ERCIM "Alain Bensoussan" Fellowship Scientific Report

Fellow : Saiful Akbar
Visited Location : Norwegian University of Science and Technology (NTNU), Trondheim
Duration of Visit : 12 months, 06.01.2009 – 05.01.2010

I - Scientific activity
(1 page at maximum)

My postdoctoral research was conducted at the Department of Computer and Information Science (IDI) NTNU on 06.01.2009-05.01.2009 with the scientific contact Prof. Øystein Nytrø and was done in close collaboration with a researcher at IDI Dr. Laura Slaughter. The main goal of the research project is applying Natural Language Processing (NLP) and Data Mining (DM) approaches for extracting useful information from medical text written either by the experts or by lay persons. Since the domains of medical/health informatics and NLP were new for me, I was started with doing literature study and discussion with other researchers in order to catch up the global view and to have deep insight about the domains[1]. In addition, in order to understand the state of the art of the research in the domain, I have also attended on several weekly seminars at Norwegian Research Centre for Electronic Patient Records (NSEP), where researchers and PhD students in the domain of medical/health informatics present their research work.

The main research activities are two folds. First, we observed how to provide clinical researchers with useful information extracted from patient health writings (PHWs), i.e. text written by patients lay persons. Our literature study has highlighted that although there is a gap between lay person expression and expert terminology, both terminologies overlap to each other. Thus, our hypothesis is, some selected PHWs should contain useful information that can be extracted, the quality of the information can be assessed using a relevant ontology, and at the end it will be valuable for clinician researchers. This work has three challenges: (a) how to collect PHWs and how to select relevant text from them, (2) how to assess the "quality" of the information delivered by lay person, and (3) what the clinical researchers need from patient text and how to provide it. In the first steps we have collected PHWs from a breast cancer mailing list, and extracted some features from the emails using UMLS terminology resources, and classified them into health- versus non health-related text using a simple DM approach, K Nearest Neighbor (KNN). The results were published on a paper listed in Section II.c.

Second, we have focused on extracting information from text produced by the experts. To be more specific, in collaboration with a PhD student I participated in the i2b2 challenge (http://www.i2b2.org). We experimented on extracting medication-related information, which is

---

[1] As a result of the study of the health/medical informatics domain, a paper on which I am the third author was published (see Section II.a). Meanwhile, during the study of problems and approaches in the NLP domain, I have built a small prototype of an NLP tool containing two modules: word stemmer and openNLP/MaxEnt based sentence detector for the Indonesian language. The work is not published yet.

useful for further analysis, from patient discharge summaries (PDSs) using NLP and DM approaches. Normally, clinicians write PDSs in a semi structured form. Moreover, they tend to use limited terms and have limited patterns of expression. Thus, extracting some medication-related information from such text, such as medication name, dosage, frequency, duration, reason of the given prescription, etc., is possible. We have built a system prototype supporting that functionality. The system made use of UMLS knowledge resources and employed some patterns to recognize and extract medication terms written in the PDSs. Furthermore, the system utilized the recognized terms and their context to extract the rest information. The detail results were published on a paper listed in Section II.b.

Ongoing research work: as the continuation of the two research works, we are working on:
(a) Collecting and observing blogs written by patients or caregivers experiencing with dementia disease. In parallel with this work, we are consulting dementia an expert in order to validate if the blogs could be useful for her clinical research and what kind of information she needs.
(b) Observing the use of medication-related information and the patterns contained in the PDSs and building a system prototype to accelerate and to improve the quality of medical documentation.

II- Publication(s) during your fellowship
Please insert the title(s), author(s) and abstract(s) of the published paper(s). You may also mention the paper(s) which were prepared during your fellowship period and are under reviewing.

(a) Wikan Danar Sunindyo, Soleh Udin Al Ayubi, Saiful Akbar, Fazat Nur Azizah, Putri Saptawati,

Hira Laksmiwati Zoro, *e-Health Implementation for Rural Areas in Indonesia -Review on Data and Technology Readiness*, International Conference on Rural Information and Communication Technology (r-ICT), Institut Teknologi Bandung, 17-18 June 2009, Bandung, Indonesia.

Abstract: e-Health can be defined as the utilization of information and communication technology (ICT) in the health sector. This paper will give a review of e-Health implementation readiness in Indonesia, especially for data and technology readiness for the rural areas in Indonesia. One of the problems in Indonesia is the health service providing still not well-distributed among the areas. It is expected, that by using ICT in the health sector, the gaps between the city and rural areas can be narrowed. The target audiences for this paper are the government that will make policies on health sector, the health service providers including hospitals and doctors, patients and people, especially in the rural areas. The challenge here is narrowing the health service gaps between the city and the rural areas, for example the unequal number of doctors, hospital and health services in both areas. The criteria for a good solution are increasing of the health rate. Our solution approach is by using information and communication technology in supporting health sector. We will review several components of e-Health both in data and technology readiness, so we can see what current condition is and what can be developed from that.

