A NEW APPROACH TO EXTRACT MEANINGFUL CLINICAL INFORMATION FROM MEDICAL NOTES

by: Awanish Ranjan & Rabindra Bista
Department of Computer Science and Engineering
Kathmandu University
Dhulikhel, Nepal
awa.ran@gmail.com, rbista@ku.edu.np

December 06, 2017

Outline

- Introduction
- Literature Review
- Computational Framework
- Results & Discussion
- Conclusion & Future Work
- Application Areas

Introduction

- Medical Notes
  - *"Pt presents with hyperlipidemia and strong family hx of CAD. Keeps active with job, kids, and softball, but no routine cardio exercise."*
- Unstructured Data
  - Can't be used as direct input for further processing
- Structured Data
  - Regular pattern and used for further processing

Introduction (cont..)

- Natural Language Processing techniques to solve this problem
- Domain
  - English Language
  - Health care data
  - Medical notes text files
  - Extracted information
    - Diagnosis, Procedure, Drug, Vital and Habits

Objectives

- Propose a new approach to extract meaningful data from clinical notes
  - Extract meaningful clinical information from notes
  - Store the information for future use
  - Compare the effectiveness of the proposed system with some existing system

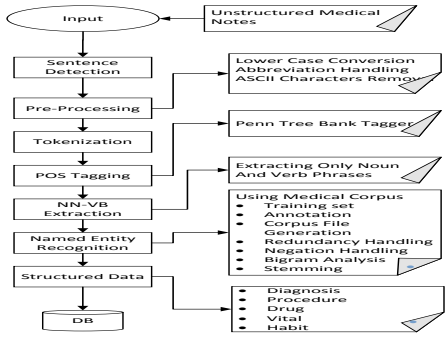
Literature Review

- MedEX - Xu, H. et.al [2009]
  - Structured data extraction
  - Input Clinical Text:
    - "acetaminophen
    - 325- 650 mg po/pr q4 -6h prn"
- Structured Output:
  - Drugname: acetaminophen
  - Strength: 325-650 mg, Route: pc/pr
  - Frequency: q4 -6h
  - Necessity: pm
- Limited to Drug Data only

Literature Review (cont..)

- Other relevant systems
  - cTAKES - Sovava, K.G. et. al [2010]
  - YTEX - Glara, V. et. al. [2011]
  - MetaMap - Aronson, et. al. [2001]
  - MEDLEE - Friedman et. al [1994]
- Problems with existing system
  - Less Accuracy
  - Too specific
  - Lot of third party dependencies

Computational Framework



Computational Framework (cont..)

- Input - Medical notes, text files

1. Sentence Detector- ["FBS & hgA1c both slightly improved, but still prediabetes (HgA1c = 5.8%).", "But did instruct on diet/exercise."]
2. Preprocessing
   - Abbreviation handling: dx -> diagnosis
   - Punctuation handling: don't" -> do not
   - Lower case conversion
   - ASCII character removal

Computational Framework (cont..)

3. Tokenizer: 'FBS', '&' , 'hgA1c', 'both'
4. Parts-Of-Speech (POS) Tagging:
   - ('FBS', 'NNS') ('&', 'CC') ('hgA1c', 'NNP') ('both', 'DT') ('slightly', 'RB') ('improved', 'VBN')
5. Noun-Verb (NN-VB) Extractor
   - Noun phrases - NN, NNS, NNP, NNPS
   - Verb phrases - VB, VBD, VBG, VBN, VBP, VBZ

Computational Framework (cont..)

6. Named Entity Recognition
   - Detection of elements – Diagnosis, Procedure, Drug, Vital, Habit
   - Medical corpus building, training and matching

Medical Corpus

6.1 Medical corpus

- Training Data Collection
  - 15000 medical notes as training data
  - Health care data

Medical Corpus (cont..)

- Manual Annotation

Pt states her labs were normal but just needs to loose ~20#'s. States she use to ba able to loose <START:vital> wt <END> fairly easy but struggles as she gets older. Pt is a RN, works night shift at St Francis. Tries to take aerobics and zumba classes 3x/week. Discussed kcal and carb intake, <START:habit> exercise <END> goals, sleep. Pt is to track intake and <START:habit> exercise <END> using \myfitnesspal\" and f/u x 1 month via phone for wt check."

Pt presents with PMHx of <START:diagnosis> diabetes <END> ~>20 yrs. <START:vital> HgbA1c <END> way above goal. Pt is on an <START:drug> insulin pump <END> and is follwed by his endo q 3 months. Pt states he struggles with elevated <START:vital> FBS<END>. Pt states that he usually always enters his carb intake and will use his bolus before meals. Pt has been through Diabetes Education many times and feels comfortable counting carbs. Pt is an Atheletic director and works long hours, eats big meal for dinner and will drink a few beers. Discussed using Carelink to download pump, and talking to doctor re wearing CGM. Will refer pt to Medtronics Rep who has worked with pt and endo in the past to adjust pump settings to get <START:vital> FBS <END> w/in goal range.

Pt presents with <START:diagnosis> diabetes<END>, currently on <START:drug> Janumet <END>. Wants to get her <START:vital> HgbA1c <END> less than 6.1. Has been using \myfitnesspal\" to track intake.

Medical Corpus (cont..)

