Maximum Entropy Modeling for Mining Patient Medication Status from Free Text

Serguei V. Pakhomov[1], PhD, Alexander Ruggieri[1,2], MD, Christopher G. Chute[1], MD, DrPH
[1]Division of Medical Informatics Research, Department of Health Sciences Research
[2]Division of Rheumatology, Department of Medicine
Mayo Clinic, Rochester MN

ABSTRACT

Using a classification scheme of patient medication status we sought to recognize and categorize medications mentioned in the unrestricted text of clinical documents generated in clinical practice. The categories refer to the patient's status with respect to the medication such as discontinuation, start or initiation, and continuation of a given medication. This categorization is performed with a machine learning technique, Maximum Entropy (ME), that is well suited to incorporating heterogeneous sources of information necessary for classifying patient's medication status. We use hand labeled training data to generate ME models and test 5 different training feature sets. Our results show that the most optimal feature set includes a combination of the following: two words preceding and following the mention of the drug, the subject of the sentence in which the drug mention occurs, the 2 words following the subject, and a binary feature vector of lexicalized semantic cues indicative of medication status or its change. The average predictive power of a model trained on these features is ~89%.

INTRODUCTION

Important decisions in medical care frequently revolve around a patient's medication status. Information recording the actions and observations surrounding patient medications is distributed in medical records that span multiple care providers , and which can involve differing documentation formats. While, even though dutifully recorded, valuable medication information is commonly dispersed and embedded in unrestricted text narrative spanning multiple sections in a medical document. For instance, a specific medical record document format may contain a sections reserved for an ordered listing of "Current Medications", while other sections might include observations of past medication use ("past history"), problems associated with past or current use (allergies or drug reactions), or forecasts of potential future use (plans for treatment). This presents daunting challenges for medication related information retrieval. Consider a use case scenario that requires identification of all patients with current or past use of a medication because of a newly discovered side effect. Although language referencing "current use" of the medication of interest might be extractable from certain dedicated sections in the document, other sections may hold important information about "past use". Machine learning techniques can potentially help in confronting the challenge retrieval of patient medication status information embedded and dispersed in medical narratives.

Mayo Clinic is a group medical practice in the United States and spans all recognized medical care settings and specialties. Currently over 50,000 patient visits occur each week that generate 40,000 medical documentation entries in Mayo electronic record that principally consists of text narratives. For this study, we extracted a small corpus of 10,000 narratives from the clinical rheumatology practice at Mayo Clinic

This paper presents an approach to information retrieval that is based on a relatively novel statistical technique - Maximum Entropy (ME) which is used in NLP for various tasks ranging from sentence boundary detection [4] to parsing [3]. In this paper we will describe the classification of patient's medication status, the technique for its identification in unrestricted text, using constructed ME models from a small corpus of hand labeled data, and validation of the technique on a reserved test set of data. The main goal of this pilot was to determine whether sentence-internal local context can be used in ME modeling for successful automated classification of patient medication status, and determination of the optimal set of features used to train the ME models. We also explore the efficacy of the proposed classification of patient medication status and lay out future directions for this research.

METHODS

Medication Status Classification

The classification consists of four categories: past (P), for past medication use, continuing (C) for medication use that continues or is continued, discontinued (D) for medication use that is discontinued, and started (S) for medication use that is initiated or begun. The "past" category includes cases with evidence that the patient has or had been taking the drug sometime in the past. "Continuing" status, means there is evidence for concurrence with the patient continuing on the current medication. "Discontinued" status means that the medication is being stopped for some reason such as a recall, an allergic reaction, or lack of effectiveness. "Started" status means that the patient is being started on a drug therapy. We also found unclassifiable mentions of a drug or medication that were ignored in this test. These involved mention of a substance that could represent a medication but was used in the context of a chemical substance being tested . An example of this is given in (5).

1. Past: she had been reluctant about disease modifying agents but had a course of **methotrexate** 7.5 mg weekly from approximately November 1992 for approximately a year
2. Continuing: I would recommend increasing his **Imuran** to 50 mg twice per day and continuing with the **prednisone** 60 mg per day
3. Stop: Intolerant of **Anaprox** (itchy skin)
4. Start: The DOCTOR started her back on **Prednisone** 20 mg every other day as of last Friday
5. Not Classified: **Glucose** is normal at 84.

