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

A Ranjan[1*], R Bista[2**]

[1] *Kathmandu University, Dhulikhel, Nepal* [2] *Kathmandu University, Dhulikhel, Nepal*

*awa.ran@gmail.com **rbista@ku.edu.np

ABSTRACT: In medical domain, one of the most important documents is the notes that doctors, nurses or other medical practitioners take during patient interview. These notes contain vital information about the patient current condition, symptoms, family history, disease diagnosed, procedures done (like x-ray, lab test etc.), medication on which the patients are and so on. These notes are taken and entered in the system in plain text English language which is in an unstructured way. An unstructured text can be defined as text which contains information not in a common structured format.

Example texts -

1. Pt. presents with hyperlipidemia and strong family hx of CAD. Keeps active with job, kids, and softball, but no routine cardio exercise.

2. Member has asthma and eczema...member on has albuterol inhaler for emergency cases and is not on any other medication.

Since these note contain such huge amount of useful information about patient, everyone are willing to extract such vital clinical information. But lack of any defined structure of these information makes these texts to be interpreted only by humans and not by any computer program. Also these note contain many abbreviated texts like pt. and family hx etc. In any medical organization, there are hundreds of thousands of notes and going manually one by one by human will cost a lot of human resource as well as a lot of time. So the ultimate solution is to device an automated computer program to read these information and present in a structured way for quick processing and further analysis.

By structure information, it means the information stored in a regular pattern not haphazardly. In structured information, we have only the useful information extracted from the text and other unnecessary information are thrown away. This makes the information to be presented in quite definite structure which can be stored in some files or database for further processing. Structured information from above example notes can be extracted as -

Note -1

- Diagnosis - Hyperlipidemia
- Family History - CAD
- Actions - Job, playing softball and being active with kids but no cardio exercise.

Note 2-

- Diagnosis1 - Asthma
- Diagnosis 2 - eczema
- Medication - Albuterol inhaler

Presenting information in the structured way mentioned above will make it easy to be interpreted by any software program and do further processing like

- Extracting the useful information with some level of accuracy and speed thus removing quite error prone manual intervention
- Presenting various patient report based on these facts like which is the widespread disease by generating diagnosis report and so on.
- Suggesting some course of action with some automated suggestion module.

Based on the state of the arts systems currently present for converting unstructured information to structured information in medical domain, clinical Text Analysis and Knowledge Extraction System (cTAKES) tool is best suited for this purpose but one of the lacking feature in cTAKES is inability to handle ambiguity.

Ambiguity is one of the inherent characteristics of natural language. By ambiguity, it means same word carrying more than one meaning. This leads to inaccuracy and eventually wrong patient information which will guide the patient to the wrong treatment area. For example lets consider following sentence,

Member has had two strokes.

In this sentence, the strokes is an ambiguous text which can carry several meaning.

- Member has played two cricket strokes (cricket shot).
- Member has written two strokes using pencil.
- Member has had heart attack.
- Member had brain stroke.

So we need to get a full information about the context in which the note is presented and extract the correct meaning out of these.

To increase the accuracy of the tool, we also need to do some pre-processing which is focused on handling the abbreviated text, like in the example presented above Pt. indicates Patient and family hx - Family History. These need to be resolved before processing the text further. This will help improving the performance of the system.

Key-words: unstructured texts, structured texts, nature language processing, ambiguity, e-health

BACKGROUND: Natural Language Processing (NLP) is quite emerging field of Artificial Intelligence. There are quite a myriad of researches going on for tackling several challenges related to the NLP systems. We'll focus on English as Natural Language for this paper. One of the challenges that we face during NLP implementation is the quite unstructured content of natural languages. Though the English language has definite structures governed by the rules of grammar, in the normal communication, it becomes quite unstructured. Use of abbreviations, ambiguity, not always following the correct grammar rules, incomplete sentences are some of the factors due to which it becomes quite unstructured. Human brains are smart enough to capture such information and convert into the meaningful information for their use. But its quite hard for the automated computerized system to extract meaningful information from such unstructured texts.

The main discussion point of this paper is focused on the healthcare and clinical sector where the use of unstructured texts are prevalent. There are several instances where the medical practitioners like doctors, pharmacists and nurses generate such unstructured texts while collecting family history, prescribing drugs and so on. These unstructured texts are rich in clinical information and those needs to be extracted as meaningful knowledge for further processing.

