A Study of Abbreviations in MEDLINE Abstracts

Hongfang Liu[1] MS, Alan R. Aronson[2] PhD, Carol Friedman[3,4] PhD
[1]Computer Science Division, Graduate School and University Center of CUNY
[2]National Library of Medicine
[3]Computer Science Department, Queens College of CUNY
[4]Department of Medical Informatics, Columbia University

Abstract

Abbreviations are widely used in writing, and the understanding of abbreviations is important for natural language processing applications. Abbreviations are not always defined in a document and they are highly ambiguous. A knowledge base that consists of abbreviations with their associated senses and a method to resolve the ambiguities are needed. In this paper, we studied the UMLS coverage, textual variants of senses, and the ambiguity of abbreviations in MEDLINE abstracts. We restricted our study to three-letter abbreviations which were defined using parenthetical expressions. When grouping similar expansions together and representing senses using groups, we found that after ignoring senses where the total number of occurrences within the corresponding group was less than 100, 82.8% of the senses matched the UMLS, covered over 93% of occurrences that were considered, and had an average of 7.74 expansions for each sense. Abbreviations are highly ambiguous: 81.2% of the abbreviations were ambiguous, and had an average of 16.6 senses. However, after ignoring senses with occurrences of less than 5, 64.6% of the abbreviations were ambiguous, and had an average of 4.91 senses.

Introduction

Abbreviations are everywhere; we read and hear them but rarely think about them, except when we do not know what they mean[1]. However, the understanding of abbreviations in a document is often a difficult task for computer systems. The abbreviation problem has been shown to affect knowledge-based systems, such as information retrieval systems and information extraction systems in biomedicine[2-4].

First, a method to associate an abbreviation to its corresponding expansion (also termed as full form or definition) in the context is needed, with an assumption that the authors define abbreviations when they are first introduced in a specific domain for the less well-known senses of abbreviations. Secondly, well-known senses of abbreviations are not always defined in documents. In order to understand these, an abbreviation database that lists abbreviations together with their senses needs to be built and updated periodically. However, manually constructing a database is time-consuming. In addition, manual maintenance and further extension are increasingly complex. But automatic construction of an abbreviation database requires a method to identify senses (i.e., expansions) in documents, a method to group textual variants of the same sense together and a method to link them to the proper sense in the corresponding sense inventory. Additionally, abbreviations are highly ambiguous: one abbreviation may represent dozens of senses. A method to resolve the sense ambiguity is needed.

Based on several previous studies[5;6], we found that the UMLS contained many abbreviations together with their expansions, and the ambiguity of abbreviations could be resolved using an automated method if the corresponding expansions occurred frequently or if they were UMLS concept names.

In this paper, we address the following issues with respect to abbreviations in MEDLINE abstracts by using three-letter abbreviations: can an abbreviation knowledge base be built from MEDLINE abstracts, hat is the UMLS concept coverage of the corresponding senses, what is the average number of textual variants for each sense, how ambiguous are the abbreviations, and what is the role of the frequency of the senses in the above issues?

Background and related work

There are several studies on matching abbreviations to their corresponding expansions in documents. Taghva et al.[7] developed an algorithm that considers strings of from 3 to 10 uppercase letters as acronyms, and looks for candidate expansions in windows of twice the number of letters in the acronym before or after. Larkey et al.[1] implemented a Web server for abbreviations, where abbreviations and their expansions were gathered automatically from a large number of Web pages. Yoshida and colleagues[8] built a workbench for the construction of a protein abbreviation dictionary. Yu and colleagues[9] developed a program to extract expansions of abbreviations from full articles. All the above studies achieved precision of over 97% when matching abbreviations to their expansions in documents. However, none of them provide a detailed analysis of characteristics of abbreviations with respect to senses.

In order to pursue our study, we developed a method, PW3, which was based on Larkey's method, for three-letter abbreviations where the associated expansions were defined in parenthetical expressions. We did not conduct an evaluation of PW3 since our primary goal was to address the characteristics of abbreviations as mentioned in the Introduction Section.

In the following, we introduce the background knowledge of resources and previous programs used in this study.

The UMLS integrates various vocabularies pertaining to biomedicine. The Metathesaurus[10] is one component of the UMLS. It contains information about biomedical concepts and terms from many controlled vocabularies. The SPECIALIST Lexicon, an English language lexicon, is another component of the UMLS.

