ELSEVIER

International Journal of Medical Informatics 67 (2002) 1–5

International Journal of Medical Informatics

www.elsevier.com/locate/ijmedinf

Introduction

The future of natural language processing for biomedical applications

1. Once upon a time

The idea to convey researchers who applied natural language processing (NLP) methods to the medical domain and others who applied such methods to the bio-informatics domain is shared at different places. The European Commission will edit a white paper on potential *synergies between medical informatics and bio-informatics* before the end of 2002. The American Medical Informatics Association (AMIA) has selected this theme for its Fall Symposium 2002. The European Federation for Medical Informatics (EFMI) congress MIE 2005 in Geneva has been announced with the sub-title: *The new challenge, Merging Medical Informatics and Bio-Informatics*. Moreover, NLP and information retrieval (IR) communities have offered forums or hosted events for reaching out to the medical informatics/bio-informatics communities: Workshop on NLP in the biomedical domain at ACL 2002 in Philadelphia or genomics pretrack investigated by the TREC community. The bio-informatics community has also a tradition of using NLP techniques and the Pacific Symposium on Bio-Informatics (PSB) conference has had a regular NLP-related session in the last 4 or 5 years (NLP, knowledge discovery, data mining). Professor Tsujii in Tokyo hosted a workshop on NLP and Ontology Building for Biology in February 2002. Many other events cannot be mentioned here. In this context, the workshop held in Nicosia, Cyprus, as a special topics conference of EFMI in March 2002, could not miss the opportunity of being one of the precursors of this promising new research direction.

Scientific meetings are numerous and adding just another one to the list is not challenging. But, trying to bring to the front the emerging trends of research, looking forwards for new techniques, tools or methods susceptible to cross-fertilize different domains, imagining synergies between researchers of different origin and different scientific culture, have been motivations for a new call for papers in October 2001. A small-scale workshop has been devised. A very constrained time frame of 5 months from the call to the conference was pushing all the actors. An excellent local organization by the Cyprus Medical Informatics Association was another ingredient for success.

2. Paper selection

An International Scientific Program Committee has been elected. 28 papers have been submitted, 15 have been accepted for oral presentation, 14 were finally presented, and finally nine papers have been retained by the Guest Editors for the present publication. The reviewing process has been efficient and without compromise: scientific quality was the

1386-5056/02/$ - see front matter
PII: S1386-5056(02)00081-3

only criterion. Most of the decisions have been unanimous. Warm thanks to the reviewers and guest editors 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.

However, the challenge of grouping people from text mining in medical informatics and in bio-informatics, is not only a matter of a decision when starting a call for papers: just the intent 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 quietness of one's own domain is more comfortable that an adventure in another domain. Indeed, such moves necessitate thought and evaluation and cannot be performed in a short time frame. This immediately raises the question of a second conference pursuing the very same objectives.

3. A guided tour

Walking through nine scientific papers and preparing a synthesis is a dangerous exercise, open for future criticisms. The challenge is to discover some convergence or trends between authors. How far any paper is from the others is the question to be answered, by adequately positioning each contribution? In the following lines, only its first author will be mentioned for each paper.

1) This tour starts with a contribution from De Bruijn, who gives a review paper on literature mining applied to biomedicine. He asks about automatic reading, seen from a NLP researcher's point of view. He provides us with an updated valuable review of the state-of-the-art literature (79 references). Judiciously, he balances between knowledge-intensive NLP methods and statistical methods, both to hold their own place. In order to structure the debate, he defines a schematic modular approach in four steps: categorization of documents, named entity recognition, fact extraction and collection wide analysis. This review helps to appropriately position any method proposed or developed elsewhere.
2) The paper from Hatzivassiloglou describes a system to automatically extract interaction relationships among proteins and genes from the published literature. To do that, it concentrates on verbs and their subjects and objects, in order to determine if it is an interaction verb or not. He exploits statistical evidences, genes or proteins names (GPN) recognition, as well as verbal frames. The results are carefully compared with manually tagged interaction verbs. The author reports that his system is able to augment and adapt manually built knowledge bases of interaction verbs and relationship patterns, which are of prime importance to conduct deeper knowledge acquisition and fact extraction.
3) The next paper from Nenadic is oriented toward terminology driven literature mining and knowledge acquisition, in order to retrieve knowledge that is 'buried' in a text and to present the distilled knowledge to users in a concise form. Heterogeneity and constant evolution of knowledge sources (mainly proceedings and journal articles) set a challenge to systems designed to assist users in locating

and integrating knowledge relevant to their needs. After an overview of related works, the author presents TIMS, a terminology driven system exploiting NLP techniques. The following steps are recognized: a collection of documents is linguistically processed, the collection is terminologically analyzed, the user formulates a query, which is executed against the collection and then relevant text is highlighted. The lack of naming standards in biomedicine brings two problems, which are term ambiguities and term variations, making automatic term recognition (ATR) a non-trivial problem. An evaluation of the method closes this paper, concluding that an efficient methodology facilitates knowledge extraction.

