The 2nd International Symposium on Languages in Biology and Medicine

6th – 7th December 2007

Creation Theatrette, Matrix, Biopolis

LBM2007

Organized with support from

WELCOME

On behalf of the committees of the symposium we welcome you to Singapore. Being a dynamic city rich in contrast and colour, you will find a harmonious blend of culture, culinary delights, arts and architecture.[1] This is an environment of traditional and modern lifestyles, fusion cuisine and innovative thinking. As such Singapore is a suitable venue for reflection on synergies between established and emerging technologies.

Language is one of our most fundamental technologies that must remain consistent and at the same time evolve to meet contemporary challenges. In biology and medicine, the importance of languages to represent knowledge, communicate and query information is immense. Likewise auxiliary tasks such as translation, summarization and information extraction play important roles supporting scientific research.

Incumbent technologies that discover, read and process language are continually stretched by the vigorous demands of bio-medical scientists and there is an ongoing need and incentive for language techniques to evolve. Despite this, the distinct communities involved in language processing rarely borrow from one another or look over the fence to see what other approaches are in use. And yet synergistic interactions across methodological disciplines and across different topics are frequently the harbingers of revolutionary technologies. In this context, it is imperative that we adopt diversification, more lateral and more creative interaction between language professionals.

The 2nd International Symposium on Languages in Biology and Medicine (LBM2007) seeks to provide a renewed opportunity for interaction between language professionals with different methodological backgrounds. LBM was established in 2005 and the remit of this event remains highly relevant today. The programme for LBM2007 comprises of (a) 3 invited keynote lectures by Olivier Bodenreider, Sophia Ananiadou, and Patrick Lambrix; (b) a panel discussion chaired by Junichi Tsujii on emerging synergies of biomedical language and biomedical knowledge; and (c) 5 sessions of oral presentations selected from 47 submitted papers.

We wish to express our deep appreciation to the programme committee members and the additional reviewers who shared their valuable time and formidable expertise in support of the LBM review process. We also wish to express our gratitude to our supporting organizations: the Institute for Infocomm Research (A*STAR), the Bioinformatics Institute (A*STAR); School of Computing and the OLS Bioinformatics Programme of the National University of Singapore; KAIST, Korea and the BK21 Project, Korea. Lastly, we look forward to seeing you again in Korea for LBM2009. Please stay tuned for forthcoming announcements on LBM2009.

Rajaraman Kanagasabai
Local Organizing Chair

Christopher J. O. Baker and Su Jian
Programme Committee Co-chairs

Jong C. Park and Limsoon Wong
General Chairs

[1] http://www.visitsingapore.com

LBM2007 ORGANIZATION

Steering Committee

See-Kiong Ng, I^2R, Singapore
Jong C. Park, KAIST, Korea
Limsoon Wong, National Univ. of Singapore

General Chairs

Jong C. Park, KAIST, Korea
Limsoon Wong, National Univ. of Singapore

Programme Committee Chairs

Christopher J. O. Baker, I^2R, Singapore
Su Jian, I^2R, Singapore

Local Organizing Chair

Rajaraman Kanagasabai, I^2R, Singapore

Programme Committee

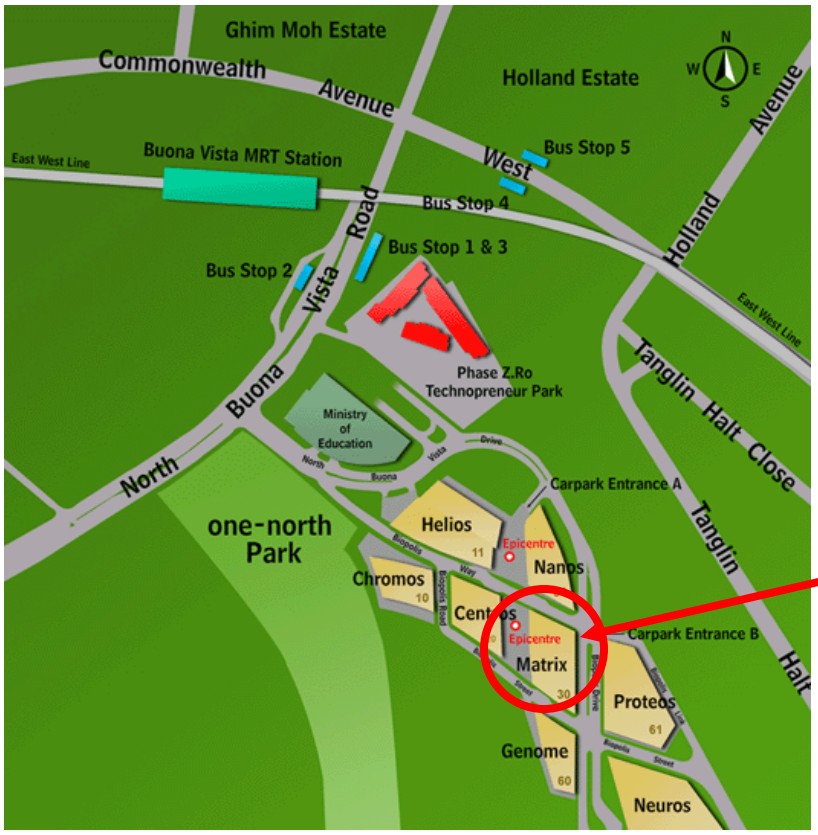
Sophia Ananiadou, Univ. of Manchester, UK
Vlad Bajic, Univ. of Western Cape, South Africa
Chitta Baral, Arizona State Univ., USA
Christian Blaschke, Bioalma, Spain
Anita Burgun, Univ. de Rennes, France
Werner Ceusters, Univ. at Buffalo NY, USA
Kevin B. Cohen, Univ. of Colorado Health Sciences Center, USA
Nigel Collier, NII, Japan
Mark Craven, Univ. of Wisconsin, USA
Rebholz Dietrich, EMBL-EBI, UK
Michel Dumontier, Carleton Univ., Canada
Julian Gough, Univ. of Bristol, UK
Volker Haarslev, Concordia Univ., Canada
Udo Hahn, Jena Univ., Germany
Lynette Hirschman, MITRE, USA
Graeme Hirst, Univ. of Toronto, Canada
Ewan Klein, Edinburgh Univ., UK
Satoshi Kobayashi, Univ. of Electro-Communications, Japan
Michael Krauthammer, Yale Univ. School of Medicine, USA
Patrick Lambrix, Linköping Univ., Sweden
Liu, Hong Fang, Georgetown Univ. Medical Center, USA
Yves Lussier, Univ. of Chicago, USA
Erik van Mulligen, Erasmus MC, Netherlands
Jinah Park, Information & Communications Univ., Korea
Tom Rindflesch, NLM, USA
Jasmin Saric, Boehringer Ingelheim Pharma GmbH & Co. KG, Germany
Neil Sarkar, Woods Hole, USA
Stefan Schulz, Freiburg Univ. Hospital, Germany
Donia Scott, Open Univ., UK
Hagit Shatkay, Queen's Univ., Canada
Margaret Anne Storey, Univ. of Victoria, Canada
Jun'ichi Tsujii, Univ. of Tokyo, Japan & Univ. of Manchester, UK
Alfonso Valencia, CNIO, Spain
W. John Wilbur, NIH, USA
Rene Witte, Univ. of Karlsruhe, Germany
Hong Yu, Univ. of Wisconsin-Milwaukee, USA
Pierre Zweigenbaum, LIMSI-CNRS, France

