Unstructured to structured information conversion for extracting meaningful clinical information from medical notes

Awanish Ranjan

Rabindra Bista

Dept. of Computer Science and Engineering

Kathmandu University

Nepal

Contents

Introduction

- **Medical notes – Rich of clinical information like**
  - diagnosis
  - Procedure
  - Family History
  - Drug etc.
- **NLP (Natural Language Processing) techniques**
- **Clinical domain**
- **Difficulty**
  - Use of abbreviations like dr., pt., dx, rx. etc.
  - Ambiguity
  - Not always following correct grammar rules

Introduction

◆Unstructured information –

- Texts not following any pre-defined structure to store information
- E.g. –
  1. Spoke with pt over the phone. Pt presents with fairly new dx of diabetes, currently not any meds. States this happened about 2 yrs ago and was able to control blood sugars with diet and exercise.
  2. Pt presents with hyperlipidemia and strong family hx of CAD. Keeps active with job, kids, and softball, but no routine cardio exercise.

Introduction

◆Disadvantages of Unstructured text

- No regular pattern and structure like the order of occurrence of information
- So many abbreviated texts
- Tedious to read all notes manually and get information
- Time consuming
- Can't be automated for further analysis

Introduction

◆Structured Information-

- the information stored in a regular and general pattern not haphazardly
- Stored in a pattern and machine readable format like in database, xml etc.
- E.g. –
  - Note 1 -
    - Diagnosis - Diabetes from past 2 years
    - Medication - Not taking any medicine
    - Actions taken - Exercise and controlled diet
    - Result - control in blood sugar
  - Note 2 -
    - Diagnosis - Hyperlipidemia
    - Family History - CAD
    - Actions - Job, playing softball and being active with kids but no cardio exercise.

Introduction

◆Advantage of Structured Information-

- Easily interpreted by computer system for further processing.
- Information extraction with accuracy and speed
- Further processing like report generation, suggesting corrective actions etc.
- No tedious manual work and can be done way much faster

Structured Output format

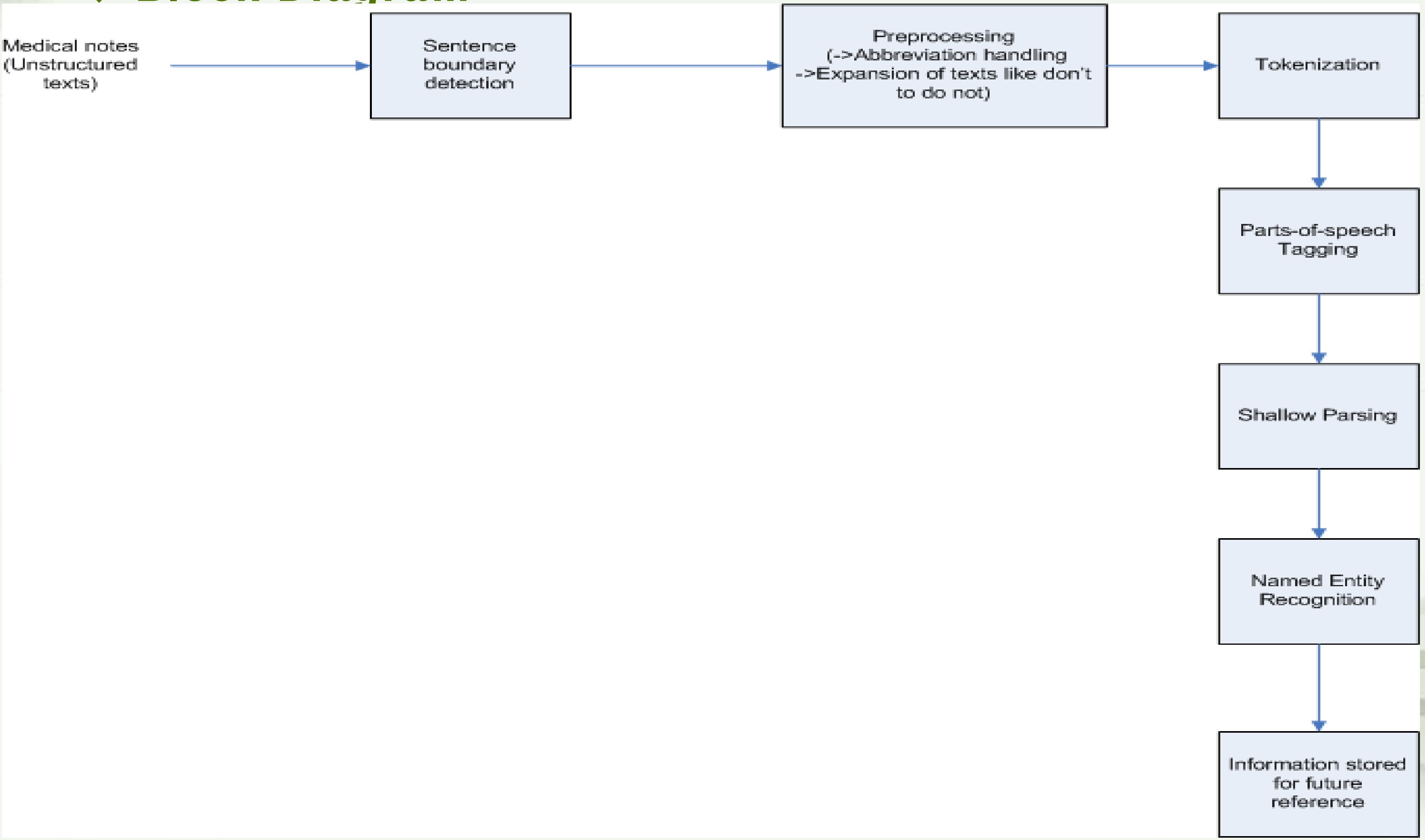
Patient Notes	Structured Info.	Examples
	PROBLEM_TIME	
		2 Yrs
	STATE	
		VITALS : Blood Sugar 150
	DIET_HABIT	
		Diet off track
		watching diet
		excercising
	DIET_COMPOSITION	
		Miracle Green
		Green Vegetables
		Water foods
		Fat Diets
		vitamin Supplements
		Fibre food
		Mono sat fats
	DIAGNOSIS	
		Diabetes
	TESTS	
		Eye Exam
		Dental Exam
		Foot care
	ADVICE	
		Followup appointment
		Take Diabetic meds
	MEDICATION	
		none