MEDLINE[11] 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 a unique MEDLINE identifier and the citation information for the corresponding journal article, and often an abstract.

The UMLS abbreviation extraction program[5] is a program that extracts expansions of abbreviations in the UMLS based on several patterns. In addition it uses the fact that abbreviations are considered synonyms of their expansions in the UMLS.

MetaMap[3] is a highly configurable program that maps biomedical text to concepts in the Metathesaurus. It contains several knowledge bases including a synonym set and several subsets of normalized concept names in the Metathesaurus. Options control MetaMap's internal behavior, such as how aggressive to be in generation of word variants, whether to respect or to ignore word order, and how aggressive to be in matching to the Metathesaurus concepts. The primary goal of the MetaMap program was to improve retrieval of bibliographic material such as MEDLINE citations.

Methods

There are several reasons we used three-letter abbreviations for this study. First, a method for pairing three-letter abbreviations with their expansions is easy to develop and has high precision according to Larkey et al.[1]. Secondly, a preliminary investigation showed that three-letter abbreviations were the most frequent in MEDLINE abstracts. In addition, unlike two-letter abbreviations, which can have several dozens of expansions, the ambiguity of three-letter abbreviations is moderate, whereas most abbreviations with more than 3 letters are not ambiguous.

The study contained several steps. The first step derived a collection of (ABBR, EXP, FREQ) tuples, where ABBR is a three-letter capitalized text string, EXP is its associated expansion derived from abstracts using PW3 (the program we developed to pair three-letter abbreviations with expansions), and FREQ is the number of abstracts in which PW3 derived the pair. The second step mapped the expansions to the UMLS using EXPMap (a program we developed based on MetaMap). The third step grouped similar expansions for the same abbreviation together using EXPGrouper (a program we developed to group similar expansions together according to several normalization criteria). The fourth step assessed results, where expansions in the same group were treated as textual variants for the same sense. In the following, we describe PW3, EXPMap, and EXPGrouper in detail. The assessment method is presented last.

PW3: is a matching method for three-letter abbreviations ABBR and is designed to search for a possible expansion from candidate text strings within a window size 6 at the left side of a parenthetical expression "(ABBR)". It applies several expansion patterns of ABBR and three groups of words that can be ignored when matching patterns.

The expansion patterns include the following several cases:

- Three letters of ABBR are initial letters of three different words in the right order: e.g. *minimum alveolar anesthetic concentration (MAC),*
- Two letters of ABBR are initial letters of two words and the remaining one appears in one of these two words in the right order followed by at least three letters*:* e,g. *procoagulant activity(PCA) or indirect immunofluorescence (IIF),*
- Three letters of ABBR appear in one word where the first one is the initial letter of the word and remaining two appear in the right order: e.g. *carboxymethyllysine (CML).*

PW3 has an additional pattern for potential chemical abbreviations, where ABBR is considered to be a chemical abbreviation if a candidate string contains a number (or a comma or right parenthesis) followed by a non-space letter or a left parenthesis preceded by a non-space letter:

- Two letters of ABBR are initial letters (or following punctuations and numbers) and the remaining letter appears in the corresponding candidate string: e.g. *n-6-(delta-2-isopentenyl)adenine (IPA)*.

The three groups of words which can be ignored when matching patterns are pre-inclusion words (i.e. a word at the beginning of an expansion, such as *department, office* etc), post-inclusion words (i.e. a word at the end of an expansion, such as *acid, protein, enzyme* etc), and a dozen of stop words (a word in the middle of an expansion, such as *of, for, and, the* etc). PW3 allows one pre-inclusion word, one post-inclusion word, one other word, or two stop words in an expansion. The number of words in the expansion is at most 6. These three groups words were learned from the three-letter abbreviations in the SPECIALIST Lexicon manually.

EXPMap: is developed based on MetaMap. EXPMap uses the following subset of concept names in the Metathesaurus: chemical names, concept names that contain less than 7 words after a normalization process, and expansions obtained by executing the UMLS extraction program based on patterns as well as expansions in the SPECIALIST Lexicon abbreviation table. All concept names are normalized by removing some patterns (e.g. *As –* in *As – Arsenic* and *(WS)* in *West syndrome (WS)*), changing to lower-case, and replacing certain punctuation by blanks. In addition, EXPMap applies a synonym-like set, which contains pairs (w_1, w_2), where w_1 and w_2 are different words in two concept names of the same UMLS concept. For example, (*hepatic, liver*) is a synonym-like pair, which is derived from two concept names of C0009714, *congenital hepatic fibrosis* and *congenital fibrosis liver*.

