

BioSearch08

HCSNet
Next-Generation Search Workshop
on
Search in Biomedical Information

Program, Information and Abstracts

Queensland University of Technology
Brisbane

30 November 2008



Contents

Welcome	1
About HCSNet.....	2
Registration and General Information	3
Venue	3
Transport	3
Public Transport.....	3
Cars and parking	4
Registration Desk	5
Delegate Materials.....	5
Travel Support	5
Refreshments and Meals	5
Program – Sunday, 30th November 2008	6
Participants with Poster Only	7
Abstracts	8
Keynote Presentation 1	8
Dina Demner-Fushman.....	8
Short Presentations - Session 1.....	8
Alexander Krumpholz	8
Jon Patrick	9
Stephen Wan	9
Wei Liu.....	9
Student Presentations - Session 1.....	10
Pooyan Asgari	10
Andrew MacKinlay	11
Juana Maria Ruiz-Martinez.....	11
Mojtaba Sabbagh-Jafari	12
Keynote Presentation 2	12
Limsoon Wong	12
Short Presentations - Session 2.....	13
Peter Ansell	13
Peter Budd.....	13
Sarvnaz Karimi.....	14
David Martinez	14
Student Presentations - Session 2.....	14
M. Asif Khawaja and Fang Chen.....	14
Stefan Pohl	15
Willy Yap	15
Sun Xiaoxun	15
Posters Only.....	16
Stefan Schaefer	16
Yeondae Kwon.....	16
Anthony Nguyen	16
Yefeng Wang.....	17
Workshop Contact List.....	18

Welcome

Welcome to BioSearch08: the HCSNet Next-Generation Search Workshop on Search in Biomedical Information.

There are many aspects of research and practice in medical environments that require searching for medical information in large information repositories of published research papers (such as PubMed), and large ontologies and other structured resources (such as UMLS). The availability of these resources and the need to find information in them has prompted intensive research on the treatment of medical text, images and videos for information extraction and retrieval.

All of these tasks require the development of techniques that go beyond the localisation of generic search and information extraction techniques, given the high level of complexity in terminology, and the need for achieving high performance.

The aim of HCSNet's Priority Research Area in Next Generation Search is to bring together researchers from a wide range of disciplines to address issues related to search and information extraction. This specialised workshop focuses on the medical and biological domains and takes place just before GIW 2008, in Brisbane.

Dr Diego Mollá-Aliod
Professor Robert Dale

Macquarie University
Macquarie University

About HCSNet

The ARC Research Network in Human Communication Science – HCSNet – was awarded five years' funding by the Australian Research Council in late 2004. HCSNet's aim is to promote and facilitate interdisciplinary research in human communication science by connecting leading researchers in the wide range of disciplines that focus on language, speech and sonics.

HCSNet's activities are centered around five Priority Research Areas:

- Speech
- Effective Interfaces
- Next-Generation Search Technology
- Human Communication Disorders
- Perception and Action

By bringing together researchers in these areas and fostering interdisciplinary research, the network aims to build Australia's reputation as a leader in communication science and technology via advances in areas as diverse as automatic speech recognition, distress call monitoring, hearing prostheses, web interfaces, and data retrieval and data mining systems.

Getting involved in HCSNet is easy: visit www.hcsnet.edu.au to sign up as a member of the network. You'll be added to our online profile database, and automatically receive our weekly electronic newsletter, HCSNet Update, which will keep you informed of HCSNet activities, including our annual week-long SummerFest, and events in the range of HCSNet disciplines. Australian-based HCSNet members can apply for funding under our various programs.

Registration and General Information

Venue

The University of Queensland (QUT), Brisbane
The Owen J. Wordsworth Room
Level 12, S Block
QUT Gardens Point Campus
2 George Street
Brisbane QLD 4000

The Gardens Point campus is located on the Brisbane River in the city centre, next to the Botanic Gardens and Parliament House. It is within easy walking distance to shops, restaurants, theatres, galleries, and public transport including buses, trains and ferries.

Transport

Public Transport

You can travel to and from QUT Garden Point campus by bus, train, taxis, river ferry or the CityCat. For more information on public transport in Brisbane, such as timetables, zone maps, information on tickets, and answers to your public transport questions, visit www.transinfo.com.au or phone 13 12 30 between 6am and 9pm weekdays, and 7am and 9pm Saturdays and Sundays (public holiday times vary). You can also collect timetables, maps or ask questions in person at the Transinfo Bus Information Kiosk in the Queen St Mall, and at Brisbane City Council (BCC) Service Centres, BCC Libraries and Ward Offices.

