

# Recent Advances in Natural Language Processing for Biomedical Applications

## 1. Introduction

**Natural Language Processing (NLP)** is a subfield of artificial intelligence. It studies the problems inherent in the processing and manipulation of natural language, and, natural language understanding devoted to making computers "understand" statements written in human languages. Historically, it covers a wide range of academic areas, including statistics, logics and linguistics, as well as different concrete applications such as information retrieval and speech recognition. From an applicative point of view, biomedicine can be seen as an ideal playground considering that knowledge-intensive resources (such as lexicons, terminologies and ontologies...) and data-intensive resources (corpora, annotated corpora, user information requests...) are available in the field. We here survey some of the most recent developments in the domain, focusing on issues which were discussed at the last NLPBA workshop (<http://www.genesis.ch/~natlang/NLPBA04/>).

## 2. A Cross-over Domain

The idea to apply natural language processing (NLP) methods to the medical and molecular biology domains is a shared vision that has emerged in several different places. Already in 2002, the European Commission edited a white paper on potential synergies between medical informatics and bio-informatics. The topic was selected by the American Medical Informatics Association (AMIA) for its Fall Symposium the same year. The next summer, the European Federation for Medical Informatics (EFMI) congress Medical Informatics Europe (MIE) 2005 announced with the sub-title: The new challenge, Merging Medical Informatics and Bio-Informatics. Moreover, the NLP and information retrieval (IR) communities have offered forums or hosted events for reaching out to the medical informatics/bio-informatics communities: the first dedicated Workshop on NLP in the biomedical domain was probably organized in 1999, soon followed by related events such as the first editions of NLPBA<sup>1</sup>, as a satellite workshop of the EFMI Special Topics conference in early 2002 and BioNLP<sup>2</sup>, as an ACL Workshop in 2003. In parallel, a track dedicated to investigating information extraction and retrieval was set up in 2003 as part of the Text Retrieval Conferences (TREC) [Hersh and Bhupatiraju 2003], with a pre-track organized in 2002.

The bio-informatics community also has a tradition of using NLP techniques and the ISMB (Intelligent Systems for Molecular Biology) and PSB (Pacific Symposium on Bio-Informatics) forum have had a regular NLP-related session in the last 6 or 7 years (NLP, knowledge discovery, data mining). Many other events could be mentioned here, but it is worth observing that following the TREC Genomics track, the biomedical text mining

---

<sup>1</sup> <http://www.genesis.ch/~natlang/NLPBA02/>

<sup>2</sup> <http://www-tsujii.is.s.u-tokyo.ac.jp/ACL03/bionlp.htm>

community started to organize shared evaluations in order to establish a consensus on standards and to share experience of resource re-use. Noticeable events such as the Knowledge Discovery and Data Mining Cup in 2002 [Yeh et al. 2002], the BioCreative challenge in 2003 [Hirschman et al. 2005] and the JNLPBA shared task in 2003 [Kim and al. 2004] have undoubtedly led to progress in our fields. The KDD Cup task 1 focused on a curation centred task by determining the weight of experimental evidence for whether an RNA and/or protein is associated with a gene. The JNLPBA shared task was focused entirely on named entity recognition in the pure tradition of the information extraction community confirming our intuitions across domains that the bio-entity mining is somehow 'harder' than in newswire. The BioCreative challenge on the other hand attempted to extend the traditional paradigm in two tasks that included entity mention normalization with respect to identifiers from an organism-specific database and assignment of gene ontology annotations. The second BioCreative task led to systems adapting text categorization methods, as well as passage retrieval in a fairly synthetic and original fashion, which aims at assisting database curation in proteomics.

Scientific meetings are numerous and adding just another one to the list, while gratifying to the organizers, is not sufficient. But, trying to bring to the front the emerging trends of research, looking forward to new techniques, tools or methods susceptible to cross-fertilize different domains, imagining synergies between researchers from different backgrounds and different scientific cultures, was the motivation for a joint NLPBA and BioNLP call for papers in Spring 2004. Our original conception was for a small-scale workshop with the idea of combining on the one hand a shared task forum, where people could discuss well-known named-entity recognition learning methods and on the other hand, a more open scientific forum dedicated to presenting some of the most promising emerging research trends in biomedical text mining. Despite the time constraints, the workshop attracted over sixty participants in two days of lively discussion. The success of the workshop owes much to the excellent local organization by the COLING steering committee, with the support of the SemanticMining<sup>3</sup> European Network of Excellence and to the help of all those who participated in the review process and the running of the shared task.

