Translating Unstructured Texts into Structured Data: A Sate of the Art

Authors: Awanish Ranjan and Rabindra Bista

Affiliation: Department of Computer Science and Engineering, Kathmandu University, Dhulikhel, Nepal

E-mail - awa.ran@gmail.com and rbista@ku.edu.np

Abstract: In our day to day communication, we use normal English language which tends to be unstructured. These unstructured texts are well interpreted by humans but for machines, it's really difficult. Such unstructured information needs to be translated into the structured format so that it could be readable by machine and further processing on that information can be done. In this paper, we explore a state of the art of the tools in place for converting unstructured texts to the structured format using Natural Language Processing (NLP) techniques. For this, we evaluate the strengths and weaknesses of the existing tools on the basis of such metrics as accuracy in order to demonstrate their ability to convert the unstructured information to the structured format. Furthermore, in this paper, we provide some insights as future directions to convert unstructured health related texts into structured ones for e-health applications.

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

1. Introduction

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 a very 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 on 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. 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. So if there is any system, which extracts information from already entered notes and make them readily available to the physicians whenever they need it, it would save a lot of time and effort. The research about converting the unstructured text to structured information will solve this problem to a great extent.

Following is the basic block diagram of the proposed system,

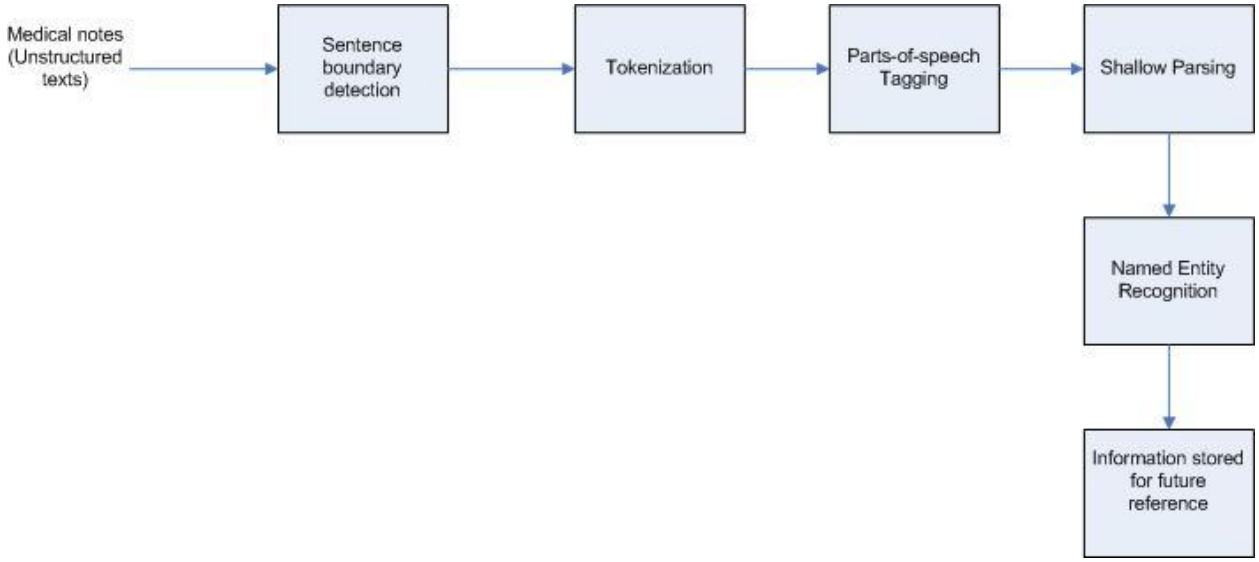

Fig 1: - Proposed block diagram

There are several research going on in this area resulting into the emergence of many tools and techniques. Out of these we are going to discuss and analyze three tools viz. cTAKES, YTEX and MedEx for the accuracy of conversion on the basis of three parameters : Precision, Recall and F-Score.

2. Application Areas

The system having the capability of converting unstructured text to structured knowledge has a huge implications in the medical sectors. Some of them are discussed here briefly,

Extracting information from family history - Family history is one of the vital document which is used in the initial phase of diagnosis. It helps identify the genetic disorders and any such inherent diseases. The system can extract such diagnosis from the family history which can be used for further diagnosis.

Extracting information from notes - There are several notes generated by doctors and nurses during diagnosis and followups in different health centers. Those are very helpful for future use in case of patient transfer from one hospital to another or from one department to another department within a hospital. Such notes serve as an eye to look into the entire health condition, diagnoses, procedures and drug conditions of the patient based on which the doctors can take further actions.

Extracting information from prescription - Drug prescription contains the important information about drug name, composition of drug, doses, strength, route, frequency, form (tablet or injection), duration, refill etc. Physicians need to check these during followups to decide on future course of action. These are also helpful for the health insurance companies which are responsible for covering these expenses for any patient.