Ferry

The CityCat and cross-river ferries (Cityferry) are operated by the Brisbane City Council between approximately 5.50am and 10.30pm daily. The CityCat runs up and down the Brisbane River from Hamilton to St Lucia, while Cityferry services run across the river between the city and South Brisbane/Kangaroo Point. Both the CityCat and Cityferry stop near Gardens Point campus. Ticketing operates in a similar way to buses and you can use the same tickets on ferries and buses, including transfers. Visit www.translink.qld.gov.au for more information.

Airport Transfers

The quickest and most economical way to get to the City from the Airport is the Airtrain; the trip is around 20 minutes. The Airtrain stops at all downtown stations including Bowen Hills, Fortitude Valley, Central Station, Roma Street, South Brisbane and Southbank.

The Airtrain stations at the Domestic and International Airports are located directly outside the terminals, with trains departing every 15 minutes to the city during peak times.

You can book Airtrain tickets in advance so your ticket is waiting when you arrive at the airport train station. This means that you don't need to carry cash at the airport, and both single and return tickets can be purchased. Airtrain offers a discount on return tickets. For more information, visit www.airtrain.com.au/products_cbd.php.

Cars and Parking

Parking at QUT is limited to staff. The closest public car park is located under the South East Freeway; the entrance is off Gardens Point Road. This car park is a 4P area, meaning that vehicles may only park for a maximum of four hours in any one day.

Transport to GIW2008

The GIW 2008 conference is being held at the Marriott Resort, Surfers Paradise. The conference Welcoming Reception will commence at 5:30pm on Sunday 30th November.

Travel from Brisbane to the Gold Coast normally takes around 90 minutes. Trains to the Gold Coast depart regularly, and there are buses or taxis from Nerang Station to the conference hotel.

Enquiries with the main taxi providers suggest that we should be able to hire a maxi taxi capable of seating 8 people with a fair bit of luggage for around \$200 for the trip. This suggests to us an easy solution of around \$25-30 per person door to door. Can you please contact Jim Hogan (j.hogan@qut.edu.au) if you are interested and we should be able to organise a booking if there are sufficient people to make it work. As a comparison, the shuttle from the airport to the Marriott is listed at around \$39. So even if we were to get only 5-6 people per taxi it should work pretty well.

In order to get to the coast in time for the reception, we will need to leave promptly at or just before the scheduled end to the workshop. There are some issues here as the GIW reception is earlier than we had anticipated. The best way of handling this is to ask for preferences as follows. Can those interested please notify Jim Hogan and choose from the following statements:

- * I would like to catch a taxi to GIW at 4PM and can't go later
- * I would like to catch a taxi to GIW at 4PM, but 5PM is ok
- * I would like to catch a taxi to GIW at 5PM, and can't go earlier
- * I would like to catch a taxi to GIW at 5PM, but 4PM is
- * I am happy with either time.

Registration Desk

The workshop registration desk is the place to go if you have enquiries related to registration, travel support reimbursement claims or the local area.

Delegate Materials

Delegate materials consist of the following: the *Program, Information and Abstracts* booklet, your name tag, a pen and notepad, a workshop evaluation form, an HCSNet Survey form, and a copy of the HCSNet brochure. If you're missing something, ask at the registration desk for a replacement.

Travel Support

Workshop participants who have been awarded travel support to attend the workshop should either provide their travel and accommodation receipts to the organizers at the workshop, or post them to:

Chris Cassidy
HCSNet Administrative Coordinator
Department of Computing
Division of Information and Communication Sciences
Macquarie University NSW 2109
Email: ccassidy@ics.mq.edu.au

Refreshments and Meals

Morning tea, afternoon tea and lunch are provided for all registrants.

Recycling

At HCSNet, we take recycling seriously. Any delegate materials you do not wish to keep, including all paper materials, your name badge and pen, can be deposited in the marked Recycling Box located on the registration desk.

Evaluation Forms

You will find various evaluation and feedback forms inside your workshop package. These provide us with useful information that helps improve future events; please take the time to fill the various forms out, and deposit them at the registration desk when you leave.

