Kraus,S., Blake,C. & West,S.L (In Press) Information Extraction from Medical Notes, Accepted to MEDINFO, Brisbane, Australia.

Information Extraction from Medical Notes

Scott Kraus[a], Catherine Blake[a], Suzanne L. West[b]

[a] School of Information and Library Science, [b] School of Public Health, University North Carolina at Chapel Hill,USA

Abstract

Health providers tend to prefer free text when describing drug dosages prescribed to their patients. However, unstructured text is not amenable to large scale analysis such as data mining. In this paper, we explore four methods that automatically identify drug, dosage, and method of delivery information from transcribed physician notes. Our final method uses just one extraction heuristic, and achieves an average precision of 96.74% on a training set and 96.70% on the test set and average recall of 69.48% on the training set and 79.72% on the test set. These results suggest that a small number of heuristics can provide accurate extraction of drug, dosage and method of delivery information in medical notes.

Keywords: Information Science, Medical; Clinical Laboratory Information Systems; Information Extraction.

Introduction

Health professionals need the flexibility of unstructured text to express their diagnoses and treatment strategies. Thus, the presence of unstructured text is inevitable, but, for data mining, information systems are required to convert the unstructured text to a structured representation for analysis.

There are two key approaches used to extract information from unstructured medical text: supervised learning, and a knowledge-based approach. Successful supervised learning models that have been demonstrated on biology include maximum entropy models[1], hidden Markov models[2], naïve Bayes, and decision trees[3]. In contrast to supervised learning, a knowledge based approach does not require labeled examples. Instead, dictionary lookups and heuristics help in extracting key facts from text. Successful implementations of knowledge-based approaches can be found for medical literature[4][5], radiology[6], and histopathology[7].

Materials and Methods

The data used in these experiments comprised narrated notes for 12,222 patients in the University of North Carolina at Chapel Hill (UNC) hospital system. The notes were transcribed and anonymized prior to our analysis as part of a parent UNC DEcIDE project. The goal in the parent project is to identify treatment and outcome patterns for newly diagnosed patients with type II diabetes mellitus, some of whom have cardiovascular comorbid conditions. The parent project had access to both structured and unstructured data, but, our experiments consider only the patient note.

Method 1, the first experiment, used a subset of diabetes drugs to trigger the text extraction. This experiment used the GNU Awk (gawk) script language. We worked closely with experts from the DEcIDE team who were familiar with drugs commonly used to treat diabetes. The output from this experiment was used for discussion with the DEcIDE team, and was instrumental in the design of future methods.

Our goal in Method 2 was to provide DEcIDE team members information to help them identify patients who already had diabetes. The script identified sentences that contained (a) the word “diabetes,” (b) drugs commonly prescribed for diabetes, or (c) the word insulin. The input to this method was the result of a program that split notes into separate sentences.

The first two experiments revealed a surprising degree of regularity in the way physicians described the drug and dosages prescribed. Our goal in Method 3 differed in two ways: (1) identify *any* drug, and (2) focus on the drug *and* the amount of drug prescribed. Thus, Method 3 moves the trigger term from a drug to an amount and dosage combination.

We used enhanced grep (egrep), a UNIX text processor for Method 3. We added a UNIX C-Shell (csh) script to the framework, which enabled us to modify and document the regular expression more easily than in previous methods.

We constructed a training set comprised of 30 notes we selected at random from the complete set. Based on the training set, a single regular expression was selected for the final experiment. It ultimately identified words at least 4 letters in length, followed by a numeral or a number, and a unit of measurement. Parameters such as how the drug is to be taken and the frequency of the dosage will be extracted.

Results and Discussion

We applied the script for Method 1 on the complete set of 12,222 patients. The script output included the patient ID, the date of treatment, and 50 characters before and after the drug that triggered the extraction. Our goal with Method 1 was to verify the text processing pipeline, which worked as expected.

Method 2 was applied to records belonging to 1,933 unique patients identified by the DEcIDE researchers. A test set of 166 patients was chosen randomly. The third author, using the criterion of successfully detecting diabetes patients, evaluated the results of this set manually. The method was found to have a precision of 61.33% and recall of 77.13%. The relatively low precision and recall were because our script did not account for negation. The most common error was family history of diabetes as opposed to diagnosis.

Our goal with Method 3 was to identify prescribed drugs and corresponding dosages. We also extracted the drug's prescribed frequency and method of delivery, if available. Drugs without dosages tended to refer to the patient's reaction, allergies, or other factors unrelated to current prescriptions.

