# Semi-Supervised Maximum Entropy Based Approach to Acronym and Abbreviation Normalization in Medical Texts

Serguei Pakhomov, Ph.D.
Mayo Foundation, Rochester, MN
pakhomov.sergey@mayo.edu

## Abstract

Text normalization is an important aspect of successful information retrieval from medical documents such as clinical notes, radiology reports and discharge summaries. In the medical domain, a significant part of the general problem of text normalization is abbreviation and acronym disambiguation. Numerous abbreviations are used routinely throughout such texts and knowing their meaning is critical to data retrieval from the document. In this paper I will demonstrate a method of automatically generating training data for Maximum Entropy (ME) modeling of abbreviations and acronyms and will show that using ME modeling is a promising technique for abbreviation and acronym normalization. I report on the results of an experiment involving training a number of ME models used to normalize abbreviations and acronyms on a sample of 10,000 rheumatology notes with ~89% accuracy.

## 1 Introduction and Background

Text normalization is an important aspect of successful information retrieval from medical documents such as clinical notes, radiology reports and discharge summaries, to name a few. In the medical domain, a significant part of the general problem of text normalization is abbreviation and acronym[1] disambiguation. Numerous abbreviations are used routinely throughout such texts and identifying their meaning is critical to understanding of the document. The problem is that abbreviations are highly ambiguous with respect to their meaning. For example, according to UMLS®[2] (2001), RA may stand for "rheumatoid arthritis", "renal artery", "right atrium", "right atrial", "refractory anemia", "radioactive", "right arm", "rheumatic arthritis," etc. Liu et al. (2001) show that 33% of abbreviations listed in UMLS are ambiguous. In addition to problems with text interpretation, Friedman, et al. (2001) also point out that abbreviations constitute a major source of errors in a system that automatically generates lexicons for medical 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." In a way, abbreviation normalization problem is a special case of the word sense disambiguation (WSD) problem. Modern approaches to WSD include supervised machine learning techniques, where some amount of training

---

[1] To save space and for ease of presentation, I will use the word "abbreviation" to mean both "abbreviation" and "acronym" since the two could be used interchangeably for the purposes described in this paper.

[2] Unified Medical Language System®, a database containing biomedical information and a tools repository developed at the National Library of Medicine to help helath professionals as well as medical informatics researchers.

data is marked up by hand and is used to train a classifier. One such technique involves using a decision tree classifier (Black 1988). On the other side of the spectrum, the fully unsupervised learning methods such as clustering have also been successfully used (Shutze 1998). 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 (Hearst 1991, Yarowski 1995). Regardless of the technique that is used for WSD, the most important part of the process is the context in which the word appears (Ide and Veronis 1998). This is also true for abbreviation normalization.

For the problem at hand, one way to take context into account is to encode the type of discourse in which the abbreviation occurs, where discourse is defined narrowly as the type of the medical document and the medical specialty, into a set of explicit rules. If we see RA in a cardiology report, then it can be normalized to "right atrial"; otherwise, if it occurs in the context of a rheumatology note, it is likely to mean "rheumatoid arthritis" or "rheumatic arthritis." This method of explicitly using global context to resolve the abbreviation ambiguity in suffers from at least three major drawbacks from the standpoint of automation. First of all, it requires a database of abbreviations and their expansions linked with possible contexts in which particular expansions can be used, which is an error-prone labor intensive task. Second, it requires a rule-based system for assigning correct expansions to their abbreviations, which is likely to become fairly large and difficult to maintain. Third, the distinctions made between various meanings 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.

The approach I am investigating falls into the hybrid category of bootstrapping or semi-supervised approaches to training classifiers; however, it uses a different notion of bootstrapping from that of Hearst (1991) and Yarowski (1995). The bootstrapping portion of this approach consists of using a hand crafted table of abbreviations and their expansions pertinent to the medical domain. This should not be confused with dictionary or semantic network approaches. The table of abbreviations and their expansions is just a simple list representing a one-to-many relationship between abbreviations and their possible "meanings" that is used to automatically label the training data.

