Family History Detection from Clinical Text

Srinivasan Iyer

Introduction

Electronic Health Record (EHR) systems are becoming highly prevalent today and have the potential[1] to add to the success of spontaneous reporting systems for post marketing surveillance of drugs. EHR systems broadly contain two forms of information viz. coded (structured) information and unstructured text. Researchers have shown[2] that the textual data have much more utility for cohort building than coded information, which, being mainly used for billing and insurance purposes, is biased and does not truly reflect the state of the patient.

The text in EHRs have already been successfully used in several areas of drug safety such as learning adverse drug reactions[3], learning drug-drug interactions and determining off-label use. The first step, in all these applications is the tagging of the unstructured text with concepts pertaining to drugs, diseases, devices and procedures. The more accurate this tagging process, the better the results of subsequent analysis. Most efforts until now[4] make use of natural language processing (NLP) methods to test if a particular sentence is about the family of the patient or about the patient himself. Several such methods[5] were developed as part of the i2b2 challenge in clinical NLP, 2010. However, such methods are usually slow and are not suitable for processing billions of text documents.

The use of machine learning to process clinical text has been somewhat limited[6] owing to the lack of a good quantity of labeled data and this applies to the problem of family history detection as well. However, in our dataset, we observe that several clinical texts explicitly define a *family history* section and a *history of present illness* section and these section headers could be used as labels for supervised machine learning algorithms. In this work, we to take a semi-supervised approach to learning to identify sentences related to a patient's family by using section headers in clinical text as labels for training. **Once a model is learned, it can then be applied to large datasets of clinical text without explicitly defined sections.**

Dataset

We have access to a corpus of over 10 million unstructured clinical notes from the Stanford Hospital, corresponding to approximately 1 million patients. These notes have been de-identified for research purposes. Some of the notes have explicitly defined sections such as "Family History", "History of present illness" etc. and we use these notes as labeled training data. Additionally, we use terms from 19 biomedical ontologies from the Unified Medical Language System (UMLS) as features in our classifiers.

Methods

Our ultimate goal is that given a document of unstructured clinical text, we wish to identify terms that supply information about the family of the patient. In this project, we choose to tokenize the document into sentences (separated by a full-stop or newlines) and treat each sentence independent of the others. This is a simplifying assumption, since sentences could refer to ideas introduced in preceding sentences, thus affecting their meaning. Also, it could be the case that some terms in a sentence refer to the patient's family and some terms refer to the patient. However, we make an assumption that all of the terms in a sentence refer to the family or refer to the patient. Thus, for this project, given a sentence, we wish to classify the entire sentence as referring to the patient's family (FH), or not (PH).

Preparation of labeled dataset

To prepare our set of positive training examples (FH), i.e. sentences that actually refer to the patient's family, we first look for a family history section within our documents. We define a family history section as a paragraph beginning with "Family History:" and we include all the text till the end of the paragraph as belonging to the section. End of paragraphs are located by the presence of a double newline or a double

carriage-return character. Following this, the section header is stripped off and the section text is tokenized into sentences. Each sentence forms a positive training example. We follow a similar approach to locate sections labeled as "History of present illness:" and these sentences form the set of negative training examples (PH) (see Table 1).

From our dataset, we obtained 1,779,264 examples for the family history class and 10,392,373 examples for the personal history class. Of this, we use 20,000 randomly chosen examples from each class for training and a different set of 20,000 randomly chosen examples from each class for testing. We do this mainly for computational tractability.

Table 1 A sample of the training data used. Text contains missing words owing to the de-identification process for Protected Health Information (PHI).

Training Example	Class
grandmother diabetes insipidis	Family History
his aunt and uncle allergy	Family History
uveitis right eye greater than left arthritis	Personal History
observed upper extremity rhinitis	Personal History

Rule-based Methods

By visual inspection, it appears that many of the sentences in the family history class contain some family member term (see Table 2). We test a simple rule-based method, which classifies a sentence as FH if it contains any term from Table 2, and classify it as PH otherwise. This method gives an accuracy of 76.47% and a specificity of 93.8% at a sensitivity of 59.2% (F-measure=0.715).

Table 2 Subset of family names used for rule-based methods. Full table contains 195 names.

