Automatic Prediction of Trauma Registry Procedure Codes from Emergency Room Dictations

William R. Hersh[a], Todd K. Leen[b], P. Steve Rehfuss[b], Susan Malveau[a]

[a] *Division of Medical Informatics and Outcomes Research, School of Medicine,*
Oregon Health Sciences University, Portland, OR, USA
[b] *Center for Information Technology, Oregon Graduate Institute, Portland, OR, USA*

Abstract

Current natural language processing techniques for recognition of concepts in the electronic medical record have been insufficient to allow their broad use for coding information automatically. We have undertaken a preliminary investigation into the use of machine learning methods to recognize procedure codes from emergency room dictations for a trauma registry. Our preliminary results indicate moderate success, and we believe future enhancements with additional learning techniques and selected natural language processing approaches will be fruitful.

Keywords:

Machine learning, natural language processing, coding

Introduction

One of the major promises of the electronic medical record (EMR) is the ability to aggregate data from individual patients for clinical research, outcomes measurement, and quality assurance. A major impediment to this vision is that most patient data is "locked" in clinical narratives. That is, much information about patient care is captured as free text, with little structure or normalization of language. While this type of information is easy for the providing clinician to generate and for his or her colleagues to read and interpret in the context of clinical care, its use for aggregation by computer has been difficult. With the growing use of EMRs and computer-based reporting of clinician dictations, the automated interpretation of clinical narratives would be a major contribution.

A number of investigators have assessed the use of natural language processing (NLP) for identifying and codifying information in clinical narratives. The Linguistic String Project of Sager et al. was one of the earliest efforts, showing that for limited domains using "cleansed" (i.e., spelling errors correction, heading and other non-clinical information removed) data, 80-90% of concepts could be recognized [1]. Some have also integrated NLP functions into EMRs for recognition of clinical events [2,3]. Others have attempted to predict diagnosis codes based on words that occur in discharge summaries [4] or exacerbations of asthma based on progress notes [5].

The goal of unrestricted NLP, however, has not been met. All of the above systems work in limited domains and, while very valuable to those domains, have not been easily scaled to other areas. We have noted a number of problems impeding large-scale NLP, such as the limitations of vocabularies that would serve as the "targets" for mapping of concepts [6] as well as practical problems, such as spelling errors or otherwise unrecognizable words [7].

Much work in NLP has focused on *deterministic* methods, whereby investigators attempt to map the concepts present in text into a normalized representation (14). The major problem with deterministic NLP is language production is not deterministic. All but the simplest language suffers from ambiguity, and when combined with the operational problems of on-line text (e.g., the spelling errors and extraneous information), it is all but impossible in large domains. Some have attempted to recast the problem in a probabilistic manner [8]. Others have abandoned most of the syntax and instead developed semantic grammars that attempt direct semantic identification with a minimum of syntactical processing based on predictable aspects of clinical language [9].

For this project, we tried a different approach; we assessed the use of *machine learning* for the assignment of procedure codes in a trauma registry. A unifying feature of different machine learning approaches is that classification and prediction rules are "learned" from existing data without the need for complex rules and knowledge bases typical of expert systems. Using emergency room dictations as input data, we trained models to predict which procedures were performed on trauma patients.

We view this initial study as our first investigation into the larger picture of whether machine learning techniques can enhance the ability to codify clinical narratives. If successful, this approach could allow the identification of medical information without identification of specific strings in the text, as is attempted in deterministic NLP.

Materials and Methods

The goal of this study was to identify procedure codes from a trauma registry. OHSU maintains a trauma registry that identifies all trauma patients in the state of Oregon

[10]. A total of 119 fields are abstracted from the trauma patient's medical record, including trauma type, vital signs, various laboratory procedures performed, diagnostic codes, and outcomes measurements (e.g., admission to hospital ward or intensive care unit, death). The coding of each record is, like most chart review processes, a very labor-intensive task. Note that the coders have access to the entire medical record, and not just the dictation.

In this study, we looked at using only words and phrases from the dictated report by the initial ER physician to predict procedures that were performed during the first 24 hours after arrival to the ER. The database codes a total of 64 procedures, for example, OXYGEN, PFP (Plain Films Pelvis), CTA (CT Abdomen), and so on. For these experiments, we selected patients entered into the registry between March 17, 1994 and October 29, 1995. We then obtained the text of each patient's ER physician's dictation from the OHSU EMR system and matched it with the record from the registry.