To disambiguate the "meaning" of abbreviations I am using a Maximum Entropy (ME) classifier. Maximum Entropy modeling has been used successfully in the recent years for various NLP tasks such as sentence boundary detection, part-of-speech tagging, punctuation normalization, etc. (Berger 1996, Ratnaparkhi 1996, 1998, Mikheev 1998, 2000). In this paper I will demonstrate using Maximum Entropy for a mostly data driven process of abbreviation normalization in the medical domain.

In the following sections, I will briefly describe Maximum Entropy as a statistical technique. I will also describe the process of automatically generating training data for ME modeling and present examples of training and testing data obtained from a medical sub-domain of rheumatology. Finally, I will discuss the training and testing process and present the results of testing the ME models trained on two different data sets. One set contains one abbreviation per training/testing corpus and the other -- multiple abbreviations per corpus. Both sets show around 89% accuracy results when tested on the held-out data.

## 2    Clinical Data

The data that was used for this study consists of a corpus of ~10,000 clinical notes (medical dictations) extracted at random from a larger corpus of 171,000 notes (~400,000 words) and encompasses one of many medical specialties at the Mayo Clinic – rheumatology. In the Mayo Clinic's setting, each clinical note is a document recording information pertinent to treatment of a patient that consists of a number of subsections such as Chief Complaint (CC), History of Present Illness (HPI), Impresssion/Report/Plan (IP), Final Diagnoses (DX)[3], to name a few. In clinical settings other than the Mayo Clinic, the notes may have different segmentation and section headings; however, most clinical notes in most clinical settings do have some sort of segmentation and contain some sort of discourse markers, such as CC, HPI, etc., that can be useful clues to tasks such as the one discussed in this paper. Theoretically, it is possible that an abbreviation such as PA may stand for "paternal aunt" in the context of Family History (FH), and "polyarthritis" in the Final Diagnoses context. ME technique lends itself to modeling information that comes from a number of heterogeneous sources such as various levels of local and discourse context.

## 3    Methods

One of the challenging tasks in text normalization discussed in the literature is the detection of abbreviations in unrestricted text. Various techniques, including ME, have proven useful for detecting abbreviations with varying degrees of success. (Mikheev 1998, 2000, Park and Byrd 2001) It is important to mention that the methods described in this paper are different from abbreviation detection; however, they are meant to operate in tandem with abbreviation detection methods.

Two types of methods will be discussed in this section. First, I will briefly introduce the Maximum Entropy modeling technique and then the method I used for generating the training data for ME modeling.

### 3.1    Maximum Entropy

This section presents a brief description of ME. A more detailed and informative description can be found in Berger (1996)[4], Ratnaparkhi (1998), Manning and Shutze (2000) to name just a few.

Maximum Entropy is a relatively new statistical technique to Natural Language Processing, although the notion of maximum entropy has been around for a long time. One of the useful aspects of this technique is that it allows to predefine the characteristics of the objects being modeled. The modeling involves a set of predefined features or constraints on the training data and uniformly distributes the probability space between the candidates that do not conform to the constraints. Since the entropy of a uniform distribution is at its maximum, hence the name of the modeling technique.

Features are represented by indicator functions of the following kind[5]:

$$(1) \qquad F(o,c) = \begin{cases} 1, & \text{if } o = x \text{ and } c = y \\ 0, & \text{otherwise} \end{cases}$$

Where "o" stands for outcome and "c" stands for context. This function maps contexts and outcomes to a binary set. For

---

[3] This format is specific to the Mayo Clinic. Probably the most commonly used format outside of Mayo is the so-called SOAP format that stands for Subjective, Objective, Assessment, Plan. The idea is the same, but the granularity is lower.

[4] This paper presents an Improved Iterative Scaling but covers the Generalized Iterative Scaling as well.

[5] Borrowed from Ratnaparkhi implementation of POS tagger.

example, to take a simplified part-of-speech tagging example, if y = "the" and x="noun", then F(o,c) = 1, where y is the word immediately preceding x. This means that in the context of "the" the next word is classified as a noun.

