# Abbreviation and Acronym Disambiguation in Clinical Discourse

Serguei Pakhomov, PhD[1], Ted Pedersen, PhD[2] and Christopher G. Chute, MD, DrPH[1]
[1]*Division of Biomedical Informatics, Mayo College of Medicine, Rochester, MN, USA*
[2] *Department of Computer Science, University of Minnesota*

*Use of abbreviations and acronyms is pervasive in clinical reports despite many efforts to limit the use of ambiguous and unsanctioned abbreviations and acronyms. Due to the fact that many abbreviations and acronyms are ambiguous with respect to their sense, complete and accurate text analysis is impossible without identification of the sense that was intended for a given abbreviation or acronym. We present the results of an experiment where we used the contexts harvested from the Internet through Google API to collect contextual data for a set of 8 acronyms found in clinical notes at the Mayo Clinic. We then used the contexts to disambiguate the sense of abbreviations in a manually annotated corpus.*

## INTRODUCTION

Many abbreviations and acronyms[i] are ambiguous with respect to their sense and constitute a significant part of the general problem of text normalization. Acronyms are used routinely throughout clinical texts and knowing their sense is critical to the understanding of the document whether we talk about automatic natural language understanding or simply human comprehension and interpretation. The acronym ambiguity is a growing problem both in the number of new acronyms and the number of new senses for existing acronyms. For example, according to the UMLS® 2001AB [1], RA had the following 8 senses: “rheumatoid arthritis”, “renal artery”, “right atrium”, “right atrial”, “refractory anemia”, “radioactive”, “right aram”, “rheumatic arthritis.” The 2005AA version of the UMLS® contains 17 additional senses: “ragweed antigen”, “refractory ascites”, “renin activity”, to name only a few. This is just an indication of the rate at which the ambiguity is proliferating. Liu et al.[2] show that 33% of acronyms listed in the UMLS in 2001 are ambiguous. In a later study, Liu et al.[3] demonstrated that 81% of acronyms found in MEDLINE abstracts are ambiguous and have on average 16 senses. In addition to problems with text interpretation, Friedman, et al. [4] also point out that acronyms constitute a major source of errors in a system that automatically generates lexicons for medical Natural Language Processing (NLP) applications.

Ideally, when looking for documents containing “rheumatoid arthritis”, we want to retrieve everything that has a mention of RA in the sense of “rheumatoid arthritis” but not those documents where RA means “right atrial.” Acronym disambiguation problem is a special case of the word sense disambiguation (WSD) problem. Approaches to WSD include supervised machine learning techniques, where some amount of training data is marked up by hand and is used to train a decision tree classifier[5]. On the other side of the spectrum, the fully unsupervised learning methods such as clustering have been also successfully used[6]. A hybrid class of machine learning techniques for WSD relies on a small set of hand labeled data used to bootstrap a larger corpus of training data[7,8]. The cornerstone of all machine learning techniques for WSD is the context[9] as this is also true for acronym disambiguation.

One way to take context into account is to consider the type of discourse in which the acronym occurs. If we see RA in a cardiology report, then it can be normalized to “right atrial”, else if it occurs in the context of a rheumatology note, it is likely to mean “rheumatoid arthritis.” This method of using global context to resolve the acronym ambiguity suffers from at least three major drawbacks. First of all, it requires a database of acronyms and their expansions linked with possible contexts in which particular expansions can be used. Second, it requires a rule-based system for assigning correct expansions. Third, the distinctions made between various senses are bound to be very coarse. We may be able to distinguish correctly between “rheumatoid arthritis” and “right atrial” since the two are likely to occur in clearly separable contexts; however, distinguishing between “rheumatoid arthritis” and “right arm” becomes more of a challenge and may require introducing additional rules to further complicate the system.

