A New Approach to Extract 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

KEYWORDS: unstructured texts, structured texts, natural language processing, health informatics, ambiguity

BACKGROUND: 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's current condition, symptoms, family history, disease, procedures (like x-ray, lab test etc.), medication on and so on. These notes are in plain text English language which is quite unstructured. An unstructured text can be defined as text which contains information not in a common structured format.

Since these notes contain very useful clinical information, everyone is willing to extract such vital clinical information. But due to lack of defined structure, it makes these texts to be interpreted only by humans and not by any computer program in most of the cases.

In structured information, we have only the useful information extracted from the text and other unnecessary information is 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.

Example –

Unstructured Text - Pt. presents with hyperlipidemia and strong family hx of CAD. Keeps active with job, kids, and softball, and routine cardio exercise.

Structured data -

- Diagnosis - Hyperlipidemia
- Family History - CAD
- Habit - exercise.

OBJECTIVE: Objective 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 using NLP techniques.
- Building medical corpus to find named entities.
- Store the information for future use

METHOD: Since the notes are entered in natural language (English is taken for this research), we used use NLP techniques like tokenization, Parts of speech tagging, Named Entity Recognition to solve the problem. To bring the accuracy, we took around 13,000 sentences from medical notes and used it as training set to generate a corpus. The corpus is then used for recognizing proper named entity.

RESULTS: The final outcome of this research is the structured data extracted from the unstructured notes. The output of the proposed system is analysed by doing repeated test in different kind of dataset, using the corpus developed during the course of this research. The analysis was focussed on finding out true positive, true negative, false positive and false negative values. With sample data and the trained corpus by this approach, we could able to achieve a precision of 0.908, recall of 0.90 and F-score as 0.904. We were also able to find the pattern of these values associated with varying number of training and test data.

CONCLUSIONS: The outcome of this research is very useful in order to get the meaningful data from medical notes. This will enable a computer system to detect the useful information automatically which brings accuracy and removes error prone human intervention. The data extracted is saved in a structured way

in some database and can be used further for different reporting systems. This structured data can be used as a basis of different machine learning approach for developing prediction algorithm in health sectors.

Authors' Bio

1. A. Ranjan

Student, MS by Research program, 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 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.