Handbook of

LBM 2009

The 3rd International Symposium on Languages in Biology and Medicine

Hyatt Regency, Seogwipo-si, Jeju Island,
South Korea
November 8-10, 2009

Hosted by LBM Steering Committee &
KIISE SIG on Human Language Technology

KOREAN INSTITUTE OF INFORMATION SCIENTISTS AND ENGINEERS

Steering Committee

Jong C. Park (*KAIST, Korea*)
Limsoon Wong (*NUS, Singapore*)
See-Kiong Ng (*NTU & I2R, Singapore*)

General Chairs

Jong C. Park (*KAIST, Korea*)
Limsoon Wong (*NUS, Singapore*)

Programme Committee Chairs

Dietrich Rebholz-Schuhmann (*EMBL-EBI, Hinxton, UK*)
Nigel Collier (*NII, Tokyo, Japan*)

Poster Session Committee

Hongfang Liu (*Georgetown University Medical Center, USA, Chair*)

Programme Committee

Sophia Ananiadou (*University of Manchester and NaCTeM, UK*)
Christopher Baker (*University of New Brunswick, Canada*)
Judith Blake (*Jackson Lab, USA*)
Olivier Bodenreider (*National Library of Medicine, USA*)
Wendy Chapman (*University of Pittsburgh, USA*)
Kevin Cohen (*University of Colorado, USA*)
Dina Demner-Fushman (*National Library of Medicine, USA*)
Juliane Fluck (*SCAI, Germany*)
Udo Hahn (*Jena University, Germany*)
Lynette Hirschman (*MITRE, USA*)
Larry Hunter (*University of Colorado, USA*)
Chun-Nan Hsu (*Academia Sinica, Taiwan*)
Su Jian (*I2R, Singapore*)
Jaewoo Kang (*Korea University, South Korea*)
Jin-Dong Kim (*University of Tokyo, Japan*)
Jung-Jae Kim (*EBI, UK*)
Martin Krallinger (*CNIO, Spain*)
Michael Krauthammer (*Yale University School of Medicine, USA*)
Hyunju Lee (*GIST, South Korea*)
Hongfang Liu (*Georgetown University Medical Center, USA*)
Goran Nenadic (*University of Manchester, UK*)
See-Kiong Ng (*NTU & I2R, Singapore*)
Jinah Park (*KAIST, South Korea*)
Fabio Rinaldi (*University of Zurich, Switzerland*)
Stefan Schulz (*Freiburg University, Germany*)
Gerold Schneider (*University of Zurich, Switzerland*)
Hagit Shatkay (*Queen's University, Canada*)
Adrian Shephard (*Birkbeck University of London, UK*)
Yoshimasa Tsuruoka (*University of Manchester, UK*)
Alfonso Valencia (*CNIO, Spain*)
Karin Vespoor (*University of Colorado, USA*)
Gwan-Su Yi (*KAIST, South Korea*)
Pierre Zweigenbaum (*LIMSI-CNRS, France*)

Local Organizing Committee

Munpyo Hong (*Sungkyunkwan University, South Korea*)
Jaewoo Kang (*Korea University, South Korea*)
Juhan Kim (*Seoul National University, South Korea*)
Gary Geunbae Lee (*POSTECH, South Korea*)
Hyunju Lee (*GIST, South Korea*)
Jinah Park (*KAIST, South Korea*)
Jong C. Park (*KAIST, South Korea, Chair*)
Gwan-Su Yi (*KAIST, South Korea*)

Local Staff

Seung-Cheol Baek *(KAIST, South Korea)*
Eunyoung Chang *(KAIST, South Korea, Head)*
Jin-Woo Chung *(KAIST, South Korea)*
Mihee Jo *(KAIST, South Korea)*
SangYoon Jung *(KAIST, South Korea)*
Youngrae Kim *(KAIST, South Korea)*
Hee-Jin Lee *(KAIST, South Korea)*
Ho-Joon Lee *(KAIST, South Korea)*
Hye-Jin Min *(KAIST, South Korea)*
Kimin Oh *(KAIST, South Korea)*
Eun-Jae Park *(KAIST, South Korea)*
Yeseul Park *(KAIST, South Korea)*

Preface

LBM is an international and interdisciplinary forum that brings together researchers in biology, chemistry, medicine, public health and informatics to discuss and exploit cutting edge language technology. Language, in its many forms, is the universal means to represent, question, and convey knowledge. Likewise applied technologies such as information extraction, summarization, and translation are key technologies for advancing biomedical research and healthcare provision. The automation of all these solutions empowers our ability to discover new knowledge.

All the available technologies are constantly being challenged by user demands in interdisciplinary research work leading into sustained growth in quantity and quality of existing language technologies. The involved communities profit from information exchange about recent progress to efficiently exploit the available solutions and to improve interdisciplinary work. This has led to the continued interest in the LBM symposium series, which offers a forum for synergistic interdisciplinary interactions as a rich source for significant advancements. After all, to efficiently manage data and knowledge requires the meeting of specialists in all relevant domains.

The International Symposium on Languages in Biology and Medicine (LBM) 2009 is one of the opportunities for interdisciplinary interaction. LBM was established in 2005 and the remit of this event remains highly relevant today. The symposium focuses on the languages that are in active use for biology and medicine.

LBM2009 has received 30 submissions, out of which 9 papers are selected as long papers, 5 papers as short papers, and 6 papers as poster papers. Among those long papers, 6 papers are invited to the special issues of BMC Journal of Biomedical Semantics and Journal of Bioinformatics and Computational Biology. We thank the PC members for their invaluable reviews of the submitted papers. We also thank the local organizing committee members and local staff for their great service to make the symposium really happen in Jeju Island, South Korea. And we thank the participants who have chosen to submit papers and to attend the symposium for its diverse activities.