- Corpus File Generation
  - Different files for different corpus
- Redundancy Handling

```
1  diabetes
2  diabetes
3  hydrocephalus
4  shunt malfunction
5  diabetes
6  diabetes
7  diabetic
8  diabetes
9  diabetes
10 diabetes
11 diabetes
12 diabetes
13 hyperlipidemia
```

Redundant information

Entity Detection

id	set_id	note_id	sent_id	detected_element	element_type	updated_date
121445	set2	Note_Number-221:	Sent_Number-4:	exercise	habit	2017-09-01 10:57:53
121433	set2	Note_Number-221:	Sent_Number-2:	chol	vital	2017-09-01 10:57:53
121377	set2	Note_Number-217:	Sent_Number-5:	lab	procedure	2017-09-01 10:57:53
121368	set2	Note_Number-217:	Sent_Number-2:	chol	vital	2017-09-01 10:57:53
121354	set2	Note_Number-215:	Sent_Number-3:	nrts	procedure	2017-09-01 10:57:53
121341	set2	Note_Number-213:	Sent_Number-4:	chewing	habit	2017-09-01 10:57:53
121315	set2	Note_Number-211:	Sent_Number-2:	chewing	habit	2017-09-01 10:57:53
121289	set2	Note_Number-210:	Sent_Number-4:	diab	diagnosis	2017-09-01 10:57:52
121283	set2	Note_Number-210:	Sent_Number-3:	diab	diagnosis	2017-09-01 10:57:52

Additional Components

6.2 Negation Handling
- Negative words, no, none, free etc

6.3 Bigram Analysis
- Bigram Generation
- Bigram Detection

6.4 Stemming
- Porter Stemmer
- Stemmed Corpus file
- Stemmed NER

Results & Discussions

- Types of Test Results

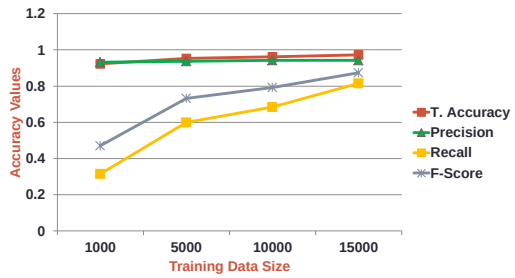
		Condition	
		Present	Absent
Test	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Results & Discussions (cont..)

- Accuracy Parameters
- $Total\ Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$
- $Precision = \frac{TP}{TP+FP}$
- $Recall = \frac{TP}{TP+FN}$
- $F - Score = \frac{2*Precision*Recall}{Precision + Recall}$

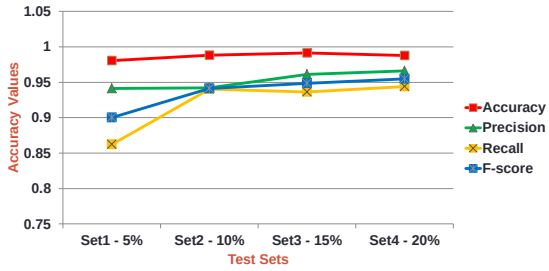
Results – Varying Corpus Size (cont..)

- Accuracy Vs. Training Data Size

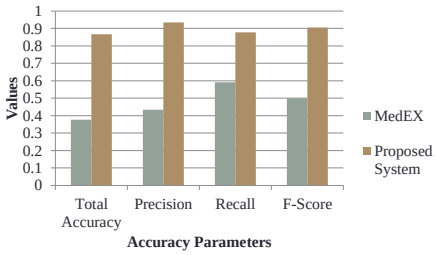


Results – Varying Test Data

- 4 different test set



Results – Comparison with MedEX



Conclusion

- Able to extract meaningful information
- Results saved in database
- Accuracy improved
- Contribution to the knowledge
  - Medical Corpus 15000+ notes.
  - Integrating the techniques of negation handling, bigram analysis and stemming in same system
  - More generalization
  - Improved accuracy

Future Works

- Corpus size can be increased
- Can be extended to detect other medical information and further parts of speech
- To build a learning system which will allow to add more undetected true positive elements in corpus
- Can be extended to work on speech and visual data
- Trigram & further n-gram analysis
- Other accuracy parameters like Specificity

Application Areas

- Extracting information from
  - Family history
  - Discharge summary
- Automatic reporting
- Developing standards
- Machine learning
  - Prediction systems
  - Suggestion systems

References

1. Xu, H.; Stenner, S.P.; Doan,S.;Johnson, K.B.; Waitman,L.R.;Denny, J.C. (2009, October 21). *MedEx: a medication information extraction system for clinical narratives.* Journal of the American Medical Informatics Association (JAMIA) pp. 19-24
2. Savova G.K.,; Masanz,J.J.; Ogren, V.P.; Zheng, J.; Sohn, S.; Kipper-Schuler, C.K.;Chute, G.C. (2010, June 29). *Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications.* Journal of the American Medical Informatics Association (JAMIA) pp. 507-513.
3. Garla, V.; Re, L.V. III; Dorey-Stein, Z.; Kidwai, F.; Scotch, M.; Womack,J.; Justice,A.; Brandt,C. (2011, April 22). *The Yale cTAKES extensions for document classification: architecture and application.* Journal of the American Medical Informatics Association (JAMIA) pp. 1-7
4. Aronson, A.R. (2001). *Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program.* AMIA
5. Friedman, C.; Alderson P.O.; Austin J.H.M.; Cimino J.J.; Johnson S.B. (1994, April). *A General Natural-Language Test Processing for Clinical Radiology.* Journal of the American Medical Informatics Association, Volume 1 Number 2.

Queries ??

Thank You !!