Preface

NEEDS AND REQUIREMENTS IN INFORMATION RETRIEVAL AND KNOWLEDGE MANAGEMENT FOR BIOLOGY AND MEDICINE

Natural language processing (NLP) is a sub-field of computational sciences which addresses the operation and management of texts, as inputs or outputs of computational devices. As such, this domain includes a large amount of distinct topics, depending which particular service is considered. Nowadays, with the Internet spreading as a worldwide tremendous reservoir of knowledge, NLP is highly solicited by various scientific communities as a worthwhile help for the following tasks:

1. Information Retrieval, Knowledge Extraction

Human produced texts are seen as a valuable input, to be processed and transformed into representations and structures directly operable by computational systems. This type of service is highly required when human need is about a set of texts relevant for a given query (information retrieval), or when the need is to build up a machine readable structure (knowledge extraction) for further computer assisted developments. In both medicine and biology, these two aspects are crucial. Scientific literature is so abundant that only a computational setup is able to browse and filter so huge amounts of information. A plain search engine is too limited to undertake queries as complex as those which meet the researchers’ requirements. From the knowledge extraction point of view, manually constructed ontologies seldom reach more than a few concepts and relationships, because of the tremendous effort necessary to achieve such a task.

For the past decades, artificial intelligence (AI) has undertaken an important endeavor in favor of knowledge management. Its result, a set of taxonomies, i.e. knowledge classifications formalized according to a graph-based representation (sometimes simplified into a tree-based representation when hierarchical ties between knowledge items are dominant) also commonly called ontologies, are obtained at a very high cost in manpower, and human involvement. As soon as statistical techniques and new programming skills have appeared through machine learning, the AI community has attempted to automate this task as much as possible, feeding systems with texts and producing, at the other end, an ontology or something similar to it. Naturally, human intervention, for validation, or re-orientation, was still needed. But the learning techniques were operated on the front filtering task, the one seen as the most tedious and the most risky. Results looked promising and a few ontologies were initiated with such processes. However, they were incomplete: Wrong interpretations were numerous (noisy aspect), structures were scarce (silent aspect), and most of all, the linguistic-conceptual relationship was totally ignored. When the community acknowledged that, it turned toward NLP works and tools to reorganize its processes, and NLP skills in separating between the linguistic and conceptual properties of words
were of a great help. This conjunction shyly began a few years ago, but is going stronger now, since NLP tools have improved with time.

2. Knowledge Integration to Existing Devices

Knowledge integration is a logical sequel to knowledge extraction. Rejecting existing ontologies and terminological classifications is not affordable, even if they do not meet expectation, since their building is definitely expensive. A new trend has emerged from such a situation: Available digital data (but not machine operable), mostly texts and figures, is regularly mined in order to complete or correct existing knowledge structures. This explains the present success of data and text mining in AI, as well as in information systems (IS) and NLP communities. Several works have been devoted to ontologies enhancement or enrichment by adding concepts or terminology related to concepts. One has to know that in natural language, unlike mathematical originated formalisms, a given item (concept) could be addressed through several different symbols (i.e. words or phrases), these symbols not being exactly identical to each other in the way that they do not stress out exactly the same concept properties. This feature is called synonymy or mostly near-synonymy, since natural language elements are not mathematically equivalent. This has been for a long time one of the most important obstacles in building and completing ontologies for knowledge management. On the other hand, not only natural language symbols are prone to synonymy. At the same time, a given symbol (again, a word or a phrase) could address different items, illustrating polysemy, or multiple meanings phenomenon. Both aspects could be seen as terminological expansion (synonymy) or contraction (polysemy) processes. They have largely impeded automated procedures that mined texts in order to find concepts and relations. Therefore, new knowledge is as complicated to integrate to existing devices as to built a classification from scratch, if NLP techniques, that have been tackling the expansion-contraction phenomenon for long, are not solicited for such a task. As it is the case for the documents coming from the general field, the main classical linguistic phenomena can be observed when considering the biomedical corpora. Among these phenomena, polysemy (e.g., “RNA” which can mean “ribonucleic acid” or “renal nerve activity”) appears, as well as synonymy (e.g., “tRNA” and “transfer RNA”). However, it should be noted that the biomedical texts contain linguistic specificities: Massive presence of acronyms (contracted forms of multiword expressions), scientific style of documents, and so forth.