Pakhomov[10] introduced a method for collecting training data for supervised machine learning approaches to disambiguating acronyms. The method is based on the assumption that the expansion (or the sense) of an acronym and the acronym itself tend to occur in similar contexts. For example, we would expect one to use the

[i] To save space and for ease of presentation, we will use the word “acronym” to mean both “abbreviation” and “acronym” since the two could be used interchangeably for the purposes described in this paper

expressions "rheumatoid arthritis" and "RA" in the sense of rheumatoid arthritis under similar contextual conditions. Clinical reports were used to find the expressions that defined the senses of ambiguous acronyms. Once the expression was identified in the corpus, its surrounding context was recorded and subsequently used for training statistical predictive models for disambiguating acronyms. That work had two main limitations. One had to do with the fact that there was no hand labeled corpus to evaluate the method. The evaluation was done using a cross-validation on the training data. The second limitation was that the method assumed the availability of large numbers of clinical reports. Due to patient confidentiality restrictions such reports cannot be shared and are available only to a few researchers. In this paper we build upon this method and extend it to include contextual data derived from a widely and publicly available resource – the Internet. One of the main goals of the work described in this paper is to explore the use of the Internet as a source of training data for acronym disambiguation.

In the rest of the paper, we will introduce some of the previous work on medical acronym disambiguation techniques. We will then give a detailed description of the manually annotated corpus of acronym data we are using for this project. We will also describe the technique we are proposing and present and discuss the evaluation results.

## BACKGROUND

There have been many proposals in the medical informatics literature for solutions to the acronym disambiguation problem. The bulk of previous work is focused primarily on supervised machine learning approaches where a corpus of text is annotated for acronyms and their sense is manually disambiguated. The disambiguation is typically cast as a classification problem where the manually annotated corpus is used for training and evaluation of statistical classifiers. A notable exception to these approaches is the work by Liu[11] who uses domain knowledge in the form of hierarchical relations between the parents and siblings of the various senses for a given acronym. The intended sense is determined by matching the terms found in the context of the acronym that is being disambiguated to the terms that instantiate the parents and siblings of the term that represents the intended sense. For example, if the acronym "RA" only had two senses "rheumatoid arthritis" and "renal artery", then according to Liu's method, we would find all the entry terms representing the siblings and parents of "rheumatoid arthritis" as well as all the entry terms representing parents and siblings of "renal artery." If the first set of terms has a greater number of matches to the terms found in the context of "RA", then "rheumatoid arthritis" is selected as the intended sense, otherwise – "renal artery."

## MATERIALS AND METHODS

### Acronyms in Clinical Discourse

By clinical discourse in this paper we mean the text found in the clinical notes repository at the Mayo Clinic. This repository represents a record of patient-physician encounters at the Mayo Clinic since 1994 and contains approximately 16 million documents. In order to address the problem of acronym ambiguity, we selected a random sample of clinical notes from a subset of ~1.7 million notes recorded in 2002. We only selected the notes where one or more of the following 8 acronyms occurred: AC, ACA, APC, CF, HA, LA, NSR and PE. These acronyms were selected pseudo-randomly from the Mayo Clinic's formulary of abbreviations and acronyms[ii]. We arbitrarily selected only those acronyms that had more than 2 senses and represented 1, 2 and 3 letter combinations. Table 1 summarizes the sizes of corpora collected for each acronym.

| | AC | ACA | APC | CF | HA | LA | NSR | PE |
|---|---|---|---|---|---|---|---|---|
| **N** | 553 | 554 | 378 | 730 | 514 | 492 | 408 | 685 |

**Table 1. Sample sizes for 8 acronyms (N = number of sample feature vectors).**

It shows the number of individual instances of the acronyms found and the number of notes in which these instances were found. A more comprehensive description of the clinical notes data can be found in a previous work by Pakhomov[10]. The identification of acronyms in the corpus was done automatically by using simple regular expression matching. Thus each of the 8 acronyms was represented by a corpus of clinical notes marked up with XML tags and presented to human experts for annotation through the Generalized Architecture for Text Engineering (GATE 3.0)[iii] interface. The manual annotation of the acronyms for their sense was done by consensus between three human experts who had substantial amount of experience in medical indexing and retrieval. Two of the experts had over 12 years and one had 3 years of experience.