Objective

- Find out the appropriate method of converting unstructured text to structured information
- Extract meaningful clinical information from notes entered by medical practitioner
- Store the information for future use
- Study of appropriate Natural Language Processing methods
- Implement the appropriate NLP technique to solve the problem

Method

- Use of NLP techniques to solve the problem
- Block Diagram



Results

◆ Result of Sentence boundary detection –

- Sample Note -

"Spoke with pt over the phone. Pt presents with fairly new dx of diabetes, currently not any meds. States this happened about 2 yrs ago and was able to control blood sugars with diet and exercise."

- Split into individual sentences enclosed within single quote and separated by comma.

['Spoke with pt over the phone.', 'Pt presents with fairly new dx of diabetes, currently not any meds.', 'States this happened about 2 yrs ago and was able to control blood sugars with diet and exercise.']

Results

◆ Result of Preprocessing

Original Sentence >>> Pt presents with fairly new dx of diabetes, currently not any meds.

Preprocessed Sentence>>>:Patient presents with fairly new diagnosis of diabetes, currently not any medication. <<<

Results

◆ Result of Tokenization

Patient
Presents
With
Fairly
New
Diagnosis
Of
Diabetes
,
Currently
Not
Any
medication

Results

◆Result of POS Tagging

*****POS Tagging [using Penn Treebank tagging]****

('Patient', 'NNP') ('presents', 'NNS') ('with', 'IN') ('fairly', 'RB') ('new', 'JJ') ('diagnosis', 'NN') ('of', 'IN') ('diabetes', 'NNS') (',', ',') ('currently', 'RB') ('not', 'RB') ('any', 'DT') ('medication', 'NN') ('.', '.')

NNP	Proper noun, singular
NNS	Noun, plural
IN	Preposition or subordinating conjunction
RB	Adverb
JJ	Adjective
NN	Noun, singular or mass
DT	Determiner
,/.	Punctuation