Additional Reviewers

Jörg Hakenberg, Tech Univ. Dresden, Germany
Catalina Hallett, Open Univ., UK
Minlie Huang, Tsinghua Univ., China
Jin-Dong Kim, Univ. of Tokyo, Japan
Martin Krallinger, CNIO, Spain
Man Lan, National Univ. of Singapore, Singapore
Robert Leaman, Arizona State Univ., USA
Goran Nenadic, Univ. of Manchester, UK
Fabio Rinaldi, Univ. of Zurich, Switzerland
Patrick Ruch, Hospital Univ. Geneva, Switzerland
Yutaka Sasaki, Univ. of Manchester, UK
Arash Shaban-Nejad, Concordia Univ., Canada
Luis Tari, Arizona State Univ., USA
Manabu Torii, Georgetown Univ., USA
Richard Tzong-Han Tsai, Yuan Ze Univ., Taiwan
Hua Xu, Columbia Univ., USA,
Wai Keong Stanley Yong, I^2R, Singapore
Yi Tao Zhang, Univ. of Sydney, Australia
Barry Haddow, Univ. of Edinburgh, UK

VENUE AND LOCATION

LBM2007 will be held in the Creation Theatrette at Matrix, Biopolis, Singapore on Thursday 6th – Friday 7th December 2007. **Address of Biopolis:** Matrix, 30 Biopolis Street, Singapore 138671. **Getting to Biopolis: By MRT:** Alight at Buona Vista MRT station and walk 8 minutes; or take the free Biopolis Shuttle Bus service. **By Taxi**: Instruct driver to take you to "Biopolis" or to "MOE" (MOE = Ministry of Education, about 2 minutes walk to Biopolis).



LBM 2007 conference venue

TRANSPORTATION LOGISTICS

Bus transfer between the conference hotel (Swissotel) and conference venue (Biopolis) has been arranged. The conference bus will leave from the Swissotel Lobby each morning at 7:45am. I**f you wish to take the conference bus, please be sure to assemble at the Swissotel Lobby by 7.45am and look out for the bus. The hotel does not allow the bus to wait beyond 5 minutes.**

On 6th December, we have arranged bus transfer from the conference venue (Biopolis) to the banquet venue (Made in China @ Haw Par Villa). We have also arranged bus transfer from the banquet venue to the conference hotel (Swissotel) after the banquet. (If you wish to go to the banquet venue yourself, the address is "Made in China, Hua Song Museum, Haw Par Villa, Pasir Panjang Road". The keyword to tell taxi cab driver is "**Haw Par Villa**".)

On 7th December, we have arranged bus transfer from the conference venue (Biopolis) to the conference hotel (Swissotel) at 5:30pm.