To find the maximum entropy distribution the Generalized Iterative Scaling (GIS) algorithm is used, which is a procedure for finding the maximum entropy distribution that conforms to the constraints imposed by the empirical distribution of the modeled properties in the training data[6].

For the study presented in this paper, I used an implementation of ME that is similar to that of Ratnaparkhi's and has been developed as part of the open source Maxent 1.2.4 package[7]. (Jason Baldridge, Tom Morton, and Gann Bierner, http://maxent.sourceforge.net). In the Maxent implementation, features are reduced to contextual predicates, represented by the variable y in (1). Just as an example, one of such contextual predicates could be the type of discourse that the outcome "o" occurs in: PA → paternal aunt | y = FH; PA → polyarthritis | y = DX. Of course, using discourse markers as the only contextual predicate may not be sufficient. Other features such as the words surrounding the abbreviation in question may have to be considered as well.

For this study two kinds of models were trained for each data set: local context models (LCM) and combo (CM) models. The former were built by training on the sentence-level context only defined as two preceding $(w_{i-2}, w_{i-1})$ and two following $(w_{i+1}, w_{i+2})$ words surrounding an abbreviation expansion. The latter kind is a model trained on a combination of sentence and section level contexts defined simply as

the heading of the section in which an abbreviation expansion was found.

## 3.2 Generating simulated training data

In order to generate the training data, first, I identify potential candidates for an abbreviation by taking the list of expansions from a UMLS database and applying it to the raw corpus of text data in the following manner. The expansions for each abbreviation found in the UMLS's LRABR table are loaded into a hash indexed by the abbreviation.

| ABBR | EXPANSIONS FOUND IN DATA |
|---|---|
| NR | normal range; no radiation; no recurrence; no refill; nurse; nerve root; no response; no report; nonreactive; nonresponder |
| PA | Polyarteritis; pseudomonas aeruginosa; polyarthritis; pathology; pulmonary artery; procainamide; paternal aunt; panic attack; pyruvic acid; paranoia; pernicious anemia; physician assistant; pantothenic acid; plasma aldosterone; periarteritis |
| PN | Penicillin; pneumonia; polyarteritis nodosa; peripheral neuropathy; peripheral nerve; polyneuropathy pyelonephritis; polyneuritis; parenteral nutrition; positional nystagmus; periarteritis nodosa |
| BD | band; twice a day; bundle |
| INF | Infection; infected; infusion; interferon; inferior; infant; infective |
| RA | Rheumatoid arthritis; renal artery; radioactive; right arm; right atrium; refractory anemia; rheumatic arthritis; right atrial |

**Table 1. Expansions found in the training data and their abbreviations found in UMLS.**

The raw text of clinical notes is input and filtered through a dynamic sliding-

---

[6] A consice step-by-step description and an explanation of the algorithm itself can be found in Manning and Shutze (2000).

[7] The ContextGenerator class of the maxent package was modified to allow for the features discussed in this paper.

window buffer whose maximum window size is set to the maximum length of any abbreviation expansion in the UMLS. When a match to an expansion is found, the expansion and it's context are recorded in a training file as if the expansion were an actual abbreviation. The file is fed to the ME modeling software. In this particular implementation, the context of 7 words to the left and 7 words to the right of the found expansion as well as the section label in which the expansion occurs are recorded; however, not all of this context ended up being used in this study.

This methodology makes a reasonable assumption that given an abbreviation and one of it's expansions, the two are likely to have similar distribution. For example, if we encounter a phrase like "rheumatoid arthritis", it is likely that the context surrounding the use of an expanded phrase "rheumatoid arthritis" is similar to the context surrounding the use of the abbreviation "RA" when it is used to refer to rheumatoid arthritis. The following subsection provides additional motivation for using expansions to simulate abbreviations.