Unstructured text and its disadvantages

As discussed above, unstructured texts don't follow any pre-defined structure to store information. It depends on person who is maintaining the note. These can be understood quite easily by humans but not by the computer systems. Here are some of the examples-

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.

3. Member has no notable health issues...says she has no pain and has not been to the ER within the last 12 months and is fine with the PCP she has been assigned...gave her overview of plan and verified address for packet information.

4. Member has asthma and eczema...member on has albuterol inhaler for emergency cases and is not on any other medication.

Disadvantages

As we can see in these notes, there is no regular patters of information present. Not all notes contain information about diagnosis or drug. Similarly there is not any structured way where the diagnosis comes first, then procedure and then drug. Also these note contain so many abbreviated texts.

Lack of any system to convert those information into structured one makes the medical practitioners' job tedious by needing to read all notes and history each time manually and extracting information from them and again storing it some other place. It consumes a considerable amount of time for them. Its more tedious when there is any transfer between one department of hospital to another like from emergency to operation theatre.

Structured information extraction & its usefulness

By structure information, it mean that the information stored in a regular and general pattern not haphazardly as we saw in previous section example notes. For this work, the information will be structured in tabular format and will be stored in a database. For above notes, the structured way of presenting information would be,

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.

Note 3 -

- Diagnosis - No pain
- Procedure - No ER

Note 4 -

- Diagnosis1 - Asthma
- Diagnosis 2 - eczema
- Medication - Albuterol inhaler

Advantages

Presenting information in the structured way mentioned above will make it easy to be interpreted by any software program and do further processing like

- Extracting the useful information with some level of accuracy and speed.
- Presenting report based on these facts.
- Suggesting some course of action with some automated suggestion module.

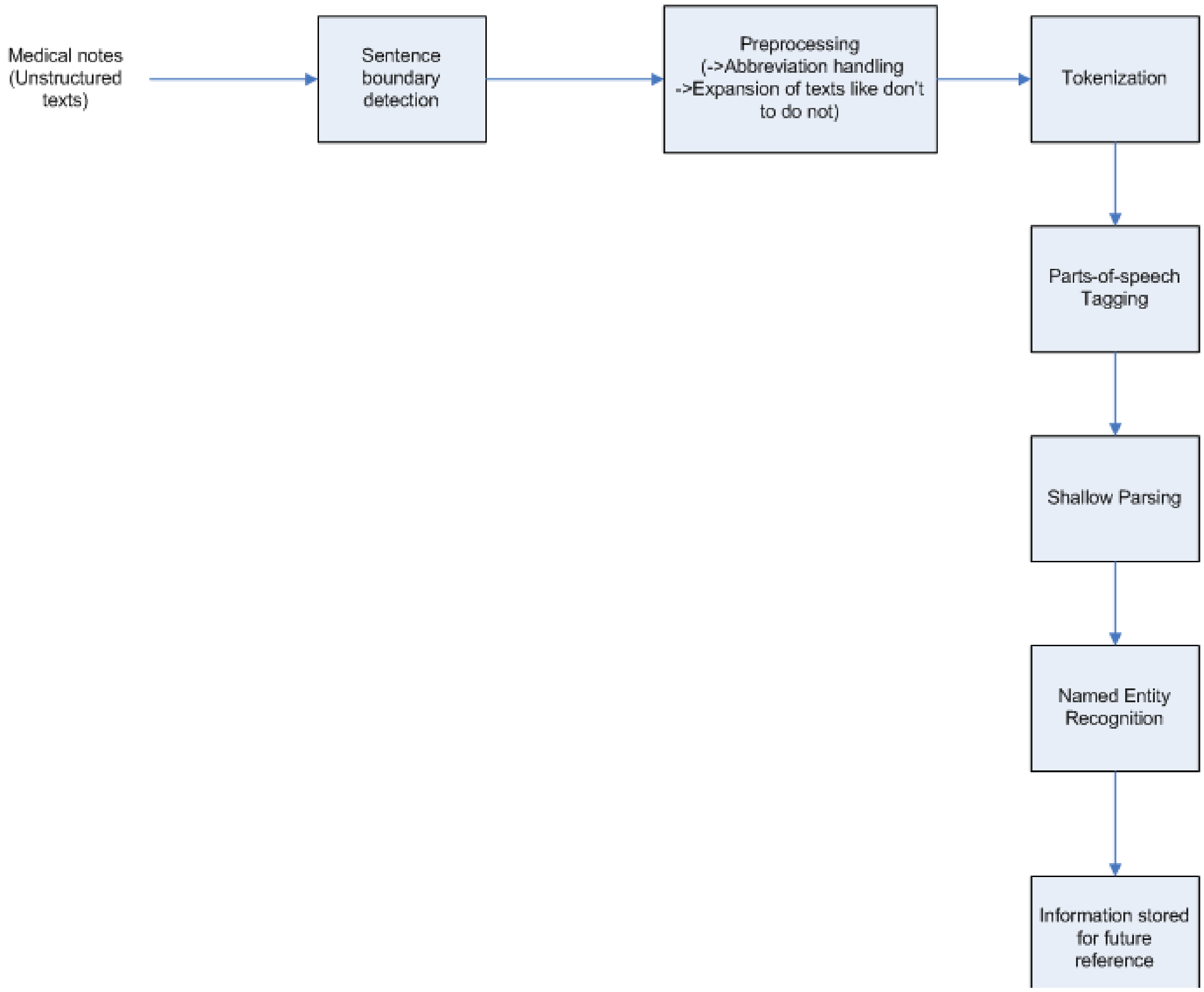
The general output format for the clinical note designed for this system looks like,