3. Using and Applying Existing Knowledge Structures for Services in Information Retrieval

Ontologies are built and extended because they are needed as basic structures for further applications, otherwise, the effort would be vain. In a kind of loop process, information retrieval appears as one of the goals of such an effort, while in previous paragraphs, it seemed to be a launching procedure for knowledge extraction. Retrieving the relevant set of texts to a complex query cannot only rely on words, it has to grasp ideas, expressed by several distinct strings of words, either phrases, or sentences. Retrieving the appropriate portion in a relevant text needs going further: It requires topical classification. Browsing a given literature in a given domain often relies on multilingual abilities: Translation has its role to play. Even if science is nowadays focused on a major tongue (English) for its publications, nevertheless, depending on domains, more than half of worldwide publications is done in other languages. Terminological translation could be of a great help also in writing down articles for non-native English speakers. All these services, knowledgeable to an important target of users, are provided by AI, statistical and NLP techniques, altogether invoked to produce the most appropriate output. Considering the
BioNLP domain, there are a lot of resources such as the corpora (e.g., PubMed: http://www.ncbi.nlm.nih.gov/pubmed/), the Ontologies/Thesaurus (e.g., GeneOntolgy: http://www.geneontology.org/, Mesh Thesaurus: http://www.nlm.nih.gov/mesh/), etc. Many chapters of this book rely on these resources in order to achieve the various NLP tasks they address.

ORIGINALITY OF OUR APPROACH

This book topic meets the knowledge extraction, integration, and application triple goal expressed in the previous description, beginning with the information retrieval task. As explained before, the extraction process for either information retrieval or knowledge management is the same, the difference lying in the output shape: Raw in information retrieval or machine operable in knowledge extraction and integration.

Two properties define this book as both totally fitting in the ongoing research trend, and original in its presentation:

1. Emphasizing Natural Language Processing as the Main Methodological Issue

NLP has appeared to a majority of researchers in the fields of complex information retrieval, knowledge extraction and integration, as the most fitting type of approach going beyond obvious statistical highways. NLP tackles linguistic data and tries to extract from linguistic properties as much information as possible. If, very obviously, non specialists think about NLP as a ‘word computing’ science, this domain goes much beyond the level of words, since language productions, i.e. texts, are complex devices structured at least at two other levels:

- **The sentence level:** Words are not haphazardly thrown in texts as dice could be on a table, unlike what would some raw statistical models in text mining implicitly assume. Words contribution to sentences determines their role as meaning conveyors. A governance principle has been defined a long time ago by scholars (going back to Aristotle), but has been rejuvenated by the works of Noam Chomsky and Lucien Tesnière in the second half of the twentieth century. Words or phrases (groups of words) may govern others, and thus interact with them, and the sentence meaning results from words interactions as much as words meanings. One of the peculiarities of this book is to demonstrate the implication of NLP in information retrieval, knowledge extraction and integration, beyond the plain terminological involvement. In other words, beyond simple words.

- **The discourse, segment or text level:** As much as sentences are intended words structures, a text is also not randomly organized. It abides by discourse rules that convey the intentions of the writer about what is written and about the potential readers. Several phenomena appear at the text level, and could highly impact the importance of text segments relatively to each other. Paragraphs positions, paragraphs beginnings, lexical markers acting as particular position flags between sets of sentences, are among the several items used by human writers to organize their texts according to their goal. These ‘clues’ are generally ignored by most computational techniques. The few researchers that take them into account, do not hesitate to stress out the difficulties in identifying them automatically. Several NLP theories have provided a formal environment for these phenomena (e.g. DRT, discourse relations theory, DRST, discourse rhetorical structures theory, speech acts theory, etc.), but experiments have not succeeded in demonstrating their efficiency on important
volumes of data. Moreover, text organization, at this level, shows the dependence between language and its nonlinguistic environment. The underlying knowledge, beliefs, social rules and conventional implicatures are present, and strongly tie linguistic data to the outside world in organized macro-structures. This is why we have tried, in this book, not to neglect this part, or else our NLP grounding would have been lame, and certainly incomplete.

