Automated Identification of Synonyms in Biomedical Acronym Sense Inventories

Genevieve B. Melton
Institute for Health Informatics & Dept of Surgery
University of Minnesota
Minneapolis, MN 55455 USA
gmelton@umn.edu

SungRim Moon
Institute for Health Informatics
University of Minnesota
Minneapolis, MN 55455 USA
moonx086@umn.edu

Bridget McInnes
College of Pharmacy
University of Minnesota
Minneapolis, MN 55455 USA
bthomson@umn.edu

Serguei Pakhomov
College of Pharmacy
University of Minnesota
Minneapolis, MN 55455 USA
pakh0002@umn.edu

Abstract

Acronyms are increasingly prevalent in biomedical text, and the task of acronym disambiguation is fundamentally important for biomedical natural language processing systems. Several groups have generated sense inventories of acronym long form expansions from the biomedical literature. Long form sense inventories, however, may contain conceptually redundant expansions that negatively affect their quality. Our approach to improving sense inventories consists of mapping long form expansions to concepts in the Unified Medical Language System (UMLS) with subsequent application of a semantic similarity algorithm based upon conceptual overlap. We evaluated this approach on a reference standard developed for ten acronyms. A total of 119 of 155 (78%) long forms mapped to concepts in the UMLS. Our approach identified synonymous long forms with a sensitivity of 70.2% and a positive predictive value of 96.3%. Although further refinements are needed, this study demonstrates the potential value of using automated techniques to merge synonymous biomedical acronym long forms to improve the quality of biomedical acronym sense inventories.

1 Introduction

Acronyms and abbreviations are increasingly used in biomedical text. This is in large part due to the expansive growth of the biomedical literature estimated to be close to one million articles annually (Stead et al. 2005). Ambiguous acronyms represent a challenge to both human readers and computerized processing systems for resolving the acronym's meaning within a particular context. For any given acronym, there are often multiple possible long form expansions. Techniques to determine the context-specific meaning or sense of an ambiguous acronym are fundamentally important for biomedical natural language processing and can assist with important tasks such as information retrieval and information extraction (Friedman 2000).

Acronym ambiguity resolution represents a special case of word sense disambiguation (WSD) with unique challenges. In particular, there are increasing numbers of new acronyms (i.e., short forms) as well as increasing numbers of new senses (i.e., long forms) for existing acronyms within biomedical text. Acronyms in biomedicine also range from those that are common, to those that are infrequent which appear to be created in an ad hoc fashion resulting essentially in neologisms distinct to small sets of biomedical discourse.

Sense inventories are important tools that can assist in the task of disambiguation of acronyms and abbreviations. The relative formal nature of biomedical literature discourse lends itself well to building these inventories because long forms are typically contained within the text itself, providing a "definition" on its first mention in an article, next to a parenthetical expression containing the short form or vice versa (Schwartz and Hearst 2003). In contrast, clinical documents are less structured and

Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents, pages 46–52,
Los Angeles, California, June 2010.

typically lack expanded long forms for acronyms and abbreviations, leaving sense inventories based on documents in the clinical domain not as well developed as the sense inventories developed from the biomedical literature (Pakhomov et al. 2005).

Compilation of sense inventories for acronyms in clinical documents typically relies on vocabularies contained in the Unified Medical Language System (UMLS) as well as other resources such as ADAM (Zhou et al. 2006). However, with the advantage of using rich and diverse resources like ADAM and the UMLS comes the challenge of having to identify and merge synonymous long form expansions which can occur for a given short form. Having synonymous long forms in a sense inventory for a given acronym poses a problem for automated acronym disambiguation because the sense inventory dictates that the disambiguation algorithm must be able to distinguish between semantically equivalent senses. This is an important problem to address because effective identification of synonymous long forms allows for a clean sense inventory, and it creates the ability for long form expansions to be combined while preserving the variety of expression occurring in natural language. By automating the merging of synonymous expansions and building a high quality sense inventory, the task of acronym disambiguation will be improved resulting in better biomedical NLP system performance.

Our approach to reducing multiple synonymous variants of the same long form for a set of ten biomedical acronyms is based on mapping sense inventories for biomedical acronyms to the UMLS and using a semantic similarity algorithm based on conceptual overlap. This study is an exploratory evaluation of this approach on a manually created reference standard.