The input to EXPMap is a pair (ABBR, EXP) and the output is (ABBR, EXP, CUI, PN, MODE), where ABBR is an abbreviation, EXP is an expansion, CUI is the resulting concept identifier, PN is the preferred name of that concept, and MODE is the matching mode, which can have one of the following four values:

Exact -- a concept name of CUI is identical to EXP, e.g., *(BAL, bioartificial liver, C0336562, artificial liver, exact)*;

SPECIALIST-normalized--a concept name of CUI is identical to EXP when normalized using the SPECIALIST Lexicon and word order is disregarded, e.g., *(CLD, chronic liver diseases, C0341439, chronic liver disease, SPECIALIST-normalized)*;

Stemmed--a concept name of CUI is identical to EXP when stemmed and disregarding word order, e.g., *(BHC, benzenehexacarboxylic, C0105581, benzenehexacarboxylate, stemmed)*;

Synonym-like-replacement--a concept name of CUI is identical to EXP after replacing one word in EXP using a synonym-like set and ignoring word order, e.g., *(HFT, hepatic function test, C0023901, liver function tests, synonym-like-replaced)*.

EXPGrouper: is a program to group similar expansions of the same abbreviation together. For an abbreviation ABBR, each expansion consists of an initial group. EXPGrouper then groups similar groups of ABBR subsequently using the following normalization phases:

Group by ignoring punctuation: after removing punctuation, if an expansion in one group is the same as an expansion in another group, two groups are merged. For example, three different expansions for IGS: *immunogold staining, immuno gold staining*, and *immuno-gold staining* are merged into the same group.

Group using the SPECIALIST Lexicon: after normalizing using the SPECIALIST Lexicon, if an expansion in one group is identical to an expansion in a different group, two groups are merged. For example, the group for IGS containing *immuno-gold stain* is merged to the group containing *immuno-gold staining*.

Group by ignoring stop words, word order, punctuation, correcting typos and expanding abbreviation: two groups are merged together if after ignoring word order and punctuation, two expansions (one from each):

- are same after ignoring stop words (e.g., the group for IMT containing *intima-media thickness* is merged to the group containing *intima and media thickness*);
- differ in one type-error operation, i.e., replacement, transposition, insertion and deletion (e.g., the group for IGR containing *insect grwoth regulator* is merged to the group containing *insect growth regulator*);
- differ in a two-letter abbreviation and its expansion (e.g. the group for MIF containing *micro-if* is merged to the group containing *micro-immunofluorescence*).

Assessment: PW3 was executed for all MEDLINE abstracts up to December 2001. For each abbreviation ABBR, the number of abstracts that contained the parenthetical expression "(ABBR)" as well as the number of abstracts that contained ABBR with expansions found by PW3 was measured.

We evaluated EXPMap using MetaMap. We used the strict mode of MetaMap to get mappings for all expansions using the following options: a) unifying adjectives with the corresponding nouns (e.g., *abdominal* VS *abdomen*), b) not expanding abbreviations, c) ignoring word order, and d) stemming candidate strings. If the resulting mappings were single concepts with a relatively high matching score (i.e., 910, where the scores range from 0 to 1000), the mappings were considered as appropriate mappings. For example, if an expansion of IGR, *intergenic region*, was mapped to a single concept C0887859 with a score 1000, the mapping result was *(IGR, intergenic region, C0887859, 1000)*. The intra-agreement of the two systems was computed. In addition, we manually checked mapping results for expansions that occurred more than 200 times, and for which MetaMap either did not have the mappings or had mappings that were different from EXPMap.

After grouping expansions using EXPGrouper, we further grouped expansions according to the mapping results since two groups with the same concept identifier have the same sense. For example, the group of IHD that contains *ischemic cardiac disease* is merged to the group that contains *ischemic heart disease* since these two expansions are concept names of the same concept.