2. Studying the Interaction between NLP and its Application Domains: Biology and Medicine

We hope to have convinced the reader in the preceding paragraph that human language entails its outside world peculiarities, and is not a simple rewriting mechanism, with inner rules, that could be applied as a mask or a filter on reality. Language interacts with what it expresses, modifies and is modified by the type of knowledge, beliefs, expectations and intentions conveyed by a given theme, topic, domain or field. This point of view drives us to consider that NLP in Biomedicine is not plainly an NLP outfit set up on biological and medical data. BioNLP (the acknowledged acronym for NLP dedicated to biology and medicine) has recently appeared as a fashionable trend in research because of the dire need of biologists and medical scientists for a computational information retrieval and knowledge management framework, as explained in the earlier paragraphs. But BioNLP has developed its own idiosyncrasies, i.e. its particular way to process biological and medical language. A domain language is not a simple subset of the general language, unlike what a common belief tends to claim. It develops its own semantics, lexical organization, lexical choice, sentence construction rules and text organization properties that translate the way of thinking of a homogeneous community. From our point of view, it was really necessary to present a set of works in BioNLP that were not plain transpositions of works done in other domain and transferred to medicine and biology without adaptation. Also, it is as important to investigate the tremendous potential of this domain, because of the following reasons:

- Biology and medicine are multifaceted, and specialists tend to specialize too narrowly. This prevents the ability to discover, or recognize common topics among them. The most obvious and cited example is the cancer issue. Oncologists, chemists, pathologists, surgeons, radiologists, molecular biologists, researchers in genetics are all concerned by this domain. But they do not publish at all in the same journals, do not read what each other write, and very often do not even communicate with each other, although they are invited to do so. Hyper-specialization prevents scientific multidisciplinary exchange. Lack of time, huge volumes of data are the most usually invoked reasons. Also, the inability to read what the neighboring community has written, because its words appear a bit foreign, its way of writing is different, etc. How much useful information is lost like this, and how precious it would be to gather it, process it and re-present it to those who need it!
- Biology and medicine are far from being the only domains suffering such a dispersion, but they are those which have the greatest impact on human public opinion. Health is nowadays the most preoccupying issue. Life sustenance and maintenance is the prime goal of every living being. Sciences that deal with this issue have naturally a priority in people minds.
- The particular linguistic features of biological and medical languages are very interesting challenges from an NLP point of view, and tackling them is an efficient way of testing the techniques robustness.
- Last, bioinformatics has been widely spreading in these last years. Algorithmical and combinatorial approaches have been very much present in the field. Statistics and data mining have also enhanced the domain. It seems that NLP had also to provide its contribution, and tackle the bioinformatics
issue from the textual side, which was either neglected or only superficially processed by the other domains.

TARGET AUDIENCE

The audience of this book could be approached from several points of views. First, the reader’s function or position: Researchers, teachers preparing graduate courses, PhD or masters students are the natural potential audience to such a work. Second, from the domain or community point of view: NLP researchers are naturally meant since they could discover through this book the particularities of the application domain language, and thus understand their fellow researchers’ work from the NLP problem solving side. AI researchers dealing with building and enhancing ontologies are also concerned when they have to deal with biological and/or medical data. Terminologists, and more generally linguists could benefit from this book since some interesting experiments are related here, and could be seen as sets of data in which they could plough to illustrate phenomena there are interested in. The health information systems community could also see this book as a set of works at their more ‘theoretical’ side. Since this book also enumerate some existing and working NLP software, medical or biological teams could be interested into browsing some of the chapters addressing more practical issues. Last but not least, the BioNLP community is itself the most concerned by this book.

A BRIEF OVERVIEW OF THE CHAPTERS AND THEIR ORGANIZATION

The book is divided into six sections.

The chapter title “Text Mining in Biomedicine”, is an introduction written by Sophia Ananiadou. She is one of the most renowned leaders in BioNLP, has edited a very interesting book about NLP, presenting the domain and its different aspects to the community of researchers in information retrieval, knowledge extraction and management, dedicated to biology and medicine. Sophia’s book has stated the grounding principles of NLP. This book complements her work by showing a progressive interaction between NLP and its target domain through text mining as a task. Her introduction emphasizes the needs, the requirements, the nature of the issues and their stakes, and the achievements of the state of the art.