Program – Sunday, 30th November 2008

08:45-09:00	<i>Opening Remarks</i>
09:00-10:00	<p>Keynote Presentation 1</p> <p>Dina Demner-Fushman, <i>Information retrieval and Natural Language Processing for Clinical Decision Support</i></p>
10:00-11:00	<p>Short Presentations 1</p> <p>Alexander Krumpholz, <i>Improving age-specific PubMed search</i></p> <p>Jon Patrick, <i>Question Answering from Clinical Information Systems</i></p> <p>Stephen Wan, <i>The Information Needs of Academic Researchers in the Wild: A Preliminary Study</i></p> <p>Wei Liu, <i>Ontology, Text Mining and a Proposed Application in BioInformatics</i></p>
11:00-11:30	Morning Break
11:30-12:30	<p>Student Presentations 1</p> <p>Pooyan Asgari, <i>Identifying for concepts in a noise prone environment: Looking up Obesity and its 15 co-morbidities In patient discharged summaries</i></p> <p>Andrew MacKinlay, <i>Information Extraction over Diverse Domains using Deep Parsing Techniques</i></p> <p>Juana Maria Ruiz-Martinez, <i>Learning non-taxonomic relationships in Biomedical Domain</i></p> <p>Mojtaba Sabbagh-Jafari, <i>Automated De-identification of the Clinical Documents</i></p>
12:30-13:30	Lunch Break

13:30-14:30	<p>Keynote Presentation 2</p> <p>Limsoon Wong, <i>Guilt by Association as a Search Principle</i></p>
14:30-15:30	<p>Short Presentations 2</p> <p>Peter Ansell, <i>Bio2RDF: Providing named entity based search with a common biological database naming scheme</i></p> <p>Peter Budd, <i>A Taxonomy of Terminology Server Desiderata</i></p> <p>Sarvnaz Karimi, <i>Ranked Search for Medical Systematic Reviews</i></p> <p>David Martinez, <i>Using Ranked Search Strategies in Combination with Supervised Text Classification</i></p>
15:30-16:00	<p>Afternoon Break</p>
16:00-17:00	<p>Student Presentations 2</p> <p>M. Asif Khawaja and Fang Chen, <i>Analysis of Bushfire Personnel's Speech Transcriptions for Linguistic Cues of Cognitive Load</i></p> <p>Stefan Pohl, <i>Query Processing in Biomedical Search</i></p> <p>Willy Yap, <i>Relation extraction for biomedical text</i></p> <p>Sun Xiaoxun, <i>Toward Privacy Preserving Microdata Publication</i></p>
17:00-17:15	<p>Closing</p>
	<p>Participants with Poster Only</p> <p>Stefan Schaefer, <i>Context Analysis in Clinical Environments using Natural Language Processing</i></p> <p>Yeondae Kwon, <i>A Proposal of a Ranking Method Based on Specificity of Biological Terms</i></p> <p>Anthony Nguyen, <i>Cancer Stage Classification from Free Text Medical Reports using Ontologies and Machine Learning</i></p> <p>Yefeng Wang, <i>Extracting and representing clinical knowledge using SNOMED CT</i></p>

Abstracts

Keynote Presentation 1

Dina Demner-Fushman

Information retrieval and Natural Language Processing for Clinical Decision Support

Information retrieval and natural language processing methods are instrumental in enhancing healthcare by providing clinicians, patients and other involved individuals with knowledge and person-specific information presented at appropriate times. Some of the specific challenges of Clinical Decision Support (CDS) are: using free-text information to drive CDS, representing clinical knowledge and CDS interventions in standardized formats, and leveraging the data available in Electronic Health Records (EHRs), which often contain narrative healthcare data.

This talk will present research on several aspects of the CDS challenges: developing strategies for automatic question and query formulation using information extracted from clinical narratives; finding adequate evidence and extracting answers to clinical and translational research questions; and retrieving images to illustrate evidence.

Dr. Dina Demner-Fushman is a Staff Scientist for the Communications Engineering Branch at the National Library of Medicine. She conducts research in clinical decision support, clinical question answering, use of natural language processing in information retrieval, human computer interaction aspects of information retrieval, and information retrieval in biomedical domain. Her interest in biomedical language processing stems from years of clinical practice (M.D. obtained from Kazan State Medical Institute in 1980) and clinical research (Doctorate (Ph.D.) in Medical Science earned from Moscow Medical and Stomatological Institute in 1989.) She earned her MS and PhD in Computer Science from the University of Maryland, College Park in 2003 and 2006, respectively.