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: Object of this research is -

- 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

METHODS: Since the notes are entered in natural language (English is taken for this research), we are going to use NLP techniques to solve the problem. The block diagram and components of the system is shown in the figure below,


Medical notes (Unstructured texts)
Sentence boundary detection
Preprocessing (->Abbreviation handling ->Expansion of texts like don't to do not)
Tokenization
Parts-of-speech Tagging
Shallow Parsing
Named Entity Recognition
Information stored for future reference

- Sentence boundary detection - In this step, the sentences from clinically rich texts is identified. It detects sentences by using sentence terminators like period (.), question mark (?) etc.
- Preprocessing - This is the very first step where we are going to handle the abbreviation and remove punctuation so that we get a complete standard text for analysis.
- Tokenization - This steps breaks the sentence into smallest unit (usually words). For detecting tokens, it uses delimiters like space, comma etc.
- Parts-of-speech tagging (POS tagging) - This step assigns parts-of-speech to each token.
- Shallow parsing - A shallow parser segments a sentence into meaningful phrases like noun phase, verb phrase etc.
- Named Entity Recognition - Named entity recognition classify tokens in text into different categories such as person, diagnosis, procedure, drugs etc. For this case, we build a medically suitable corpus using some standard medical library (UMLS - Unified Medical Language System)
- Information stored for future reference - The structured text extracted then can be stored in database or xml format for future use like generating reports, suggesting appropriate path of action for patient and so on.

RESULTS: For this research, the tool used is Natural Language ToolKit (NLTK) based on python programming language. The results of each phase are,

- 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.']
```

- 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. <<<
```

- Result of Tokenization -

Tokens of this sentence are as follows

```
----
Patient
----
presents
----
with
----
fairly
----
new
----
diagnosis
----
of
----
diabetes
----
,
----
currently
----
not
----
any
```

```
----
medication
----
.
```