Dietrich Rebholz-Schuhmann and Nigel Collier, PC Chairs
Jong C. Park and Limsoon Wong, General Chairs

Contents

LBM2009 Programme

Programme
Sunday, 8 November
Monday, 9 November (Morning)
Monday, 9 November (Afternoon)
Tuesday, 10 November

Programme

	Sunday, 8 November	Monday, 9 November	Tuesday, 10 November
8:15		Registration	
9:00		**Invited Talk 1** Thérèse Vachon	**Invited Talk 2** Pierre Zweigenbaum
~			
10:00		Break	Break
~		**Paper Session 2**	**Paper Session 4**
11:00			
15			
30			
45			**Wrap-Up**
12:00	Registration		Lunch
15		Lunch	
45			
1:00	**Tutorial**		Bus Excursion
15		**Poster Spotlight Presentations**	
30			
45		**Poster Session**	
2:00			
15			
30			
45	Break		
3:00	**Opening Remarks**		
15	**Paper Session 1**	Break	
30		**Paper Session 3**	
45			
4:00			
~			
5:00			
15		Symposium Banquet	
30			
45			
6:00			
~			
7:00	Welcome Party		
~			
8:00			
~			
9:00			

Sunday, 8 November

12:00-1:00	Registration
1:00-2:45	**Tutorial** Exploitation of Ontological Resources for Information Retrieval and Information Extraction (Antonio Jimeno-Yepes and Dietrich Rebholz-Schuhmann)
2:45-3:00	Break
3:00-3:15	**Opening Remarks**
3:15-5:15 Session 1: Event Extraction	
3:15-3:45	*** Event Extraction with Complex Event Classification using Rich Features** (Makoto Miwa, Rune Sætre, Jin-Dong Kim and Jun'ichi Tsujii)
3:45-4:15	**Biological Event Recognition with Textual Induction** (Yutaka Sasaki, Paul Thompson, John McNaught and Sophia Ananiadou)
4:15-4:45	**Automatic Annotation by BioExcom for categorizing prior and new speculations in biological papers** (Julien Desclés, Motasem Alrahabi, Jean-Pierre Desclés and Antoine Blais)
4:45-5:15	*** Analysis of syntactic and semantic features for fine-grained event-spatial understanding in outbreak news reports** (Hutchatai Chanlekha and Nigel Collier)
7:00-	Welcome Party

Monday, 9 November (Morning)

8:15-9:00	Registration
9:10-10:00	**Invited Talk 1** Thérèse Vachon
10:00-10:15	Break
10:15-12:15 Session 2: Identifying Semantics	
10:15-10:45	*** The application of an ontology design pattern for functional abnormalities to phenotype ontologies and the extraction of an ontology of anatomical functions** (Robert Hoehndorf, Axel-Cyrille Ngonga Ngomo and Janet Kelso)
10:45-11:15	*** The Value of an In-Domain Lexicon in Genomics QA** (Yutaka Sasaki, John McNaught and Sophia Ananiadou)
11:15-11:45	**Automatic Extraction of the Usage Information from the Component Words in Gene Ontology Terms to Enhance Consistency and Predictability** (Seung-Cheol Baek and Jong C. Park)
11:45-12:15	**A Re-evaluation of Biomedical Named Entity - Term Relations** (Tomoko Ohta, Sampo Pyysalo, Jin-Dong Kim and Jun'ichi Tsujii)
12:15-1:15	Lunch

Monday, 9 November (Afternoon)

1:15-1:45	**Poster Spotlight Presentations**
1:45-3:15	**Poster Session** **A Thesaurus and an Application Ontology for the Juvenile Arthritis Domain** (Ernesto Jim´enez-Ruiz, Rafael Berlanga-Llavori, Victoria Nebot, Antonio Jimeno-Yepes and Dietrich Rebholz-Schuhmann) **Comparison of methods for topic template queries in the biomedical domain** (Antonio Jimeno-Yepes, Rafael Berlanga-Llavori and Dietrich Rebholz-Schuhmann) **Inference for bio-IE: GENIA meets EKOSS** (Jin-Dong Kim, Steven Kraines, Weisen Guo and Jun'ichi Tsujii) **ONER: Tool for Organization Named Entity Recognition from Affiliation Strings in PubMed Abstracts** (Siddhartha Jonnalagadda, Philip Topham and Graciela Gonzalez) **Bio-medical Term Extraction on Simple Rule Language** (Takashi Sinnou, Koichi Takeuchi and Nigel Collier) **Literature mining for protein acetylation** (Youngrae Kim, Hodong Lee and Gwan-Su Yi)
3:15-3:30	Break
3:30-5:30 Session 3: Building Resources	
3:30-4:00	**Demystifying protein annotations: toward increasing the compatibility of different corpora** (Yue Wang, Jin-Dong Kim, Rune Saetre, Sampo Pyysalo, Tomoko Ohta and Jun'ichi Tsujii)
4:00-4:30	*** The CALBC Silver Standard Corpus - Harmonizing multiple semantic annotations in a large biomedical corpus**

	(Dietrich Rebholz-Schuhmann, Antonio José Jimeno Yepes, Erik M. van Mulligen, Ning Kang, Jan Kors, David Milward, Peter Corbett and Udo Hahn)
4:30-5:00	* **De-identifying Swedish Clinical Text - Refinement of a Gold Standard and Experiments with Conditional Random Fields** (Hercules Dalianis and Sumithra Velupillai)
5:00-5:30	
5:30-	Symposium Banquet

Tuesday, 10 November

9:10-10:00	**Invited Talk 2** Pierre Zweigenbaum
10:00-10:15	Break
10:15-11:45 Session 4: Information Extraction	
10:15-10:45	**Enabling Recognition of Diseases in Biomedical Text with Machine Learning: Corpus and Benchmark** (Robert Leaman, Christopher Miller and Graciela Gonzalez)
10:45-11:15	**Sentence Simplification Aids Protein-Protein Interaction Extraction** (Siddhartha Jonnalagadda and Graciela Gonzalez)
11:15-11:45	**Effective Mining of Protein Interactions** (Fabio Rinaldi, Gerold Schneider, Kaarel Kaljurand and Simon Clematide)
11:45-12:00	**Best Paper Award and Wrap Up**
12:00-1:00	Lunch
1:00-	Bus Excursion

* invited for special issues (Journal of Biomedical Semantics and Journal of Bioinformatics and Computational Biology)

Invited Talks

Invited Talk 1: Thérèse Vachon
Invited Talk 2: Pierre Zweigenbaum

Invited Talk 1

Text mining, data integration and semantic technologies supporting and accelerating drug discovery

Thérèse Vachon
Head of Text Mining Services,
Novartis Institutes for Biomedical Research, NITAS, France

Abstract

The Text Mining Services group at Novartis Institutes and Biomedical Research (NIBR) develops and implements information integration platforms, enterprise search solutions and a set of tools in various areas e.g. information retrieval, information extraction, development of terminologies, text mining, data integration, contextual navigation and semantic enrichment. Our group develops formal models for describing the data independently from their physical representation with the aim to allow interoperability of data and systems. We are using the technologies that we implement to partner with business organizations and support their efforts in the area of target identification, target validation and confirmation of Research strategies. The presentation will outline some of the technologies and systems developed and their applications in the biomedical and biochemical fields.