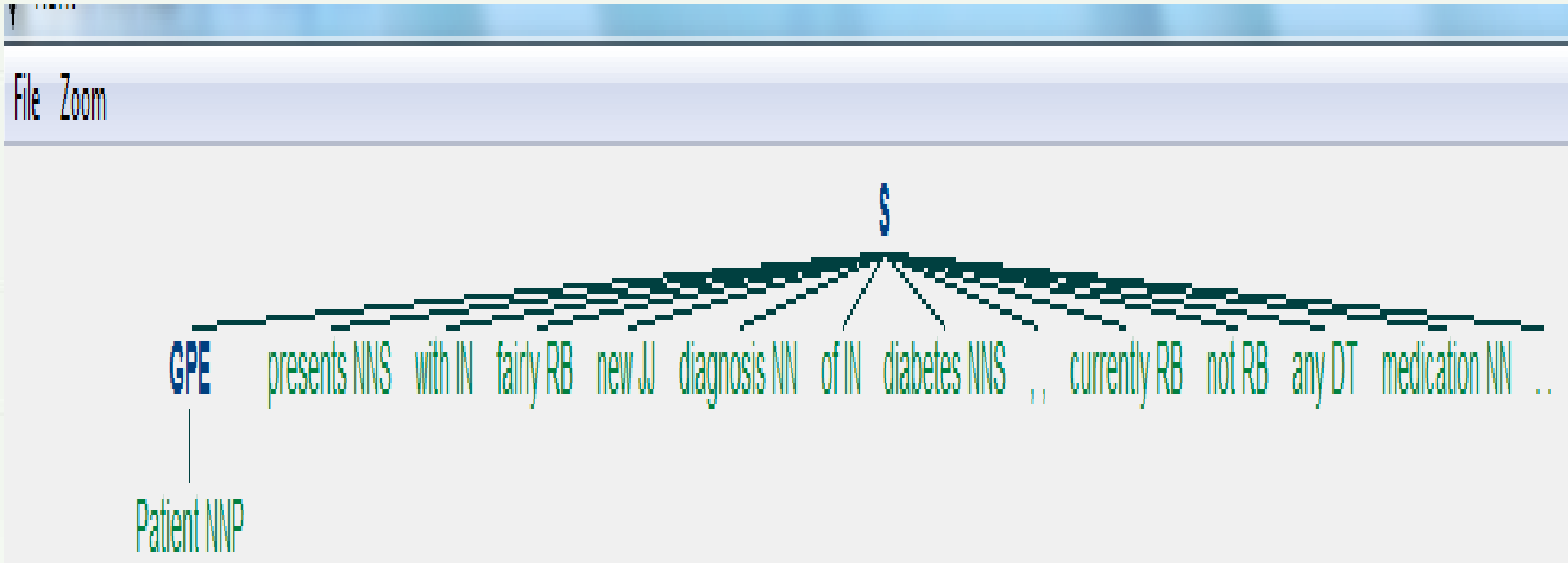
Results

- **Result of Shallow parsing and Named Entity Recognition –**

(S
(GPE Patient/NNP)
presents/NNS
with/IN
fairly/RB
new/JJ
diagnosis/NN
of/IN
Diseases diabetes/NNS
,/,
currently/RB
not/RB
any/DT
Drug medication/NN
./.)

Results

◆ Parse Tree



Discussion and Conclusion

Ambiguity

- One text carrying multiple meanings
- E.g. - Member has had two strokes.
  - Member has played two cricket strokes (cricket shot).
  - Member has written two strokes using pencil.
  - Member has had heart attack.
  - Member had brain stroke.
- Need to analyze the context of sentence
- Probabilistic approach –
  - Conditional Probability
  - Probability of occurrence of text based on previous text and finds the highest probability of occurrence

Discussion and Conclusion

- **Lack of suitable medical corpus**
  - Need to build a well defined corpus to refine Medical Named Entity Recognition

- **Training –**
  - Whole dataset is divided into 80-20 ratio
  - First 80% of dataset is used for training data and refining the algorithm
  - The next 20% data is used for test data

References

1. Xu, H.; Stenner, S.P.; Doan,S.;Johnson, K.B.; Waitman,L.R.;Denny, J.C MedEx: a medication information extraction system for clinical narratives; Journal of the American Medical Informatics Association (JAMIA), 2009; pp. 19-24
2. Savova G.K,; Masanz,J.J.; Ogren, V.P.; Zheng, J.; Sohn, S.; Kipper-Schuler, C.K.;Chute, G.C. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications; ; Journal of the American Medical Informatics Association (JAMIA), 2010; pp. 507-513.
3. Garla, V.; Re, L.V. III; Dorey-Stein, Z.; Kidwai, F.; Scotch, M.; Womack,J.; Justice,A.; Brandt,C. The Yale cTAKES extensions for document classification: architecture and application Journal of the American Medical Informatics Association (JAMIA), 2011; pp. 1-7
4. Liddy, E.D. Natural Language Processing; Encyclopedia of Library and Information Science 2nd Edition,2001

References

5. Ronan Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu,K.; Kuksa, P. Natural Language Processing (Almost) from Scratch; Journal of Machine Learning Research 12, 2011
6. Wolniewicz, R. Auto-Coding and Natural Language Processing; 3M Health Information Systems
7. Madnani, N.; Getting Started on Natural Language Processing with Python
8. Wu, Y.; Denny, C.J; Rosenbloom, S.T.; Miller, R.A.; Giuse, D.A.;Dr.Ing; Xu, H. A comparative study of current clinical natural language processing systems on handling abbreviations in discharge summaries; Department of Biomedical Informatics, Department of Medicine, School of Medicine, Vanderbilt University, Nashville, TN
9. Bodenreider, O.; Willis, J.; Hole, W. The Unified Medical Language System; National Library of Medicine, 2004
10. Klassen, P. Gate Overview and Demo; University of Washington CLMA treehouse Presentation, 2010

References

11. OpenNLP, URL - https://opennlp.apache.org/ (visited on December 2014)
12. NLTK, http://www.nltk.org/book/ (visited on August 2015)
13. http://sujitpal.blogspot.com/2013/04/language-model-to-detect-medical.html (visited on August 2015)
14. http://nlp-mentor.com/ambiguities/ (visited on September 2015)
15. https://www.packtpub.com/books/content/python-text-processing-nltk-20-creating-custom-corpora (visited on August 2015)
16. Coffman, A.; Wharton, N.; Clinical Natural Language Processing Auto-Assigning ICD-9 Codes;2007
17. Jurafsky, D.; Martin, J.H.; Speech and Language Processing; second edition

Thank You !!