2 Background

2.1 Similarity measures in biomedicine

The area of semantic similarity in biomedicine is a major area within biomedical NLP and knowledge representation research. Semantic similarity aids NLP systems, improves the performance of information retrieval tasks, and helps to reveal important latent relationships between biomedical concepts. Several investigators have studied conceptual similarity and have used relationships in controlled biomedical terminologies, empiric statistical data from biomedical text, and other knowledge sources (Lee et al. 2008; Caviedes and Cimino 2004). However, most of these techniques focus on generating measures between a single pair of concepts and do not deal directly with the task of comparing two groups of concepts.

Patient similarity represents an important analogous problem that deals with sets of concepts. The approach used by Melton et al. (2006) was to represent each patient case as a set of nodes within a controlled biomedical terminology (SNOMED CT). The investigators then applied several measures to ascertain similarity between patient cases. These measures ranged from techniques independent of the controlled terminology (i.e. set overlap or Hamming distance) to methods heavily reliant upon the controlled terminology based upon path traversal between pair of nodes using defined relationships (either IS-A relationships or other semantic relationships) within the terminology.

2.2 Lesk algorithm for measuring similarity using sets of definitional words

A variety of techniques have been used for the general problem of WSD that range from highly labor intensive that depend upon human data tagging (i.e. supervised learning) to unsupervised approaches that are completely automated and rely upon non-human sources of information, such as context and other semantic features of the surrounding text or definitional data.

The Lesk algorithm (Lesk 1986) is one example of an unsupervised method that uses dictionary information to perform WSD. This algorithm uses the observation that words co-occurring in a sentence refer to the same topic and that dictionary definition words will have topically related senses, as well. The classic form of this algorithm returns a measure of word overlap. Lesk depends upon finding common words between dictionary definitions. One shortcoming of Lesk, however, it that it can perform worse for words with terse, few word definitions.

As a modification of Lesk, researchers have proposed using WordNet (Felbaum 1998) to enhance its performance. WordNet has additional semantic information that can aid in the task of disambiguation, such as relationships between the term of interest and other terms. Banerjee and Pe-

dersen (2002) demonstrated that modifications to Lesk improved performance significantly with the addition of semantic relationship information.

2.3 Biomedical literature sense inventories

A number of acronym and abbreviation sense inventories have been developed from the biomedical literature using a variety of approaches. Chang et al. (2002) developed the Stanford biomedical abbreviation server[1] using titles and abstracts from MEDLINE, lexical heuristic rules, and supervised logistic regression to align text and extract short form/long form pairs that matched well with acronym short form letters. Similarly, Adar (2004) developed the Simple and Robust Abbreviation Dictionary (SaRAD)[2]. This inventory, in addition to providing the abbreviation and definition, also clusters long forms using an *N*-gram approach along with classification rules to disambiguate definitions. This resource, while analogous with respect to its goal of merging and aligning long form expansions, is not freely available. Adar measured a normalized similarity between *N*-gram sets and then clustered long forms to create a clustered sense inventory resource.

One of the most comprehensive biomedical acronym and abbreviation databases is ADAM (Zhou et al. 2006) an open source database[3] that we used for this study. Once identified, short form/long form pairs were filtered statistically with a rule of length ratio and an empirically-based cut-off value. This sense inventory is based on MEDLINE titles and abstracts from 2006 and consists of over 59 thousand abbreviation/long form pairs. The authors report high precision with ADAM (97%) and up to 33% novel abbreviations not contained within the UMLS or Stanford Abbreviation dictionary.

2.4 MetaMap resource for automated mapping to the UMLS

An important resource for mapping words and phrases to the UMLS Metathesaurus is MetaMap. This resource was developed at the National Library of Medicine (Aronson 2001) to map text of biomedical abstracts to the UMLS. MetaMap uses a knowledge intensive approach that relies upon computational linguistic, statistical, and symbolic/lexical techniques. While MetaMap was initially developed to help with indexing of biomedical literature, it has been applied and expanded successfully to a number of diverse applications including clinical text.

With each mapping, an evaluation function based upon centrality, variation, coverage, and cohesiveness generates a score for a given mapping from 0 to 1000 (strongest match). A cut-off score of 900 or greater is considered to represent a good conceptual match for MetaMap and was used in this study as the threshold to select valid mappings.

3 Methods

Ten randomly selected acronyms with between 10 to 20 long forms were selected from the ADAM resource database for this pilot study.