Extracting information from discharge summary - Whenever patient gets discharged from inpatient department, a summary information is generated. This contains the information about the inpatient stay duration, diagnosis/procedures done during the inpatient stay etc. Again such information are used during follow-ups. This information is also useful to the insurance company while covering the inpatient costs for any patient.

Document Classification - There are several documents generated during medical treatment of a patient. Thos needs to be classified quickly for further screening. One example could be radiology documents. Reading the texts of such document and extracting information from them, the physicians should be quickly able to identify the type of symptoms and can quickly proceed further. In one of the tools YTEX, this classification system was used to quickly and accurately identify the patient who are having hepatic de-compensation by extracting information from radiological documents.

Besides the clinical use, the knowledge extraction system has uses in other sectors too. Like in business sector, such system can help extracting information some business

articles which will be helpful for business firms. Similarly it helps extract some political information from news texts and help identify the changing political trends overtime.

3. Classification of tools

For this study, we've studied three different tools available for converting unstructured information to structured form in clinical domain.

MedEX - Xu, H. *et al* designed a natural language processing (NLP) based tool which is used for extracting the medication or drugs information from clinical narratives. It was initially developed using discharge summary. It does well in identifying not only the drug names but other useful information as well like strength, route and frequency etc. A clinical text undergoes three steps in MedEX system to obtain the structured information, 1) Preprocessing in which the sentence boundary detection is done in the clinical texts. 2) Semantic tagging in which each clinical sentence is broken down in tokens and each token is labeled with a semantic category like Drug name, drug strength etc. 3) Parsing then uses context-free grammar to parse the textual sentences into structured form using a chart parser.

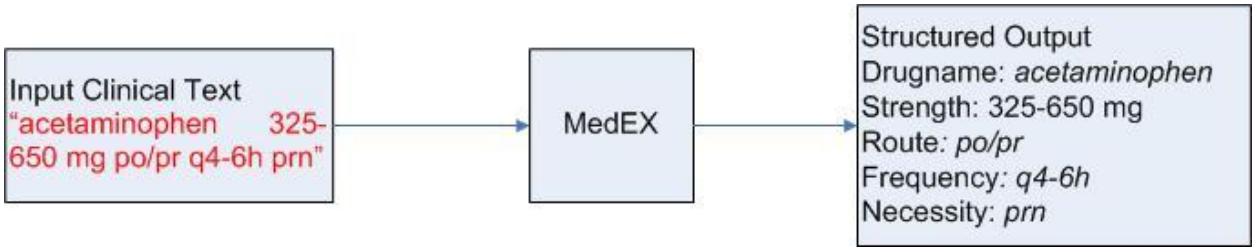

Fig 2: - MedEX block diagram

cTAKES - Savova K.G. *et al* developed system Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES) based on open source NLP for information extraction from Electronic Medical Record (EMR). The system is built on existing open source NLP technologies like the Unstructured Information Management Architecture (UIMA) framework and OpenNLP natural language processing toolkit. This system is trained on clinical domain for creating rich linguistic and semantic annotations. The dataset used for cTAKES is a subset of clinical notes from Mayo Clinic EMR. Following is one example of knowledge extraction using cTAKES system.