Sophia Ananiadou is the most appropriate writer for this chapter because of her extensive knowledge about the domain and the ongoing research.

The core sections of the book are four: Three that follow the NLP granularity scope, ranging from the lexical level to the discourse level as explained in an earlier paragraph, and one devoted to selected existing software. Chapters belonging to these sections are numbered from II to XIX.

Section I, named “Works at a Lexical Level, Crossroads Between NLP and Ontological Knowledge Management”, is the most abundant, since research has reached here a recognized maturity. It is composed of nine chapters. The order in which chapters have been organized is set up by the following pattern:

1. **Using existing resources to perform document processing tasks**: Indexation (Chapter II), categorization (Chapter III) and information retrieval (Chapter IV). Indexation and categorization could be seen as previous tasks to an intelligent information retrieval, since they pre-structure textual data, according to topics, domain, keywords or centers of interest.
2. **Dealing with the cross-linguistic terminological problem**: From a specialists language to general language within one tongue (Chapter V), or across different tongues (Chapter VI).
3. **Enriching terminology: The beginning of a strong lexical NLP involvement** (Chapter VII).

4. **Increasing lexical NLP involvement in biomedical application** (Chapter VIII).

In a more detailed version, these chapters are very representative of the state-of-the-art. Most works are devoted to words, word-concepts relations, word-to-word relations.

Chapter II titled, “*Lexical Granularity for Automatic Indexing and Means to Achieve It: The Case of Swedish M*ē*SH®*”, by Dimitri Kokkinakis from the University of Gothenburg, Sweden, is one of the three articles in this section involving MēSH, the Medical terminological classification that is currently one of the backbone of research and applications in BioNLP. Kokkinakis’s paper emphasizes the indexing function, and clearly demonstrates the impact of lexical variability within a dedicated technical language (medical language). The authors explain the challenges met by researchers when they go beyond indexation with broad concepts (introducing vagueness, and thus noise), and try to tackle the fine grained level where the complex term-concept relationship creates mismatches, thus jeopardizing precision.

Chapter III named, “*Expanding Terms with Medical Ontologies to Improve a Multi-Label Text Categorization Systems*”, by M. Teresa Martín-Valdivia, Arturo Montejo-Ràez, M.C. Diaz-Galiano, José Perea Ortega and L. Alfonso Ureña-Lopez from the University of Jaén, in Spain, tackles the categorization issue, which is in spirit, very close to indexation. An index highlights the importance of the word and its possible equivalents. A category is more adapted to needs, and conveys pragmatic knowledge. The chapter focuses on terminological expansion, that facet of the expansion-contraction linguistic phenomenon that troubles the ontological world so much. By trying to resolve multiple labels in different medical classification sets, the authors fruitfully complement Kokkinakis’ approach.

A third representative of this type of work is Chapter IV, “*Using Biomedical Terminological Resources for Information Retrieval*”, by Piotr Pezik, Antonio Jimeno Yepes and Dietrich Rebholz-Schuhmann, from the European Bioinformatics Institute at Cambridge, UK. It could be seen as the third volume of a trilogy: Previous chapters deal with introductory (as well as highly recommended) tasks to information retrieval, this one directly tackle the task in itself. Chapter II clearly states the complexity of the word-concept relationship and its variety, Chapter III focuses on expansion, Chapter IV focuses on contraction (ambiguity). Chapter II was centered on MēSH, Chapter III enlarged this focus to other medical thesauri, Chapter IV provides an extensive account of biomedical resources, highlighting their linguistic properties. One of the added values of this chapter is that it largely describes queries properties and reformulation, thus shedding the light on the specific issues of information retrieval processes. This chapter is of a great help to understand why general search engines could not be efficient in biomedical literature.

Chapter V, “*Automatic Alignment of Medical Terminologies with General Dictionaries for an Efficient Information Retrieval*”, by Laura Dioşan, Alexandra Rogozan, and Jean-Pierre Pecuchet (a collaboration between the Institut National des Sciences Appliquées, France, and the Babeş-Bolyai University, Romania), tackles the delicate issue of the neophyte-specialist linguistic differences. This aspect is crucial and often remains in the shadow, because the attention of researchers in information retrieval mostly focuses on the specialists’ language, and addresses specialists. Mapping general language and technical language is a necessity since all technical texts contain general words. Moreover, meanings variations may introduce misinterpretations. The authors offer an automatic alignment system which classifies specialized and general terminology according to their similarity.