3.1 Long form mappings to UMLS

Each acronym long-form was mapped to the UMLS with MetaMap using two settings. First, MetaMap was run with its default setting on each long form expansion. Second, MetaMap was run in its "browse mode" (options "-zogm") which allows for term processing, overmatches, concept gaps, and ignores word order.

Processing each long form with MetaMap then resulted in a set of Concept Unique Identifiers (CUIs) representing the long form. Each CUI with a score over 900 was included in the overall set of CUIs for a particular long form expansion. For a given pair of long form expansions the two sets of CUIs that each long form mapped to were compared for concept overlap, in an analogous fashion to the Lesk algorithm. The overlap between concept sets was calculated between each pair of long form expansions and expressed as a ratio:

$$\frac{\#\ overlapping\ concepts\ shared\ between\ long\ forms}{\#\ concepts\ for\ the\ long\ form\ with\ least\ \#\ concepts}.$$

For this study, an overlap of 50% or greater was considered to indicate a potential synonymous pair.

Now let us assume that we have two concept sets: The first one is {A, B} and the second one is {A, B, C}, with each CUI having a score over 900. In this example, the overlap of concepts for the first concept set between it and the other is 100%, and for the second that is 66.7%. Because overlaps

[1] http://abbreviation.stanford.edu
[2] http://www.hpl.hp.com/shl/projects/abbrev.html
[3] http://arrowsmith.psych.uic.edu

are greater than 50%, they are a potential synonymous pair, and the overlap ratio is calculated as $\frac{\#\{A,B\}}{\#\{A,B\}} = \frac{2}{2} = 1$ (100%).

3.2 Expert-derived reference standard

Two physicians were asked to judge the similarity between each pair combination of long forms expansions on a continuous scale for our initial reference standard. Physicians were instructed to rate pairs of long forms for conceptual similarity. Long forms were presented on a large LCD touch-screen display (Hewlett-Packard TouchSmart 22" desktop) along with a continuous scale for the physicians to rate long form pairs as dissimilar (far left screen) or highly similar (far right screen). The rating was measured on a scale from 1 to 1500 pixels representing the maximum width of the touch sensitive area of the display (along the x-coordinate). Inter-rater agreement was assessed using Pearson correlation.

Expert scores were then averaged and plotted on a histogram to visualize expert ratings. We subsequently used a univariate clustering approach based on the R implementation of the Partitioning Around Medoids (PAM) method to estimate a cut-off point between similar and dissimilar terms based on the vector of the average responses by the two physicians. The responses were clustered into two and three clusters based on an informal observation of the distribution of responses on the histogram showing evidence of at least a bimodal and possibly a trimodal distribution.

As a quality measure, a third physician manually reviewed the mean similarity ratings of the first two physicians to assess whether their similarity judgments represented the degree of synonymy between long form expansions necessary to warrant merging the long form expansions. This review was done using a binary scale (0=not synonymous, 1=synonymous).

3.3 Evaluation of automated methods

Long form pair determinations based on the mappings to the UMLS were compared to our reference standard as described in Section 3.2. We calculated overall results of all long form pair comparisons and on all long form pairs that mapped to the UMLS with MetaMap. Performance is reported as sensitivity, specificity, and positive predictive value.

4 Results

A total of 10 random acronyms were used in this study. All long forms for these 10 acronyms were from the sense inventory ADAM (Zhou et al., 2006). This resulted in a total of 155 long form expansions (median 16.5 per acronym, range 11-19) (Table 1).

Acronym	N of LF expansions	LF expansions mapped by MetaMap
Total	155	119 (78%)
ALT	13	9 (70%)
CK	14	9 (64%)
CSF	11	7 (74%)
CTA	19	14 (74%)
MN	19	17 (89%)
NG	17	15 (88%)
PCR	17	8 (47%)
PET	17	15 (88%)
RV	16	14 (88%)
TTP	12	11(92%)

Table 1. Number of acronym long forms in ADAM and mapping to the UMLS

4.1 Long form mappings to UMLS

The default mode of MetaMap resulted in 119 (78%) long forms with mappings to the UMLS with MetaMap (Table 1). Use of MetaMap's browse mode did not increase the total number of mapped long forms but did change some of the mapped concepts returned by MetaMap (not depicted).

