COLING 2004 Post-Conference Workshop

JNLPBA

Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications

Edited by
Nigel Collier
Patrick Ruch
Adeline Nazarenko

Supported by
The SemanticMining Network
of Excellence (EU-IST 507505 - OFES 03.3399)

August 28-29, 2004
University of Geneva, Switzerland

Preface

Recent years have seen a growing interest in the application of NLP techniques to texts in the domains of biology and medicine. The problem of information overload that has resulted from the massive growth in the scientific literature has clearly shown the necessity to automatically locate, organize and manage facts relating to experimental results. At the same time clinicians have experienced greatly improved access to the medical literature and clinical repositories which needs to be matched by the development of enhanced information access tools. This year NLPBA (http://www.genisis.ch/~natlang/NLPBA02/) and BioNLP (http://www-tsujii.is.s.u-tokyo.ac.jp/ACL03/bionlp.htm) have merged to form a joint workshop with the aim of bringing together researchers from natural language processing, bio-informatics, medicine and ontologies who are concerned with developing methods and resources for solving these problems.

Over the last five years we have seen significant steps forward in the development of language technology and large-scale resources for the Bio-Medical domain such as linguistically annotated corpora (e.g. GENIA POS and NE corpora), ontologies (e.g. Gene Ontology), thesauri (e.g. UMLS Metathesaurus), lexicons and term lists (e.g. UMLS SPECIALIST) as well as information retrieval collections (e.g. TREC Genomics track). At the application level we see development of question answering systems, event recognition, zone (rhetorical region) identification, as well as term and bio-entity recognition. The demand for information access tools from domain users is increasing to support literature survey, often integrated into online ‘portals’ where scientists can navigate through related information resources such as genetics and disease databases. Ongoing challenges relate to the growing and ambiguous nomenclature, the need to integrate deep knowledge sources into machine learning, a need to scale up methods for processing full text articles etc.

The objective of the workshop is to bring together researchers in this area, to establish common themes and goals between different groups. We have seen from previous experience in the natural language learning and information retrieval communities the benefits of sharing resources and developing common evaluation criteria. In this workshop we are introducing a special shared task to promote discussion of these issues as well as the objective of integrating machine learning with knowledge resources.

In getting the workshop program finalized we are very grateful to our program committee for their many efforts under a short time schedule. Also we acknowledge the kind support of the COLING-2004 workshop and local organizers as well as the GENIA group at University of Tokyo for their hard work organizing the shared task. Finally, we would like to thank all the authors who submitted papers to the workshop and for helping to give us such a wealth of choice in the final program.

Nigel Collier
Patrick Ruch
Adeline Nazarenko

JNLPBA Committees

Workshop Co-Chairs

- Nigel Collier (National Institute of Informatics, Japan)
- Patrick Ruch (University Hospital of Geneva and EPFL, Switzerland)
- Adeline Nazarenko (LIPN, France)

Steering Committee

- Alfonso Valencia (Centro Nacional de Biotecnologia, Spain)
- Carol Friedman (CUNY/Columbia University, USA)
- Donia Scott (University of Brighton, UK)
- Udo Hahn (Albert-Ludwigs University, Freiburg, Germany)
- Junichi Tsujii (University of Tokyo, Japan)

JNLPBA Committees (continued)

Program Committee

- Sophia Ananiadou (University of Salford, UK)
- Alan Aronson (National Library of Medicine, USA)
- Robert Baud (University Hospital of Geneva, Switzerland)
- Christian Blaschke (CNB, Spain)
- Oliver Bodenreider (National Library of Medicine, USA)
- Berry de Bruijn (National Research Center, Canada)
- Marc Craven (University of Wisconsin, USA)
- Robert Gaizauskas (University of Sheffield, UK)
- Eric Gaussier (Xerox, XRCE, France)
- Vasileios Hatzivassiloglou (Columbia University, USA)
- Lynette Hirschman (MITRE, USA)
- Dimitar Hristovski (University of Ljubljana, Slovenia)
- Jerry Hobbs (USC/ISI, USA)
- Aravind Joshi (University of Pennsylvania, USA)
- Su Jian (Institute for Infocomm Research, Singapore)
- Asao Fujiyama (National Institute of Informatics, Japan)
- Arne Jönsson (University of Linköping, Sweden)
- Frédérique Lisacek (GeneBio SA, Switzerland)
- Yuji Matsumoto (NAIST, Japan)
- Claire Nédellec (INRA, France)
- Kousaku Okubo (Kyushu University, Japan)
- Jong C. Park (KAIST, Korea)
- Thierry Poibeau (LIPN, France)
- Denys Proux (Xerox, XRCE, France)
- James Pustejovsky (Brandeis University, USA)
- Dietrich Rebholz-Schuhmann (European Bioinformatics Institute, EU)
- Irena Spasic (UMIST, UK)
- Ben Stapley (UMIST, UK)
- Padmini Srinivasan (University of Iowa, USA)
- Hirotoshi Taira (NTT Communication Science, Japan)
- Toshihisa Takagi (University of Tokyo, Japan)
- Yuka Tateishi (University of Tokyo, Japan)
- Anne-Lise Veuthey (SIB, Switzerland)
- Limsoon Wong (Institute for Infocomm Research, Singapore)
- Pierre Zweigenbaum (AP-HP, INSERM & INaLCO, France)