We computed the average number of variants for each group, and the number of groups with expansions having mappings associated with eight frequency thresholds: 1, 5, 10, 50, 100, 200, 500, and 1000, where the frequency of each group is the summation of occurrences of all expansions in that group. The ambiguity was measured considering the number of groups of each abbreviation with respect to five frequency thresholds: 1, 2, 5, 10, and 100.

Results

We excluded four capitalized text strings (i.e., III, VII, XII, XXI) from the result since they usually represented numbers. Among 4,839,200 unique occurrences of the parenthetical expression "(ABBR)", PW3 extracted 1,793,479 (ABBR, EXP) pairs, where 206,964 unique (ABBR, EXP, FREQ) tuples were derived (FREQ is the number of abstracts that have EXP as expansion of ABBR). The tuples with a FREQ value larger than 15,000 were:

- *(PCR, polymerase chain reaction, 19,067),*
- *(HIV, human immunodeficiency virus, 15,232).*

For the 35,981 expansions with mappings found by both EXPMap and MetaMap, the intra-agreement between the two systems was 99.6%. MetaMap matched 1,280 expansions for which EXPMap failed to find a match. EXPMap matched 14,230 expansions for which MetaMap failed to find a match. Among 39 expansions that we manually checked, 36 expansions were correct mappings (including 28 exact mappings for chemical names).

Among 50,211 expansions with mappings found by EXPMap, 31,223 of them were exact mappings, 13,871 were SPECIALIST-normalized mappings, 880 were stemmed mappings and the remaining 4,237 mappings were associated with synonym-like-replacement.

FTV	NG	FREQ	AV	Mapping	
				PG (%)	PO(%)
1	155,302	1,793,479	1.33	23.5	77.8
5	23,841	1,601,503	2.72	50.8	84.6
10	13,445	1,534,198	3.45	59.3	86.5
50	3,850	1,336,424	6.01	77.1	91.2
100	2,187	1,221,825	7.74	82.8	93.1
200	1,168	1,078,704	10.06	88.9	95.4
500	505	872,035	13.66	96.2	98.0
1000	247	695,902	17.23	98.8	99.1

Table I. The number of variants and the UMLS coverage with respect to eight thresholds.

FTV	NA	NG	POA (%)	PANA (%)	AAMB
1	11,328	155,302	99.6	81.2	16.6
2	9,299	64,338	98.2	74.5	8.95
5	6,767	23,841	94.0	64.6	4.91
10	5,297	13,445	87.7	55.4	3.78
100	1,683	2,187	42.3	22.0	2.36

Table II. The ambiguity assessment result.

The number of groups was 155,302. The average number of variants (i.e., expansions) for each group was 1.33. The group with the largest number of variants was *(FDG, 18f-fluorodeoxyglucose)* with 170 variants. 23.5% of the groups had expansions with mappings found by EXPMap, and covered 77.8% of the total occurrences. Table I lists the results with respect to different thresholds: FTV is the minimal number of occurrences for an group to be considered, NG is the number of groups considered, FREQ is the total number of occurrences of groups considered, AV is the average number of variants for each group, PG and PO are the percentages of the number of groups and the number of occurrences, respectively, for groups considered that had mappings. For example, after disregarding groups with occurrences of less than 500, there were 505 groups, with the total number of occurrences of 872,035; the average number of variants for each considered group was 13.66; 96.2% of the considered groups had expansions with mappings found by EXPMap, which covered 98.0% of the total occurrences of considered groups.

Among 11,328 different abbreviations, 81.2% had multiple groups with an average of 16.6 groups for each abbreviation, which means that they were ambiguous. These ambiguous abbreviations occurred 99.6% of the total occurrences of three-letter abbreviations in MEDLINE. The most ambiguous abbreviation was CAP, which had 191 groups. Table II lists the results with respect to five thresholds: FTV and NG are the same as Table I, NA is the number of different abbreviations which include at least one considered group, POA is the percentage of occurrences of ambiguous abbreviations in considered groups, PANA is the ratio of the number of ambiguous abbreviations in consideration to the total of considered abbreviations, and AAMB is the average ambiguity for ambiguous abbreviations in consideration, i.e., abbreviations with more than one group. For example, after disregarding groups with occurrences of less than 5, there were 6,767 different abbreviations with a total of 23,841 groups that had occurrences of not less than 5; 64.6% of the considered abbreviations appeared in more than one considered group, with an average of 4.91 groups for ambiguous considered abbreviations.

