EDITORS:

Sophia Ananiadou (University of Salford, UK)
Jun'ichi Tsujii (University of Tokyo, Japan)

SPONSOR:

Microsoft

INVITED SPEAKER:

Carol Friedman (CUNY / Columbia University, USA)

ORGANIZING COMMITTEE:

Sophia Ananiadou (University of Salford, UK)
Udo Hahn (University of Freiburg, Germany)
Jerry Hobbs (USC/ISI, USA)
James Pustejovsky (Brandeis University, USA)
Jun'ichi Tsujii (University of Tokyo, Japan)

PROGRAM COMMITTEE:

Carol Friedman (CUNY / Columbia University, USA)
Robert Gaizauskas (University of Sheffield, UK)
Aravind Joshi (University of Pennsylvania, USA)
Goran Nenadić (UMIST, UK)
Kousaku Okubo (Kyushu University, Japan)
Christos Ouzounis (EMBL, UK)
Irena Spasić (University of Salford, UK)
Toshihisa Takagi (University of Tokyo, Japan)
Alfonso Valencia (Centro Nacional de Biotecnologia, Spain)
Limsoon Wong (Institute for Infocom Research, Singapore)
Pierre Zweigenbaum (Universite Paris 6, France)

ADDITIONAL REVIEWERS:

Christian Blaschke (Centro Nacional de Biotecnologia, Spain)
Asako Koike (University of Tokyo, Japan)
Yasunori Yamamoto (University of Tokyo, Japan)
Takako Takai (University of Tokyo, Japan)

WORKSHOP WEBSITE:

http://www-tsujii.is.s.u-tokyo.ac.jp/ACL03/bionlp.htm

PREFACE

Recently, the need for the fields of biomedicine and natural language processing to collaborate and exchange ideas has been demonstrated through the rapid emergence of SIGs, dedicated workshops and tutorials. This workshop follows on from workshops with similar aims and objectives such as ACL (2002), NLP and Ontology Building (Tokyo, 2001), ISMB (2001, 2002), PSB (2001, 2003). Our aim was to bring together NLP researchers in biomedicine to discuss recent advances in text data mining in the field.

The vast majority of knowledge in biomedicine is found in scientific literature. In biomedicine, a large information repository, Medline, contains over 11 million abstracts, and approximately 40,000 new abstracts are added each month. Although there are growing numbers of sequence databases and other hand-constructed databases, most new information is unstructured text in Medline and full text journals. The ability to have access to natural language techniques and tools that automate and facilitate the process of knowledge discovery consistently is of paramount importance.

For NLP researchers, processing biomedical texts presents many challenges such as in the areas of terminology (named entity), information extraction from texts, knowledge discovery or ontology building from large collection of documents, sharing and integrating knowledge from factual and textual data bases, semantic annotation, to name a few.

The task of named entity recognition (NER) occupies a central role in this workshop. Detecting gene names, protein names, etc., in texts is necessary for knowledge discovery in biomedicine especially as new names are constantly discovered. NER is nevertheless a challenging task, as terminology in biomedicine does not follow predefined naming patterns. The lack of standardisation, terminological ambiguity and diversity of term variation add to the problems of heterogeneous information found in biomedical databases and thesauri.

NE recognisers are based on approaches such as corpus-based, rule-based, supervised machine learning techniques using HMMs, support vector machines and decision trees, with varying degrees of performance. A step further from traditional NER involves the organisation of scientific knowledge, that is term classification. Reported classification techniques are based on statistical knowledge, HMMs, naive Bayesian learning, support vector machines, decision trees, inductive rule learning, genetic algorithms, etc. Selecting the most appropriate features, contextual or internal terminological, either automatically or manually for classification is not trivial since many features are shared by different types of entities. NER relies on existing dictionaries and annotated corpora. Researchers in biomedicine have already constructed large scale linguistic resources such as UMLS, SNOMED, Mesh, Gene Ontology (GO), etc., which can be used for knowledge-based NLP applications, intelligent IR, knowledge-triggered discovery of new scientific knowledge, etc. Nevertheless, the integration and linking of various terminological resources is a necessary step for the construction of common resources.

Past experience has shown that sharing of common resources, not only domain specific dictionaries and thesauri but also properly annotated common test/training corpora, plays a significant role in the systematic development of NLP techniques in a specific domain. Ongoing efforts in building large-scale corpora annotated with biomedical information include the semantically annotated corpus GENIA. Semantic annotation of biomedical texts for information extraction, updating of ontologies and merging of heterogeneous sources of information have used techniques such as machine learning, rule-based approaches and approximate string matching.

NLP in biomedicine combines theoretical issues with applications. Many researchers have worked on protein interactions based on NER, information retrieval and ontology development. Medical applications of NLP technology reported in this workshop include the detection of nosocomial pneumonia in infants based on clinical rules extracted from narrative reports, a medical question answering system based on role identification, and the identification of patients with heart failure by automatically classifying medical notes.

What is lacking in NLP in biomedicine is standardisation of terminological resources, agreement on the annotation standards, evaluation metrics and initiatives similar to TREC and MUC for the biomedical domain. For these purposes interaction, between the two research fields is crucial and we hope that this workshop contributed to this goal.

We would like to thank the authors who submitted their work, the PC members and reviewers for their time and invaluable suggestions to the authors. We also thank J.Kim, T. Ohta, I. Spasic, Y. Tateishi, and Y. Tsuruoka who compiled the proceedings and set up the web site for the workshop.

The Organizing Committee
June 2003