A machine learning approach inspired our evaluation of the regular expression used in Method 3. We created a training set by selecting 30 random patient notes from the complete set. We then wrote the regular expression based on notes in the training set. Once the expression was complete, we selected an additional 40 patient notes at random to form a test set. We checked to see that the training and test sets were similar and found an average of 4.83 drug dosages per patient. A test set was derived from a random set of 40 patient notes, with an average of 3.78 drugs per patient.

The error analysis of the training set in Method 3 revealed that the most important factor is detecting a number followed by an SI unit. The analysis also revealed several legitimate dosages in the notes that Method 3 did not detect. The primary error in recall was due to unexpected units of measurement in the text. The second most important factor was correctly extracting the drug name. Three key reasons caused the drug extraction to fail:

1. The drug name did not precede the dosage.
2. The drug contained multiple words.
3. Short drug names. Method 3 requires that a drug name be 3 or more characters long.

We developed a fourth and final regular expression based on the underlying causes of error in Method 3. We made the following changes to Method 4:

1. Added floating point (decimal) numbers,
2. Improved the set measurement units used to trigger the heuristic, by expanding the list of SI units and added apothecary units, and
3. Removed unnecessary delivery abbreviation.

Without any background knowledge of drug names, Method 4 performed surprisingly well on drug extraction for the training and test sets (see Table 1). An informal analysis showed that drugs paired with dosages were usually associated with current prescriptions rather than drugs used previously.

Table 1 – Method 4 Detailed Training and Test Set Results

	Training		***Test***	
	Precision	***Recall***	***Precision***	***Recall***
Drug	86.96%	82.76%	86.81%	82.78%
Amount	100.00%	91.39%	100.00%	96.08%
Method of delivery	100.00%	62.03%	100.00%	91.89%
Prescribed frequency	100.00%	41.73%	100.00%	48.15%
Average	**96.74%**	**69.48%**	**96.70%**	**79.72%**

Precision for drug amount was 100% for both training and test sets because the method uses an explicit pattern match on the unit of measurement and its quantifier. The prescribed frequency recall, while low (41.73%/48.15%), is on par with other methods, as is the method of delivery in the training set (62.03%). This is due to variances in how the notes report frequency. We are continuing to work on methods to improve the currently low retrieval rate of prescribed frequency.

Conclusion

Our results show that a single extraction rule achieves precision of 96.72% and recall of 74.60% on a set of randomly selected notes. Precision was lowest for drug extraction (86.81%), and consistently high (100%) for amount, method of delivery and frequency. Recall was 44.94% for the prescribed frequency, but remained high for the other data.

Our final method ran at a rate of 24-26 notes per second with an Intel Pentium 4 2.0 GHz processor. This suggests that a single well defined extraction rule would scale well to larger numbers of notes, and result in precision and recall that are competitive with systems that employ deep natural language processing methods. Further research that characterizes high precision rules is critical if we are to move towards a completely generalizable and scalable model of information extraction from medical records.

References

1. Raychaudhuri, S., et al., Associating genes with Gene Ontology codes using a maximum entropy analysis of biomedical literature. Genome Research, 2002. 12: p. 203-14.

2. Zhou, G., et al., Recognizing names in biomedical texts: a machine learning approach. Bioinformatics, 2003. 20(7): p. 1178-90.

3. Hatzivassiloglou, V., P. Duboue, and A. Rzhetsky, Disambiguating proteins, genes, and RNA in text: a machine learning approach. Bioinformatics, 2001. 17(Suppl 1): p. S97–S106.

4. Fukuda, K., et al., Toward Information Extraction: Identifying protein names from biological papers. Pacific Symposium on Biocomputing, 1998. 3: p. 707-18.

5. Rindflesch, T., C. and J.V. Rajan. Extracting Molecular Binding Relationships from Biomedical Text. in Proceedings of the 6th Applied Natural Language Processing Conference. 2000.

6. Friedman, C., et al., A general natural-language text processor for clinical radiology. JAMIA, 1994(1): p. 161-174.

7. Hahn, U., M. Romacker, and S. Schulz. Creating knowledge repositories from biomedical reports: the MEDSYNDIKATE text mining system. in Pacific Symposium on Biocomputing. 2002.

Acknowledgments

This project was funded in part under Contract No. 290-05-0040 from the AHRQ, US DHHS as part of the Developing Evidence to Inform Decisions about Effectiveness (DEcIDE) program. The authors of this paper are responsible for its content. Statements in the presentation should not be construed as endorsement by the AHRQ or the US DHHS. Thanks to R.Akers, B.Schwartz & Z.Liu for pre-processing.

Address for correspondence: **Scott Kraus, scott@unc.edu**