Subsequent to the annotation, we analyzed the distribution of different senses. The distribution happens to be highly skewed towards a single sense for four of the 8 acronyms: ACA, HA, LA and NSR. Table 2 shows the predominant senses for each of the 8 acronyms. The boldfaced acronyms have a single predominant sense where no other sense has more than 10% representation. For the other 4 acronyms the sense distribution is also

[ii] Currently, the Mayo Clinic is in the process of restricting the use of certain acronyms and abbreviations as part of the patient safety regulations.

[iii] http://gate.ac.uk/

skewed; however, the majority of the instances include at least two senses. CF is a borderline case with only 14% of the sense inventory represented by 'cold formula' and the rest of the space taken up by 'cystic fibrosis.'

| | **N senses** | **Predominant senses** |
|---|---|---|
| AC | 13 | 'acid controller' (22%),<br>'acromyoclavicular' (30%),<br>'antitussive with codeine' (29%) |
| **ACA** | 6 | 'adenocarcinoma' (85%) |
| APC | 10 | 'activated protein C' (12%)<br>'adenomatous polyposis coli' (25%)<br>'argon plasma coagulation' (42%) |
| CF | 9 | 'cold formula' (14%)<br>'cystic fibrosis' (72%) |
| **HA** | 5 | 'headache' (92%) |
| **LA** | 8 | 'long acting' (79%) |
| **NSR** | 2 | 'normal sinus rhythm' (99%) |
| PE | 11 | 'pressure equalizing' (32%)<br>'pulmonary embolism' (48%) |

**Table 2. Partial sense inventory for the 8 acronyms. Only predominant senses are listed.**

We have also compared the sense inventories derived empirically from the manually annotated corpus to those provided by the Mayo Clinic's approved senses and the senses listed in the UMLS 2005AA LRABR table for each of the 8 acronyms. The findings of the comparison are summarized in Table 3.

| | Number of Senses | | | Mayo Clinic source | | UMLS source | |
|---|---|---|---|---|---|---|---|
| | U | M | C | over | extra | over | extra |
| AC | 48 | 3 | 13 | 1 | 12 | 6 | 7 |
| ACA | 1 | 3 | 6 | 1 | 6 | 0 | 6 |
| APC | 2 | 2 | 10 | 0 | 10 | 0 | 10 |
| CF | 2 | 11 | 9 | 4 | 6 | 1 | 8 |
| HA | 26 | 3 | 5 | 3 | 2 | 4 | 1 |
| LA | 1 | 4 | 8 | 3 | 5 | 0 | 8 |
| NSR | 0 | 2 | 2 | 2 | 2 | 0 | 2 |
| PE | 36 | 3 | 11 | 3 | 8 | 5 | 6 |
| Totals | 116 | 31 | 64 | 17 | 51 | 16 | 48 |

**Table 3. Quantitative comparison of sense inventories. (U stands for 'N UMLS senses", M – 'N Mayo approved senses', C – 'N Mayo corpus/empirical senses.' The values in the 'over' columns indicate the number of overlapping senses and the values in the 'extra' columns show the number of senses missing from the source inventory.)**

These results show that only 26% (17/64) of the empirically found senses for the 8 acronyms overlap with those provided by the Mayo Clinic list of approved acronyms and their senses. About the same number 25% (16/64) represents the overlap in sense inventories between the empirically found senses and those provided by the UMLS 2005AA. These observations indicate that established sources of acronyms and abbreviations may not be entirely suitable as sources of sense inventories for acronyms in clinical notes.

## Experiments

We have experimented with two approaches to acronym sense disambiguation: fully supervised and semi-supervised. The latter is the focal point of this paper and the former was used to establish an upper bound for subsequent experimentation. The main difference between these two approaches is in how training data is collected. For the fully supervised approach, both the training and the testing data are manually annotated. The data collection for the semi-supervised approach is automated but only for the training data. The evaluation is still performed on manually annotated data. We describe the two approaches in the following two subsections in further detail.