Short Bio

Thérèse Vachon graduated in Computer Science from the Institut National des Sciences Appliquees in Lyon (France). She held positions with the Paris Stock Exchange, the University of Compiegne and an IT consulting company before joining Novartis. She developed end-user search tools in the biomedical and chemistry field. Since 1998, she has been very active in the field of information retrieval, information extraction, text mining, development of terminologies and integration of Life Sciences data using semantic capabilities. She is currently Head of Text Mining Services in NITAS (IT in Research).

Invited Talk 2

Multilingualism for Medical Information Processing

Pierre Zweigenbaum
LIMSI-CNRS, France

Abstract

Whereas most of the international biomedical literature is written in English, a number of medically relevant texts are written in other languages. This includes clinical texts in hospitals, recommendations, patient-oriented documents, scientific articles, and a variety of Web documents. Medical language processing, e.g., controlled indexing, information extraction, question-answering, and more generally text mining, is therefore needed in languages for which resources are less developed than those for English Building such resources is a task researchers and developers have to address for each target language (e.g., French, German, Dutch, Swedish, Korean, Chinese, Japanese, etc.). It can nevertheless leverage existing resources in English or in the target language. On the one hand, transfer methods help convert existing English resources, e.g., terminologies or language processors, to other languages. On the other hand, automatic mono- or multilingual acquisition techniques can help to bootstrap language resources more quickly. This talk will present methods and experiments to build language resources for processing medical texts in non-English languages, taking advantage where possible of existing English resources.

Short Bio

Pierre Zweigenbaum graduated from École Polytechnique, Paris, France, and received a PhD in Computer science from the École Nationale Supérieure des Télécommunications (ENST). He has been a researcher at Assistance Publique- Hôpitaux de Paris, the largest hospital group in Europe. He is currently Senior Researcher at LIMSI (Computer Sciences Laboratory for Mechanics and Engineering Sciences), a unit of CNRS, the French National Center for Scientific Research, and is Associate Professor at the National Institute of Oriental Languages and Civilizations, where he teaches Natural Language Engineering. His research interests lie in Natural Language Processing, from morphology to question answering, with a focus on specific domains (medicine and medical terminology) and multilingual issues (alignment in parallel and comparable corpora).

LBM2009 Abstracts

Long Papers

Event Extraction with Complex Event Classification using Rich Features

Makoto Miwa, Rune Sætre, Jin-Dong Kim and Jun'ichi Tsujii

To capture biomedical phenomena more deeply, it is required to extract relations that are more complex than binary relations. To extract such complex relations, the BioNLP'09 shared task provided complex events; binding and regulation were provided as complex relations. To improve the biomedical event extraction systems, finding these complex events automatically is important; thus, we focus on the extraction of the complex events. In this paper, we propose an automatic event extraction system, which contains a model for complex events, by solving a classification problem with rich features. Our complex event detector performed better than the top system (in the shared task), and in overall performance, our system outperformed the top system.

Automatic Annotation by BioExcom for categorizing prior and new speculations in biological papers

Julien Desclés, Motasem Alrahabi, Jean-Pierre Desclés and Antoine Blais

Biological research papers are replete with speculative sentences. This paper presents the BioExcom software, an adaptation of EXCOM to the biology and biomedical fields, which annotates automatically all speculative sentences in full texts papers by the

means of the Contextual Exploration processing. This annotation process is based on a fine semantic analysis of the multiple ways to express speculation in biology. Furthermore, BioExcom enables the automatically distinguishing of prior and new speculations in a biological paper. We argue that these annotations are useful for biologists' work, regardless of their domains of interest, helping them to evaluate quickly the content and new output of a paper. We discuss also some possible future applications of speculative sentences extraction and the CE processing in biology.

Analysis of syntactic and semantic features for fine-grained event-spatial understanding in outbreak news reports

Hutchatai Chanlekha and Nigel Collier

Previous studies have suggested that epidemiological reasoning needs a fine-grained modeling of events, especially their spatial and temporal attributes. While the temporal analysis of events has been intensively studied, far less attention has been paid to their spatial analysis. This article aims at filling the gap concerning automatic event-spatial attribute analysis in order to support health surveillance and epidemiological reasoning. In this work, we propose a methodology that provides a detailed analysis on each event reported in news articles to recover the most specific locations where it occurs. Various features for recognizing spatial attributes of the events were studied and incorporated into the models which were trained by several machine learning techniques. The best performance for spatial attribute recognition is very promising; 85.9% F-score (86.75% precision / 85.1% recall).

The application of an ontology design pattern for functional abnormalities to phenotype ontologies and the extraction of an ontology of anatomical functions

Robert Hoehndorf, Axel-Cyrille Ngonga Ngomo and Janet Kelso

Functions play an important role throughout biology. Although molecular functions are covered in the Gene Ontology, there is

currently no publicly available ontology of anatomical functions. Ontological considerations on the nature of functional abnormalities and their representation in current phenotype ontologies show that we can automatically extract a skeleton for such an ontology of anatomical functions by using a combination of process, phenotype and anatomy ontologies. We provide an ontological analysis of the nature of functions and functional abnormalities. From this analysis, we derive an approach to the automatic extraction of anatomical functions from existing ontologies using a combination of natural language processing, graph-based analysis of the ontologies and formal inferences. Alternatively, we introduce a new relation to relate material objects to processes that realize the function of the object to avoid a needless duplication of processes already present in the Gene Ontology in a new ontology of anatomical functions. We discuss several limitations of the current ontologies that still need to be addressed to ensure a consistent and complete representation of anatomical functions and functional abnormalities.