After stressing out the need of translating words within one language, depending on its specialization, Chapter VI named, “*Translation of Biomedical Terms by Inferring Rewriting Rules*”, by Vincent Claveau (IRISA-CNRS, France), addresses the cross-linguistic aspect. The author ambitiously deals with several languages: Czech, English, French, Italian, Portuguese, Spanish, Russian. The idea is to provide an automatic translation between a pair of languages for unknown medical words (i.e. not existing in
the available bilingual dictionaries), and to use this not only for terminological resources enhancement (which is its natural recipient) but also for a cross-linguistic information retrieval tasks. Results achieved look highly promising.

After the translation issue, two other chapters go deeper within the NLP classical issue of lexical functions, variety and ambiguity. Here the reader dives into the genuine ‘NLP culture’.

Chapter VII, “Lexical Enrichment of Biomedical Ontology”, by Nils Reiter and Paul Buitelaar, respectively from the universities of Heidelberg and Saarbrücken in Germany, is one of the most representative of what NLP research could offer to knowledge enrichment. Enhancing domain ontologies with semantic information derived from sophisticated lexica such as WordNet, using Wikipedia as another source of information (very fashionable in lexical NLP nowadays), and mostly selecting the most appropriate interpretations for ambiguous terms (the case of those broad concepts evoked beforehand), are but a few of the several contributions of this chapter to knowledge integration.

Chapter VIII, “Word Sense Disambiguation in Biomedical Application: A Machine Learning Approach”, by Torsten Schiemann, Ulf Leser (both from the Humboldt-Universität zu Berlin, Germany) and Jörg Hackenberg (Arizona State University, USA) is the ideal complement to Chapter VII. Going deep in the ambiguity phenomenon, not only as it might appear in ontologies, but also as it happens in the existing texts (specialized texts have proven to convey ambiguity almost as much as non specialized literature, despite the common belief that specialization is logically assorted with disambiguation!)

Section II titled, “Going Beyond Words: NLP Approaches Involving the Sentence Level”, groups three chapters. If the lexical level has been explored to a certain extent, broader linguistic granularities are yet to be investigated. A such, those three chapters could be considered as representative attempts to go beyond the lexical horizon. The need is certain: Chapters VII and VIII have shown the effects of ambiguity: Intrinsic ambiguity (localized in ontologies themselves) and induced ambiguities, detected in texts. A too fine-grained division might introduce such an effect, therefore, some researchers have turned their attention to the next complete linguistic unit, the sentence. The sentence context might erase lexical ambiguities effects in some cases. Sentences and not words might be queries or answers to queries.

Chapter IX named, “Information Extraction of Protein Phosphorylation from Biomedical Literature”, by M. Narayanaswamy, K. E. Ravikumar (both from Anna University in India), Z. Z. Hu (Georgetown University Medical Center, USA), K. Vijay-Shanker (Universiteit of Delaware, USA), and C. H. Wu (Georgetown University Medical Center, USA), describes a system that captures “the lexical, syntactic and semantic constraints found in sentences expressing phosphorylation information” from MEDLINE abstracts. The rule-based system has been designed as such because isolated words or phrases could possibly be thematic clues, but can by no means account for the different stages of a process. This means that event-type or procedure-type information cannot be encapsulated in a linguistic shape smaller than the significant text unit, the sentence. And according to the authors, a fair amount of the biomedical literature contains such information, which lexical approaches sometimes fail to capture.

Chapter X, “CorTag: A Language for a Contextual Tagging of the Words Within Their Sentence”, by Yves Kodratoff, Jérôme Azé (both from the University Paris-Sud (Paris XI), France) and Lise Fontaine (Cardiff University, UK) complements the preceding one, which presented an application and a need, by offering a design to extract contextual knowledge from sentences. Knowledge extraction from text is generally done with part-of-speech taggers, mostly with lexical categories. Higher level tags such as noun or verb phrases, adjectival groups, and beyond, syntactic dependencies, are generally either neglected, or wrongly assigned by taggers in technical texts. CorTag corrects wrong assignments, and a few examples have been given by the authors, among other examples, about the same protein phosphorylation process tackled by Chapter IX. If Chapter IX has amply focused on knowledge extraction, Chapter X also deals with knowledge discovery, and how to induce relations between concepts recognized in texts,
thanks to the services of syntactic and semantic information provided by sentences. In fact, sentences are the linguistic units that stage concepts and their relationships. Words or phrases capture concepts, but grammar expresses their respective roles.