Short Presentations: Session 1

Alexander Krumpholz

Improving age-specific PubMed search

PubMed allows age related search via a filter, which lets the user specify age-related Medical Subject Heading (MeSH) terms. The Medline abstracts often contain age-related terms that specify a narrower age-range than the Mesh terms available. For example the closest matching age-related MeSH term for a 104 year old person is "aged, 80 and over". Over 250 publications contain the term "centenarian" - a person between 100 and 109 years of age.

In order to retrieve publications that best match the case of a particular patient, we use the patient's age and map it to its related age-terms in analogy to fuzzy membership

functions. We use the degree of membership as term weights to rank the closer matching publications higher.

Jon Patrick

Question Answering from Clinical Information Systems

Clinical Information Systems typically have no search capability over the narrative notes that staff write about their patients. This creates the opportunity to invent useful language technologies that serve their various needs for both daily operational work and for their research. The notes collected in the course of the care of patient are important to the on-going care of the patient. Each day clinical staff need to ask questions of the patient record so that they can check the care administered by other staff. Resolving such questions has to run the gamut of poor spelling, poor grammar, technical terminology with variable morphology and phrasal structure, neologisms, and homespun abbreviations & mnemonics. Current research has concentrated on two related technologies, firstly producing information extraction for the routine process of daily care by a Smart Notes tool. Second, by developing a clinical data analytics language CliniDAL which is intended to allow expression of all questions that can be answered from the contents of the clinical database, and to compute the answers to those questions. It has the following features: Question formation using a controlled natural language; allowable usage of local sociolect terminology; retrieval of all components of the clinical record and search of text components in any combination; and formation and evaluation of statistical hypotheses using any retrievable content.

Stephen Wan

The Information Needs of Academic Researchers in the Wild: A Preliminary Study

With the near exponential growth in the available academic literature, staying up-to-date with the latest advances in research is a challenging task. In this paper, we describe the design and development of smarter tools to support researchers in navigating the literature and deciding whether something is relevant or not. Our work is directed by a preliminary user study which ascertains the information needs of academic researchers (primarily in the bio-medical science domain) as they read a publication. We are interested in kinds of information needs that users self-report when engaging in tasks that require further reading of bio-medical literature. We use the results of a preliminary user study to develop a next generation Aggregated Search system which includes query- and task-focused automatic text summarisation capabilities. In this paper, we present our analysis of the user study and outline the design implications for our research and development.

Wei Liu

Ontology, Text Mining and a Proposed Application in BioInformatics

Most current ontology management systems concentrate on detecting usage-driven changes and representing changes formally in order to maintain the consistency. In this work, we present a semi-automatic approach for measuring and visualising data-driven changes through ontology learning. Terms are first generated using text mining techniques using an ontology learning module, and then classified automatically into clusters. The clusters are then manually named, which is the only manual process in this system. Each cluster is considered as a sub-concept of the root concept, and thus one dimension of the feature space describing the root concept. The changes of terms in each cluster contributes to the change of the root concept. Using our system, Web documents are collected at different time periods and fed into the system to generate different versions of the same ontology for each time period. The paper presents several ways of visualising and analysing the changes. Initial experiments on online media data have demonstrated the promising capabilities of our system. BioInformatics is a much better domain in terms of detecting changes in concepts as it is highly dynamic. The purpose of attending this workshop is to learn more about the BioInformatics domain and see how our work can be applied.

Student Presentations: Session 1

Pooyan Asgari

Identifying for concepts in a noise prone environment: Looking up Obesity and its 15 co-morbidities In patient discharged summaries

Developing a tool for identifying clinical terms and concepts within a noise prone collection of clinical notes has its own requirements and issues. The specific nature of a noisy data collection raises at least two major issues. The first issue comes from a scattered matrix of evidence for a specific concept in which however has many common thereby confounding attributes with other concepts. Considering more features or patterns with the hope of covering more rare situations may lead to the absorption of more noise by the system and impact the identification of other major terms and concepts, and therefore the overall performance of the system. The second issue comes from the nature of the data collection and the necessary process for gathering evidence about existence/absent of a specific concept. Assuming 4 possible answers for a search concept namely Exists/Not Exists/Questionable/Unmentioned, biases the decision algorithm towards the two more frequent classes: Unmentioned and Exists which have unlike characteristics. The Unmentioned label has to be identified based on lack of evidence for a given search concept while an Exists label should only be assigned in presence of clear indication of a given concept. Adding more features to the feature list in the machine learner leads the system to a better and more confident classification for Exists class but at same time may lead to inaccurate results for the Unmentioned answer due to an increase in the level of the noise. We designed a customized system to address the common challenge of both issues, which is Noise reduction. Using a mixture of rules, different techniques in language processing algorithms, a decision tree classifier and some innovative solutions, a system was developed specifically for these types of noise prone corpora. We kept the number of features to monitor as low as possible based on the proposition that concepts are best defined in a few features and many features would add noise to the

