Developing Supervised Machine Learning Methods for the Extraction of Maternal History Data from Neonatal Clinical Notes

Samuel Maynard
Dr. Demner-Fushman, National Library of Medicine

Abstract

Maternal history information such as blood type is very important for newborn care, but pediatricians do not have access to this information due to privacy rules. The purpose of this research is to investigate machine learning methods for extraction of maternal prenatal history from newborn notes. This information is included only within prenatal history notes that nurses take by hand. Currently, the whole document is blocked to the pediatrician, however if the specific values of information were processed out of the document those could be sent to the child care provider. Unfortunately, the information is located within free form text and is difficult to automatically process. Our research attempted to develop a system that could pull these key values out of prenatal notes. After manually annotating 500 documents we used a classification toolkit MALLET to train a classifier to sort sentences and a Conditional Random Fields (CRF) algorithm to recognize the maternal history. We used the classifier to determine which sentences were the most likely to contain the maternal information, and the CRF to pick out the specific clinical variables. In the end we achieved a lower level of precision and recall than we wanted. However, the classifier was left unoptimized as the main focus was the CRF during the trial runs. In the future results may be improved through the use of classifiers, like a Support Vector Machine.

Background

Maternal health information is important data for pediatricians; knowing the mother's test results allows physicians to make better choices for the child. Unfortunately, physicians are unable to access that information due to privacy constraints. If automatic extraction of the data were to be developed then this information could be given to the physician. Difficulties arise because the text is not in any consistent structured form that makes it easy to parse automatically. The style of writing varies from notetaker to notetaker. This makes automated retrieval of mother's health information difficult.

A previous approach by Abhyankar and Demner-Fushman used regex and pattern matching tools. Patterns were manually developed for each individual variable by going through a selection of 289 neonatal notes and looking for common words or symbols that would appear before or after the information in question. These were synthesised it into a regex string. This manual approach achieved a recall of .91 to .99 and a precision of .95 to 1.0, high scores for a machine extraction method over a non-determined system. However, this approach was only tested on a limited set of documents, and one of the weaknesses of pattern matching is that it does not scale upwards. It is predicted that as this method is applied to larger and larger document sets the previously seen high scores will fall considerably.

This investigation used a more scalable method of extracting maternal history information. A system was created that relied on using supervised machine learning techniques. These techniques are more scalable because they can be trained over a larger set of documents with no additional human intervention. In our system, both a Conditional Random Fields (CRF) model and a Maximum Entropy classifier trained over sentences were used. This method is attractive, because although it may underperform the pattern matching approach over a limited set of test documents, the machine learning methods are envisioned to do better as the number of documents being processed grows.

Results

Table 1: Recall and Precision of the established method (Rules Based) versus a combined sentence classifier and CRF (Machine Learning).

Maternal Variable	Recall (Rules Based)	Precision (Rules Based)	Recall (Machine Learning)	Precision (Machine Learning)
Age	.854	.992	.220	.790
Gravida	.833	.997	.457	.854
Blood type	.851	.981	.383	.857
Rh_antigen	.857	.981	.417	.962
Ab	.814	.955	.528	.986
GBS	.785	.997	.450	.935
HepB	.826	1	.543	.991
RPR	.838	1	.506	1
Rubella	.847	1	.471	1

The Annotation

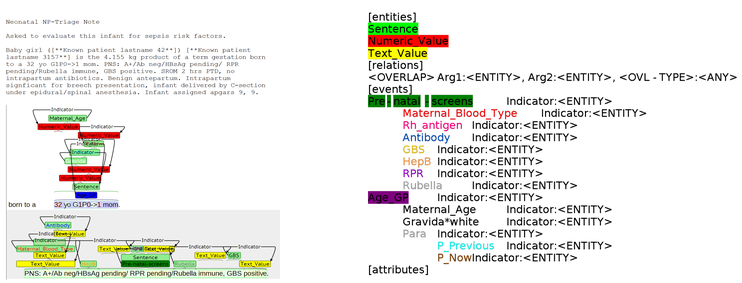

Figure 1. On the right are the tags and associated colors used to manually annotate prenatal notes. On the left is an example not being tagged within the BRAT annotator.