### **3. Paper selection**

31 papers were submitted in response to the workshop's call for paper from which the international scientific committee selected 7 to be accepted for oral presentation, with 5 others presented as posters. From this paper set, 8 papers have been extended and re-reviewed for the present publication. The reviewing process has been efficient and selective with a 25% acceptance rate. Most of the decisions on selection of papers have been unanimous.

However, the challenge of grouping people from text mining in medical informatics and in molecular biology is not only a matter of a decision when starting a call for papers: just

---

<sup>3</sup> <http://www.semanticmining.org/>

having the intention is not enough. Each author was asked to consider what in her or his domain of activity could be of interest to other authors; the idea was to foster opportunities for communications between different areas. The result or the lessons from this event are positive but not sufficient. They are positive because the papers presented and published here draw a roadmap for synergies and common developments. They are not sufficient, because the familiarity of one's own domain is more comfortable than an adventure into another domain. Indeed, such moves necessitate thought about such basic issues as evaluation criteria which cannot be performed in a short time frame. This immediately raises a few questions, which came up repeatedly at the workshop and that we'll try to synthesize after surveying the content of the special issue.

#### **4. A guided tour**

Walking through eight scientific papers and preparing an opinionated synthesis is a dangerous exercise, open to future criticisms. The challenge is to discover some convergence or trends between authors, and if possible to see beyond the current trends to what will come next. How far any paper is from the others is the question to be answered, by adequately positioning each contribution.

The first paper, by Pakhomov, Coden and Chute, is the only paper in our selection which applies NLP to clinical corpora. This can be partly explained by the intensive focus recently on molecular biology, and partly by the usual constraints on confidentiality of research applied to sensitive clinical corpora. Interestingly, the main goal is to construct a corpus of clinical texts manually annotated with part-of-speech information. The annotation process and its challenges regarding inter-annotator agreement are discussed. The authors underline the importance of developing domain-specific NLP tools to process clinical narratives. More specifically, they observe that the medical language is far from homogeneous, so that discriminating between different types of discourse represented by different sections of clinical text may be very beneficial to improve the correctness of POS tagging.

Pyysalo, Ginter, Pahikkala, Boberg, Järvinen and Salakoski also study the specificity of the biomedical language, but concentrate on proteomics and protein interactions rather than on clinical corpora. Like Pakhomov and al., they also test general-purposes syntactic tools –including commercial solutions. Although they are interested in deep parsing rather than in part-of-speech tagging, they also conclude that specific adaptations are necessary to achieve state-of-the-art performances with biomedical corpora. They highlight a number of challenges including the need to re-use bio-entity taggers as tokenizers, missing domain-specific senses of words in the lexicon, unknown grammatical structures and also a high than expected ungrammatical sentences .

Huang, Zhu and Li present an application of dynamic programming for pattern extraction to identify relations in molecular biology. They study the impact of combining a shallow syntactic parser together with simpler pattern matching methods to identify long distance

dependencies, such as coordination phenomena, which are particularly complex to handle in biomedical articles.

Also incorporating morpho-syntactic features, Zhou proposes an interesting learning method which combines a maximal-margin classifier with a linear language model. The resulting combination, achieves high performance classification effectiveness for gene and protein named-entity recognition.

Going beyond syntactic or intra-sentence models, Mizuta, Korhonen, Mullen and Collier introduce a new annotation scheme, based on previous work initiated by Teufel and Moens [Teufel and Moens 1999] for automatic text summarization, by describing rhetorical zones in scientific texts. They demonstrate the relevance of their proposed argumentation typology for describing the novelty characteristics of result mentions in biomedical articles on a carefully analyzed set of 20 articles. The authors also provide a detailed analysis of the salient features attached to each rhetorical class.

Directly related to Mizuta and al.'s work on discourse analysis, Tbahriti, Chichester, Lisacek and Ruch's paper presents a practical application of an induced four-class argumentative model to perform information retrieval in MEDLINE. The authors report that retrieval effectiveness can be significantly improved by boosting implicit purpose and conclusion's sentences in scientific abstracts. In this paper, co-citation networks are used to automatically built a set of relevance judgements.

Kirsch and Rebholz-Schuhmann presents an end-user integration effort to unite available databases in genomics and proteomics via NLP, which can be seen as a hypothetical testbed to incorporate various NLP services. In the proposed system, the user navigation starts with a text -for instance a short abstract- which will be then automatically linked to various knowledge repositories, including the Gene Ontology or the Uni-Prot database via a cascaded set of NLP filters. The system can detect protein-protein interactions, mutations and should help speeding up the database curation process.