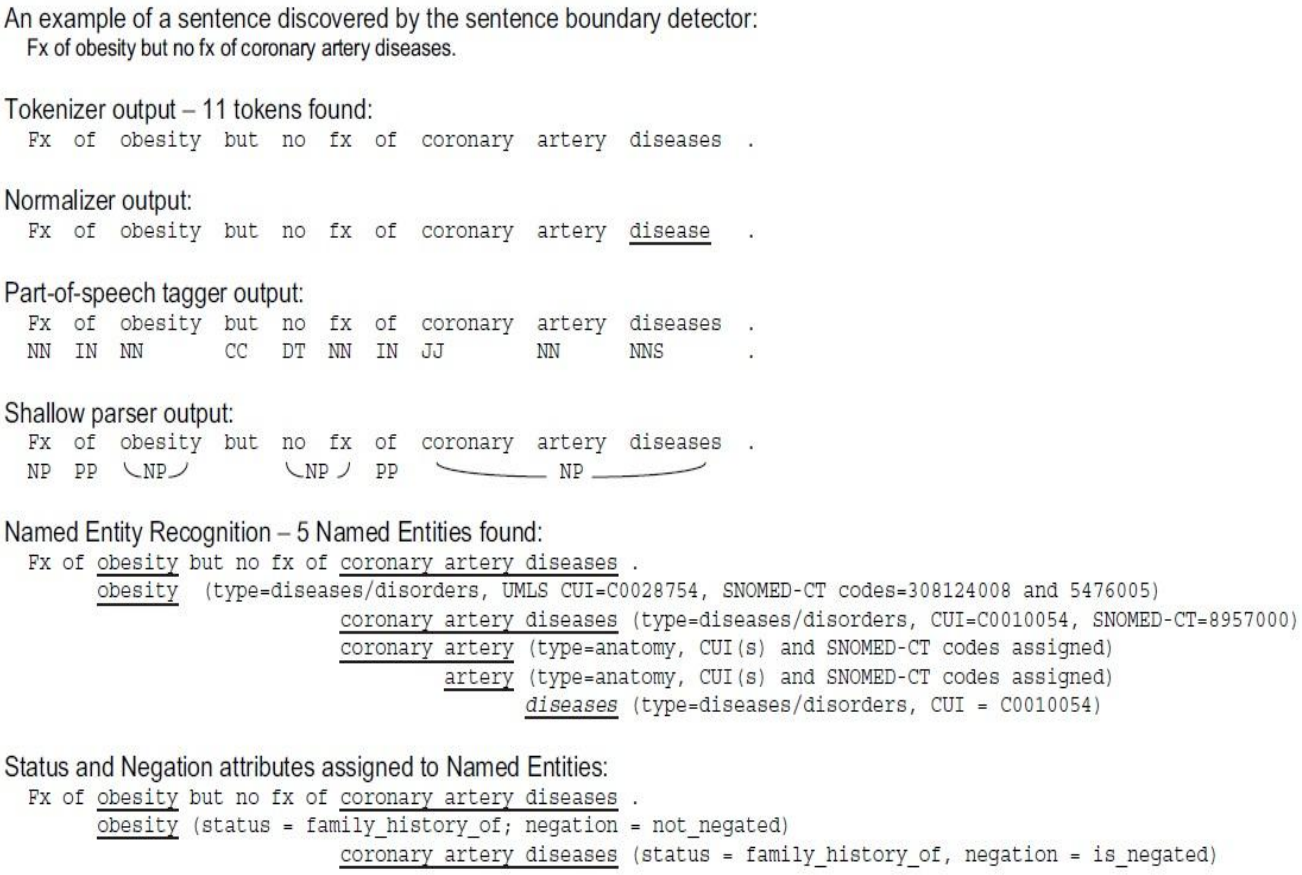

Fig 3: - cTAKES system output

YTEX - Glara, V. *et al* developed a system based on cTAKES for the classification of radiology reports that contains the findings leading to suggest that the case is of hepatic decompensation. The system is called The Yale cTAKES extensions for document classification. YTEX exploits the knowledge extracted using cTAKES system and stores the annotations in a relational database. YTEX modified cTAKES by using 1) A regular expression based named entity recognition. 2) Latest version of NegEx algorithm to detect Negation 3) a database module to store the annotations in the database. The figure below shows the workflow of YTEX system.

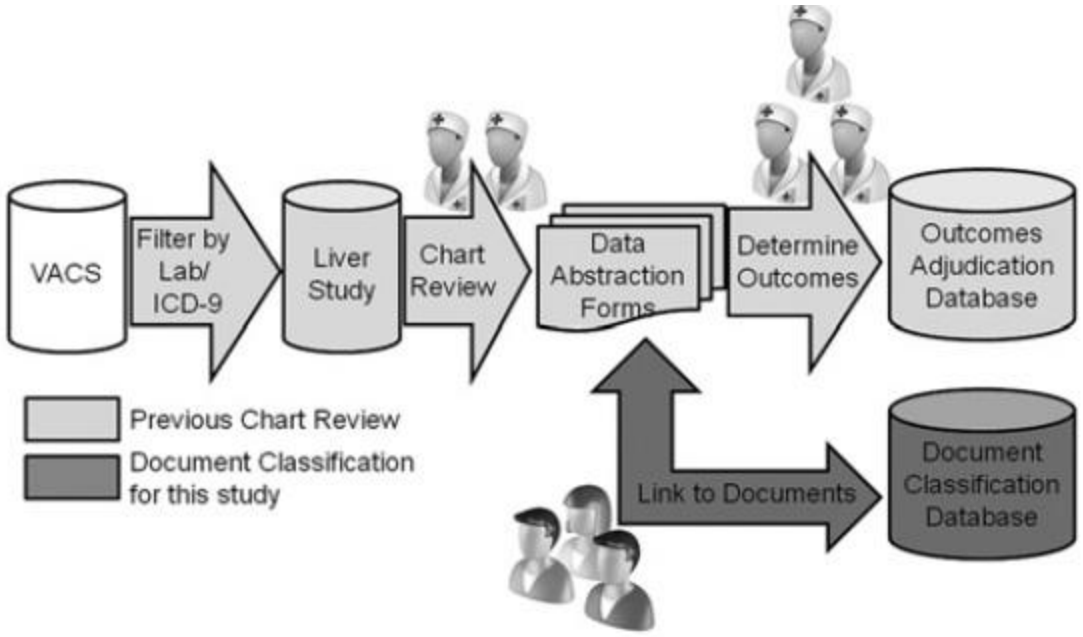

Fig 4: - YTEX block diagram

4. Comparative study of NLP systems

In this section, we are going to present comparison of these NLP systems based on following metrics

Type of data (ToD): This parameter presents an idea about what sort of data is used for evaluating these systems. There are different variety of data on which these systems have been evaluated.