LBM 2007 PROGRAM

Thursday, 6 Dec 2007

9:00 – 9:15 Welcome address	
9:15 –10:15	Keynote 1 (Chaired by Jong C. Park) **Terminological systems in biomedicine: From terminology integration to information integration** Olivier Bodenreider
10:15 – 10:45 Tea Break	
10:45 – 12:15 Session 1A: Terminology and Named Entity, chaired by Nigel Collier	
10:45 – 11:15	**Normalizing biomedical terms by minimizing ambiguity and variability** Yoshimasa Tsuruoka, John McNaught, Sophia Ananiadou
11:15 – 11:45	**Assessment of diseases named entity recognition on a corpus of annotated sentences** Antonio Jimeno, Ernesto Jimenez-Ruiz, Vivian Lee, Sylvain Gaudan, Rafael Berlanga-Llavori, Dietrich Rebholz-Schuhmann
11:45-12:00	**Analysis and enhancement of conditional random fields gene mention taggers in BioCreative II Challenge Evaluation** Yu-Ming Chang, Cheng-Ju Kuo, Han-Shen Huang, Yu-Shi Lin, Chun-Nan Hsu
12:00-1:00 Lunch Break	
1:00-2:00	Keynote 2 (Chaired by Su Jian) **Delivering text mining services for the biosciences** Sophia Ananiadou
2:00-3:00 Session 1B: Text Classification (I), chaired by Hongfang Liu	
2:00-2:30	**Exploiting and integrating rich features for biological literature classification** Hongning Wang, Minlie Huang, Shilin Ding, Xiaoyan Zhu
2:30-2:45	**The integration of multiple feature representations for protein-protein interaction classification task** Man Lan and Chew Lim Tan
2:45-3:00	**Protein-protein interaction abstract identification with contextual bag of words** Richard Tzong-Han Tsai, Hsieh-Chuan Hung, Hong-Jie Dai and Yi-Wen Lin
3:00-3:30 Tea Break	
3:30-5:00 Session 1C: Text Mining, chaired by Dietrich Rebholz-Schuhmann	
3:30-4:00	**New challenges for text mining: Mapping between text and manually curated pathways** Kanae Oda, Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, Jun'ichi Tsujii
4:00-4:30	**A comparative analysis of five protein-protein interaction corpora**

	Sampo Pyysalo, Antti Airola, Juho Heimonen, Jari Björne, Filip Ginter ,Tapio Salakoski
4:30-4:45	**Syntactic features for protein-protein interaction extraction** Rune Saetre, Kenji Sagae and Jun'ichi Tsujii
4:45-5:00	**Recognition of multisentence n-ary subcellular localization mentions in biomedical abstracts** Gabor Melli, Martin Ester, Anoop Sarkar
19:00-10:00 Conference Banquet @ Made In China Museum, Haw Par Villa Bus leaves from Biopolis to Haw Par Villa at 5:20 pm. Delegates are invited to tour the museum before banquet	

Friday, 7 Dec 2007

9:00-10:00	Keynote 3 (Chaired by Christopher Baker) **Aligning biomedical ontologies** Patrick Lambrix
10:00-10:30 Tea Break	
10:30-12:15 Session 2A: Ontologies and Logic, chaired by Mark Schreiber	
10:30-11:00	**Monitoring the evolutionary aspect of the Gene Ontology to enhance predictability and usability** Jong C. Park, Tak-eun Kim, Jinah Park
11:00-11:30	**Structuring an event ontology for disease outbreak detection** Ai Kawazoe, Hutchatai Chanlekha, Mika Shigematsu, Nigel Collier
11:30-12:00	**Combining Gene Ontology and argumentative features for automatic geneRIF extraction** Julien Gobeill, Patrick Ruch
12:00-12:15	**Decentralised clinical guidelines modelling with lightweight coordination calculus** Bo Hu, Srinandan Dasmahapatra, David Robertson, Paul Lewis
12:15-1:30 Lunch Break	
1:30-2:45 Session 2B: Text Classification (II), chaired by Minlie Huang	
1:30-2:00	**Automatic construction of rule-based ICD-9-CM coding systems** Richard Farkas, Gyorgy Szarvas
2:00-2:30	**Identification of transcription factor contexts in literature using machine learning approaches** Hui Yang, Goran Nenadic
2:30-2:45	**Classifier ensemble for biomedical document retrieval** Manabu Torii, Hongfang Liu
2:45-3:30 Tea Break	
3:30-5:00 Panel Discussion	**Biomedical language and biomedical knowledge: Emerging synergies** Panel: Jun-ichi Tsujii (chair), Olivier Bodenreider, Sophia Ananiadou, Patrick Lambrix
5:00-5:15 Closing address	

INVITED KEYNOTES

Keynote 1

Terminological systems in biomedicine: From terminology integration to information integration

Olivier Bodenreider
National Library of Medicine, USA

Abstract: The use of different names and codes for the same entity in different terminologies has long been identified as a barrier to integrating biomedical information sources. By integrating terms from many disparate biomedical sources, terminological systems help bridge across terminologies. And because terms and codes constitute an entry point into information sources, terminology integration represents a key element to information integration. As an example of terminological system, we briefly introduce the Unified Medical Language System (UMLS), a resource integrating more than one hundred biomedical terminologies. We show how integrative resources such as the UMLS can play a role to bridge across namespaces in the Semantic Web. Translational research requires the integration of information between the "bench" (basic research) and the "bedside" (clinical practice). Using the example of oncology, we show practical issues in the integration of cancer terminologies.

Biography: Olivier Bodenreider is a Staff Scientist in the Cognitive Science Branch of the Lister Hill National Center for Biomedical Communications at the U.S. National Library of Medicine. His research interests include terminology, knowledge representation and ontology in the biomedical domain, both from a theoretical perspective and in their application to natural language understanding, reasoning, information visualization and integration. Dr. Bodenreider is a Fellow of the American College of Medical Informatics. He received a M.D. degree from the University of Strasbourg, France in 1990 and a Ph.D. in Medical Informatics from the University of Nancy, France in 1993. Before joining NLM in 1996, he was an assistant professor for Biostatistics and Medical Informatics at the University of Nancy, France, Medical School.