The Value of an In-Domain Lexicon in Genomics QA

Yutaka Sasaki, John McNaught and Sophia Ananiadou

This paper demonstrates that a large-scale lexicon tailored for the biology domain is effective in improving question analysis for genomics Question Answering (QA). We use the TREC Genomics Track data to evaluate the performance of different question analysis methods. It is hard to process textual information in biology, especially in molecular biology, due to a huge number of technical terms which rarely appear in general English documents and dictionaries. To support biological Text Mining, we have developed a domain-specific resource, the BioLexicon. Started in 2006 from scratch, this lexicon currently includes more than four million biomedical terms consisting of newly curated terms and terms collected from existing biomedical databases. While conventional genomics IR/QA systems provide query expansion based on thesauri and dictionaries, it is not clear to what extent a biology-oriented lexical resource is effective for question pre-processing for genomics QA. Experiments on the genomics QA data set show that question analysis using the BioLexicon performs slightly better than that using *n*-grams and the UMLS Specialist Lexicon.

Automatic Extraction of the Usage Information from the Component Words in Gene Ontology Terms to Enhance Consistency and Predictability

Seung-Cheol Baek and Jong C. Park

The Gene Ontology (GO) is a controlled vocabulary that has gone through constant changes, motivated primarily by the need to reflect the dynamic nature of knowledge it addresses and the need for usability improvement. A good policy on such changes would be to maintain consistency across terms and structures so as to highlight the missing parts that are likely to be added afterwards, or the unchanged parts to which a policy on usability improvement might not have yet applied. In particular, we argue that the component words inside terms must be used consistently across terms, in order to enhance the predictability of such terms, thus their usability as well. For this purpose, we propose a representation for word usage and a method for extracting it from GO and show its utility in identifying the direction of future changes readily as well as in enhancing the consistency of terms.

The CALBC Silver Standard Corpus - Harmonizing multiple semantic annotations in a large biomedical corpus

Dietrich Rebholz-Schuhmann, Antonio José Jimeno Yepes, Erik M. van Mulligen, Ning Kang, Jan Kors, David Milward, Peter Corbett and Udo Hahn

The CALBC initiative aims to provide a large-scale biomedical text corpus that contains semantic annotations for tagged named entities of different kinds. The generation of this corpus requires that the annotations from different automatic annotation systems are harmonized.

In the first phase, the annotation systems from 5 participants (EMBL-EBI, EMC Rotterdam, NLM, JULIE Lab Jena, and Linguamatics) were gathered. All annotations were delivered in a common annotation format that included concept ids in the boundary assignments and that enabled comparison and alignment of the results.

During the harmonization phase, the produced results from different systems have been integrated into a single harmonised

corpus ("silver standard" corpus) by applying a voting scheme. We give an overview of the processed data and the principles of harmonization – formal boundary reconciliation and semantic matching of named entities. Finally all submissions of the participants have been evaluated against the silver standard corpus. We found that species and disease annotations are better standardised amongst the partners than the annotations of genes and proteins.

The raw corpus is now available for additional named entity annotations. Part of the annotated corpus will be made available later for a public challenge. We expect that we can improve corpus building activities both in terms of the numbers of named entity classes being covered, as well as the size of the corpus in terms of annotated documents.

De-identifying Swedish Clinical Text -Refinement of a Gold Standard and Experiments with Conditional Random Fields

Hercules Dalianis and Sumithra Velupillai

In order to perform research on the information contained in Electronic Patient Records (EPRs), access to the data itself is needed. This is often very difficult due to confidentiality regulations. The data sets need to be fully de-identified before they can be distributed to researchers. De-identification is a difficult task where the definitions of annotation classes are not self-evident. We present work on the creation of two refined variants of a manually annotated Gold standard for de-identification, one created automatically, and one created through discussions among the annotators. These are used for the training and evaluation of an automatic system based on the Conditional Random Fields algorithm. Evaluating with four-fold cross-validation on sets of around 4-6 000 annotation instances, we obtained very promising results for both Gold Standards; F-score around 0.80 for a number of experiments, with higher results for certain annotation classes. Moreover, 49 false positives that were verified true positives were found by the system but missed by the annotators. Our intention is to make this Gold standard available for other research groups in the future. Despite being slightly more time consuming we believe the manual consensus gold standard is the most valuable for further research. We also propose a set of annotation classes to be used for similar de-identification tasks.

Enabling Recognition of Diseases in Biomedical Text with Machine Learning: Corpus and Benchmark

Robert Leaman, Christopher Miller and Graciela Gonzalez

Many lines of inquiry in biomedicine lead directly or indirectly to the prevention, diagnosis or treatment of disease. Utilizing text mining to further these lines of inquiry typically involves applying an extraction system including the recognition and identification (normalization) of the diseases mentioned as early steps in the pipeline. In recent years there has been a trend away from dictionary-based systems for the recognition of biomedical entities in favor of named entity systems based on machine learning for sequence tagging, such as conditional random fields. However, this trend has not yet extended to tagging diseases, despite a strong interest in disease entities, perhaps because of the difficulty in obtaining adequate corpora for training a machine learning system. We therefore introduce a new corpus (the Arizona Disease Corpus, or AZDC), derived from biomedical research abstracts, containing the necessary annotations for both named entity recognition and normalization of disease entities. We utilize this corpus to explore the performance of machine-learning based systems and dictionary match. We anticipate that this resource will prove valuable for mining disease-related knowledge from biomedical text, supporting the ability to translate our ever-increasing biomedical understanding into clinical applications and improved quality of life. The Arizona Disease Corpus (AZDC) can be freely downloaded*.