classifier. In a second stage, an effective noise reduction algorithm which filtered suspicious noisy features was applied to the dataset to suppress possible noise. The primary goal was to evaluate a proposed approach for processing a collection of 724 discharge summaries with a noise prone nature. Evaluation has been done against given human performance as a gold standard with precision and recall of 0.969 and 0.969 respectively.

Andrew MacKinlay

Information Extraction over Diverse Domains using Deep Parsing Techniques

We present a preliminary system for evaluating semantic similarity of documents using machine learning techniques over diverse genres -- specifically online technical forums and biomedical abstracts. The gold-standard similarity judgements are in some cases hand-annotated but we also present an automated method for determining semantic similarity over the GENIA event annotation document set. The present system uses fairly naive feature vectors based on applying transformations on the bag-of-words statistics for the documents inspired by well-known metrics such as TF-IDF and skew divergence, but we plan to add more sophisticated features based on domain-specific named entity recognition as well as the outputs of shallow and deep parsing.

Juana Maria Ruiz-Martinez

Learning non-taxonomic relationships in Biomedical Domain

Semantic technologies are becoming more and more important in biomedical domains. Ontologies provide vocabulary standardization, allowing for reasoning mechanisms and supporting semantic interoperability issues between computer systems or between experts and computer systems, which are basic for tackling the problem of information overload in biomedicine. However, the construction and the update of biomedical ontologies is a problematic issue, since it is a time and resource consuming task. In this sense, Textual Knowledge Acquisition from electronically accessible bio-literature has become an important application area in order to create and manage biomedical ontologies automatically (ontology learning). However, an important drawback of most existing approaches is that they are only capable for extracting taxonomies or a very reduced set of relations. With the aim of overcoming these limitations a set of semantic relationships compatibles with OBO (Open Biomedical Ontologies) has been proposed. By means of information extraction techniques noun phrase candidates which can form part of a relationships are identified. Verbs, which are considered in this approach the key to identify non-taxonomic relationships between concepts, are also identified. This can be combined with a MCRDR (Multiple Classification Ripple Down Rules) module by which new relationships are proposed automatically according to the stored relationships. This module could be a semi-automatic aid of validation of the acquired relationships and candidates by an expert.

Mojtaba Sabbagh-Jafari

Automated De-identification of the Clinical Documents

Removing protected health information (PHI) from clinical documents is a required task and should be done before clinical documents can be used for research or other text processing systems. If this process is performed manually, it is tedious and prone to error, therefore computer support is valuable.

The purpose of this system is to find PHI objects in the free clinical texts and replace them with proper surrogate information, in order to retain their interpretability and usefulness for research. To develop the de-identification approach, the system uses several gazetteers, regular expressions for pattern matching, heuristic rules and local context features. These lists of words are proper names, medical terminology, common English words and list of locations used to find PHIs. In many cases there is ambiguity between PHI and non-PHI elements as well as some foreign names or misspelt words which cannot be recognized. In these cases local context features and heuristic rules help this system to classify correctly. POS and syntactic bigram as local context features are extracted from words which are located in a window of one or two word from the target word.

Keynote Presentation 2

Limsoon Wong

Guilt by Association as a Search Principle

The exploitation of fundamental invariants is among the most elegant solutions to many computational problems in a wide variety of domains. One of the more powerful approaches to exploit invariants is the principle of "guilt by association". In particular, the principle of guilt by association is the foundation of remote homolog detection, protein function prediction, disease subtype diagnosis, treatment plan prognosis, and other challenges in computational biology. The principle suggests that two entities are in a specific relationship if they exhibit invariant properties underlying that relationship. For example, a protein is predicted to have a particular biological function if it exhibits the underlying invariant properties of that functional group --- viz., guilty by association to other members of that functional group through the shared invariant properties.