Keynote 2

Delivering text mining services for the biosciences

Sophia Ananiadou
University of Manchester, UK

Abstract: The UK National Centre for Text Mining is providing text mining services for the Biosciences. These services range from terminology management, to advanced information retrieval, semantic querying, relation mining and are customised for different users. NaCTeM provides users with a coherent interoperable set of core text mining tools through adoption of UIMA. The Centre is also building bio-resources and annotated corpora. Last, NaCTeM's current and future benefits to the UK research community and its future vision will be presented.

Biography: Sophia Ananiadou is Reader in Text Mining in the School of Computer Science at the University of Manchester and Deputy Director of the National Centre for Text Mining (NaCTeM). She is the main developer of the terminology management services provided by NaCTeM. Her current research includes building bio-resources, advanced IR systems, the text-mining-based visualisation of the provenance of biochemical networks and text mining for systematic reviews. She is recipient of the 2004 Daiwa Adrian prize for her research in Knowledge Mining for Biology, and in 2006 of the IBM UIMA innovation award for her work on the interoperability of text-mining tools.

Keynote 3

Aligning biomedical ontologies

Patrick Lambrix
Linköpings Universitet, Sweden

Abstract: The use of ontologies is a key technology for the Semantic Web and in particular in the biomedical field many ontologies have already been developed. Many of these ontologies, however, contain overlapping information and to make full use of the advantages of ontologies it is important to know the inter-ontology relationships, i.e. we need to align the ontologies. Knowledge of these alignments would lead to improvements in search, integration and analysis of biomedical data. It has been realized that this is a major issue and some organizations have started to deal with it. In this talk we give an overview of techniques for ontology alignment with a focus on approaches that compute similarity values between terms in the different ontologies. Further, we discuss the results of evaluations of these techniques using biomedical ontologies. Finally, we discuss the recent development of approaches for providing recommendations of ontology alignment strategies for a given alignment task.

Biography: Patrick Lambrix is a professor of bioinformatics/knowledge engineering at Linköpings universitet, Sweden. His current research interests relate to the areas of semantic web, ontologies, databases and bioinformatics and he leads projects on alignment of biomedical ontologies and grouping of biological data. He received MSc degrees in mathematics (1988) and computer science (1990) from KU Leuven, Belgium and a PhD degree in computer science from Linköpings universitet (1996).

SESSION 1A

TERMINOLOGY AND NAMED ENTITY

Normalizing biomedical terms by minimizing ambiguity and variability

Yoshimasa Tsuruoka, John McNaught and Sophia Ananiadou

Background: One of the difficulties in mapping biomedical named entities, e.g. genes, proteins, chemicals and diseases, to their concept identifiers stems from the potential variability of the terms. Soft string matching is a possible solution to the problem, but its inherent heavy computational cost discourages its use when the dictionaries are large or when real time processing is required. A less computationally demanding approach is to normalize the terms by using heuristic rules, which enables us to look up a dictionary in a constant time regardless of its size. The development of good heuristic rules, however, requires extensive knowledge of the terminology in question and thus is the bottleneck of the normalization approach.

Results: We present a novel framework for discovering a list of normalization rules from a dictionary in a fully automated manner. The rules are discovered in such a way that they minimize the ambiguity and variability of the terms in the dictionary. We evaluated our algorithm using two large dictionaries: a human gene/protein name dictionary built from BioThesaurus and a disease name dictionary built from UMLS.

Conclusions: The experimental results showed that automatically discovered rules can perform comparably to carefully crafted heuristic rules in term mapping tasks, and the computational overhead of rule application is small enough that a very fast implementation is possible. This work will help improve the performance of term-concept mapping tasks in biomedical information extraction especially when good normalization heuristics for the target terminology are not fully known.

Assessment of diseases named entity recognition on a corpus of annotated sentences

Antonio Jimeno, Ernesto Jimenez-Ruiz, Vivian Lee, Sylvain Gaudan, Rafael Berlanga-Llavori and Dietrich Rebholz-Schuhmann

Background: In recent years, the recognition of semantic types from the biomedical scientific literature has been focused on named entities like protein and gene names (PGNs) and gene ontology terms (GO terms). Other semantic types like diseases have not received the same level of attention. Different solutions have been proposed to identify disease named entities in the scientific literature. While matching the terminology with language patterns suffers from low recall (e.g., Whatizit) other solutions make use of morpho-syntactic features to better cover the full scope of terminological variability (e.g., MetaMap). Currently, MetaMap that is provided from the National Library of Medicine (NLM) is the state of the art solution for the annotation of concepts from UMLS (Unified Medical Language System) in the literature. Nonetheless, its performance has not yet been assessed on an annotated corpus. In addition, little effort has been invested so far to generate an annotated dataset that links disease entities in text to disease entries in a database, thesaurus or ontology and that could serve as a gold standard to benchmark text mining solutions.

Results: As part of our research work, we have taken a corpus that has been delivered in the past for the identification of associations of genes to diseases based on the UMLS Metathesaurus and we have reprocessed and re-annotated the corpus. We have gathered annotations for disease entities from two curators, analyzed their disagreement (0.51 in the kappa-statistic) and composed a single annotated corpus for public use. Thereafter, three solutions for disease named entity recognition including MetaMap have been applied to the corpus to automatically annotate it with UMLS Metathesaurus concepts. The resulting annotations have been benchmarked to compare their performance.

Conclusions: The annotated corpus is publicly available and can serve as a benchmark to other systems. In addition, we found that dictionary look-up already provides competitive results indicating that the use of disease terminology is highly standardized throughout the terminologies and the literature. MetaMap generates precise results at the expense of insufficient recall while our statistical method obtains better recall at a lower precision rate. Even better results in terms of precision are achieved by combining at least two of the three methods leading, but this approach again lowers recall. Altogether, our analysis gives a better understanding of the complexity of disease annotations in the literature. MetaMap and the dictionary based approach are available through the Whatizit web service infrastructure.