Conference Program

Saturday, August 28th, 2004

8:30-9:15 On site Registration
9:15-9:30 Introduction

Regular session 1

9:30-10:00 *Recognizing Names in Biomedical Texts using Hidden Markov Model and SVM plus Sigmoid*
GuoDong Zhou

10:00-10:30 *Using Argumentation to Retrieve Articles with Similar Citations from MEDLINE*
Imad Tbahriti, Christine Chichester, Frédérique Lisacek and Patrick Ruch

10:30-11:00 *Analysis of Link Grammar on Biomedical Dependency Corpus Targeted at Protein-Protein Interactions*
Sampo Pyysalo, Filip Ginter, Tapio Pahikkala, Jorma Boberg, Jouni Järvinen, Tapio Salakoski and Jeppe Koivula

11:00-11:30 BREAK

Regular session 2

11:30-12:00 *Discovering Patterns to Extract Protein-Protein Interactions from Full Biomedical Texts*
Minlie Huang, Xiaoyan Zhu, Donald G. Payan, Kunbin Qu and Ming Li

12:00-12:30 *Zone Identification in Biology Articles as a Basis for Information Extraction*
Yoko Mizuta and Nigel Collier

12:30-14:00 LUNCH

14:00-15:00 **Invited talk**

15:00-16:15 **Poster session**

Distributed Modules for Text Annotation and IE Applied to the Biomedical Domain
Harald Kirsch and Dietrich Rebholz-Schuhmann

Support Vector Machine Approach to Extracting Gene References into Function from Biological Documents
Chih Lee, Wen-Juan Hou and Hsin-Hsi Chen

Improving the Identification of Non-Anaphoric it using Support Vector Machines
José Carlos Clemente Litrán, Kenji Satou and Kentaro Torisawa

Creating a Test Corpus of Clinical Notes Manually Tagged for Part-of-Speech Information

Serguei Pakhomov, Anni Coden and Christopher Chute
Classification from Full Text: A Comparison of Canonical Sections of Scientific Papers
Gail Sinclair and Bonnie Webber

Regular session 3

16:15-16:45 *Assessing the Correlation between Contextual Patterns and Biological Entity Tagging*
M. Krallinger, M. Padr?n, C. Blaschke and A. Valencia

16:45-17:15 *Event-Based Information Extraction for the Biomedical Domain: the Caderige Project*
Erick Alphonse, Sophie Aubin, Philippe Bessières, Gilles Bisson, Thierry Hamon, Sandrine Lagarrigue, Adeline Nazarenko, Alaine-Pierre Manine, Claire Nédellec, Mohamed Ould Abdel Vetah, Thierry Poibeau and Davy Weissenbacher

17:15-17:45 **Round table and closing**

Sunday, August 29th, 2004

8:30-9:30 On site registration

9:30-10:00 *Introduction to the Bio-entity Recognition Task at JNLPBA*
Nigel Collier and Jin-Dong Kim

Shared task session 1

10:00-10:15 *Incorporating Lexical Knowledge into Biomedical NE Recognition*
Kyung-Mi Park, Seon-Ho Kim, Ki-Joong Lee, Do-Gil Lee and Hae-Chang Rim

10:15-10:30 *Annotating Multiple Types of Biomedical Entities: A Single Word Classification Approach*
Chih Lee, Wen-Juan Hou and Hsin-Hsi Chen

10:30-10:45 *Named Entity Recognition in Biomedical Texts using an HMM Model*
Shaojun Zhao

10:45-11:00 *Exploiting Context for Biomedical Entity Recognition: From Syntax to the Web*
Jenny Finkel, Shipra Dingare, Huy Nguyen, Malvina Nissim, Christopher Manning, and Gail Sinclair

11:00-11:30 BREAK

	Shared task session 2
11:30-11:45	*Adapting an NER-System for German to the Biomedical Domain* Marc Rössler
11:45-12:00	*Exploring Deep Knowledge Resources in Biomedical Name Recognition* Zhou GuoDong and Su Jian
12:00-12:15	*POSBIOTM-NER in the Shared Task of BioNLP/NLPBA2004* Yu Song, Eunju Kim, Gary Geunbae Lee and Byoung-kee Yi
12:15-12:30	*Biomedical Named Entity Recognition using Conditional Random Fields and Rich Feature Sets* Burr Settles
12:30-13:00	**Discussion and closing**

Table of Contents

Regular Papers

Short Papers

Shared Task Papers

Author Index