The Process

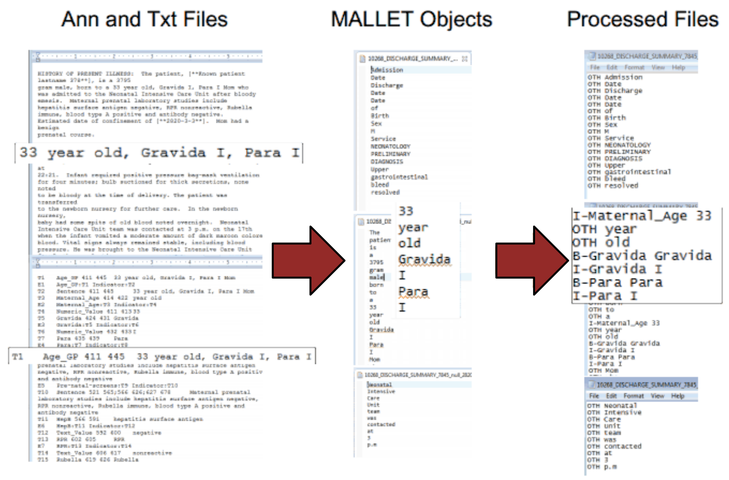

Figure 2. Flow of data through the program. First the data start from Brat files and is converted to MALLET objects. Then these objects are classified by mallet and the words are assigned categories.

Methods

Annotation

The training set was 289 newborn records, and the testing set was 464 unique newborn records. All annotations were done with the open source tool Brat (v1.3 Crunchy Frog).

A schema was developed and used to manually annotate all of the training set. Two primary types of sentences were seen, Age_GP and Pre-natal-screens. Every occurrence of the maternal health information was tagged within these sentences. The tags were: Maternal_Blood_Type, Rh_Antigen, Antibody, GBS, HepB, RPR, Rubella, Maternal_Age, Gravida, Para, P_Previous, and P_Now (see Figure 1). Finally, tagged words were linked to their surrounding words.

Machine Learning

Brat annotated files were converted into Machine Learning for LanguagE Toolkit (MALLET) readable formats. A Java method to parse the Brat annotation into a java object was developed. This object contained information about what each tag was and what it referenced.

The classifier was developed by reformatting data to suit MALLET's CLI (Command Line Interface) processing. The entire training set of sentences were processed. All of the Age_GP and Pre-natal-screen sentences were used, and for every document two randomly selected 'other' sentences were used to prevent classification bias. A total of 1053 positive examples and 867 negative examples were trained.

In addition, a CRF classifier for clinical variables was trained. Sentences were divided into words and labeled with their tag, a before/after tag, or an OTH tag. This was then processed on every positive sentence example (1053).

Testing

Finally we built the system that would extract the information from an untagged document. This was done by creating a Java program that would parse the document into sentences. The sentences were fed to the classifier which would determine if they had Age_GP or Pre-natal-screens data in them or not. If the sentence had data, then it was separated into words and fed into the CRF, where every occurrence of the tag information would be picked out and labeled. This is demonstrated in Figure 2.

Conclusions

The F1 scores for the machine learning algorithm were not as high as expected, however, there is still promise with the approach. Because the recall of the system is much lower than the precision, the sentence classifier, may be at fault. Because the precision was higher, in most of the identified sentences the prenatal terms were correctly identified. Identification of individual words is purely the job of the CRF, so this indicates that the CRF trained better than the sentence classifier.

Our results indicate that the overall performance of the system may be improved by forgoing the sentence classifier, at the cost of calculation time. Because a simple Maximum Entropy classifier was used, upgrading to a Support Vector Machine (SVM) or other more complex classifier may increase the recall of the system.

Future Research

In the future investigations should proceed without the sentence classifier. This work does not completely rule out the possibility of using machine learning techniques for the automatic processing of this information, but it does not strongly point in this direction either.