The examples (1-5) show that there are a number of linguistic cues that speakers use to signal the patient's medication status.

Medication status appears to be encoded by a variety of linguistic entities and can be viewed as being expressed using sets of linguistic elements[1] such as the lexical items surrounding the mention of a drug and the grammatical aspect of the sentence in which the drug name appears. It is these sets of elements that we will try to isolate to be used for Maximum Entropy modeling discussed in the following section.

[1] These are really sets of linguistic features; however, we don't want them to be confused with the features used for ME modeling, so we use the word "elements" instead.

Maximum Entropy Modeling

In this section we give a brief description of the machine learning approach used to address the problem. A more detailed and informative description can be found in [1] [2], [4], [5].

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 one 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, the modeling technique borrows its name.

Features are represented by indicator functions of the following kind[3]:

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

Where o stands for outcome and c – for context. This function maps contexts and outcomes to a binary set. For example, if y = "increase" and x="C", then F(o,c) = 1. In other words, if Amoxicillin is mentioned and we classify that mention as C in the context of the word "increase", then the mapping from "increase" to C status is set to 1. For the particular task at hand, the outcomes are constrained to 4 possible choices: D,P.S and C.

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 modeled properties in the training data[4].

For the study presented in this paper, we used an implementation of ME that is similar to that of Ratnaparkhi's and has been developed as part of the OpenNLP initiative (Jason Baldridge, Tom Morton, and Gann Bierner http://maxent.sourceforge.net). In the OpenNLP implementation, features are reduced to contextual predicates, represented by the variable y. Just as an

[2] Berger et al.'s paper presents an Improved Iterative Scaling but covers the Generalized Iterative Scaling as well.

[3] Borrowed from Ratnaparkhi's implementation of a POS tagger.

[4] A complete and concise description and explanation of the algorithm can be found in Manning and Shutze (2000).

example, one of such contextual predicates could be the word that is immediately to the left (w_{i-1}) of the event whose outcome we are trying to predict: w_{i-1} = "discontinue" | y = D(discontinue), w_{i-1} = "increase" | y = C (continue). Of course, using w_{i-1} as the only contextual predicate may not be sufficient.

Other features such as w_{i-2}, w_{i+1}, w_{i+2}, the first word in the clause, presence/absence of auxiliary verbs such as "is, has, had, been", etc. are used in this study as well. An example in (6) illustrates some of the features that were predefined for this study.

6. she has been on **Prednisone** since that time and had tapered down to 5 mg per day

The following list of features is recorded and used for training for this example:

{w_i = "Prednisone"; w_{i-1} = "on"; w_{i-2} = "been"; w_{i+1} = "since"; w_{i+2} = "that"; s = "she"; s_{i+1} = "has"; s_{i+2} = "been";}

The S (subject) features are motivated by the fact that the auxiliary material following the subject may indicate the tense and aspect[5] of what is being predicated of the subject as well as the modality of the predication. If the subject is a personal pronoun such as "he" or "she", in a medical dictation, we can safely assume that it refers to the patient. Since the drug name appears as part of the predicate in the same sentence, we can assume that the aspect indicated by the auxiliary verbs such as "has been" may be relevant to determining the patient's medications status with respect to the drug name. Thus, by using the S features, we are attempting to capture tense and aspectual information encoded in the sentence and use it to classify the drug name along with the immediately surrounding context.

Another source of information that we used to classify drug mentions is composed of lexicalized semantic cues such as "increase", "decrease", "start", "stop", etc. The list is derived by empirical observation of the data and is displayed in Table 1. The cues are arranged in a feature vector where each feature is a binary value indicating presence or absence of a particular word in the context of the drug mention. It is important to set the proximity boundaries for the occurrence of such words with respect to the drug mention. If the cue is set too far off the drug mention, one probably should not use that cue as an indicator of medication status. It is unclear at this point how far is "too far"; therefore, for this first pass, we picked an arbitrary limit of 14 words – 7 prior to mention and 7 post mention of the drug.