### Fully-supervised Approach

The fully supervised approaches were used here in order to establish the upper bound for subsequent evaluation of the semi-supervised learning. For the fully supervised approach we used two well established machine learning algorithms, Maximum Entropy and C5.0 Decision Trees. We used words that occur in the same sentence as the acronym as a set of features for both algorithms. This technique is also known as the 'bag-of-words' approach. At this stage, no feature selection was performed on the manually tagged acronym samples apart from the standard exclusion of stop words. C5.0 algorithm uses information gain in order to select relevant attributes to split on. Maximum Entropy determines the relevant features through a Generalized Iterative Scaling (GIS) procedure. For more detailed information on these algorithms, see Manning and Shutze (2000).

### Semi-supervised approach

The semi-supervised approach is the main contribution of this paper and consists of the following steps:

*Step 1: Sense Inventory.* We developed a sense inventory for each acronym. In this step, we used the empirical sense inventory derived from the manually annotated samples.

*Step 2: Data Collection* For each sense of each acronym, we collected the 'contexts' from corpora of textual data by finding exact matches in the corpus of the entire character sequence that represented the sense and recording the surrounding lexical items within a specified horizon. For our experiments, we selected the following three corpora:

1. Unrestricted World Wide Web (Web)
2. Medline abstracts (Med)
3. Mayo Clinic corpus of 1.7M notes (Mayo)

The first two corpora were accessed and manipulated via the Google Java API. For example, in order to collect samples for 'cystic fibrosis' from unrestricted WWW space, we would make a simple call to the Google API and collect the words contained in the 'snippets' and the titles of the pages returned by the API for the first 100 hits. In order to collect samples from Medline abstracts, we restricted the search space through the API to the 'ncbi.nlm.nih.gov' domain. If a sense produced no hits in the restricted space, the search was opened up to the general web. The third corpus was accessed via a Perl script that scanned the text of clinical notes and harvested 'bags-of-words' in the +-20 word vicinity of the detected string representing the sense of an acronym. Crossing of sentential boundaries was permitted. This step resulted in three sets of training samples for each acronym: *Web (general)*, *Med (medline)* and *Mayo (clinical corpus)*.

*Step 3: Data merging.* In this step, we created additional sets by merging the data generated in the previous step, which resulted in two additional sets: *Mayo+Web* and *Mayo+Med*. The merging alleviated the problem where the data gathered from the clinical notes was missing contexts for some of the senses of an acronym because the character strings representing the senses were not detected anywhere in the corpus. The merging compensated for the missing data.

*Step 4: Context vectors generation.* Each training sample was represented as a context vector of lexical items and their frequency. Each test sample was represented in the same way.

We experimented with two ways of using the automatically generated training data. One was to use a standard technique such as the C5.0 Decision Trees and the other was to use the cosine between the training and the test vectors as a measure of their similarity. A method similar to the latter technique has been used successfully by Pedersen et al.[12] for determining semantic similarity between concepts. Their approach consisted of generating context vectors for each concept by averaging the vectors representing the words that make up the definition of the concept. We use a simplified version of this technique where we normalize the training and the testing vectors that represent the various senses of a given acronym and then compute the cosine between all training vectors for a given acronym and the test vectors. The training vector with the largest cosine is selected to represent the sense of the acronym represented by the test sample.

## RESULTS AND DISCUSSION

Evaluation of all methods was performed using the standard accuracy measure where accuracy is computed as the ratio of correctly disambiguated test samples to the total number of test samples.