Analysis and enhancement of conditional random fields gene mention taggers in BioCreative II Challenge Evaluation

Yu-Ming Chang, Cheng-Ju Kuo, Han-Shen Huang, Yu-Shi Lin and Chun-Nan Hsu

Background: Tagging gene and gene product mentions in scientific text is an important initial step of literature mining. In BioCreative 2 challenge, the conditional random fields model (CRF) was the most prevailing method in the gene mention task. In this paper, we analyze two best performing CRF-based systems in BioCreative 2. We examine their key claims and propose enhancement based on the analysis results. Their key claims include that a rich feature set is required, post-processing is necessary, backward parsing is superior, and integrating divergent and high performance models always improve the performance.

Results: We implemented their systems in MALLET as specified in their report and in CRF++, a different CRF package, to empirically analyze their claims. We found that their feature set is effective for models trained by MALLET, but a smaller set works better for those by CRF++, and the selection of best features depends on the CRF package. We confirmed the effectiveness of pairing parentheses as a post processing step. We found that backward parsing is not always superior to forward parsing. The benefit of applying bidirectional parsing is the creation of a wider variety of complementary models. We created four models by applying MALLET, CRF++ and YamCha and compared their individual performance and the performance of their unions. Many of the unions outperform the best system in BioCreative 2, confirming that integrating divergent models improves the performance. We elaborated the notion of divergent models by relating it to the difference of the increments of ture positives and false positives of the union model.

Conclusions: To further enhance the performance, we can integrate more models based on the elaborated notion of divergent models that we derived to minimize the number of models required. It is not clear why MALLET and CRF++ respond differently with regards to feature selection and parsing direction. This cannot be elucidated unless we actually trace their implementation to figure out the real cause.

SESSION 1B

TEXT CASSIFICATION (I)

Exploiting and integrating rich features for biological literature classification

Hongning Wang, Minlie Huang, Shilin Ding and Xiaoyan Zhu

Background: Efficient features play an important role in automated text classification, which definitely facilitates the access of large-scale date. In the field of bioscience, where biological structures and processes are described by a large number of features, domain dependent features significantly improve the classification performance. How to effectively select features and integrate different types of features to improve the performance is the major issue studied in this paper.

Results: To efficiently classify biological literatures, we propose a probabilistic feature value estimation schema, novel features from low level domain independent "sring feature" to high level domain dependent "semantic template feature", and proper integrations among the features. Compared to our previous results, the performance is improved by 11.5% and 8.8% in the term of AUC and F-Score respectively, which outperforms the best performance achieved in BioCreAtIvE 2006.

Conclusions: Different types of features possess different description capability in literature classification; proper integration of domain dependent and independent features would significantly improve the performance and overcome the sensitivity on the data's distribution.

The integration of multiple feature representations for protein-protein interaction classification task

Man Lan and Chew Lim Tan

Background: Protein protein interaction (PPI) information is crucial for both biological and biomedical researchers. In order to extract and retrieve protein-interaction information from text, automatic detecting first protein interaction relevant articles for database curation is a crucial task for the subsequent steps. Generally, this detection system has been realized through traditional text classification techniques. The vast majority of this research uses the "bag-of-words" representation of text, where each feature corresponds to a single word or stem. For the sake of capturing more information left out from this simple bag-of-word representation, we examined alternative ways to represent text based on semantic knowledge from the whole corpus level and biological domain experts' knowledge, i.e. protein named entities, trigger keywords.

Results: The representations are evaluated using SVM classifier on the BioCreAtIvE II benchmark corpus, but on their own the new representations are not found to produce a significant performance improvement based on the statistical significance tests. On the other hand, the performance achieved by integration of 70 trigger keywords and 4 protein named entities features is comparable with that achieved by using 900 bag-of-words. Finally, we try combining classifiers based on different representations using a majority voting technique.

Conclusions: In general, our work supports the emerging consensus in the information retrieval community that more sophisticated Natural Language Processing techniques need to be developed before better text representations can be produced.

Protein-protein interaction abstract identification with contextual bag of words

Richard Tzong-Han Tsai, Hsieh-Chuan Hung, Hong-Jie Dai and Yi-Wen Lin

Background: In recent years, the recognition of semantic types from the biomedical scientific literature has been focused on named entities like protein and gene names (PGNs) and gene ontology terms (GO terms). Other semantic types like diseases have not received the same level of attention. Different solutions have been proposed to identify disease named entities in the scientific literature. While matching the terminology with language patterns suffers from low recall (e.g., Whatizit) other solutions make use of morpho-syntactic features to better cover the full scope of terminological variability (e.g., MetaMap). Currently, MetaMap that is provided from the National Library of Medicine (NLM) is the state of the art solution for the annotation of concepts from UMLS (Unified Medical Language System) in the literature. Nonetheless, its performance has not yet been assessed on an annotated corpus. In addition, little effort has been invested so far to generate an annotated dataset that links disease entities in text to disease entries in a database, thesaurus or ontology and that could serve as a gold standard to benchmark text mining solutions.

Results: As part of our research work, we have taken a corpus that has been delivered in the past for the identification of associations of genes to diseases based on the UMLS Metathesaurus and we have reprocessed and re-annotated the corpus. We have gathered annotations for disease entities from two curators, analyzed their disagreement (0.51 in the kappa-statistic) and composed a single annotated corpus for public use. Thereafter, three solutions for disease named entity recognition including MetaMap have been applied to the corpus to automatically annotate it with UMLS Metathesaurus concepts. The resulting annotations have been benchmarked to compare their performance.