A priori, any combination of procedures may apply to a particular case. The task is thus detection (where several procedure codes may occur), rather than classification (where only one of a number of codes occurs). A total of 600 dictations were used for this study. The data were divided into separate parts for model fitting and evaluation. The remainder of the data is being held out for future work.

The preliminary representation used was simply a vector of the frequency of occurrence of each word in the dictation. After removing stop words, and discarding those words occurring only once in the data, we were left with a lexicon of 3186 words from which to build detectors. Hence each dictation is described as a word-frequency vector of dimension 3186.

Building models with a limited amount of data in such a large feature space is impractical, so a dimension-reduction technique is required to obtain an input of reasonable size. We used principal component analysis (PCA) [11] to perform the reduction.

In PCA, the original features (word-frequency vectors for each dictation) of dimension *N* are mapped to a lower-dimensional vector space by projecting onto the *M (M<N*) eigenvectors of the data correlation matrix corresponding to the highest eigenvalues. This choice minimizes the expected squared error between the original vector and its dimension-reduced representation, and retains the maximal variance directions in the original space. The eigenvalues and corresponding eigenvectors were extracted by singular value decomposition (SVD) [12] performed on the matrix whose rows consist of the word frequency vectors. The SVD used only the dictations reserved for model fitting, so that the model generation is carried out without use of the evaluation data. Hence, the starting point is a feature space of dimension *N=360*, reduced to some *M<N* by PCA.

Our preliminary detectors were built by logistic regression [13]. A separate detector was built for each of the chosen procedure codes, using the PCA-transformed word frequency vectors as input.

Each detector maps the input $x \in R^M$ to output $y \in [0,1]$ through a logistic function

$$y(x) = 1/\left(1+\exp-\left(w^T x + w_0\right)\right), \qquad (1)$$

where $w \in R^M$ and $w_0 \in R$ are the model parameters.

The output is regarded as the probability that the code occurred, i.e., that target *t=1*, for that particular input. That is,

$$y(x) = P\left(t=1 \mid x\right).$$

Equivalently, the detector output is the conditional mean of a binomial distribution on the target values. This distribution can be written

$$P\left(t \mid x\right) = y(x)^{t(x)} \left(1-y(x)\right)^{(1-t(x))}.$$

Maximum likelihood estimation of the model parameters w and w_0 for this distribution (applied to all the training data) is equivalent to minimization of the cross-entropy cost function [14, for example]

$$E(w,w_0) = -\sum_{\{x,t\}} t(x)\ln y(x) + \left(1-t(x)\right) \ln\left(1-y(x)\right)$$

,

where the sum is over all input-target pairs in the training data. The cost *E* can be minimized by any of a number of standard function minimization algorithms. We used the Broyden Fletcher Goldfarb Shanno quasi-Newton algorithm [15] as implemented in the MATLAB numerical package.

Even with the preliminary dimension reduction, we have a relatively sparse amount of data with which to build and evaluate models.[1] Thus, as with most statistical modeling procedures, we require some means of balancing model complexity to the available training data. Regularization techniques such as ridge regression [13, for example], early stopping in recursive optimization procedures [14], and pruning techniques based on cost function curvature (Hessian) [16,17, for example] all provide complexity control. For the purpose of this study, we pruned the model size by successively eliminating input variables (principal components of the word frequency data) starting with those of least variance. This is a basic form of the pruning technique developed in [17].

Results

For these experiments, we selected the subset of 23 (of the total 64 procedure codes) that occur in between 5 and 95%

[1] More data is available, but our preliminary experiments were kept small until confidence in the representation and approach merits increased data preparation.

of the dictations. We trained separate logit model detectors for these procedure codes using between 2 and 60 principal components as inputs. We reserved 1/5 of the data, chosen at random, for testing. The remaining “training” data was used for model size selection and parameter fitting. Performance for a model of a given size was estimated by fitting a model of that size to a randomly selection of ¾ of the training data, and then measuring performance on the remaining ¼. This was done 5 times. The model size with the best average performance was selected, and a model of that size fit to the entire training set and tested on the reserved 1/5. This entire procedure was done for 8 random partitions of the data into training and testing sets; the result reported is the average over the 8 partitions of the performance on the testing set for that partition.