*http://diego.asu.edu/downloads/AZDC

Short Papers

Biological Event Recognition with Textual Induction

Yutaka Sasaki, Paul Thompson, John McNaught and Sophia Ananiadou

This paper describes a supervised approach to the recognition of biological events, which combines statistical sequential labeling and symbolic event extraction rules. Bottom-up textual induction has been applied to generating event extraction rules. As an evaluation data set, we use a corpus of biomedical abstracts, in which biological events concerning *gene regulation* in *E. coli* and *H. Sapiens* have been annotated by a group of biologists. The event instance extraction performance has been evaluated using 10-fold cross validation. The experimental results show that *named entity recognition (NER)* and *semantic role labeling (SRL)* performance are close to annotator performance, as indicated by the inter-annotator agreement *(IAA)* scores, whereas automatic event extraction performance is around 28%, as compared to 40% IAA for exact manual event extraction.

A Re-evaluation of Biomedical Named Entity - Term Relations

Tomoko Ohta, Sampo Pyysalo, Jin-Dong Kim and Jun'ichi Tsujii

Recent developments in biomedical text mining include advances at the reliability of named entity recognition as well as movement toward richer representations of the associations of named entities. We argue that this shift in representation should be accompanied by the adoption of a a more detailed model of the relations holding between named entities and other relevant domain terms. As a step toward this goal, we study named entity – term relations with the aim of defining a detailed, broadly applicable set of relation types based on accepted domain standard concepts for use in corpus annotation and domain information extraction approaches.

Demystifying protein annotations: toward increasing the compatibility of different corpora

Yue Wang, Jin-Dong Kim, Rune Sætre, Sampo Pyysalo, Tomoko Ohta and Jun'ichi Tsujii

While there are a number of corpora with protein annotations, the annotations in different corpora are not compatible with each other. It is, however, not yet well understood how they are different and how the incompatibilities can be overcome. The situation discourages utilization of the corpora in a united way. It also indicates that even within individual corpora, the actual annotations are not well understood. We first compare the protein annotations of two corpora, GENIA and GENETAG. Based on the result, we propose several strategies to increase the cross-corpus compatibility. Experimental results show that the proposed strategies are effective and the incompatibility of the protein annotations between the two corpora can be removed if we properly consider their differences.

Sentence Simplification Aids Protein-Protein Interaction Extraction

Siddhartha Jonnalagadda and Graciela Gonzalez

Accurate systems for extracting Protein-Protein Interactions (PPIs) automatically from biomedical articles can help accelerate biomedical research. Biomedical Informatics researchers are collaborating to provide meta-services and advance the state-of-art in PPI extraction. One problem often neglected by current Natural Language Processing systems is the characteristic complexity of the sentences in biomedical literature. In this paper, we report on the impact that automatic simplification of sentences has on the performance of a state-of-art PPI extraction system, showing a substantial improvement in recall (8%) when the sentence simplification method is applied, without significant impact to precision.

Effective Mining of Protein Interactions

Fabio Rinaldi, Gerold Schneider, Kaarel Kaljurand, and Simon Clematide

The detection of mentions of protein-protein interactions in the scientific literature has recently emerged as a core task in biomedical text mining. We present effective techniques for this task, which have been developed using the IntAct database as a gold standard, and have been evaluated in two text mining competitions.

Posters

A Thesaurus and an Application Ontology for the Juvenile Arthritis Domain

Ernesto Jim´enez-Ruiz, Rafael Berlanga-Llavori, Victoria Nebot, Antonio Jimeno-Yepes and Dietrich Rebholz-Schuhmann

This paper is intended to present our experiences in the creation of a light weight thesaurus for the Arthritis domain and the reuse of this terminological resource to create an ontology to classify patients suffering different subtypes of Rheumatoid Arthritis, which is part of the Health-e Child project.

Comparison of methods for topic template queries in the biomedical domain

Antonio Jimeno-Yepes, Rafael Berlanga-Llavori and Dietrich Rebholz-Schuhmann

Topic template queries are focused on a facet of a structured user information need. Examples of these topic templates are: the role of gene G in disease D and the interaction of proteins P1 and P2. These templates allow for multiple instances and some commonalities might be found which might provide improved retrieval on unseen instance queries of a template.

In this paper, we have analyzed two possible solutions that integrate the analysis of existing results based on query reformulation and the boosting of documents based on text categorization.

We show that both approaches produce interesting results when enough example queries are provided and that the boosting of retrieved document based on text categorization has a better performance.

Inference for bio-IE: GENIA meets EKOSS

Jin-Dong Kim, Steven Kraines, Weisen Guo and Jun'ichi Tsujii

Information extraction for molecular biology (bio-IE) aims to find useful pieces of bio-molecular knowledge (bio-knowledge, hereafter) from natural language expressions in the literature, and to store them in a structured form accessible by computers. One example is protein-protein interaction (PPI) extraction (Bunescu et al., 2004),

which has long been a primary task of bio-IE. Usually, a PPI is expressed by a pair of proteins. For example, from the text, "Secretion of TNF was abolished by BHA …," the following PPI can be extracted:

P1: (TNF, BHA)

Recently, as the need grows for semantically rich bio-knowledge – e.g. Gene Ontology annotation (GOA) (Camon et al., 2004), pathways (Bader et al., 2006) – the structure of bio-knowledge to be extracted is becoming more complex. BioNLP'09 Shared Task (BioNLP'09, hereafter) (Kim et al., 2009) addressed IE for bio-molecular events (bio-events). In the task, a bio-event is expressed by a predicate-argument structure, where the predicate specifies the type of event, and the argument expresses various aspects of the event, e.g. *theme*, *cause*. From the sample text above, the following events can be extracted according to BioNLP'09:

E1:(Localization, T:TNF)

E2:(Neg regulation, T:E1, C:BHA)

As the structure of the target bio-knowledge becomes complex, a more elaborate language is required to describe extracted knowledge pieces. An elaborate description language can encode a considerable amount of information, allowing useful computation over the knowledge descriptions, e.g. inferences. For example, initially, the relation between TNF and BHA is not explicit by E1 and E2, but if the description language defines the *theme* relation to be transitive, then the relation can be induced:

E3:(Neg regulation, T:TNF, C:BHA),

which holds the meaning, "*BHA negatively regulates (a unspecified activity of) TNF*". This paper reports our preliminery implementation to show that if we properly define the semantics of description language, we can find implicit knowledge descriptions which are implied by existing ones, through inferences over those semantics.