In my talk, I plan to present several facets of guilt by association in the computational prediction of protein function and draw parallels of these facets in information retrieval. Specifically, I plan to touch on the following facets: (a) the issue of chance associations; (b) novel generalizable forms of association; (c) fusion of multiple heterogeneous sources of evidence; (d) the dichotomy of knowing to a high degree of reliability that two entities are in some relationship and yet not knowing what that relationship is. I hope this talk will be, for the informational retrieval community, a window to the opportunities in computational biology that may benefit from the depth and variety of solutions information retrieval has to offer.

Limsoon Wong is Professor and Head of Computer Science and Professor of Pathology at the National University of Singapore. He currently works mostly on knowledge discovery technologies and is especially interested in their application to biomedicine. He has written about 150 research papers, a few of which are among the best cited of their respective fields. He serves on the editorial boards of Journal of Bioinformatics and Computational Biology (ICP), Bioinformatics (OUP), and Drug Discovery Today (Elsevier). He is chairman of Molecular Connections and scientific advisor to CellSafe International. Limsoon received his BSc(Eng) from Imperial College London and his PhD from University of Pennsylvania.

Short Presentations: Session 2

Peter Ansell

Bio2RDF: Providing named entity based search with a common biological database naming scheme

The Bio2RDF project provides effective cross-database biomedical search functionality through the use of a common representation format, RDF, and common query mechanism, HTTP. Although we provide mostly biological databases, we also provide dbpedia, the RDF form of Wikipedia, and, in the future, WordNet, to provide for complete throughput from vocabularies to biological databases. We focus on the linked database aspects, although basic text-searches are supported. Text-mining on data sources which are included in Bio2RDF, such as PubMed, can be used together with current knowledge about the links between biological databases to both enrich the text-mining process and to make the text search results applicable in a larger context. In addition, the results of dynamic searches can be stored and tagged, to be included as part of a dynamic data source for future Bio2RDF users.

Peter Budd

A Taxonomy of Terminology Server Desiderata

The use of terminology servers in the health domain will lead to the standardisation of the organisation of medical thesauri, terminologies, ontologies and classifications (TTOCs). By using mappings between TTOCs, users will be able to search the semantic content of medical files using their own TTOC, and still have the meaning of their search terms preserved across TTOCs and by implication across the clinical information systems that use the TTOCs.

The actual use of terminology servers in the field however is sporadic at best and the functionality is usually implemented within an individual clinical information system, leading to inconsistency in record keeping and data representations. This research canvases the literature from more than a decade of research into the problem. Our research defines the role of a terminology server and details the desiderata for the use of a terminology server. The limit of these desiderata are discussed and a functional taxonomy is produced that specifies the features a terminology server must possess to provide for indexation, storage and retrieval of medical concepts based on semantic rather than lexical features. A prototype implementation of a terminology server built on these desiderata has been produced by the Health Information Technology

Research Laboratory at the University of Sydney and currently serves numerous applications, including; the GCIMS project, a generic ontology viewer, a ward round information system, a clinical data analytics engine, and an automated medical concept identification engine for use on text.

Sarvnaz Karimi

Ranked Search for Medical Systematic Reviews

Searching and selecting articles to be included in systematic reviews is a real challenge for healthcare agencies responsible for publishing these reviews. The current practice of manually reviewing all papers returned by complex hand-crafted Boolean queries is human labour-intensive and difficult to maintain. We demonstrate a searching system that takes advantage of ranked queries to assist in the retrieval of relevant articles, and to restrict results to higher-quality documents.

David Martinez

Using Ranked Search Strategies in Combination with Supervised Text Classification

One of the goals of the project BioTALA (Biomedical Text And Language Applications), is to address the search needs for building systematic clinical reviews for medicine, an increasingly growing area that can benefit the way medical treatments are applied throughout the world. This problem is especially difficult to solve with standard search strategies, because of the very high recall required for the medical research questions. This results in complex Boolean queries that are time-consuming to produce and difficult to maintain. Our approach is to rely on ranked search strategies in combination with supervised text classification. We present our initial results over systematic reviews from the Agency for Healthcare Research and Quality (AHRQ), showing that our system can significantly contribute to the state of the art.

Student Presentations: Session 2

M. Asif Khawaja and Fang Chen

Analysis of Bushfire Personnel's Speech Transcriptions for Linguistic Cues of Cognitive Load

In complex, time-critical and data-intense situations users of a system can experience extremely high cognitive demands imposed on their limited working memory which can interfere with their ability to perform and complete the task at hand efficiently. Intelligent adaptive user interface systems which are aware of the users' current level of cognitive load could in fact, alleviate these problems by implementing strategies to adjust the behavior, support, user interaction material, and resources needed as per users' current cognitive burden to help them complete the task effectively.