The point of this study was to address whether or not anything of use can be learned from even a naive representation of the dictations. The results indicate that the word frequencies *do* carry information enabling one to do better than chance for many of the procedure codes.

The results are shown in Table 1. To assess how well one can do by guessing, consider the following. If the frequency of occurrence of a procedure is thought to be greater (less) than 0.5, then, in the absence of information from the dictation, one should always guess that the procedure did (did not) occur. This produces the best guess. The table gives the frequency of occurrence (in the training data) in the column labeled "Prior”, and the fraction of correct guesses on the test samples in the column labeled "Guess". (Note that the prior frequency, and even correct guess, can differ between the training and test sets and between partitions.) The fraction of correct detections for the optimal logit models is in the fourth column. The last column gives the fractional improvement in detection rate between the best guess, and the logit model calculated as

(*logit* - *guess*) / *logit*

All values are averages over the 8 partitions.

For the procedures with very low or very high priors, the logit models are not able to improve the already good performance available by guessing based on the prior frequencies. One expects the logit models, or any other model, to show the most improvement for priors near 0.5, where guessing is most difficult. This is indeed the trend.

Discussion and Future work

We find it promising that even our simple representation can extract some information from the dictations, leading to prediction performance beyond the chance level.

Table 1 – Detector performance

Code	**Prior**	**Guess**	**Logit**	**Increase**
PFANK	0.054	0.943	0.942	-0.001
VECURON	0.062	0.933	0.925	0.-009
PFLS	0.070	0.920	0.921	0.001
PFTIB/FIB	0.072	0.931	0.931	0.000
PFTS	0.069	0.935	0.935	0.000
PFK	0.084	0.929	0.926	-0.003
SUCC	0.088	0.902	0.892	-0.012
PFSHOULD	0.069	0.919	0.918	-0.001
BVM	0.091	0.913	0.927	0.016
ET	0.111	0.885	0.892	0.008
EKG	0.120	0.896	0.896	0.000
CTP	0.161	0.820	0.818	-0.002
CTA	0.202	0.774	0.794	0.026
PFP	0.405	0.577	0.767	0.330
OXYGEN	0.507	0.483	0.623	0.299
DT	0.525	0.543	0.590	0.092
CTH	0.552	0.576	0.852	0.484
CARDIAC	0.719	0.717	0.717	0.000
PFCS	0.746	0.768	0.902	0.175
CRYST	0.895	0.894	0.894	0.000
PFC	0.906	0.926	0.924	-0.003

Evaluating our algorithm’s performance is difficult since the procedure codes have been assigned to cases by human coders with access to more information than is given to our algorithm. If future results warrant, we plan to measure the performance of professional coders given only the dictation, as a more accurate benchmark.

Even without an accurate benchmark, it is clear that our naive dimension-reduced word frequency vector approach has some potential problems. PCA is carried out without regard for the detection task; it is only concerned with capturing variance in the input data.[2] There is no guarantee that the high variance directions in the input data are those most suitable for discrimination. For example, suppose a rare procedure is "marked" by a word that occurs if and only if that procedure has been performed, as may be the case for procedures consisting of the administration of a particular drug. Then the marker word is also rare, and it will be nearly orthogonal to the reduced-dimention PCA representation. In this case, the magnitude of its PCA projection can be very small, and the projection of a document containing it may be dominated by other irrelevant but more commonly occurring words. Hence, even though perfect prediction is possible in this case, the necessary input has gotten ‘lost in the noise’, and learning may be difficult due to spurious correlations in the training set. We plan to look at more directly word-based input features to improve performance on rare procedures (see

[2] Dimension reduction/feature selection techniques that concentrate on discrimination, rather than input variance, can aid detection. One simple extension of PCA, proposed by Fukunaga and Koontz [18] offers improved discrimination and may be useful for this task.

below).