ONER: Tool for Organization Named Entity Recognition from Affiliation Strings in PubMed Abstracts

Siddhartha Jonnalagadda, Philip Topham and Graciela Gonzalez

Automatically extracting organization names from the affiliation sentences of articles related to biomedicine is of great interest to the pharmaceutical marketing industry, health care funding agencies and public health officials. It will also be useful for other scientists in normalizing author names, automatically creating citations, indexing

articles and identifying potential resources or collaborators. Today there are more than 18 million articles related to biomedical research indexed in PubMed, and information derived from them could be used effectively to save the great amount of time and resources spent by government agencies in understanding the scientific landscape, including key opinion leaders and centers of excellence. Our process for extracting organization names involves multi-layered rule matching with multiple dictionaries. The system achieves 99.6% f-measure in extracting organization names.

Bio-medical Term Extraction on Simple Rule Language

Takashi Sinnou, Koichi Takeuchi and Nigel Collier

For disease surveillance system, bio-medical term extraction is a key technology for a surveillance system of epidemic disease news from the Web. In the previous work we applied statistical learning model to extract terms from the Web site. The previous approach is good at extracting terms with high precision rates; however it is weak at extracting new terms that do not exist in the training data. Since we usually have new disease names a new term extraction approach with high coverage for unknown or low-frequent terms is needed. Recently, Simple rule Language (SRL), a rule-based word extraction language, is freely available. The SRL also has an developing environment called SRL editor. Thus we are constructing rules of bio-medical terms on the several language (such as English, Japanese, Thai and Vietnam) for the multilingual disease surveillance system. In this manuscript we confirm how we construct rules to extract Japanese bio-medical terms from Japanese news articles.

Literature mining for protein acetylation

Youngrae Kim, Hodong Lee and Gwan-Su Yi

This paper presents a method of text mining to extract information of acetylation. Acetylation is known to be involved in epigenetic pathways for cancer, stem cell, and neural disease. However, previous effort that gathers information about acetylation only relies on experimental data, excluding the epigenetic mechanisms reported in the literature. To compile the epigenetic effects on biological pathways, we developed a preliminary method to extract information of acetylation target and site information from the PubMed abstracts.

Tutorial

Exploitation of Ontological Resources for Information Retrieval and Information Extraction

Antonio Jimeno-Yepes and Dietrich Rebholz-Schuhmann
EMBL-EBI, UK

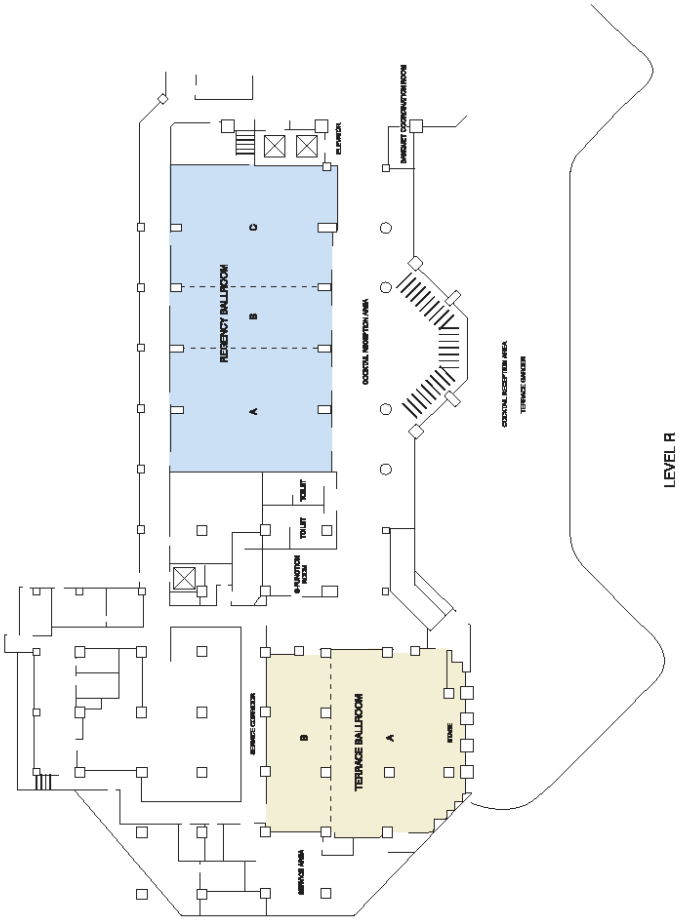
Ontological resources such as controlled vocabularies, taxonomies and ontologies from the OBO foundry are used to represent biomedical domain knowledge. The development of such resources is a time consuming task. Once they are finished they contribute to standardization of information representation, interoperability of IT solutions, literature analysis and knowledge discovery.

Text mining comprises IT solutions for information retrieval (IR) and information extraction (IE). IR technology exploits ontological resources to select documents that fit best to the processed query, for example, through indexing of the literature content with concept ids or through disambiguation of terms in the query. IE solutions make use of the ontological labels to identify concepts in the text. The text passages that denote conceptual entries are then used either to annotate named entities or to relate the named entities to each other. For knowledge discovery (KD) solutions the identified concepts in the scientific literature are used to relate entities to each other, e.g. to identify gene-disease relations based on shared molecular functions.