Term
father
dad
mother
mum
mummy
mom
parent
parents
child
children
son
daughter
brother
sister
grandfather
granddad
grandpa
grandmother
grandma
granny

Bernoulli and Multinomial Naïve Bayes

For a baseline estimate of performance using Machine Learning methods, we use multivariate Bernoulli (BNB) and Multinomial Naïve Bayes (MNB), using a simple bag of words model. We therefore tokenize the sentences into words and treat each unique word as a feature. We construct a 40000 x 16846 matrix in which every row is a sentence and every column is a feature. MNB seems to perform slightly better than BNB in terms of accuracy (see Table 3). Note that the training error is very close to the test error, convincing us about the absence of over-fitting.

Lib-linear SVM

Using the same matrix of word frequencies as MNB, we use the Lib-linear SVM[7] classifier and it delivers a greater accuracy than MNB (see Table 3). Training error seems to be lesser than test error, and this classifier may be subject to over-fitting.

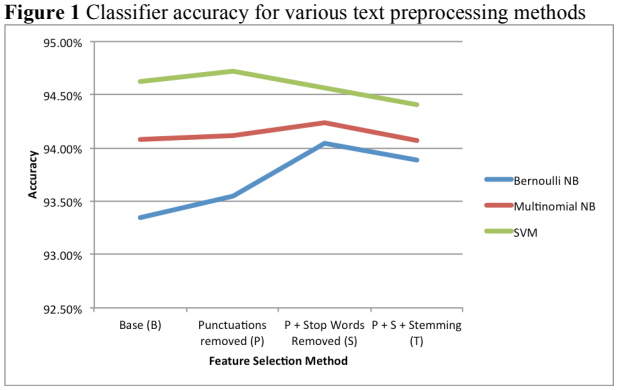
Data preprocessing

We use three methods: a) Elimination of punctuations, b) Removal of stop words, c) Stemming[8], incrementally in order to reduce the number of features and choose more useful features. In general, accuracy improved on removal of punctuations and stop words (resulting in 11009 features). However, the stemming procedure reduced accuracy. These results are summarized in Figure 1.

Table 3 Baseline values for three classifiers.

	BNB	MNB	SVM
Training Accuracy	93.92%	94.67%	97.66%
---Test Results---			
# True Positives	18825	18630	18776
# False Positives	1653	1168	1279
# False Negatives	985	1180	1096
# True Negatives	18191	18676	18748
# Correctly classified	37016	37306	37524
#Total (39654)			
Sensitivity	95.03%	94.04%	94.47%
Specificity	91.67%	94.11%	93.61%
Accuracy	93.34%	94.08%	94.63%
F-measure	93.452%	94.07%	94.03%

Figure 1 Classifier accuracy for various text preprocessing methods



We now look at misclassified examples to get insights into additional feature engineering that can be used to increase the accuracy. Table 4 lists certain examples that were misclassified by MNB.

Table 4 Some examples that were misclassified by Multinomial Naïve Bayes. The actual examples cannot be disclosed (PHI). However, these examples are modifications of the originals, keeping the essential anomaly intact.

Test Example	Misclassified as
Patient with hypoglycemic conditions in the company of this mother	Family History
Family reports that patient is having hallucinations	
Her daughter was not at home and she developed X	
She followed the advice of her mother and took Y	
His son was diagnosed with arthritis.	Personal History
Patient is living with sister who has Z	
Father strange behavior at night	
Mother has a history of Z, with X and Y	

Feature Selection

To reduce the amount of over-fitting, we attempt to keep only the most useful features with respect to information gain. Table 5 shows the top 20 most useful features. It agrees with our intuition, for example we would expect that sentences containing *mother* are mostly FH and sentences containing *diagnosed* are mostly PH. We find that retaining only the top 500 or 1000 features reduces accuracy. However, removing the last 1000-2000 features increases our accuracy. Figure 2 shows the variation of accuracy with the number of features.