First, we established three benchmarks. The first benchmark was established by taking the ratio between the most frequent sense for a given acronym found in the test data to the total number of test samples. Since the data is highly skewed for some of the acronyms, we would expect to see a fairly high benchmark with this approach. The other two benchmarks were established by doing a 10-fold cross-validation test on the test data with two standard machine learning algorithms: C5.0 Decision Tree and Maximum Entropy classifiers. Table 4 contains the benchmark accuracies.

| | **Accuracy (%)** | | |
|---|---|---|---|
| | **Majority Sense** | **C5.0** | **Max Ent** |
| AC | 31.40 | 94.60 | 96.70 |
| ACA | 87.40 | 93.10 | 97.00 |
| APC | 42.30 | 90.70 | 95.90 |
| CF | 76.30 | 95.80 | 94.20 |
| HA | 92.30 | 94.70 | 95.80 |
| LA | 88.50 | 92.60 | 94.60 |
| NSR | 99.00 | 98.80 | 99.00 |
| PE | 48.30 | 90.80 | 93.30 |
| **Mean (all)** | 70.70 | 93.90 | 95.80 |
| **Mean (bal)** | 49.60 | 92.90 | 95.40 |
| **Mean stdev. (X-validation)** | | 0.96 | 2.97 |

**Table 4. Benchmark accuracies (two means are reported – 'all' across all acronyms and 'bal' only across acronyms with more or less balanced distribution)**

The results in Table 4 show that even a straightforward 'bag-of-words' approach to disambiguation of acronyms gets fairly good results where the training data is manually annotated.

| **Acr.** | **Base** | **Web** | **Med** | **Mayo** | **Mayo Med** | **Mayo Web** |
|---|---|---|---|---|---|---|
| AC | 0.2 | 92.4 | 91.1 | 38.4 | 81.9 | 82.7 |
| **ACA** | **87.4** | **62.2** | **72.2** | **89.1** | **89.3** | **89.6** |
| APC | 12.5 | 46.0 | 36.2 | 76.9 | 64.9 | 56.1 |
| CF | 0.1 | 55.2 | 56.9 | 67.6 | 80.6 | 80.3 |
| **HA** | **92.3** | **50** | **22.4** | **71.1** | **74.3** | **74.1** |
| **LA** | **0.2** | **31.3** | **9.4** | **6.2** | **8.9** | **13.1** |
| **NSR** | **99.0** | **99.0** | **99.0** | **85.7** | **85.7** | **85.7** |
| PE | 4.4 | 61.2 | 47.0 | 59.5 | 57.2 | 58.4 |
| **Mean (all)** | 37 | 62.2 | 54.3 | 61.8 | 67.8 | 67.5 |
| **Mean (bal)** | 4.0 | 63.7 | 57.8 | 60.6 | 71.1 | 69.4 |

**Table 5. Experimental % accuracy results with 5 sources of data (bold indicates highly skewed distribution, two means are reported – 'all' across all acronyms and 'bal' only across acronyms with more or less balanced distribution).**

Except for NSR, in all other cases both C5.0 and Maximum Entropy outperform the "most frequent" approach. NSR is so highly skewed (only 3 out of 405 samples are different from the rest) that it is unlikely that any algorithm will be able to 'beat' the "most frequent" approach.

Table 5 contains the results of experimenting with collecting training data from various sources: WWW, MEDLINE, MAYO, MAYO+WWW and MAYO+MEDLINE. The results in Table 5 indicate that a combination of the data derived from clinical notes and MEDLINE (Mayo+Med) generates the best accuracy with the context vector matching approach. The baseline accuracy numbers were derived by taking the most frequent sense in the

Mayo+Med training data as the suggestion for each test sample.

The majority sense in the training data happened to coincide with the majority sense in the test data for ACA (adenocarcinoma), HA (headache) and NSR (normal sinus rhythm). This fact, coupled with the skewness of sense distribution, resulted in a very high baseline for these 3 acronyms. For the rest of the acronyms, the skewness of the data had the opposite effect resulting in very low baselines. LA presents an interesting case. The majority sense in the Mayo+Med training data is 'left arm' while the predominant majority sense in the test data is 'long acting.' The fact that the Web-only approach (31.3%) outperforms all others is probably due to the fact that only the data used to generate vectors for this approach contained the sense 'long acting.'