Our study presents a speech content analysis approach to the measurement of cognitive load which employs users' linguistic features of speech to determine their

experienced level of cognitive load. We present the detailed analyses of several linguistic features extracted from the live speech data collected from the subjects, the members of a bushfire incident management team, involved in highly time-critical and data-intense bushfire management tasks around Australia. We discuss the results for nine selected linguistic features showing significant differences between the speech from the low load tasks and the high load tasks.

Despite the fact that the study focuses on bushfire operators' speech transcriptions, we believe that the proposed method can be used with any clinical or medical transcriptions of patients' speech for the purpose of cognitive load measurement of those patients in order to help the clinicians and/or doctors better understand the mental state of the patients.

Stefan Pohl

Query Processing in Biomedical Search

Query processing becomes costly when large collections are involved, or long, and complex queries are to be answered. Biomedical search has to deal with both, because high recall requirements lead to long, expanded queries and medical publication archives are ever growing. Recent trends in computer architecture are ambivalent: In 64-bit architectures, high amounts of memory become available so that more data can readily be held in main-memory. This shifts query processing costs from being dominated by disk to memory accesses and computation. At the same time, processors stopped becoming faster. Instead, more of them are suddenly available, and new ways have to be found to use them efficiently in order to reduce query processing times.

Willy Yap

Relation extraction for biomedical text

Relation extraction is a sub-task of Information Extraction (IE) that is concerned with extracting semantic relations between word pairs based on corpus data. Past work on relation extraction has concentrated on creating a small set of patterns that are good indicators of whether a given word pair contains a semantic relation. In recent years, there has been work on using machine learning to automatically learn these patterns from English corpus text. We build on this research in applying a generic relation extraction algorithm to the biomedical domain. However, instead of extracting word pairs with semantic relations as already been done to English corpus, we are interested in extracting the interaction between proteins in biomedical documents.

Sun Xiaoxun

Toward Privacy Preserving Microdata Publication

High quality and useful knowledge is to be found in the integrated data from various organizations, and the discovered knowledge is essential for building intelligent

systems such as business analysis and health surveillance. However, concern about breaching privacy is a major obstacle of this process. This project aims to develop new efficient and effective techniques for privacy protection in data sharing and data mining by combining techniques in data mining and security research. We focus primarily on notions of anonymity that are defined with respect to individual identity, or with respect to the value of a sensitive attribute. Our goal is to propose a variety of techniques to anonymize original data sets, while preserving the utility of the input data. We adopt extensive evaluations to indicate that it is possible to distribute high-quality data that respects several meaningful notions of privacy. Further, it is possible to do this efficiently for large transactional data sets. The developed cutting edge techniques will advance and facilitate data mining within many organizations and businesses and lead to the better utilization of information.

Posters

Stefan Schaefer

Context Analysis in Clinical Environments using Natural Language Processing

Medical records comprise of a variety of detailed documents written in natural language such as clinical case studies, patient profiles and treatment reports. Extracting the clinical information from these documents is crucial as this makes clinical data fit for automated processing. This poster introduces a new concept of clinical contexts and proposes a new approach to context analysis in clinical environments which allows information retrieval from clinical documents.

Yeondae Kwon

A Proposal of a Ranking Method Based on Specificity of Biological Terms

There are a lot of interests on extracting associations between diseases and genes from literatures such as MEDLINE abstracts. For a given disease, a search engine returns a ranked list of candidate genes according to some criterion. In this research, we propose a ranking algorithm that focuses on specificity to a particular disease. A specificity-based ranking method should rank a gene that causes a given disease but does not cause other diseases at the top. This is important for drug developments because users can find relevant genes that do not have side effects quickly. We describe a specificity-based baseline algorithm using term dictionaries and co-occurrence data of terms in MEDLINE abstracts and discuss future directions.

Anthony Nguyen

Cancer Stage Classification from Free Text Medical Reports using Ontologies and Machine Learning

Cancer staging is the process of classifying the extent of the primary tumour and metastatic spread to other parts of the body using the TNM (Tumour-Nodes-Metastasis) standard. This process is conducted through a multidisciplinary team

(MDT) conference, which is time and resource-intensive. As a result stage data is not routinely collected. Tools to retrospectively collect stage data are therefore needed to fill in gaps from their cancer stage collection efforts.