Feature identifier	Keywords
incr	Increase, increased, increasing, increases
decr	Decrease, decreased, decreasing, decreases
cont	Continue, continues, continuing, continued
take	Take, took, taken, taking, takes
toler	Tolerate, tolerating, tolerates, tolerated
use	Use, used, using, uses
red	Reduce, reducing, reduces, reduced
ref	Refuse, refused, refuses, refusing
bmp	Bump, bumped, bumping
stop	Stop, stopped, stopping, stops
disc	Discontinue, discontinues, discontinued, discontinuing
all	Allergy, allergic, allergies,
naus	Nauseated, nauseates, nausea
tape	Taper, tapers, tapered, tapering
star	Start, starts, started, starting
swit	Switch, switches, switched, switching
chng	Change, changed, changing, changes
try	Try, tries, tried, trying
add	Add, adds, added, adding
res	Resume, resumes, resumed, resuming
rest	Restart, restarts, restarted, restarting
rx	Prescribe, prescribed, prescription, prescribing
beg	Begin, began, begun, beginning, begins
neg	No, not, any

Table 1. Lexicalized Semantic Cues Feature Vector and the morphological forms corresponding to the cues

DATA

The data for this study were compiled from a set of 10,000 of ~171,000 rheumatology clinical notes collected from the clinical notes repository at the Mayo Clinic[6] and consists of 814 single sentence (or clause) chunks containing one or more drug names. The sentences and clauses were extracted from the raw data based on cues such as personal pronouns in subjective case, punctuation

[5] By tense and aspect here we mean grammatical tense and aspect. For example, future tense (... will start her on ...) may indicate S status, whereas perfective/resultative aspect (... has stopped taking ...) may indicate D status.

[6] Careful attention was paid to the patient's privacy and information security. No patient identifying information made it into the data set used for this study.

and formatting of the clinical notes (in many cases the text appears in one sentence per line format) to determine the leftmost clause boundary.

Each instance of a drug name was classified into one of the 4 categories (C,S,D,P) by a trained rheumatologist and the data set was split 10 times into ~697 (80%) training and ~154 (20%) testing items at random resulting in 10 training-testing sets. The ME models were trained on 100 iterations with no frequency cutoff (all data samples participated in the training) for each of the 10 sets, resulting in a total of 30 training/test sets. The decision to use no frequency cutoff was based on the relatively small size of the training data.

ME Models

We trained 5 types of ME models. The different types of ME models reflected different feature sets. Type I includes the following set: $\{w_i, w_{i-1}, w_{i-2}, w_{i+1}, w_{i+2}, s, s_{i+1}, s_{i+2}\}$. Type II excludes all S features, which eliminates aspectual information as a predictive feature composed of $\{s, s_{i+1}, s_{i+2}\}$. Type III does include all S features but extends the contexts by one word: $\{w_i, w_{i-1}, w_{i-2}, w_{i-3}, w_{i+1}, w_{i+2}, w_{i+3}, s, s_{i+1}, s_{i+2}, s_{i+3}\}$. Type IV includes all features used in Type I as well as the feature vector shown in Table 1. Type V includes only features from the feature vector in Table 1 and excludes all other contextual features.

The five types can be separated into two groups: Types I-III and Types IV,V. The latter two are different from the former in that they are using predefined lexical cues as binary features. The former three use multi-valued features whose values are filled with lexical items.

RESULTS AND DISCUSSION

A 10-fold cross-validation test was performed on all three types of models. The Type I ME model yielded an average of 77.80 % accuracy of predicting the right category for a drug mention in the patient's record. The mean accuracy for 10 test iterations are shown in Table 2. The Type II ME model tested with a considerably lower accuracy – 74.82%. A paired t-test for means confirmed that the difference between Type I and Type II result sets is statistically significant at 0.002 p-value. The results for each iteration of cross-validation on Type II ME model are displayed in Table 2. The difference between Type I and Type III was negligible – 0.12% in favor of Type III. This difference is not statistically significant.

The results shown in Table 2 lump all 4 medication status categories together. Table 3 displays the averaged results separated by category. It is evident from the results in Table 3 that the ME models do very well on predicting the C category, but very poorly on predicting other categories, with P being the worst.

Including the keyword feature vector (see Table 2) improves the results dramatically. Table 4 displays the average accuracy for the ME models of Type IV and V. Type V model was used as a sanity check to ensure that the keywords feature vector is not the only thing that is responsible for the medication status determination, but that it indeed works in tandem with other contextual features, which is what makes the choice of ME models for this task ideal.

Iter.	Type I	Type	Type III
Mean	77.80	74.82	77.92

Table 2. Accuracy results for Type I,II and III models across 10 iterations of cross-validation.