## ***5. Potential for convergence and limitations between medical informatics and bio-informatics***

This event has brought together different scientists from separate domains and multiple locations and continents. It has shown two points: first, the methods we use are largely similar and are nearly all candidates for migration from one group to another; second, the scientific community is waiting more and more for intelligent text analysis, data mining and knowledge representation.

Taking advantage of our editorial position, we believe that a major challenge for future achievements in biomedical NLP is to share clinical contents<sup>4</sup>. Indeed, apart from some rare groups located in large healthcare institutions, access to clinical narratives and

---

<sup>4</sup> We observe that in a scientific world becoming more and more global, multilingual and cross lingual issues, which are traditional when processing clinical data, were remarkably absent from the debates.

patient data is usually impossible. Therefore, NLP in the medical domain tends to remain both a protected area and a field where state-of-the-art techniques might penetrate more slowly than elsewhere. In contrast, NLP for bioinformatics, and genomics/proteomics in particular, became recently a very attractive research area even for non-domain experts, so that the field can be legitimately regarded as a privileged domain for developing and testing the most advanced methods and the most speculative hypothesis, leading to the impressive achievements.

Regarding further in the future, and trying to draw a parallel with what has happened with the World Wide Web, the next steps will demand us to seriously address semantic interoperability and data interchange. Interoperability will become of major importance to link textual contents and online biomedical databases, such as Swiss-Prot/TrEMBL or OMIM, which could in return also serve to foster synergy between clinical and biological interests. This means that working with heterogeneous contents, such as structured data and non-structured texts, but also with annotated or semi-annotated images [Clough and al. 2005], will become a higher priority issue for researchers. In such an integrated perspective, controlled vocabularies and related knowledge structuring efforts, such as the Gene Ontology, the Medical Subject Headings or the SNOMED classification, which are now all part of the Unified Medical Language System (UMLS) are rapidly becoming strategic resources for text miners.

From a data analysis point-of-view, the gap between Medical Informatics and Bioinformatics could be merged more easily by using dedicated terminology and entity recognition tools. There is certainly a necessity to build a basic domain-specific infrastructure at first. Named entities acquisition and issues related to terminological variation are basic prerequisites, but should not as such absorb a too large fraction of the efforts in the community. More cognitive and semantically driven tasks must follow to satisfy users' information access needs, with an obvious priority to be given to information retrieval from different perspectives: functional genomics and proteomics, entity interactions, phenotypes and clinical findings... It is clear that these will require the emergence of new trends such as using discourse analysis to perform deeper extraction tasks. In parallel, the development of elaborated concept representations, the quest for useful and broad-coverage ontologies is all strong concerns for both disciplines. Here lies the place for convergence. Finally, the study of medical processes in the human body is dependent on published papers and patient records, which are the two sources of medical texts. Only when a convenient coverage of both aspects is realized, will we be in a position to mix in a single record the bio-medical information as a causative, informative and therapeutic active agent, and the medical observations of individuals, their prognosis and outcomes. This should benefit both researchers and patients.

### ***Acknowledgements***

Warm thanks to the reviewers and co-organizers for their determination to accomplish a good job in a very constrained time, thus allowing this special issue to be published within 1 year of the seminal call for papers: Alfonso Valencia (Centro Nacional de Biotecnología, 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), 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), Hirotoishi 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).

## References

W Hersh, R Bhupatiraju (2004) TREC genomics track overview, *The Twelfth Text Retrieval Conference - TREC 2003*, 14-23.

L Hirschman, A Yeh, C Blaschke, A Valencia (2005) Overview of BioCreAtIvE: critical assessment of information extraction for biology *BMC Bioinformatics* 2005, 6(Suppl 1)

J Kim, T Ohta, Y Tsuruoka, Y Tateisi and N Collier (2004) Introduction to the Bio-Entity Task at JNLPBA. Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA).

S Teufel and M Moens (1999) Argumentative classification of extracted sentences as a first step towards flexible abstracting. In: I. Mani, M. Maybury (eds.), *Advances in automatic text summarization*, MIT Press.

P Clough, H Müller and Mark Sanderson (2004) The CLEF Cross Language Image Retrieval Track (ImageCLEF) 2004, CLEF Proceedings - Springer Lecture Notes in Computer Science, 2005.

A S Yeh, L Hirschman and A A Morgan (2002) Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup, *J. Bioinformatics*, v. 19 Suppl 1, i331-i339, Oxford University Press.