### 3.2.1 Distribution of abbreviations compared to the distribution of their expansions

Just to get an idea of how similar are the contexts in which abbreviations and their expansions occur, I conducted the following limited experiment. I processed a corpus of all available rheumatology notes (171,000) and recorded immediate contexts composed of words in positions $\{w_{i-1}, w_{i-2}, w_{i+1}, w_{i+2}\}$ for one unambiguous abbreviation – DJD (degenerative joint disease). Here $w_i$ is either the abbreviation DJD or its multiword expansion "degenerative joint disease." Since this abbreviation has only one possible expansion, we can rely entirely on finding the strings "DJD" and "degenerative joint disease" in the corpus without having to disambiguate the abbreviation by hand in each instance. For each instance of the strings "DJD" and "degenerative joint

disease", I recorded the frequency with which words (tokens) in positions $w_{i-1}$, $w_{i-2}$, $w_{i+1}$ and $w_{i+2}$ occur with that string as well as the number of unique strings (types) in these positions.

It turns out that "DJD" occurs 2906 times , "degenerative joint disease" occurs 2517 times. Of the 2906 occurrences of DJD, there were 204 types that occurred immediately prior to mention of DJD ($w_{i-1}$ position) and 115 types that occurred immediately after ($w_{i+1}$ position). Of the 2517 occurrences of "degenerative joint disease", there were 207 types that occurred immediately prior to mention of the expansion ($w_{i-1}$ position) and 141 words that occurred immediately after ($w_{i+1}$ position). The overlap between DJD and its expansion is 115 types in $w_{i-1}$ position and 66 types in $w_{i+1}$ position. Table 2 summarizes the results for all four $\{w_{i-1}, w_{i-2}, w_{i+1}, w_{i+2}\}$ positions.

| Context | Context overlap | N of unique contexts | Context similarity (%) |
|---|---|---|---|
| **$W_{i-1}$** | | | |
| DJD | 115 | 204 | 56 |
| degen. joint dis | 115 | 207 | 55 |
| **Mean** | | | **55.5** |
| **$W_{i+1}$** | | | |
| DJD | 66 | 115 | 50 |
| degen. joint dis | 66 | 141 | 46 |
| **Mean** | | | **48** |
| **$W_{i-2}$** | | | |
| DJD | 189 | 371 | 50 |
| degen. joint dis | 189 | 410 | 46 |
| **Mean** | | | **48** |
| **$W_{i+2}$** | | | |
| DJD | 126 | 245 | 51 |
| degen. joint dis | 126 | 301 | 41 |
| **Mean** | | | **46** |
| **Total** | | | **49.37** |

**Table 2. DJD vs. "degenerative joint disease" distribution comparison.**

On average, the overlap between the contexts in which DJD and "degenerative

joint disease" occur is around 50%, which is a considerable number because this overlap covers on average 91% of all occurrences in $w_{i-1}$ and $w_{i+1}$ as well as $w_{i-2}$ and $w_{i+2}$ positions.

### 3.2.2 Data sets

One of the questions that arose during implementation is whether it would be better to build a large set of small ME models trained on sub-corpora containing context for each abbreviation of interest separately or if it would be more beneficial to train one model on a single corpus with contexts for multiple abbreviations.

This was motivated by the idea that ME models trained on corpora focused on a single abbreviation may perform more accurately; even though such approach may be computationally expensive.

| ABBR | N OF UMLS EXPANSIONS | N OF OBSERVED EXPANSIONS |
|------|---------------------|--------------------------|
| NR | 23 | 10 |
| PA | 72 | 15 |
| PN | 28 | 11 |
| BD | 30 | 3 |
| INF | 13 | 7 |
| RA | 28 | 8 |
|  |  |  |
| Mean | 32.33 | 9 |

**Table 3. A comparison between UMLS expansions for 6 abbreviations and the expansions actually found in the training data.**

For this study, I generated two sets of data. The first set (Set A) is composed of training and testing data for 6 abbreviations (NR, PA, PN, BD, INF, RA), where each training/testing subset contains only one abbreviation per corpus. resulting in six subsets. Table 1 shows the potential expansions for these abbreviations that were actually found in the training corpora.

Not all of the possible expansions found in the UMLS for a given abbreviations will be found in the text of the clinical notes. Table 3 shows the number of expansions actually found in the rheumatology training data for each of the 6 abbreviations listed in Table 1 as well as the expansions found for a given abbreviation in the UMLS database.