The results show that Type IV model outperforms all other models and it is also the model that uses the greatest variety of information sources. The general conclusion that can be reached so far is that the optimal set of features for training ME models for automatic patient medication status classification consists of a combination of three types of features: immediate context (2 words to the left and right), subject context (the subject of the sentence + 2 words following it), a set of keywords that indicate patient medication status or its change. Another conclusion is that a set of lexicalized semantic cues, even as impoverished and incomplete as the one in Table 1 produces dramatic improvements in the classification accuracy.

Category	Type I	Type II	Type III
D	31.18%	27.64%	25.65%
C	94.17%	92.31%	94.99%
P	24.08%	18.51%	26.22%
S	43.08%	35.57%	41.41%

Table 3. Accuracy results for Type I, II and III ME models separated by status category

The vector elements presented in Table 1 are roughly based on the underlying lexical items and their morphological forms. Clearly, one could also group the lexical items according to their semantic properties, in which case "stop" feature, for example, will contain "stop, stopped, stopping, stops" as well as "discontinue, discontinues, etc.", "refuse, refuses, etc." and others that would fit the

semantic category of "stopping." The impact of such regrouping is unclear at the moment and needs to be investigated.

Overall, the results of this exploratory study are encouraging. We believe that the accuracy of the models can be improved in a number of ways. First of all, more training data is needed, which will enable us to experiment with pruning the ME models of potentially spurious data by using a frequency cutoff. A greater number of training samples is also likely to give better results with using larger sets of contextual features, so we expect the Type III model to perform better compared to Type I models proportionately with the size of the training corpus.

Category	Type IV	Type V
D	87.18%	24.98%
C	98.19%	79.14%
P	83.62%	0%
S	86.04%	20.20%

Table 4. Accuracy results for Type IV and V ME models separated by status category

It is also possible that our classification of the patient medication status does not generalize well enough across the data. An inter-rater agreement test with a number of physicians making classification judgements will enable us to improve on the classification which may also lead to better accuracy on the part of ME modeling.

An interesting property of this classification is that its elements stand in a unidirectional entailment relationship and can be arranged on the following hierarchy[7]:

(2) **D >> C >> P >> S**

If a drug has been discontinued (D), it necessarily implies that the patient has been continuing the drug (C), which in turn implies that there was some past use (P) and finally that it had been started at some point (S). If the drug is being continued (C), that implies that there was some past use (P) and that it had to have been started (S). Going the opposite direction along this hierarchy presents a different picture: if a patient is being started on a drug (S), it does not necessarily imply that there was any past use of the drug (P), or that it is being continued (C) or discontinued (D); past use of the drug (P) does not imply continuing (C) or discontinuing (D) and, obviously, continuing a drug (C) does not imply its discontinuing (D).

The unidirectional entailment relationship inherent in this hierarchy means, for example, that if a patient is told to increase the dose of Coumadin, it is assumed that the patient was taking the drug in the past and had to have been started on it. Also, when the patient is told to increase the dose, it means that the medication status of the drug is at most continuing (C). An expression such as "increase Coumadin to 500 mg" may signal that the highest status in this case is C and not D; however, that does not exclude other statuses, in fact, P and S are implied. The flip side of this is that all mentions of a drug can be viewed as having the S status. Even if the drug is being discontinued, it is still implied that it must have been started at some point.

Given this view of the relationships between the categories, the task of automatic classification becomes a task of determining the highest possible medication status for a particular drug mention. We can leverage this entailment relationship within the hierarchy of patent medication statuses to reduce the number of categories. The reduction can be achieved by folding lower categories into higher ones. For example, for some purposes, we could reduce the classification to a binary set {discontinued, continuing} by folding the past use and the started categories into the continuing category. The effects of such reduction are to be further investigated.

[7] This has been inspired by Gundel et al.'s Givenness Hierarchy [2].

REFERENCES

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

[2] Gundel, J., Hedberg, N. and Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. Language 69(2), pp. 274-307.

[3] Charniak, E. (2000). A Maximum-Entropy-Inspired Parser. In Proceedings of NAACL-2000.

[4] 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

[5] Manning, C. and Shutze H. (1999). Foundations of Statistical Natural Language Processing. MIT Press, Cambridge,Mass