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

**Awanish Ranjan and Rabindra Bista**

Department of Computer Science and Engineering, Kathmandu University, Dhulikhel, Nepal; E-mail - awa.ran@gmail.com and rbista@ku.edu.np

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

**Abstract:**  In medical domain, one of the most important document 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 (like x-ray, lab test etc.) done, 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 occurrence makes these texts to be interpreted only by humans and not by any computer program. Also these note contain so many abbreviated texts like pt. and family hx. But 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 further processing.

By structure information, it mean that 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 stored in quite definite structure in some files or database for further processing. Structured information from above example note 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 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. It means same word carry 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,

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.

This paper will discuss the effects of implementing ambiguity handling algorithm in cTAKES tool to enhance the accuracy.

To increase the accuracy of the tool, we also need to do some preprocessing 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.

**References**

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.

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

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