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, e-health

Theme : 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 the information but not in a common structured format due to which it is quite difficult for any automated system to extract meaningful information without converting those text in structured format. These notes need to be processed to convert into structured format and extracting valuable information.

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.

Since these note contain such huge amount of useful information about patient, everyone are willing to extract such vital clinical information. Lack of any defined structure of these information occurrence makes these texts to be interpreted only by humans and not by any computer program. 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.

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.

Problems tackled by this research

The task of converting medical unstructured notes to structured form is not that easy. There are several challenges and problems that we will need to tackle to come to the desired output. Those are -

1. **Abbreviated texts** - As we see from the unstructured notes presented in above section, those are full of abbreviated medical texts like
   - pt. - Patient
   - dx - Diagnosis
   - med - Medication or drug
   - family hx - Family History

   These need to be resolved accurately.

2. **Medical named entity recognition** - The notes as rich in many medical terms which can't be handled by simply using the English vocabulary. We need to have some medical terms vocabulary to identify those clinically significant information maintaining a high percent of accuracy.
3. **Negation handling** - As you can see in some of the notes above, the terms like 'no pain', 'no meds', 'no exercise' actually denoting negation. This means that the patient should not be identified as having pain or taking medicine or doing exercise. We need to introduce some negation handling technique here to handle such situation.
4. **Ambiguity handling** - Ambiguity is an inherent feature of English feature where 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.

Biography

1. Ranjan Awanish

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

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. Bista Rabindra

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

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.