- 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') ('.', '.')
```

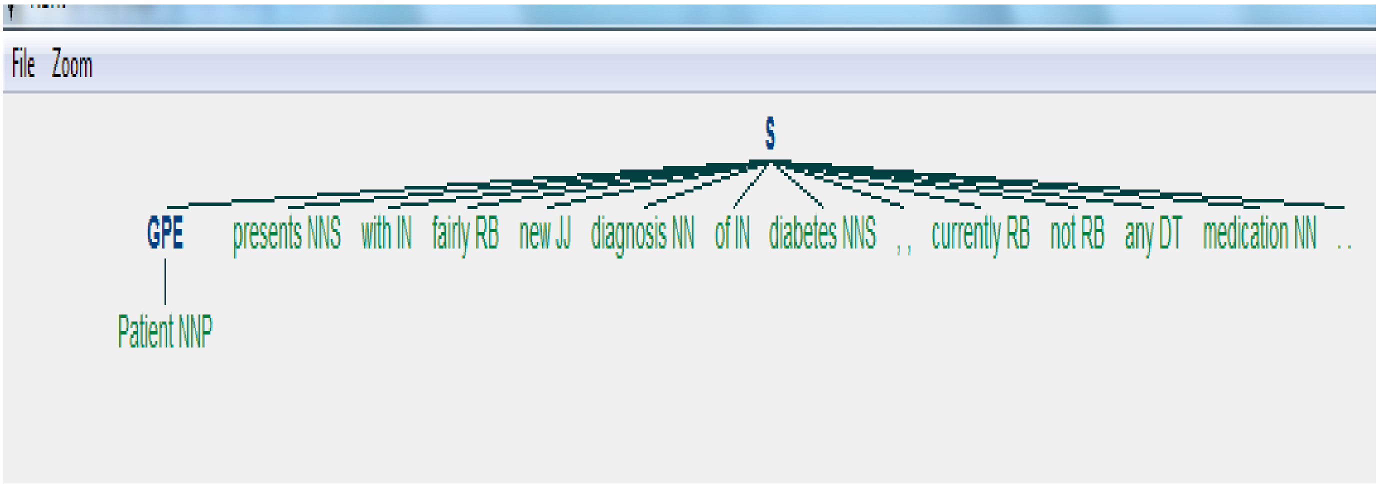
Here the parts of speech used are,

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

- Result of Shallow parsing and Named Entity Recognition -

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

Parse Tree -



DISCUSSION AND CONCLUSIONS: The medical notes written in natural language tend to be very ambiguous. One text carries multiple meanings. For example,

Member has had two strokes.

In this sentence, the strokes is an ambiguous text which can carry several meaning.

- Member has played two cricket strokes (cricket shot).
- Member has written two strokes using pencil.
- Member has had heart attack.
- Member had brain stroke.

In order to find out the most appropriate meaning, we need to analyse the context of sentence. For this, we need to use probabilistic approach. Conditional probability is the tool which helps us reduce the ambiguous situation and derive the most probable meaning. Here the conditional probabilistic approach determines the probability of occurrence of text based on previous text and finds the highest probability of occurrence. Based on that, we can determine the context and eventually the most probable meaning.

The challenges in taking out structural information from medical notes is the lack of any suitable medical corpus against which we can determine the appropriate medical terms and meaning. This needs building a well defined medical corpus and corpus reader which can be implemented for finding the NER.

Also the training is one of the most vital part of recognition. The whole dataset is generally 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. We then find the accuracy based on the result of the system ran on this unseen sample of 20%.

Future work of this research involves -

- Defining the well defined medical corpus and corpus reader.
- Implementing the probabilistic approach to reduce ambiguity.

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

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

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

Authors' Bio

1. A. Ranjan

MS by Research student, Department of Computer Science and Engineering, Kathmandu University, Nepal

Mr. Ranjan has completed his Bachelor in Computer Engineering from Khwopa Engineering College (Affiliated to Purbanchal University), Nepal, in 2006. He is currently pursuing his MS by Research in Computer Engineering from Kathmandu University, Nepal. He has worked as Software Quality Engineer in Verisk Information Technology, Nepal from 2007 to 2012. Currently, he is working as Director of Production Engineering in Deerwalk Services, Nepal, which deals with US healthcare analytics.

2. R. Bista

Asst. Professor, Department of Computer Science and Engineering, Kathmandu University, Nepal

Mr. Bista completed his B.Sc. IT from Sikkim Manipal University, India, in 2004. He has completed his MS and PhD in Computer Engineering from Chonbuk National University, S. Korea, in 2007 and 2011 respectively under S. Korea Government Research Funds. His research interest areas are wireless sensor networks, software engineering and health informatics. He has reviewed papers for many international journals and conferences. He is author of many international conferences and journal papers including book chapter. He is also associated with many organizing committees of international conferences. Since 2011, he is working as an Assistant Professor in Computer Science and Engineering, Kathmandu University, Nepal.