The UMLS database has on average 3 times more variability in possible expansions that were actually found in the given set of training data. This is not surprising because the training data was derived from a relatively small subset of 10,000 notes.

The other set (Set B) is similar to the first corpus of training events; however, it is not limited to just one abbreviation sample per corpus. Instead, it is compiled of training samples containing expansions from 69 abbreviations. The abbreviations to include in the training/testing were selected based on the following criteria:

a.    has at least two expansions
b.    has 100-1000 training data samples

The data compiled for each set and subset was split at random in the 80/20 fashion into training and testing data. The two types of ME models (LCM and CM) were trained for each subset on 100 iterations through the data with no cutoff (all training samples used in training).

## 4    Testing

To summarize the goals of this study, one of the main questions in this study is whether local sentence-level context can be used successfully to disambiguate abbreviation expansion. Another question that naturally arose from the structure of the data used for this study is whether more global section-level context indicated by section headings such as "chief complaint", "history of present illness" , etc., would have an effect on the accuracy of predicting the

abbreviation expansion. Finally, the third question is whether it is more beneficial to construct multiple ME models limited to a single abbreviation. To answer these questions, 4 sets of tests were conducted:

1. Local Context Model and Set A
2. Combo Model and Set A
3. Local Context Model and Set B
4. Combo Model and Set B

## 4.1    Results

Table 3 summarizes the results of training Local Context models with the data from Set A (one abbreviation per corpus).

| ABBR | Acc. (%) | Test Event | Train Events | Out. | Predic. |
|---|---|---|---|---|---|
| | | | | | |
| NR | 87.87 | 139.6 | 495.7 | 10.8 | 580.4 |
| PN | 77.05 | 166.2 | 612.7 | 11 | 722.5 |
| BD | 98.49 | 174.4 | 724.6 | 3 | 704.8 |
| PA | 86.45 | 182.8 | 653.3 | 13.9 | 707.1 |
| INF | 87.33 | 196.2 | 819.3 | 6.9 | 950.3 |
| RA | 97.67 | 924.6 | 2535 | 7.6 | 1549.4 |
| | | | | | |
| Mean | 89.14 | 297.3 | 973.43 | 8.87 | 869.08 |

**Table 3. Local Context Model and Set A results**

The results in Table 3 show that, on average, after a ten-fold cross-validation test, the expansions for the given 6 abbreviations have been predicted correctly 89.14%.

| ABBR | Acc. (%) | Test Event | Train Events | Out. | Predic. |
|---|---|---|---|---|---|
| | | | | | |
| NR | 89.515 | 139.6 | 504.6 | 10.8 | 589.4 |
| PN | 78.739 | 166.2 | 618.7 | 11 | 746.1 |
| BD | 98.39 | 174.4 | 736.6 | 3 | 713.8 |
| PA | 86.193 | 182.8 | 692.2 | 13.9 | 717 |
| INF | 87.409 | 196.2 | 842.3 | 7 | 959.8 |
| RA | 97.693 | 924.6 | 2704 | 7.6 | 1559.4 |
| | | | | | |
| Mean | 89.66 | 297.3 | 1016.4 | 8.88 | 880.92 |

**Table 4. Combo Model and Set A results**

Table 3 as well as table 4 display the accuracy, the number of training and testing events/samples, the number of outcomes (possible expansions for a given abbreviation) and the number of contextual predicates averaged across 10 iterations of the cross-validation test.

Table 4 presents the results of the Combo approach with the data also from Set A. The results of the combined discourse + local context approach are only slightly better that those of the sentence-level only approach.

Table 5 displays the results for the set of tests performed on data containing multiple abbreviations – Set B but contrasts the Local Context Model with the Combo Model.

| | Acc. (%) | Test Event | Train Event | Out. | Pred. |
|---|---|---|---|---|---|
| LCM | 89.169 | ~4791 | ~21999 | ~250 | ~9400 |
| CM | 89.015 | ~4792 | ~22000 | ~251 | ~9401 |

**Table 5. Local Context Model performance contrasted to Combo model performance on Set B**

The first row shows that the LCM model performs with 89.17% accuracy. CM's result is very close: 89.01%. Just as with Tables 3 and 4, the statistics reported in Table 5 are averaged across 10 iterations of cross-validation.