4) Franzen then comes with the necessity of detecting named entities (GPN) as a first step towards higher levels of analysis. He devises a system with combination of heuristic pattern matching techniques and full syntactic analysis. He then insists on the design and application of a convenient evaluation system. In bio-molecular biology, the named entities present a number of problems: variant structural characteristics, somewhat unclear status of names, specific text domain, absence of standard for coining of names, proliferation of synonym names, multiple word names. The proposed system makes use of trigger terms (more than 50 found) as indicators of the presence of protein names, lexical analysis of core terms, filtering and parsing for noun phrases considered as potential protein sites. The good results are claimed to be due to the syntactic analysis capabilities, generally not present in other systems.
5) The paper of Hahn is targeted at the automatic extraction of relevant information directly from documents. He achieves an open architecture, not restricted by human fed templates. Wanting to automate the acquisition process, he develops incremental concept learning routines, by integration of comprehensive medical knowledge base, like UMLS. Incremental learning is based on linguistic indicators like syntactic case, apposition or comparative, leading to a ranked list of concept hypotheses. He takes care to provide us with an evaluation of performance of the system. Reengineering UMLS has been the source for a large terminological knowledge base in anatomy–pathology.
6) The next paper from Ruch considers that spelling errors are a major challenge for most information retrieval systems: any word-based systems are affected by data corruption. The quality of a published paper cannot be compared with the text of the patient records, most of them not being read outside of the institution. Real information retrieval and text mining tasks conducted on patient records imply the design of a system able to handle misspellings. While basic spelling correction is realized by computing a string edit distance between a given token and the items of a lexical list, the author capitalizes on a fully-automatic spelling checker, which returns the good candidate at the top of the list with a probability of 96%. Results show that with this improved spelling correction, the retrieval degradation is limited to a few percent only (less than 5%).
7) Improving the consistency of existing terminologies using linguistic phenomena to represent similar lexical or semantic features in the constituent terms of a vocabulary is another challenge addressed by Bodenreider. The idea is to use adjectival modifiers, which usually introduce a hyponymic relationship. He proposes an

unsupervised method to detect inconsistencies, which can support and focus the effort of human editors of a medical vocabulary. The author develops a method based on pairs of frequently co-occurring modifiers of terms, like acute–chronic, unilateral–bilateral, primary–secondary and acquired–congenital and reports on its usefulness. The discussion shows that the method is effective at automatically identifying potential inconsistencies in terminological resources.

8) Volk presents a concept-driven approach for cross-language IR, which exploits semantic information (thesauri) to bridge the gap between surface linguistic form and meaning. It is appropriate for domains and languages for which extensive multilingual semantic resources are available, such as UMLS in the medical domain. The linguistic tools are: Part of speech tagging, morphological analysis, compound word resolution, chunking of noun phrases and phrase recognition. The author provides evaluations of different approaches, working in English and German. It emerges that the MeSH codes are the most useful indexing features for both languages. High quality linguistic analysis is crucial for a good retrieval performance.
9) Finally, Zweigenbaum addresses the problem of unaccented words in diacritic languages, like French, Spanish or German, giving rise to spurious ambiguities, which are already pervasive in NLP systems. The issue is to find a method for (semi-) automatic accentuation. As a general statement, the error rate resulting from leaving unknown words unaccented accounts for one half of the total error rate. Of course, if lexicon-based accentuation is relatively simple, nearly no reference about accentuation of unknown words has been found. Methods may be either heuristic methods or statistical methods. The latter have been examined by the author, who reports useful results on a target data, after careful tuning of the system on a training data set. Future improvements are planned.

On a global map of this NLPBA workshop, De Bruijn gives a review paper oriented toward bio-informatics. Hatzivassiloglou, Nemadic and Franzen examined the specific aspects of bio-informatics, whereas Hahn and Volk addressed the counterpart in medical informatics. Bodenreider, Ruch and Zweigenbaum present useful techniques applicable to both domains.

Syntactic and lexical approaches are shared by both disciplines. Semantic resources are more advocated in medical informatics, because there exist resources like UMLS. But similar or new sources are expected to spread in the domain of bio-informatics (for example the Gene Ontology recently published). Finally, multilingual and data quality issues are an important matter in the medical domain.

4. Potential for convergence between medical informatics and bio-informatics

This first event of NLPBA has brought together different scientists from separate domains and multiple locations and continents. It has shown two points: first, the used methods 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. It turns out that heuristic approaches are still popular, but statistical methods based on large corpora and auto-

matic knowledge extraction is a need for future developments.

From a data analysis point-of-view, the gap between Medical Informatics and Bio-informatics is really there, because so much is new today. There is certainly a necessity to build a basic domain-specific infrastructure at first. Named entities hunting or working with misspelling errors or absence of accents are prerequisite, but are not a goal as such. Later, more cognitive and semantic driven tasks are profiled. The hope of discovering elaborated concept representations, the quest for useful ontologies, the search of proximities of meaning, are all strong concerns for both disciplines. Here lies the place for the junction. Finally, the study of medical processes in the human body is dependant 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.

Robert Baud, Patrick Ruch
Medical Informatics Division, University Hospitals of Geneva, Geneva, Switzerland
E-mail: robert.baud@dim.hcuge.ch