(b) Saiful Akbar, Thomas Brox Røst, Laura Slaughter, Øystein Nytrø. *Extracting Medication Information from Patient Discharge Summaries*, Third i2b2 Shared-Task and Workshop Challenges in Natural Language Processing for Clinical Data Medication Extraction Challenge, 13 November, 2009, San Francisco

Abstract: This paper describes our system for extracting medication-related information from Patient Discharge Summaries (PDS). The system consists of four modules. The first module is the medication extractor (MDXtor), which extracts expected medication using UMLS Metamap Transfer (MMTx) and is the base for other modules. The second module is the DMFDR extractor that extracts dosage, mode of administration, frequency, duration, and reason for giving the medication. The DMFDR extractor consists of a partial parser and an EBNF (Extended Backus-Naur Form) grammar covering medication-related information. Module ETCL determines type (class) of event, temporal marker, certainty, and textual form (list or narrative) where the medication is found in the text. It extracts features relevant to specific classifications and employs a KNN classifier for determining the class. The last module, called "collector", is responsible for merging the output of the three modules and building a single output for the individual PDS.

(c) Saiful Akbar, Laura Slaughter, Øystein Nytrø . *Collecting Health Related Text from Patient Health Writings.* the 2nd International Conference on Computer and Automation Engineering (ICCAE 2010), 26-28 February 2010, Singapore (to appear)

Abstract: The Internet has been a huge resource for sharing and collecting information including health related information. Some health related information is written by patients (lay persons) discussing their experience about health problems and treatments. This paper introduces our initial work on providing physicians with clinically useful patient health writings. More specifically, the paper presented our experiments, as a part of the whole research work, on filtering health related text from patient health writings. We focused on selecting possible feature for classifying text from breast cancer mailing list into health and non health related text. Using KNN classification method, we experimented with various features, i.e. all terms, all terms except most frequently used terms, UMLS terms, health related UMLS terms, and health related UMLS semantic types. The experiments showed that UMLS terms extracted from the text is a good feature, compared to the other features.

III -Attended Seminars, Workshops, and Conferences
Please identify the name(s), date(s) and place(s) of the events in which you participated during your fellowship period.

During the fellowship program, I have participated in:
- Some weekly seminars/workshops at Norwegian Research Centre for Electronic Patient Records (NSEP)
- some PhD defences at NTNU
- ITOvation at NTNU
- Nokia QT seminar for developers in Trondheim, **NTNU**.

I will present the third paper as shown in Section II.c at the 2nd International Conference on Computer and Automation Engineering (ICCAE 2010), 26-28 February 2010, http://www.iccae.org/.

Other activities:
I have been a scientific reviewer for the following conferences:

- IEEE Africon 2009, 23-25 September 2009, Nairobi, Kenya, http://www.africon2009.co.za/
- National Conference on System and Informatics (KNS&I2009), 14 November 2009, Bali, Indonesia, http://knsi.stikom-bali.ac.id/

IV – Research Exchange Programme (12 month scheme)

Please identify the name(s), date(s) and place(s) of your Research Exchanges during your fellowship period and detail them .

In term of Research Exchange Program, I have visited the following institutes:

(a) ICS-FORTH, Crete, Greece, 23.06.2009-30.06.2009, Scientific Contact: Prof. Manolis Tsiknakis.
    In this visit, I gave an introduction about my ongoing research. The scientific contact provided me with a global introduction about research topics and his research teams and the research environment in ICS-FORTH. We had interesting discussion about research projects in ICS-FORTH especially their involvement in ACGT project which is relevant to my research area. I also consulted many of the colleagues there to know about the research environment. A colleague provides me with a demonstration about the ACGT workflow editor. It is a system to compose different data access, knowledge extraction, and analytical services into complex workflow in order to assist physicians in their research.

(b) Information Engineering Group, IFS TU Vienna, Austria, 15.07.2009-31.07.2009, Scientific Contact: Prof. A Min Tjoa.
    In this visit, I presented my research project in extracting useful information from medical text, and collected many valuable feedbacks from some colleagues there. I also had an intensive discussion especially about practical know how on medical-information extraction with some colleagues from, IFS TU Vienna. A colleague there arranged a meeting with the group leader in order to provide me with a rich discussion about their research area and how to manage their research. We also had agreed in exploring some research proposal calls in Austria for our future cooperation.

In general, through the visits I had broadened my scientific network with new valuable scientific contacts.