| Acr. | Context Vectors | Max Ent |
|---|---:|---:|
| AC | 81.85 | 41.25 |
| ACA | 89.25 | 88.88 |
| APC | 64.89 | 63.56 |
| CF | 80.6 | 70.08 |
| HA | 74.26 | 91.35 |
| LA | 8.85 | 3.2 |
| NSR | 85.67 | 98.76 |
| PE | 57.22 | 40.07 |
| **Mean (all)** | 67.82 | 62.1438 |
| **Mean(bal)** | 71.14 | 53.74 |

**Table 5. Experimental % accuracy results comparing the performance of context vectors and maximum entropy (two means are reported – 'all' across all acronyms and 'bal' only across acronyms with more or less balanced distribution).**

We also compared the context vector matching method for disambiguating acronyms with a standard maximum entropy based classifier. We trained maximum entropy classifiers using MAYO+MED samples for all 8 acronyms. The test results are presented in Table 6 and indicate a substantial advantage of the context vectors over maximum entropy particularly where the meanings of acronyms have a distribution balanced at least between 2 meanings.

This study has a number of limitations. Currently, we do not control for the fact that MEDLINE abstracts typically have acronyms defined upon their first appearance in text while clinical notes and general WWW derived acronyms are not likely to be defined anywhere in the text. Another obvious limitation is that we have a relatively small set of testable acronyms whose sense inventory distribution is highly skewed. We intend to address these limitations in future work by developing a more principled approach to selecting the acronyms and the raw data used to generate training samples.

## CONCLUSION

This paper presents preliminary results suggesting that using the WWW in conjunction with clinical corpora can be used for generating training data for acronym disambiguation. These results are encouraging as they suggest that there is a potential in leveraging very large amounts of publicly available data for disambiguating acronyms found in clinical discourse. The results of this study also indicate that a disambiguation method based on the vector space model may be more effective in conjunction with the proposed data generation approach than standard classification methods such as maximum entropy.

## References

1. *UMLS Knowledge Sources* [computer program]. Version 2001 AB, 2004 AA. Bethesda, MD: National Library of Medicine; 2004.
2. Liu H, Lussier Y, Friedman C. A Study of Abbreviations in UMLS. Paper presented at: American Medical Informatics Association (AMIA), 2001.
3. Liu H, Aronson A, Friedman C. A study of abbreviations in MedLINE Abstracts. Paper presented at: American Medical Informatics Association (AMIA), 2002.
4. Friedman C, Liu H, Shagina L, Johnson SB, Hripcsak G. Evaluating the UMLS as a Source of Lexical Knowledge for Medical Language Processing. Paper presented at: American Medical Informatics Association, 2001.
5. Balck E. An experiment in computational discrinmination of English word senses. *IBM Journal of Research and Development.* 1988;32(2):185-194.
6. Manning C, H. S. *Foundations of Statistical Natural Language Processing.* Cambridge, MA: MIT Press; 1999.
7. Hearst M. Noun homograph disambiguation using local context in large text corpora. Paper presented at: 7th Annual Conference of the University of Waterloo Center for the new OED and Text Research, 1991; Oxford.
8. Yarowski D. Unsupervised word sense disambiguation rivaling supervised methods. Paper presented at: Association for Computational Linguistics (ACL), 1995.
9. Ide N, Veronis J. Word sense disambiguation: the state of the art. *Computational Linguistics.* 1998;24(1).
10. Pakhomov S. Semi-Supervised Maximum Entropy Based Approach to Acronym and Abbreviation Normalization in Medical Texts. Paper presented at: Association for Computational Linguistics (ACL), 2003.
11. Liu H, Johnson SB, Fridman C. Automatic Resolution of Ambiguous Terms Based on Machine LEarning and Conceptual Relations. *Journal of American Medical Informatics Association.* 2002;9:621-636.
12. Pedersen T, Patwardhan S, Michelizzi J. WordNet::Similarity - Measuring the Relatedness of Concepts. Paper presented at: Nineteenth National Conference on Artificial Intelligence (AAAI-04), 2004.