Maps

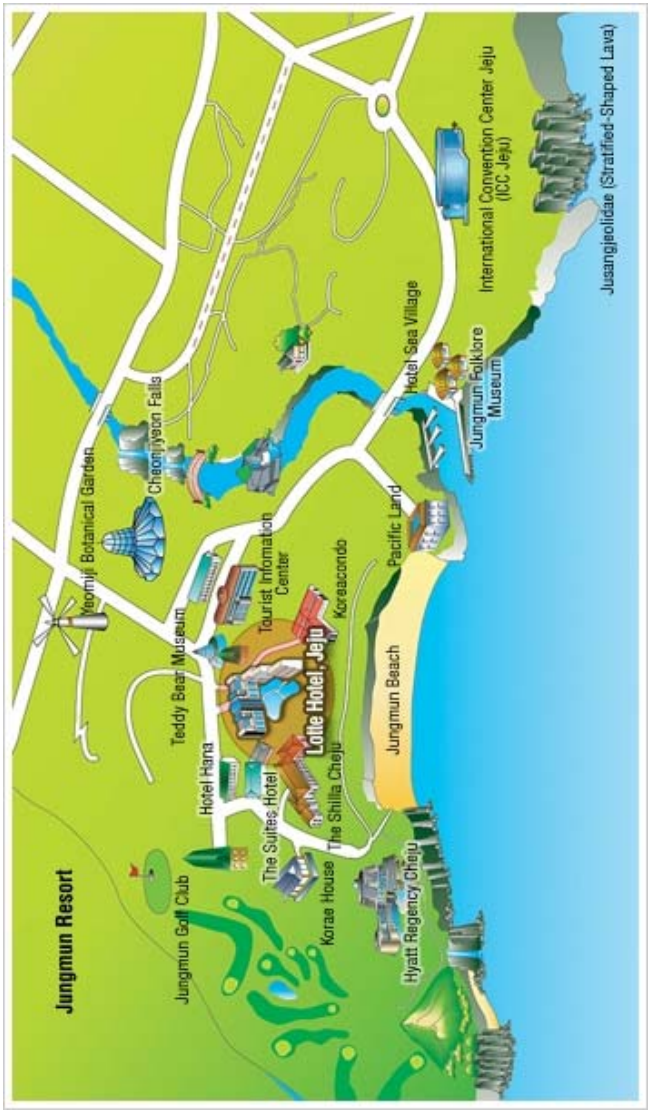
Symposium Venue
Jungmun Resort
Jeju Island
Excursion: Jeju Olle Trail Route 8
Excursion: Jeju Jeolmul Recreation Forest

Symposium Venue

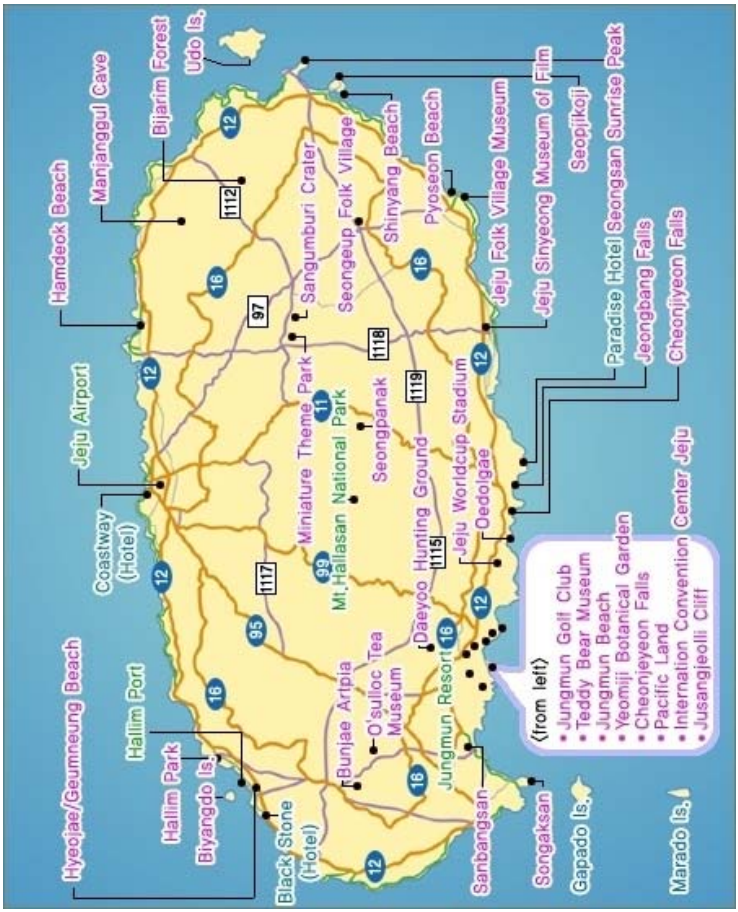


Paper Presentations & Poster Session: Terrace Ballroom

Jungmun Resort



Jeju Island


Hyeojae/Geumneung Beach
Hamdeok Beach
Manjanggul Cave
Bijarim Forest
Udo Is.
Jeju Airport
Coastway (Hotel)
Hallim Port
Hallim Park
Biyangdo Is.
Black Stone (Hotel)
Sangumburi Crater
Seongeup Folk Village
Shinyang Beach
Pyoseon Beach
Seopjikoji
Jeju Folk Village Museum
Jeju Sinyeong Museum of Film
Paradise Hotel Seongsan Sunrise Peak
Jeongbang Falls
Cheonjiyeon Falls
Miniature Theme Park
Mt. Hallasan National Park
Seongpanak
Daeyoo Hunting Ground
Jeju Worldcup Stadium
Oedolgae
Bunjae Artpia
O'sulloc Tea Museum
Jungmun Resort
Sanbangsan
Songaksan
Gapado Is.
Marado Is.
(from left)
• Jungmun Golf Club
• Teddy Bear Museum
• Jungmun Beach
• Yeomiji Botanical Garden
• Cheonjeyeon Falls
• Pacific Land
• Internation Convention Center Jeju
• Jusangjeolli Cliff

Excursion: Jeju Olle Trail Route 8

Olle [Ole] is the Jeju word for a narrow pathway that is connected from the street to the front gate of a house. Hence, Olle is a path that comes out from a secret room to an open space and a gateway to the world. If the road is connected, it is linked to the whole island and the rest of the world as well. It has the same sound as “Would you come?” in Korean, so Jeju's Olle sounds the same as “Would you come to Jeju?”.

This route continues along the seashore through Jusangjeolli which is a formation of stone pillars piled up along the coast. The sight of the abundant pampas grass makes your walk even more enjoyable. A pathway made of numerous rocks on the coastline was built by the marine corps for Jeju Olle, so it is called ‘The Marine Corps Trail’. The pathway used to be used only by local divers.