In an effort to interpret the models directly in terms of word-based inputs, we transformed the discriminatory direction *w* in equation (1), from the PCA representation back to the word space. One finds that it has large components along a few words, and smaller components on the remainder. This would suggest that one can build good predictors using a subspace of the original word space, and thereby obtain discriminatory models that have obvious linguistic interpretation. Our efforts to do this have failed -- models built in word subspaces perform poorly. We suspect that the problem lies with the original PCA representation, which uses coordinates (the eigenvectors) that are very diffuse mixtures of the words. This again suggests that an input representation closer to the original word base would be a better choice.

The second problem with our naive representation is that the use of single words as features loses all contextual information. In particular, it breaks up semantically meaningful units such as noun phrases into potentially less predictive pieces, destroys the relation of negation to the thing negated, and loses disambiguating syntactic information.

We experimented with a simple phrase generation algorithm, taking a phrase to be a longest sequence of non-stop words, with stop words chosen from a standard list, but this gave no essential improvement. There are two related possible reasons for this: first, the simplicity of the phrase determination algorithm causes many essentially similar phrases to be viewed as different, leading to diminished correlation between the phrases and the procedures. Second, the increased number of phrases, compared to the number of single words, exacerbates the sparse data problem.

We plan to explore two approaches here. First, we'll explore the use of NLP techniques to construct noun phrases and part-of-speech tags, and to "match" negation to its object. Second, we plan to modify Cohen & Singer's *sleeping experts* algorithm to our task [19]. This algorithm constructs context-bearing 'sparse phrases' on a statistical rather than NLP basis.

A more informative representation of the dictations, incorporating the NLP techniques discussed above, should improve performance considerably. When such a representation is established, it is likely that we will be able to leverage more sophisticated detection technology to further improve performance. We have applied neural networks to the present word frequency representation, both for individual code detection, and for detection of clusters of correlated codes. Those experiments showed no gain relative to the simple logistic model presented here; though we expect that with more data, and with more sophisticated representations, we will see improvement beyond the logistic regression model.

In general, virtually any statistics-based attempt at free text classification has a sparse data problem, due to the large size of any usable vocabulary relative to the number of available labelled examples. The trauma registry now contains several years worth of entries, and future experiments will use larger amounts of data. Dimension reduction of the input space will still be needed to ensure that the different input features are sampled sufficiently often that the relevant statistics can be reliably estimated. Unreliable estimation, when uncorrected, leads to increased model variance, i.e. to classifiers that generalize differently from the same training data. We plan to explore reducing input dimension using semantic information, in three ways. First, by using a dictionary to filter out words unlikely to be relevant. Second, by using a thesaurus (e.g., WordNet [20] to canonicalize the input space by reducing all synonyms, or more generally all words of a particular category, to a single feature. Third, by mapping phrases into a controlled vocabulary [21]. We also plan to deal more directly with model variance effects by use of bootstrapping and other data resampling techniques to improve the statistical estimates [22], and by the use of committees of models to directly reduce model variance.

We also intend to broaden our approach to use other, structured, data in conjunction with narrative reports. We plan to add vital signs and laboratory data into our models to determine whether they can, alone or in concert with narrative data, improve our predictive performance.

Finally, our long-term plan is to assess the optimal use of techniques in the real-world setting. It is unlikely that any automated approach to coding will ever achieve 100% accuracy. Therefore some human input will be required to choose the proper codes. It may be that a practical role for these techniques will be to assist human coders in identifying terms faster and/or with more accuracy. We will enlist colleagues in the human-computer interface field to develop optimal interfaces that use our techniques. One possible use of text categorization in general is the construction of high-level overviews of a collection of documents, for example, a medical record [23]. If successful, we will move closer to realizing one of the promises of the EMR, which is to aggregate data from individual patients for clinical research, outcomes measurement, and quality assurance.

Acknowledgments

This work was supported by Cooperative Agreement LM05879 from the U.S. National Library of Medicine and Contract DE-FG03-94ER61918 from the U.S. Department of Energy.

References

[1] Sager N, Lyman M, Bucknall C, Nhan N, Tick L. Natural language processing and the representation of clinical data. *JAMIA* 1994:1:142-160.

[2] Zingmond D, Lenert L. Monitoring free-text data using medical language processing. *Computers and Biomedical Research* 1993:26:467-481.