Application Domain (AD) : This parameter will present the information about the domain in which these tools are applicable.

Pre-processing required (PRE): This is one of the basic steps for NLP systems. We'll see how pre-processed data effects the accuracy of the system. Some of these tools require pre-processing of data before being fed in the system. The pre-processing generally involves the tasks of detecting the sentences from the clinical text based on the operators like period (.), question marks (?) etc.

True Positive (TP): It indicates the result as positive when actually positive. Like detecting cancer from the text when actually cancer is present in that patient.

False Positive (FP): It indicates the result as positive but in reality it is not. Like detecting cancer from the text when actually its not present.

True Negative (TP): It indicates the result as negative when actually negative. Like not detecting cancer from the clinical text and in reality cancer not being present in the patient.

False Negative (FN): It indicates the result as negative when actually its positive. Like not detecting cancer from the clinical text but in reality its being present.

Precision (P): It is the ability of system to detect the actual positive results. It shows how accurately the systems detects the positive results i.e. the result what is actually expected to be detected by the system. Following is the formula for calculating precision-

$P = TP / (TP + FP)$

Recall (R): Its also called the sensitivity and defined as,

$R = TP / (TP + FN)$

F-Score (F): It shows the relationship between Precision and Recall and is defined as,

$F = (2*P*R) / (P+R)$

Ambiguity Resolution (AR): Ambiguity is an inherent characteristics of NLP where same word can have different meaning based on different scenarios. This parameter shows if the tool is capable of handling of ambiguity or not.

Redundancy Handling (RH): This parameter shows if the tool is able to detect the redundancy or not. In clinical texts, same information may be repeating again and again and its expected that if there is same scenarios with same information, those should be treated as redundant information and further processing should not be done in order to save the system from doing redundant tasks again and again.

Approach (A): There are two approaches of system being used 1) general (G) - if the system can be used for any purpose and any type of data 2) specific (S) - if the system can used for only some specific type of data.

Processing Type (PT): Processing type could be 1) Centralized (C) - processing capability of single platform only. 2) Distributed (D) - Processing capability across multiple platform simultaneously hence enhancing the speed and performance of the system.

Table 1 - Comparison results of NLP tools for converting unstructured texts to structured information

Tools\Metrics	ToD	AD	PRE	P	R	F	A	AR	RH	PT
MedEX	Medication	Clinical (Drug)	Y	0.97	0.88	0.92	S	Y	N	C
cTAKES	Mayo clinical EMR	Clinical (EMR)	N	0.80	0.64	0.71	G	N	N	C
YTEX	Radiology document	Clinical (Radiology)	N	0.96	0.92	0.94	S	N	N	C

Legend: Y = Yes; N=No; S=Specific; G=Generic; C=Centralized

5. Future Directions

From this study we found that following are some of the areas which needs to be discovered in future -

Distributed Processing capability: The tools right now are lacking the capability to be processed in distributed environment. These days, we need to process huge amount of data and in that case the centralized processing will be huge bottleneck for performance and speed. Distributed processing makes the processing spread across different processing unit chunking the data in smaller piece. Extracting information from the smaller piece would be faster. After processing completes in the smaller piece, the extracted information needs to be aggregated to find the overall result.

Mobile platform: The NLP tools studied so far are also not used for the handheld devices like Mobile platform. In this changing world of technology where everything is available in the handheld devices, such NLP tools should also be able to be used easily there.

Redundancy Handling: The systems lack the capability of handling the redundant information processing right now. In huge clinical texts, there are a lot of redundant sentences. The system should be smart enough to detect the redundancy and prevent itself from processing such information repeatedly time and again. This capability needs to be added.

Ambiguity Resolution: So far the NLP systems are poor in handling ambiguity. The semantic portion of the tools should be enhanced in order to increase the ambiguity

handling capacity. There should be enough training and context based analysis to improve the ambiguity resolution.

6. Concluding Remarks

This study of tools provided us a great insight of the currently available tools in the area of extraction of structured information from unstructured text written in some Natural language in clinical domain. What actually we need is to extract the major attributes like diagnosis, procedure and drug information from the clinical text and stored somewhere as structured information. The tool that is close to this research is cTAKES. YTEX looks too specific to the document classification where as MedEX to the medication field. We continue to research on how to improve ambiguity resolution and redundancy handling capability of this tool.

7. 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)