Acronym	N pairs	Pearson r
Total	1125	0.78*
ALT	78	0.79*
CK	91	0.77*
CSF	55	0.80*
CTA	136	0.92*
MN	171	0.69*
NG	136	0.68*
PCR	136	0.89*
PET	136	0.78*
RV	120	0.67*
TTP	66	0.76*

Table 2. Pearson correlation coefficient for ratings overall and for individual acronyms. *$p<0.0001$

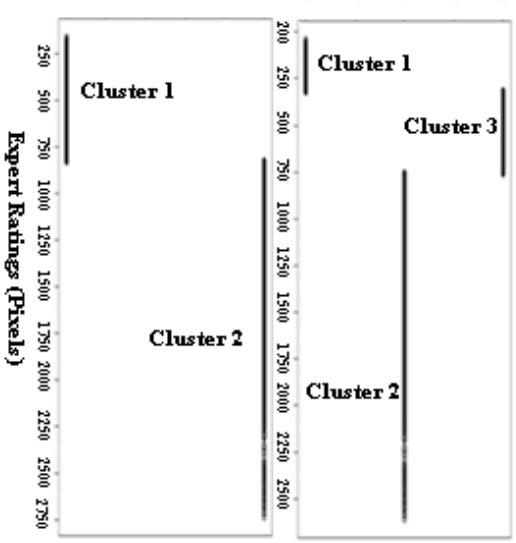

Figure 1. Two-way and three-way clustering solution of expert ratings of long form pairs.

4.2 Expert-derived reference standard

For the 1125 total comparison pairs, two raters assessed similarity between long form pairs on a continuous scale. The overall mean correlation between the two raters was 0.78 (standard deviation 0.08). Pearson correlation coefficients for each acronym are depicted in Table 2.

Two-way and three-way clustering demonstrated an empirically determined "cutoff" of 525 pixels from the left of the screen. This separation point between clusters (designated as "low cutoff") was evident on both the two-way and three-way clustering approaches using the PAM method to estimate a cut-off point between similar and dissimilar terms based on the vector of the average responses by the two physicians (Figure 1). Intuitively this low cutoff includes manual ratings indicative of moderate to low similarity (as 525 pixels along a 1500 pixel-wide scale is approximately one-third of the way from the left "dissimilar" edge of the touch-sensitive screen). To isolate terms that were rated as highly similar, we also created an arbitrary "high cutoff" of 1200 pixels.

CTA:	
"CT hepatic arteriography"	"CT angiography"
MN:	
"median nerve"	"motor neuron"
RV:	
"rabies virus"	"rotavirus"
"right ventricular free wall"	"right ventricle"
TTP:	
"thiamine triphosphate"	"thymidine triphosphate"

Figure 2. Examples of terms originally rated as highly similar but not synonymous by the curating physician.

Expert curation of the ratings by the third physician demonstrated that conceptual similarity ratings were sometimes not equivalent to synonymy that would warrant the collapse of long form pairs. Of 1125 total pairs of long forms, 70 (6%) origi-

	Default Mode: MetaMap		Browse Mode: MetaMap	
	All LF	Mapped LF only	All LF	Mapped LF only
High Cutoff				
Sensitivity	21.6%	39.6%	23.8%	43.8%
Specificity	98.1%	96.8%	99.4%	99.0%
PPV	48.7%	48.7%	77.8%	77.8%
NPV	93.6%	95.5%	93.9%	95.9%
Expert Curation				
Sensitivity	34.3%	64.9%	37.1%	70.2%
Specificity	98.6%	97.7%	99.9%	99.8%
PPV	61.5%	61.5%	96.3%	96.3%
NPV	95.8%	98.0%	96.0%	98.3%

Table 3. Performance of automated techniques for merging biomedical long form senses for all long forms and for long forms that mapped to the UMLS only. PPV, positive predictive value; NPV, negative predictive value.

nally classified as similar were re-classified as conceptually different by the third physician. Several examples of long form pairs that were originally rated as highly similar but were judged as not synonymous are contained in Figure 2.

4.3 Evaluation of automated methods

The performance of our algorithm is shown in Table 3 using MetaMap in the default mode and browse mode and then applying our reference standard using the "low cutoff", "high cutoff", and expert curation (Table 3). Performance is reported for all 155 long forms (All LF) and for the subset of 119 long forms that mapped to the UMLS (Mapped LF only). Compared to the "low cutoff" reference standard, the "high cutoff" and expert curation were positively associated with more consistent performance. The browse mode identified fewer potential terms to merge and had higher accuracy than the default MetaMap mode.

