in the biomedical domain or those available in the general English domain. We are currently working on incorporating a semantic category classification system so that users can limit their LF search to certain semantic categories.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Manabu Torii – Overall development and implementation of the software for the study and preparation of substantial portions of the manuscript. Zhang-zhi Hu – Input regarding to the conceptual design and help with the manuscript preparation. Min Song – Input to the conceptual design and assist with the manuscript preparation. Cathy Wu – Input regarding to the conceptual design and help with manuscript preparation. Hongfang Liu – Overall conceptual design and supervision of the project, and preparation of the final manuscript.

Acknowledgements

The project was supported by IIS-0639062 from the National Science Foundation. We thank the authors of the three systems used in the comparison study (Hiroko Ao, Jeffrey Chang, Ariel Schwartz, and each of their colleagues) for making their systems available. We also thank for the authors, developers, or maintainers for making their search engines available.

This article has been published as part of *BMC Bioinformatics* Volume 8 Supplement 9, 2007: First International Workshop on Text Mining in Bioinformatics (TMBio) 2006. The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/8?issue=S9.

References

1. Hirschman L, Morgan AA, Yeh AS: **Rutabaga by any other name: extracting biological names.** *Journal of Biomedical Informatics* 2002, **35(4):**247-259.
2. Hunter L, Cohen KB: **Biomedical language processing: what's beyond PubMed?** *Mol Cell* 2006, **21(5):**589-594.
3. Ao H, Takagi T: **ALICE: an algorithm to extract abbreviations from MEDLINE.** *J Am Med Inform Assoc* 2005, **12(5):**576-586.
4. Schwartz AS, Hearst MA: **A simple algorithm for identifying abbreviation definitions in biomedical text.** *Pac Symp Biocomput* 2003:451-462.
5. Chang JT, Schutze H, Altman RB: **Creating an online dictionary of abbreviations from MEDLINE.** *J Am Med Inform Assoc* 2002, **9(6):**612-620.
6. Pustejovsky J, Castaño J, Cochran B, Kotecki M, Morrell M, Rumshisky A: **Extraction and Disambiguation of Acronym-Meaning Pairs in Medline.** *Medinfo* 2001, **10:**371-375.
7. Bodenreider O: **The Unified Medical Language System (UMLS): integrating biomedical terminology.** *Nucleic Acids Res* 2004:D267-270.
8. Zahariev M: **A (Acronyms).** In *Unpublished PhD thesis School of Computing Science, Simon Fraser University, USA*; 2004.
9. Liu H, Lussier YA, Friedman C: **Disambiguating ambiguous biomedical terms in biomedical narrative text: an unsupervised method.** *J Biomed Inform* 2001, **34(4):**249-261.
10. Luxton T, Al-Qassab H: **Better use of abbreviations - a lesson from a stroke.** *Medical Education* 2000, **34(11):**965.
11. Bloom DA: **Acronyms, abbreviations and initialisms.** *BJU International* 2000, **86(1):**1-6.
12. Eyre TA, Ducluzeau F, Sneddon TP, Povey S, Bruford EA, Lush MJ: **The HUGO Gene Nomenclature Database, 2006 updates.** *Nucleic Acids Res* 2006:D319-321.
13. Taghva K, Gilbreth J: **Finding Acronyms and Their Definitions.** *Int Journal on Document Analysis and Recognition* 1999, **1(4):**191-198.
14. Yu H, Hatzivassiloglou V, Rzhetsky A, Wilbur WJ: **Automatically identifying gene/protein terms in MEDLINE abstracts.** *J Biomed Inform* 2002, **35(5–6):**322-330.
15. Yoshida M, Fukuda K, Takagi T: **PNAD-CSS: a workbench for constructing a protein name abbreviation dictionary.** *Bioinformatics* 2000, **16(2):**169-175.
16. Nadeau D, Turney P: **A Supervised Learning Approach to Acronym Identification.** *18th Conference of the Canadian Society for Computational Studies of Intelligence: 2005; Victoria, BC, Canada* 2005:319-329.
17. Park Y, Byrd RJ: **Hybrid Text Mining for Finding Abbreviations and Their Definitions.** *Conference on Empirical Methods in Natural Language Processing: 2001; Pittsburgh, PA* 2001.
18. Wren JD, Garner HR: **Heuristics for identification of acronym-definition patterns within text: towards an automated construction of comprehensive acronym-definition dictionaries.** *Methods Inf Med* 2002, **41(5):**426-434.
19. Adar E: **SaRAD: a Simple and Robust Abbreviation Dictionary.** *Bioinformatics* 2004, **20(4):**527-533.
20. Liu H, Friedman C: **Mining terminological knowledge in large biomedical corpora.** *Pac Symp Biocomput* 2003:415-426.
21. Okazaki N, Ananiadou S: **Building an abbreviation dictionary using a term recognition approach.** *Bioinformatics* 2006.
22. Zhou W, Torvik VI, Smalheiser NR: **ADAM: another database of abbreviations in MEDLINE.** *Bioinformatics* 2006, **22(22):**2813-2818.
23. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, *et al.*: **The Universal Protein Resource (UniProt).** *Nucleic Acids Res* 2005:D154-159.