Discussion

EXPMap, which is based on MetaMap, is comparable to it: the two systems only disagreed on 0.4% of the mappings. Despite the use of different subsets of the Metathesaurus and the use of different synonym sets, one major difference between the two systems is that EXPMap normalizes terms in the Metathesaurus as well as terms that are being mapped while MetaMap generates different textual variants of terms that are being mapped and checks the existence of these textual variants in the Metathesaurus.

In this study, we grouped similar expansions of the same abbreviation together to assess coverage and ambiguity. We believed similar expansions would have similar senses. We found that frequency of senses plays an important role in the assessment:

- the UMLS coverage: those with higher frequency were more likely to have a mapping concept. For example, from Table I, we can see that 23.5% of the senses with occurrences of at least 1 were mapped to the UMLS; while for senses with occurrences of at least 100, 82.8% had mappings.
- the number of textual variants: senses with higher frequency had more textual variants, which follows Zipf's law, i.e., senses with high frequency tend to have many synonyms. For example, from Table I, we can see that senses with occurrences of at least 1 had an average of 1.33 variants; while senses with occurrences of at least 100 had an average of 7.74 variants.
- the ambiguity of abbreviations: abbreviations were less ambiguous when ignoring rarely occurring senses. For example, from Table II, we can see that 81.2% of abbreviations were ambiguous with an average of 16.6 senses; after ignoring senses with frequency of less than 10, 55.4% of abbreviations were ambiguous with an average of 3.78 senses.

We did not evaluate EXPGrouper because of a lack of a gold standard. Some expansions with different senses were incorrectly grouped together. For example, (*IMN*, *intramedullary nail)* and (*IMN*, *intramedullary nailing*) were merged together by EXPGrouper, but had different concept identifiers: C0348001 for the former one and C0021885 for the latter. About 630 out of 155,302 groups were mapped to different concept identifiers by different expansions in the same group using the first three modes of EXPMap.

Conclusion

From the above study and previous studies[5;6], we conclude that automatic understanding of abbreviations can be achieved for abbreviations that occur frequently with their definitions. After ignoring senses with less than 100 occurrences, over 80% of the senses matched the UMLS, with 7.74 textual variants for each sense; 22.0% of the abbreviations were ambiguous, with an average of 2.36 senses for ambiguous ones, which can be resolved based on our previous studies. These results suggest that an automatic method that constructs an abbreviation database from documents should take account of textual variants (i.e., expansions) with the same sense in order to be useful for NLP systems. In addition, authors should define abbreviations when the corresponding senses are not well-known.

Acknowledgement: We would like to thank to Jim Mork at National Library of Medicine for helping with the use of MetaMap. This study was supported in part by grants LM06274 from the NLM and ILS-9817434 from the NSF.

Reference

(1) Larkey L, Ogilvie P, Price A, Tamilio B. *Acrophile: An Automated Acronym Extractor and Server.* ACM Digital Libraries 2000;205-214.

(2) Friedman C *A Broad Coverage Natural Language Processing System*. Proc AMIA Symp 2000 270-274

(3) Aronson A *Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program*. Proc. AMIA Symp 2001. 17-21

(4) Nadkarni P, Chen R, Brandt C. *UMLS Concept Indexing for Production Databases*. J Am Med Inf Assoc 2001; 8:80-91.

(5) Liu H, Lussier Y, Friedman C *A study of the UMLS abbreviations.*2001 Proc. AMIA Symp. 393-397

(6) Liu H, Johnson SB, Friedman C. *Automatic Resolution of Ambiguous Terms Based on Machine Learning and Conceptual Relations in the UMLS*. J Am Med Inf Assoc 2002.

(7) Taghva K, Gilbrech J. Recognizing Acronyms and their Definitions. 95-03. 1995.

(8) Yoshida M, Fukuda K, Takagi T. *PNAD-CSS: a workbench for construction a protein name abbreviation dictionary*. Bioinformatics 2000; 16(2):169-175.

(9) Yu H, Hripcsak G, Friedman C. *Mapping abbreviations to full forms in electronic articles*. JAMIA 2002

(10) UMLS Knowledge Sources, 2000 Edition.

(11) MEDLINE. http://www.nlm.nih.gov . 2001.