## 5    Discussion

The results of this study suggest that using Maximum Entropy modeling for abbreviation disambiguation is a promising avenue of research as well as technical implementation for text normalization tasks involving abbreviations. Several observations can be made about the results of this study. First of all, the accuracy results on the small pilot sample of 6 abbreviations as well as the larger sample with 69 abbreviations are quite encouraging in light of the fact that the training of the ME models is largely unsupervised[8].

---

[8] With the exception of having to have a database of acronym/abbreviations and their expansions which has to be compiled by hand. However, once such list is compiled, any amount of data can be used for training with no manual annotation.

Another observation is that it appears that using section-level context is not really beneficial to abbreviation expansion disambiguation in this case. The results, however, are not by any means conclusive. It is entirely possible that using section headings as indicators of discourse context will prove to be beneficial on a larger corpus of data with more than 69 abbreviations.

The abbreviation/acronym database in the UMLS tends to be more comprehensive than most practical applications would require. For example, the Mayo Clinic regards the proliferation of abbreviations and acronyms with multiple meanings as a serious patient safety concern and makes efforts to ensure that only the "approved" abbreviations (these tend to have lower ambiguity) are used in clinical practice, which would also make the task of their normalization easier and more accurate. It may still be necessary to use a combination of the UMLS's and a particular clinic's abbreviation lists in order to avoid missing occasional abbreviations that occur in the text but have not made it to the approved clinic's list. This issue also remains to be investigated.

## 6    Future Work

In the future, I am planning to test the assumption that abbreviations and their expansions occur in similar contexts by testing on hand-labeled data. I also plan to vary the size of the window used for determining the local context from two words on each side of the expression in question as well as the cutoff used during ME training. It will also be necessary to extend this approach to other medical and possibly non-medical domains with larger data sets. Finally, I will experiment with combining the UMLS abbreviations table with the Mayo Clinic specific abbreviations.

## References

Baldridge, J., Morton, T., and Bierner, G URL: http://maxent.sourceforge.net

Berger, A., Della Pietra, S., and Della Pietra, V. (1996). A maximum entropy approach to natural language processing. Computational Linguistics, 22(1):39-71.

Black, E. (1988). An experiment in computational discrinmination of English word senses. IBM Journal of Research and Development, 32(2), 185-194.

Friedman, C., Liu, H., Shagina, L., Johnson, S. and Hripcsack, G. (2001) Evaluating the UMLS as a Source of Lexical Knowledge for Medical Language Processing. In Proc AMIA 2001.

Hearst, M. (1991). Noun homograph disambiguation using local context in large text corpora. In Proc. 7th Annual Conference of the University of Waterloo Center for the new OED and Text Research, Oxford.

Ide, N and Veronis, J. (1998). Word sense disambiguation: the state of the art. Computational Linguistics, 24(1).

Liu, H., Lussier, Y., and Friedman, C. (2001) A Study of Abbreviations in UMLS. In Proc. AMIA 2001.

Mikheev, A. (2000). Document Centered Approach to Text Normalization. In Proc. SIGIR 2000.

Mikheev, A. (1998). Feature Lattices for Maximum Entropy Modeling. In Proc. ACL 1998.

Manning, C. and Shutze H. (1999). Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA.

Park, Y and Byrd, R. (2001). Hybrid text Mining for Finding Abbreviations and their Definitions. In Proc. EMNLP 2001.

Ratnaparkhi A. (1996). A maximum entropy part of speech tagger. In Proceedings of the conference on empirical methods in natural language processing, May 1996, University of Pennsylvania

Ratnaparkhi A. (1998). Maximum Entropy Models for Natural Language Ambiguity Resolution. Ph. D. Thesis, U of Penn.

Jurafski D. and Martin J. (2000). Speech and Language Processing. Prentice Hall, NJ.

Yarowski, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In Proc. ACL-95, 189-196.

UMLS. (2001). UMLS Knowledge Sources (12th ed.). Bethesda (MD) : National Library of Medicine.