Conclusions: The annotated corpus is publicly available and can serve as a benchmark to other systems. In addition, we found that dictionary look-up already provides competitive results indicating that the use of disease terminology is highly standardized throughout the terminologies and the literature. MetaMap generates precise results at the expense of insufficient recall while our statistical method obtains better recall at a lower precision rate. Even better results in terms of precision are achieved by combining at least two of the three methods leading, but this approach again lowers recall. Altogether, our analysis gives a better understanding of the complexity of disease annotations in the literature. MetaMap and the dictionary based approach are available through the Whatizit web service infrastructure.

SESSION 1C

TEXT MINING

New challenges for text mining: Mapping between text and manually curated pathways

Kanae Oda, Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi and Jun'ichi Tsujii

Background: Associating literature with pathways poses new challenges to the Text Mining (TM) community. There are three main challenges to this task: (1) the identification of the mapping position of a specific entity or reaction in a given pathway, (2) the recognition of the causal relationships among multiple reactions, and (3) the formulation and implementation of required inferences based on biological domain knowledge.

Results: To address these challenges, we constructed new resources to link the text with a model pathway; they are: the GENIA pathway corpus with event annotation and NF-kB pathway. Through their detailed analysis, we address the untapped resource 'bio-inference,' as well as the differences between text and pathway representation. Here, we show the precise comparisons of their representations and the nine classes of 'bio-inference' schemes observed in the pathway corpus.

Conclusions: In this study, we present a rudimentary research through challenging efforts for seeking the technology to advance beyond the conventional Information Extraction (IE). We believe that the creation of such rich resources and their detailed analysis is the significant first step for accelerating the research of the automatic construction of pathway from text.

A comparative analysis of five protein-protein interaction corpora

Sampo Pyysalo, Antti Airola, Juho Heimonen, Jari Björne, Filip Ginter and Tapio Salakoski

Background: Growing interest in the application of natural language processing methods to biomedical text has led to an increasing number of corpora and methods targeting protein protein interaction (PPI) extraction. However, there is no general consensus regarding PPI annotation and consequently resources are largely incompatible and methods are difficult to evaluate.

Results: We present the first comparative evaluation of the diverse PPI corpora, performing quantitative evaluation using two separate information extraction methods as well as detailed statistical and qualitative analyses of their properties. We find that the F-score performance of a state-of-the-art PPI extraction method varies on average 19 percentage units and in some cases over 30 percentage units between the different evaluated corpora, suggesting definite limits on the ability to compare methods evaluated on different resources. We analyse a number of potential sources for these differences and identify factors explaining approximately half of the variance. We further suggest ways in which the difficulty of the PPI extraction tasks codified by different corpora can be determined to advance comparability. Our analysis also identifies points of agreement and disagreement in PPI corpus annotation that are rarely explicitly stated by authors of the corpora.

Conclusions: Our comparative analysis uncovers key similarities and differences between the diverse PPI corpora, thus taking an important step towards standardization. In the course of this study we have created a major practical contribution in merging the corpora under a

shared format. The conversion software is freely available at http://mars.cs.utu.fi/PPICorpora.

Syntactic features for protein-protein interaction extraction

Rune Sætre, Kenji Sagae and Jun'ichi Tsujii

Background: Extracting Protein-Protein Interactions (PPI) from research papers is a way of translating information from English to the language used by the databases that store this information. With recent advances in automatic PPI detection, it is now possible to speed up this process considerably. Syntactic features from different parsers for biomedical English text are readily available, and can be used to improve the performance of such PPI extraction systems.

Results: A complete PPI system was built. It uses a deep syntactic parser to capture the semantic meaning of the sentences, and a shallow dependency parser to improve the performance further. Machine learning is used to automatically make rules to extract pairs of interacting proteins from the semantics of the sentences. The results have been evaluated using the AImed corpus, and they are better than earlier published results. The F-score of the current system is 69.5% for cross-validation between pairs that may come from the same abstract, and 52.0% when complete abstracts are hidden until final testing. Automatic 10-fold cross-validation on the entire AImed corpus can be done in less than one hour on a single server. People interested in using the system for their own research can obtain a free academic license by contacting the corresponding author. The system was first implemented in Perl, but later also translated to C++, and wrapped in Java as a UIMA component. We also present some previously unpublished statistics about the AImed corpus, and a short analysis of the AImed representation language.

Conclusions: We present a PPI extraction system, using different syntactic parsers to extract features for SVM with Tree Kernels, in order to automatically create rules to discover protein interactions described in the molecular biology literature. The system performance is better than other published systems, and the implementation is freely available to anyone who is interested in using the system for academic purposes. The system can help researchers quickly discover new PPIs, and increase the speed at which databases can be populated and signaling pathways constructed in the future.

Recognition of multisentence n-ary subcellular localization mentions in biomedical abstracts

Gabor Melli, Martin Ester and Anoop Sarkar

Background: Research into semantic relation recognition from text has focused on the identification of binary relations that are contained within one sentence. In the domain of biomedical documents however relations of interest can have more than two arguments and can also have its entity mentions located on different sentences. An example of this scenario is the ternary relation of subcellular localization which relates whether an organism's (O) protein (P) has subcellular location (L) as one of its target destinations. Empirical evidence suggests that approximately one half of the mentions for this ternary relation reside on multi-sentence passages.