5 Conclusions

The results of this pilot study are promising and demonstrate high positive predictive value and moderate sensitivity for our algorithm, which indicates to us that this technique with some additional modifications has value. We found that mapping long form expansions to a controlled terminology to not be straightforward. Although approximately 80% of long forms mapped, another 20% were not converted to UMLS concepts. Because each long form resulted in multiple paired comparisons, a 20% loss of mappings resulted globally in a 40% loss in overall system performance. While long form expansions were entered into MetaMap using a partially normalized representation of the long form, it is possible that additional normalization will improve our mapping.

An important observation from our expert-derived reference standard was that terms judged by physicians as semantically highly similar may not necessarily be synonymous (Figure 2). While semantic similarity is analogous, there may be some fundamentally different cognitive determinations between similarity and synonymy for human raters.

The current technique that we present compares sets of mapped concepts in an analogous fashion to the Lesk algorithm and other measures of similarity between groups of concepts previously reported. This study did not utilize features of the controlled terminology nor statistical information about the text to help improve performance. Despite the lack of additional refinement to the presented techniques, we found a flat overlap measure to be moderately effective in our evaluation.

6 Future Work

There are several lines of investigation that we will pursue as an extension of this study. The most obvious would be to use semantic similarity measures between pairs of concepts that capitalize upon features and relationships in the controlled terminology. We can also expand upon the type of similarity measures for the overall long form comparison which requires a measure of similarity between *groups* of concepts. In addition, an empiric weighting scheme based on statistical information of common senses may be helpful for concept mappings to place more or less emphasis on important or less important concepts. We plan to determine the impact of automatically reduced sense inventories on the evaluation of WSD algorithms used for medical acronym disambiguation.

Finally, we would like to utilize this work to help improve the contents of a sense inventory that we are currently developing for acronyms and abbreviations. This sense inventory is primarily based on clinical documents but incorporates information from a number of diverse sources including ADAM, the UMLS, and a standard medical dictionary with abbreviations and acronyms.

Acknowledgments

This work was supported by the University of Minnesota Institute for Health Informatics and Department of Surgery and by the National Library of Medicine (#R01 LM009623-01). We would like to thank Fairview Health Services for ongoing support of this research.

References

Eytan Adar (2004) SaRAD: A simple and robust abbreviation dictionary. Bioinformatics 20:527–33.

Alan R Aronson (2001) Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001:17-21.

Satanjeev Banerjee, Ted Pedersen. 2002. An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet, Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing, p.136-145, February 17-23.

Jorge E. Caviedes JE, James J Cimino. (2004) Towards the development of a conceptual distance metric for the UMLS. J Biomed Inform. Apr;37(2):77–85.

Jeffrey T Chang, Hinrich Schutze, Russ B Altman (2001) Creating an online dictionary of abbreviations from Medline. J Am Med Inform Assoc 9:612–20.

Christiane Fellbaum, editor. 1998. *WordNet: An electronic lexical database*. MIT Press.

Carol Friedman. 2000. A broad-coverage natural language processing system. Proc AMIA Symp., 270–274.

Wei-Nchih Lee, Nigam Shah, Karanjot Sundlass, Mark Musen (2008) Comparison of Ontology-based Semantic-Similarity Measures. AMIA Annu Symp Proc. 2008. 384–388.

Michael E. Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from a ice cream cone. In *Proceedings of SIGDOC '86*.

Genevieve B. Melton, Simon Parsons, Frances P. Morrison, Adam S. Rothschild, Marianthi Markatou, George Hripcsak. 2006. Inter-patient distance metrics using SNOMED CT defining relationships, Journal of Biomedical Informatics, 39(6), 697-705.

Serguei Pakhomov, Ted Pedersen, Christopher G. Chute. 2005. Abbreviation and Acronym Disambiguation in Clinical Discourse. American Medical Informatics Association Annual Symposium, 589-593.

Ariel S Schwartz and Marti A. Hearst. 2003. A Simple Algorithm for Identifying Abbreviation Definitions in Biomedical Text. Pacific Symposium on Biocomputing p451-462.

William W Stead, Brian J Kelly, Robert M Kolodner. 2005. Achievable steps toward building a National Health Information infrastructure in the United States. J. Am. Med. Inform. Assoc., 12, 113–120.

Wei Zhou, Vetle I Torvik, Neil R Smalheiser (2006) ADAM: Another database of abbreviations in Medline. Bioinformatics 22:2813– 8.