This poster presents the use of SNOMED CT or UMLS SPECIALIST Lexicon ontologies and Support Vector Machines (SVM) for the automatic classification of cancer stages from free text medical reports. Preliminary experiments on the classification of a clinical M (Metastasis) stage for lung cancer patients by analysing their free text radiology reports have achieved promising results with sensitivity-specificity (SS) break-even points of approximately 0.89, area under the SS curves of 0.95, and precisions of approximately 0.70.

Yefeng Wang

Extracting and representing clinical knowledge using SNOMED CT

Automatic indexing of clinical concepts in free text patient records using a standard medical terminology will enhance semantic retrieval, which can then be used for important applications such as decision support and disease outbreak detection. SNOMED CT is a rich terminology that provides standardisation of knowledge and language in the clinical domain. Two important challenges are identification of the concepts in clinical reports and then using the identified concepts to construct an integrated representation of the patient case. Although most clinical words found in the patient notes are present in the terminology, the rich set of relationships between the words and concepts cannot be fully represented. Lexical and concept verification is error prone due to the variance of the clinical language used in different departments in hospitals and the ungrammatical nature of the narrative reports. To integrate the recognised concepts extensions need to be made to the ontology or it needs to be placed in a wider ontological model to fully represent all matters relevant to the patient case.

This research aims to address the concept extraction and concept representation issues, by classifying medical concepts into the 17 SNOMED CT semantic categories, and representing their relationships using SNOMED CT 60+ relationship categories. The experiments will be conducted on a subset of a 44 million token Intensive Care corpus from the Royal Prince Alfred Hospital, Sydney. Through the classification, an extended ontology will be built to represent the ICU terms for use in data retrieval activities.

Workshop Contact List

Surname	First Name	Email Address	Institution
Ansell	Peter	p.ansell@qut.edu.au	Queensland University of Technology
Anthony	Stephen	s.anthony@unsw.edu.au	University of New South Wales
Asgari	Pooyan	pooyan@it.usyd.edu.au	University of Sydney
Budd	Peter	pbud3427@mail.usyd.edu.au	University of Sydney
Chun	Woo-Hoo	chun@dbcls.rois.ac.jp	Database Centre for Life Science, Japan
Demner-Fushman	Dina	ddemner@mail.nih.com	National Library of Medicine
Hogan	James	jamesmichaelhogan@gmail.com	Queensland University of Technology
Jarvelin	Kal	Kalervo.jarvelin@uta.fi	University Tampere, Finland
Khawaja	M. Asif	asif.khawaja@nicta.com.au	NICTA, University of NSW
Krumpholz	Alexander	alexander@krumpholz.com	ANU/CSIRO
Kwon	Yeondae	yekwon@lab.nig.ac.jp	National Institute of Genetics
Liu	Wei	wei@csse.uwa.edu.au	University of Western Sydney
MacKinlay	Andrew	amack@csse.unimelb.edu.au	University of Melbourne
Martinez	David	davidm@csse.unimelb.edu.au	University of Melbourne
Molla- Aliod	Diego	diego@comp.mq.edu.au	Macquarie University
Nguyen	Anthony	Anthony.Nguyen@csiro.au	CSIRO
Patrick	Jon	jonpat@it.usyd.edu.au	University of Sydney
Pohl	Stefan	spohl@csse.unimelb.edu.au	NICTA, The University of Melbourne
Ruiz Matínez	Juana María	jmruymar@um.es	University of Murcia
Sabbagh-Jafari	Mojtaba	ssab8677@mail.usyd.edu.au	University of Sydney
Sarvnaz	Karimi	skarimi@unimelb.edu.au	University of Melbourne
Schaefer	Stefan	stefans@it.usyd.edu.au	University of Sydney
Shimizu	Shogo	shimizu-syogo@aait.ac.jp	Advanced Institute of Industrial Technology
Smith	Timothy	tim@variome.org	University of Melbourne
Tann	Aaron	aaron@itee.uq.edu.au	University of Queensland
Wan	Stephen	stephen.wan@csiro.au	CSIRO
Wang	Yefeng	ywang1@it.usyd.edu.au	University of Sydney
Wise	Michael	Michael.Wise@uwa.edu.au	University of Western Australia
Wong	Limsoon	wongls@comp.nus.edu.sg	National University of Singapore
Xiaoxun	Sun	sunx@usq.edu.au	University of Southern Queensland
Yap	Willy	willy@csse.unimelb.edu.au	University of Sydney