Results: We introduce a relation recognition algorithm that can detect n-ary relations across multiple sentences in a document, and use the subcellular localization relation as a motivating example. The approach uses a text-graph representation of the entire document that is based on intrasentential edges derived from each sentence's predicted syntactic parse trees, and on intersentential edges based on either the linking of adjacent sentences or the linking of coreferents, if reliable coreference predictions are available. From the text graph state-of-the-art features such as named-entity features and syntactic features are produced for each argument pairing. We test the approach on the task of recognizing, in PubMed abstracts, experimentally validated subcellular localization relations that have been curated by biomedical researchers. When tested against several baseline algorithms, our approach is shown to attain the highest F-measure.

Conclusions: We present a method that naturally supports the recognition of semantic relations with more than two arguments and whose mentions can reside across multiple sentences. The algorithm accelerated the extraction of experimentally validated subcellular localizations. Given that the corpus is based on abstracts, not copyrighted papers, the data is publicly available from koch.pathogenomics.ca/pplre/. Significant work remains to approximate human expert levels of performance. We hypothesize that additional features are required that provide contextual information from elsewhere in the document about whether the relation refers to an experimentally validated finding.

SESSION 2A

ONTOLOGIES AND LOGIC

Monitoring the evolutionary aspect of the Gene Ontology to enhance predictability and usability

Jong C. Park, Tak-eun Kim and Jinah Park

Background: Much effort is currently made to develop the Gene Ontology (GO). Due to the dynamic nature of information it addresses, GO undergoes constant updates whose results are released at regular intervals as separate versions. Although there are a large number of computational tools to aid the development of GO, they are operating on a particular version of GO, making it difficult for GO curators to anticipate the full impact of particular changes along the time axis on the larger scale. We present a method for tapping into such an evolutionary aspect of GO, by making it possible to keep track of important temporal changes to any of the terms and relations of GO and by consequently making it possible to recognize associated trends.

Results: We develop visualization methods for viewing the changes between two different versions of GO by constructing a colour-coded layered graph. The graph shows both versions of GO with highlights to those GO terms that are added, removed and modified between the two versions. Focusing on a specific GO term or terms of interest over a period, we demonstrate the utility of our system that can be used to make useful hypotheses about the cause of the evolution and to provide new insights into more complex changes.

Conclusions: GO undergoes fast evolutionary changes. A snapshot of GO, as presented by each version of GO alone, overlooks such evolutionary aspects, and consequently limits the utilities of GO. The method that highlights the differences of consecutive versions or two different versions of an evolving ontology enhances the utility of GO for users as well as for developers.

Structuring an event ontology for disease outbreak detection

Ai Kawazoe, Hutchatai Chanlekha, Mika Shigematsu and Nigel Collier

Background: This paper describes the design of an event ontology being developed for machine understanding of disease-related events reported in natural language text. This event ontology is designed 1) to bridge a gap between layman's language used in disease outbreak reports and public health experts' deep knowledge and 2) to make multi-lingual information available, in order to support timely detection of disease outbreaks and rapid judgment of their alerting status.

Construction and Content: This event ontology integrates a model of experts' knowledge for disease surveillance, and at the same time a structured vocabulary of linguistic expressions which denote disease-related events, and formal definitions of event classes. In this ontology, rather general event classes, which are suitable for application to language-oriented tasks such as recognition of event expressions, are placed on the upper-level, and more specific events of the experts' interest are in the lower level. Each class is related to other classes which represent participants of events, and linked with multi-lingual synonym sets and axioms.

Conclusions: We consider that the design of the event ontology and the methodology introduced in this paper are applicable to other domains which require integration of natural language information and machine support for experts to assess them. The first version of the ontology, with about 40 concepts, will be available in November 2007.

Combining Gene Ontology and argumentative features for automatic geneRIF extraction

Julien Gobeill and Patrick Ruch

Background: This paper describes and evaluates a summarization system that extracts the gene function textual descriptions (called GeneRIF in ENTREZ-Gene) based on a MEDLINE record. Inputs for this task include both a locus (a gene in the LocusLink database), and a pointer to a MEDLINE record supporting the GeneRIF. In the suggested approach we merge two independent phrase extraction strategies. The first proposed strategy (LASt) uses argumentative, positional and structural features in order to suggest a GeneRIF. The second extraction scheme (GOPEx) incorporates statistical properties of the Gene Ontology to select the most appropriate sentence as the GeneRIF.

Results: Based on the TREC-2003 Genomics collection for GeneRIF identification, the LASt extraction strategy is already competitive (52.78%). When used in a combined approach, the extraction task clearly shows improvement, achieving a Dice score of over 57%.

Conclusions: Argumentative representation levels and Gene Ontology features are complementary for functional annotation in proteomics.

Decentralised clinical guidelines modelling with lightweight coordination calculus

Bo Hu, Srinandan Dasmahapatra, David Robertson and Paul Lewis

Background: Clinical protocols and guidelines have been considered as a major means to ensure that cost-effective services are provided at the point of care. Recently, the computerisation of clinical guidelines has attracted extensive research interest. Many languages and frameworks have been developed. Thus far, however, an enactment mechanism to facilitate decentralised guideline execution has been a largely neglected line of research. It is our contention that decentralisation is essential to maintain a high-performance system in pervasive health care scenarios. In this paper, we propose the use of Lightweight Coordination Calculus (LCC) as a feasible solution. LCC is a light-weight and executable process calculus that has been used successfully in multi-agent systems, peer-to-peer (p2p) computer networks, etc. In light of an envisaged pervasive health care scenario, LCC, which represents clinical protocols and guidelines as message-based interaction models, allows information exchange among software agents distributed across different departments and/or hospitals.