Excursion: Jeju Jeolmul Natural Recreation Forest

(The English versions of the map will be available on site.)

Jeju Island's Jeju Jeolmul Natural Recreation Forest is located northeast of Hallasan Mountain. Jeolmul Ascension is famous for its beautiful Japanese cedar forest. Jeolmul Ascension's is about 650 meters long and created for volcanic activity. The peak of Jeolmul Ascension boasts amazing views. It is even possible, during the clear weather, to see Ilchulbong Sacred Mountain.

Information

Internet connectivity

– 8-10 November, during the symposium
A desk with 5 Ethernet cables will be available at the foyer.

Social Events

Welcome Party

– Sunday, 8 November, 19:00 to 21:00 @ foyer

Banquet

– Monday, 9 November, 17:30
A Hanjeongsik (Korean table d'hôte) will be served together with folk performance.
The bus will depart from Hayatt Regency Front Door at 17:30 and come back to the symposium hotel in about 3 hours after the departure.

Excursion

– Tuesday, 10 November

Time	Place
2:00 PM	Hayatt Regency Front Door (Departure)
	Olle Trail Route 8
3:20 PM	Jusangjeolli (Pillar-Shaped Joint)
	Bus
4:30 PM	Jeolmul Natural Recreation Forest Parking Lot
	Jeolmul Natural Recreation Forest Route 2
5:40 PM	Jeolmul Natural Recreation Forest Parking Lot
	Bus
7:00 PM	Hyatt Regency Front Door (Arrival)

Local Information

Helpful Telephone Numbers

[Emergency Call]

- Dial 112 for the police, 119 for the fire department, or 1339 for medical emergencies (though most Korean operators speak little or no English). A hotel staff member or the hotel manager can arrange for a doctor or an ambulance.

[Local Telephone Number Guide]

- Local area code + 114

[Tourist Complaint Center]

- 02-735-0101

[International Emergency Rescue]

- 02-790-7561, A 24-hour emergency rescue service for foreigners

International Calls

[How to call a number in Korea from overseas]

- When you make a phone call to Korea from abroad, first dial 82 (country code for Korea), then the area code (except for the first number 0), and finally, dial the phone number you wish to call.
- For example: in order to call Jeju (Area Code 064) with 777-7777 as the phone number, dial +82-64-777-7777.

[How to call overseas from Korea]

- First dial any of the following international call company numbers and then enter the country code, the area code and finally the number you are calling.

*Regular International Phone Call Carriers: 001, 002, 008
*Cell Phone Carriers: 00345, 00365, 00388, 00700, 00727,

00766, 00770 and so on.
*Pre-paid Phone Cards: Available at any convenience stores or newsstands.

Useful Korean Phrases

Greetings and Common Courtesy

How do you do? 처음 뵙겠어요 [cheo-eum boep-get-seo-yo.]	You are welcome 천만에요. [choenmaneyo.]
I'm glad to meet you 만나서 반가워요 [mannaseo ban-ga-wo-yo.]	Excuse me. 실례합니다. [sillye-hamnida.]
Good bye. 안녕히 가세요. [annyeonghi-gaseyo.]	I am sorry. 미안합니다. [mian-hamnida.]
Yes./No. 예/아니요 [ye]/[aniyo]	Please help me. 도와주세요. [dowa-juseyo.]

Transportation

Will you show me the way to (Deoksugung Palace)?
(덕수궁) 가는 길을 가르쳐주세요.
[(Deoksugung) ganeun gireul gareucheo juseyo.]

Where can I take a taxi?
어디서 택시를 탈 수 있을까요?
[eodiseo taxireul talsu isseulkkayo?]

Where is …? … 이 어디 있습니까? […i eodi isseumnikka?]	Please take me to the …. … 으로 가주세요. […euro gajuseyo.]

Does this bus go to …?
이 버스 …갑니까?
[Ee beoseu … gamnikka?]

Haw far is it to …?
…까지 얼마나 멉니까?
[…kkaji eolmana meomnikka?]

Please stop here.
여기서 세워주세요.
[yeogiseo sewo-juseyo.]

Shopping

Please show me this.
이것을 보여 주세요.
[igeoseul boyeo juseyo.]

I want this.
이것을 주십시요.
[igeoseul jusipsiyo.]

How much is it?
그것은 얼마입니까?
[geugeoseun eolma-imnikka?]

Do you take credit card?
신용카드 받습니까?
[sinyong kadeu batseumnikka?]

Eating Out

May I see the menu, please?
메뉴 좀 보여주세요.
[menyu jom boyeo-juseyo.]

I would like to have (bulgogi).
(불고기) 주세요.
[(bulgogi) juseyo.]

What is your specialty here?
이 집에서 잘하는 음식이 무엇이죠?
[ee jibeseo jalhaneun eumsigi mueosijiyo?]

Could you bring me some more of this?
이것 조금 더 주세요.
[igeot jogeum deo juseyo.]

Numbers

0	영	[yeong]
1	일/하나	[il/hana]
2	이/둘	[i/dul]
3	삼/셋	[sam/set]
4	사/넷	[sa/net]
5	오/다섯	[o/daseot]
6	육/여섯	[yuk/yeoseot]
7	칠/일곱	[chil/ilgop]
8	팔/여덟	[pal/yeodeol]
9	구/아홉	[gu/ahop]
10	십/열	[sip/yeol]
20	이십	[isip]
100	백	[baek]
1000	천	[cheon]

Miscellaneous

Lost & Found Center	분실물보관소	[bunsilmul bogwanso]
Hospital	병원	[byeongwon]
Police Station	경찰서	[gyeongchalseo]
Toilet	화장실	[hwajangsil]
Pharmacy	약국	[yakguk]
Inn	여관	[yeogwan]
Market	시장	[sijang]
Restaurant	식당	[sikdang]
Airport	공항	[gonghang]
Subway	지하철	[jihacheol]
Railroad Station	기차역	[gichayeok]

Sponsors


NRF
National Research Foundation of Korea
한국연구재단


두뇌한국
BK21


iTRC

Notes