Tutorial title: BioNLP

Speaker: K. Bretonnel Cohen

Description of the tutorial topic and content

This tutorial will provide general natural language processing specialists with an introduction to the field of “BioNLP”—natural language processing in the fields of medicine and biology. This field has long roots in the history of natural language processing, but has been an absolutely burgeoning area of interest in recent years. The past few years have been characterized by an unusual mixing of bioinformatics and NLP specialists at the conferences of both communities: ACL or NAACL has now hosted workshops on BioNLP every year since 2002, with excellent attendance numbers, and bioinformatics and medical informatics meetings have featured NLP papers, sessions, and SIG meetings since the late 1990s. Recent MUC-like and TREC-sponsored shared tasks have had some unusual results, and the implications of these findings should make for an interesting tutorial for the general NLP specialist.

BioNLP presents unique challenges in a number of areas, ranging from low-level processing tasks—tokenization and sentence boundary detection are demonstrably different tasks in biomedical publications than in newswire text—to high-level conceptual issues, such as theoretical issues in predicate-argument structure representation, which have been a topic of much discussion in recent work in the field. Despite the many challenges that are unique to biomedical text, most of the sub-topics of NLP are the subject of current research in the BioNLP community—information retrieval, named entity recognition, information extraction, text classification, semantic role labelling, coreference resolution, question-answering, parsing, morphological analysis, and discourse analysis. Thus, there are interesting challenges in BioNLP for almost anyone working in natural language processing.

One unique advantage to the field of BioNLP is the wide availability of resources, including an enormous body of freely available text. The tutorial will include an overview of a variety of publicly available BioNLP resources, including:

- A variety of domain-specific ontologies, including the popular Gene Ontology
- Corpora, including the popular GENIA corpus and a number of less-well-known but valuable corpora and text collections, some of them featuring full text

One potential stumbling block in the field of BioNLP is the requirement for domain knowledge. The tutorial will include a brief overview of just enough biology to enable the NLP specialist to comprehend the topics under discussion in typical biomedical texts, if not enough to follow the specifics of the discussion.

The core of the tutorial will be an overview of current “hot topics” in BioNLP, including:

- Recent shared tasks and their results
- High-arity relations in information extraction, semantic representation, and semantic role labelling
- Modelling certainty and speculation
- Portability from general-domain to domain-specific tasks, and between sub-domains

Brief outline of the tutorial structure

1. Just enough biology for BioNLP

- genes, proteins, and cells

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL pages 1–2,
Rochester, April 2007. ©2007 Association for Computational Linguistics

- genotypes, phenotypes, and high-level structures

2. Why bioscientists fund and publish research in BioNLP

- clinical versus genomic NLP
- the genomic and high-throughput revolutions in biology
- three different biomedical user types and their different text mining needs

3. Some things that make BioNLP different

- issues in tokenization, semantic normalization, word sense disambiguation, and named entity recognition, and how to approach them successfully
- special challenges in biomedical corpus construction

4. Getting up to speed: 10 essential papers and resources that will allow you to read most other papers in BioNLP

- Fukuda, Collier, and Hirschman on named entity recognition
- Blaschke, Craven, and Friedman on information extraction
- The Gene Ontology, Entrez Gene, the GENIA corpus, and the PubMed/MEDLINE database

5. Shared tasks: recent MUC-like and TREC evaluations and their sometimes surprising results

- the KDD Cup in 2002
- JNLPBA in 2004
- BioCreative 2004 & 2006
- the TREC Genomics track, 2003-2006

6. Current hot topics in BioNLP:

- the right model for biomedical semantic representation
- modelling certainty, speculation, and other aspects of discourse structure in scientific texts
- trends in Open Access publishing and their implications for BioNLP: issues in dealing with "full text"
- portability of tools, grammars, and resources
- true integration of NLP into laboratory data interpretation

Intended audience

The intended audience is general NLP specialists. No prior background in the biomedical domain is assumed.

Speaker bio

Kevin Bretonnel Cohen
kevin.cohen@gmail.com
http://compbio.uchsc.edu/Hunter_lab/Cohen
Phone number: 303-916-2417

Kevin leads the Biomedical Text Mining Group at the Center for Computational Pharmacology in the University of Colorado School of Medicine. He has been involved in biomedical NLP in the industrial and academic worlds since 1997. He has worked in both the clinical and the genomic fields, on technologies including information extraction, corpus construction, statistical language modelling for speech recognition, named entity recognition, and computational lexical semantics. He has organized several workshops and conference sessions on BioNLP at ACL, NAACL, and bioinformatics meetings, and has presented tutorials on BioNLP for non-NLP specialists at the Pacific Symposium on Biocomputing, the University of Colorado at Denver Center for Computational Biology, and (this spring) at the Medical Library Association Annual Meeting.