Results: We outlined the syntax and semantics of LCC; proposed a list of refined criteria against which the appropriateness of candidate clinical guideline modelling languages are evaluated; and presented two LCC interaction models of real life clinical guidelines.

Conclusions: We demonstrated that LCC is particularly useful in modelling clinical guidelines. It specifies the exact partition of a workflow of events or tasks that should be observed by multiple "players" as well as the interactions among these "players". It presents

the strength of both process calculus and Horn clauses pair of which can provide a close resemblance of logic programming and the flexibility of practical implementation.

SESSION 2B

TEXT CLASSIFICATION (II)

Automatic construction of rule-based ICD-9-CM coding systems

Richard Farkas and Gyorgy Szarvas

Background: In this paper we focus on the problem of automatically constructing ICD-9-CM coding systems for radiology reports. ICD-9-CM codes are used for billing purposes by health institutes and are assigned to clinical records manually following clinical treatment. Since this labeling task requires expert knowledge in the field of medicine, the process itself is costly and is prone to errors as human annotators have to consider thousands of possible codes when assigning the right ICD-9-CM labels to a document. In this study we use the datasets made available for training and testing automated ICD-9-CM coding systems by the organisers of an International Challenge on Classifying Clinical Free Text Using Natural Language Processing in spring 2007. The challenge itself was dominated by entirely or partly rule-based systems that solve the coding task using a set of 'hand crafted expert rules'. Since the feasibility of the construction of such systems for thousands of ICD codes is indeed questionable, we decided to examine the problem of automatically constructing similar rule sets that turned out to achieve a remarkable accuracy in the shared task challenge.

Results: Our results are very promising in the sense that we managed to achieve comparable results with purely 'hand-crafted rule-based"' ICD-9-CM classifiers. Our best model got a (90.26\%) F measure on the training dataset and an (88.93\%) F measure on the challenge test dataset, using the micro-averaged F measure, the official evaluation metric of the International Challenge on Classifying Clinical Free Text Using Natural Language Processing. This result would have come second in the challenge, with a hand-crafted system achieving slightly better results.

Conclusions: Our results demonstrate that hand-crafted systems -- which proved to be successful in ICD-9-CM coding -- can be reproduced by replacing several laborious steps in their construction with statistical learning models. These hybrid systems preserve the favourable aspects of rule-based classifiers and thus achieve a good performance, and their development can be performed rapidly and requires less human effort. Hence the construction of such hybrid systems can be feasible for a set of labels one magnitude bigger, and with more labeled data.

Identification of transcription factor contexts in literature using machine learning approaches

Hui Yang and Goran Nenadic

Background: Transcription factors (TFs) play a central role in regulation of gene expression. Manual literature curation of TF databases is expensive and labour intensive. Development of text mining support is hindered by unavailability of training data to build effective recognisers. There have been no studies on how existing data sources (e.g. TF-related data from the MeSH thesaurus and GO ontology) or noisy example data (e.g. protein-protein interaction, PPI) could be used to provide training data for the task.

Results: In this paper we described a text classification system designed to automatically recognise sentences describing transcription factors in the literature. A learning model is

based on a set of biological features that are deemed relevant for the task. We exploited background knowledge from existing biological resources (MeSH and GO) to engineer such features. Three machine-learning methods have been investigated, along with a vote-based result-merging of individual approaches and/or different training datasets. The training and testing data sets have been collected from weak and noisy examples originated from descriptions of TF-related terms in MeSH and GO, PPI data and data showing non gene function descriptions. The system achieved highly encouraging results, with most classifiers achieving F-measure above 90% (for the combined approach).

Conclusions: The experiments have shown that the biological model proposed can be used for identification of TF-related sentences with high accuracy, with a significantly reduced set of features when compared to traditional bag-of-word approach. We have also shown that existing knowledge sources are useful both as features and as a source of noisy positive training data. The analysis of the classification results when PPI data is used suggest that there is no as high similarity between TF and PPI contexts as we have expected.

Classifier ensemble for biomedical document retrieval

Manabu Torii and Hongfang Liu

Background: Due to rich information embedded in published articles, literature review has become an important aspect of research activities in the biomedical domain. Machine Learning (ML) techniques have been successfully used in retrieving relevant articles from a large literature archive (i.e., classifying articles into relevant and irrelevant classes), and thus accelerating the literature review process. Meanwhile, an ensemble classifier, a system that assigns classes to instances based on the outputs of multiple classifiers, tends to be more robust and usually has better performance than each constituent classifier. Ensemble classifiers are often composed of classifiers trained on different training sets (e.g., obtained by sampling techniques) or of those using different ML algorithms. In this paper, we propose a simple ensemble approach where an ensemble is composed of classifiers using different feature sets and the same ML algorithm. We evaluated the approach using Support Vector Machine (SVM) on two publicly available document collections, the Post-translational modification (PTM) data sets and the Immune Epitope Database (IEDB) data sets, which resulted from biomedical database curation projects.

Results: The evaluation showed that ensemble classifiers outperformed their constituent classifiers as measured by both Area under ROC curve (AUC) and Precision/recall break-even-point (BEP), provided with enough training data. We observed that AUC measures for SVM ensembles were competitive or better than the best results previously reported for the data sets used.

Conclusions: The proposed ensemble approach was found to be effective in improving performance of SVM classifiers. The approach is also simple and easy-to-deploy in document classification/retrieval tasks. In future work, we plan to explore different ways to derive and combine constituent classifiers, and continue our investigation over other data sets.