Table 5 Top features by information gain

term
mother
diagnosed
breast
cancer
age
negative
malignancy
significant
history
family
father
died
emphysema
otherwise
grandmother
asthma
chronic
headaches
insect
sensitivities

N-gram Features using Biomedical Ontologies

Instead of using all bigrams and trigrams as features, we take an approach that uses existing knowledge of biomedical text. The UMLS group of biomedical ontologies contain sets of phrases typically found in biomedical text. We recognize these phrases in the sentences and use them as features. This results in 19228 features (see Table 6). We also add the length of the sentence as the number of words as another feature. Figure 3 shows the variation in accuracy of MNB and SVM with feature selection, using these more elaborate features. SVM performs the best with 95.38% accuracy.

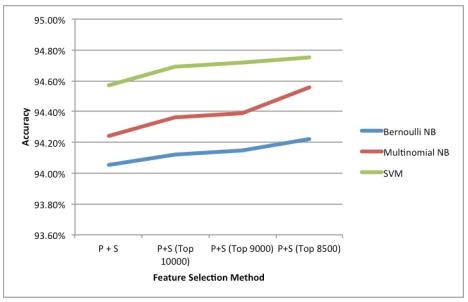


Figure 2 Accuracy vs #Features for various classifiers

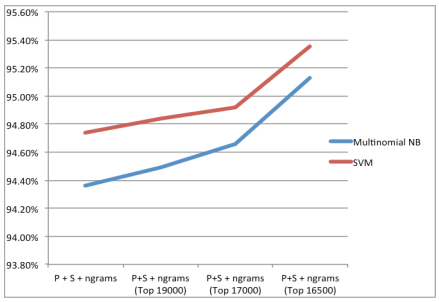


Figure 3 Accuracy vs #Features for n-gram features.

Discussion/Conclusion

The best performance was achieved by SVM with features from biomedical ontologies, and feature selection by information gain. This achieved an accuracy of 95.38%. Since the training set was build artificially, there were many cases where personal history items were actually labeled as family history and vice versa. This puts an upper bound on the maximum accuracy that can be obtained for algorithms that predict family history perfectly. It would be interesting to build a manually curated training set and test whether the same methods perform well. Also, several gaps were introduced in the sentences owing to the PHI de-identification process. These gaps were ignored in this project, but nevertheless, they could possess significant predictive power.

Overall, simple machine learning methods seem to do much better than rule-based methods. More improvement can perhaps be obtained by using features from preceding and following sentences and by using language-modeling methods like Hidden Markov Models and Conditional Random fields.

Table 6 Features derived from biomedical ontologies (Total: 19228)

term
is a
history of
physical examination
evidence of
blood pressure
status post
review of
does not
review of systems
medical history
no evidence of
past medical history
follow up
last name
family history
history of present illness
due to
social history
consistent with
vital signs
secondary to
prior to
final report
children's hospital
normal limits
ct scan
per day
was a
reason for
by mouth

Acknowledgements

We thank the CS229 teaching staff for their help, resourcefulness and guidance for this project. We thank Prof. Nigam Shah's group at the Stanford Dept. of Bioinformatics for access to the STRIDE dataset.

References

1. Schuemie MJ, Coloma PM, Straatman H, et al. Using Electronic Health Care Records for Drug Safety Signal Detection: A Comparative Evaluation of Statistical Methods. *Medical care.* 2012.
2. Classen DC, Resar R, Griffin F, et al. 'Global trigger tool' shows that adverse events in hospitals may be ten times greater than previously measured. *Health Aff (Millwood).* 2011;30(4):581-589.
3. Lependu P, Iyer SV, Fairon C, Shah NH. Annotation Analysis for Testing Drug Safety Signals using Unstructured Clinical Notes. *Journal of biomedical semantics.* 2012;3 Suppl 1:S5.
4. Harkema H, Dowling JN, Thornblade T, Chapman WW. ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. *J Biomed Inform.* 2009;42(5):839-851.
5. Roberts K, Harabagiu SM. A flexible framework for deriving assertions from electronic medical records. *Journal of the American Medical Informatics Association : JAMIA.* 2011;18(5):568-573.
6. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet.* 2012;13(6):395-405.

7. Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ. LIBLINEAR: A library for large linear classification. *The Journal of Machine Learning Research.* 2008;9:1871-1874.
8. Porter MF. An algorithm for suffix stripping. *Program: electronic library and information systems.* 2006;40(3):211-218.