[3] Jain N, Knirsch C, Friedman C, Hripcsak G. Identification of suspected tuberculosis patients based on natural language processing of chest radiograph reports. In: Cimino J, ed. *Proceedings of the 19th Annual AMIA Fall Symposium*. Washington, DC: Hanley-Belfus, 1996:542-546.

[4] Larkey L, Croft W. Combining classifiers in text categorization. In: Frei H, Harman D, Schauble P, Wilkinson R, eds. *Proceedings of the 19th Annual International ACM Special Interest Group in Information Retrieval*. Zurich: ACM Press, 1996:289-297.

[5] Aronow D, Soderland S, Ponte J, Feng F, Croft W, Lehnert W. Automated classification of encounter notes in a computer based medical record. In: Greenes R, Peterson H, Protti D, eds. *MEDINFO 95*. Vancouver, BC: Healthcare Computing & Communications Canada, 1995:8-12.

[6] Hersh W, Campbell E, Evans D, Brownlow N. Empirical, automated vocabulary discovery using large text corpora and advanced natural language processing tools. In: Cimino J, ed. *Proceedings of the 19th Annual AMIA Fall Symposium*. Washington, DC: Hanley-Belfus, 1996:159-163.

[7] Hersh W, Campbell E, Malveau S. Assessing the feasibility of large-scale natural language processing in a corpus of ordinary medical records: a lexical analysis. In: Masys D, ed. *Proceedings of the 21st Annual AMIA Fall Symposium*. Nashville, TN: Hanley-Belfus, 1997:in press.

[8] Charniak E. *Statistical Language Learning*, Cambridge, MA: MIT Press, 1993

[9] Friedman C, Alderson P, Austin J, Cimino J, Johson S. A general natural-language txt processor for clinical radiology. *JAMIA* 1994:1:161-174

[10] Mullins R, Veum-Stone J, Hedges J, Zimmer-Gembeck M, Mann C, Helfand M. An analysis of hospital discharge index as a trauma data base. *Journal of Trauma* 1995:39:941-948.

[11]I.T. Jolliffe, *Principal Component Analysis*, Springer-Verlag, 1986.

[12] Gene H. Golub and Charles F. van Loan, *Matrix Computations*, The Johns Hopkins University Press, 1983.

[13] G.A.F. Seber and C.J. Wild, *Nonlinear Regression*, John Wiley & Sons, 1989.

[14] C. M. Bishop, *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford, 1995.

[15] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery, *Numerical Recipes in C*, 2nd edition, Cambridge University Press, 1992.

[16] Babak Hassibi, David Stork, and Gregory J. Wolff, *Optimal Brain Surgeon and Network Pruning*, RICOH California Research Center, CRC-TR-9235, 1992.

[17] A. Levin, T. Leen, and J. Moody, Fast Pruning Using Principal Components, in *Advances in Neural information Processing Systems 6*, Cowan, Tesauro, and Alspector (ed.), Morgan Kaufmann Publishers, 1994.

[18]Keinosuke Fukunaga and Warren Koontz. Application of the Karhunen-Loeve expansion to feature selection and ordering. *IEEE Transactions on Computers*, C-19:311-318, 1970.

[19] W. Cohen and Y. Singer, Context-Sensitive learning Methods for Text Categorization. In *Proceedings of the 19th Annual ACM SIGIR Conference*, Zurich, 1996

[20]Voorhees E. Using Wordnet to disambiguate word senses for text retrieval. In: Korfhage R, ed. *Proceedings of the 16th Annual International ACM Special Interest Group in Information Retrieval*. Pittsburgh: ACM Press, 1993:171-180.

[21]Hersh W, Leone T. The SAPHIRE server: a new algorithm and implementation. In: Gardner R, ed. *Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care*. New Orleans, LA: Hanley-Belfus, 1995:858-862.

[22]B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall, 1993

[23] L. Delcambre, P. Gorman, D. Maier, R. Reddy, S. Rehfuss, Precis-Based Navigation for Familiarization. Submitted to: *MEDINFO '98*.

Address for correspondence

William Hersh, M.D.; Division of Medical Informatics and Outcomes Research, School of Medicine, Oregon Health Sciences University, 3181 SW Sam Jackson Park Rd., Portland, OR, USA, 97201; Email: hersh@ohsu.edu