Chapter XI titled, “Analyzing the Text of Clinical Literature for Question Answering”, by Yun Niu and Graeme Hirst from the University of Toronto, Canada, is an ideal example of the usefulness of the sentence level information in a question-answer task, one of the most representative tasks in information retrieval. Focusing on clinical questions, and the need to retrieve evidences from corpora as automatically as possible, the authors propose here a method, a design and a system that not only deal with complex information at the sentence level, but also inchoate a possible argumentative articulation between fragments of sentences. As such, this chapter is the most adapted one to the intersection between the sentence level and the discourse level investigations. The use of semantic classes introduces a thematic, and thus a pragmatic approach to meaning (the authors are not frightened to use the word ‘frame’) completing lexical and sentence semantics. This attitude finds an echo in Chapter XIV, ending next section, but the latter mostly emphasizes the ‘script’ approach (isn’t it bold to reuse frames and scripts in the late two thousands?) with a considerable highlight on situation pragmatics, whereas Chapter XI is closer to the heart of NLP literature in terms of linguistic material analysis. This chapter is also highly recommended to readers who need an accurate survey of question answering systems state-of-the-art.

As a natural sequel to Section III titled, “Pragmatics, Discourse Structures and Segment Level as the Last Stage in the NLP Offer to Biomedicine”. It also groups three chapters.

Chapter XII, “Discourse Processing for Text Mining”, by Nadine Lucas (GREYC CNRS, Université de Caen Basse-Normandie Campus 2, France), has to be considered as an educational chapter about discourse, the nature and effects of text structures, as well as the broad panel of inter-sentences relations available to readers and writers in order to organize information and knowledge in texts. The author is a linguist, and has deeply studied biomedical corpora from the linguistics point of view. This chapter is an ‘opinion chapter’: We have taken the risk to integrate such a text because it was important for us to offer a book which is not only a patchwork of present trends. We felt committed to research in NLP, IR, knowledge management, as well as to our target domain, biomedicine. Our commitment has been to give voice to criticism to approaches representatives of which been described in the same book. This chapter plays this role. It reveals the complexity of language that cannot be circumscribed by statistics, it shows that text mining scientific claims and trends still need to be improved, linguistically speaking. It emphasizes the fact that academic genre is not subsumed by its own abstracts features, and if texts are not ‘bags of words’, they are not ‘bags of sentences’ either. The longer the text, the more organized and the more conventional it is. The corpora analysis in this chapter is an interesting counter-argument to our gross computational attitudes toward text mining.

Chapter XIII titled, “A Neural Network Approach Implementing Non-linear Relevance Feedback to Improve the Performance of Medical Information Retrieval Systems”, by Dimosthenis Kyriazis, Anastasios Doulamis and Theodora Varvarigou from the National Technical University of Athens, Greece, acknowledges the content reliability issue. In other words, whatever the technique is, the main stake is either not to loose information, or not to provide false tracks. Sole lexical approaches discard information, retrieved as much as possible by sentence level approaches. But both might provide wrong answers to users’ needs. Including the user in the IR loop is a mandatory attitude. The authors attempt here to mathematically ground their information reliability hypothesis. Their contribution is in this section in the sense that they make room for pragmatics, if not for discourse structuration.

The increasing pragmatics involvement is clear in Chapter XIV, “Extracting Patient Case Profiles with Domain-Specific Semantic Categories”, by Yifao Zhang and Jon Patrick, from The University of Sydney, Australia. Here the retrieved part is a complex structure and entails domain pragmatics. Authors
tackle a particular issue, the role of fragments of patients medical records as diagnosis or symptoms clues. This naturally involve a discourse-type relationship. But if discourse relationships are difficult to retrieve with most means, according to the authors, the medical domain provides the researchers with ‘sentences types’ representing patterns to a given role. Acknowledging the existence of a conventional script for texts describing individual patients, authors have annotated a corpus and by acting at both the lexical and sentence level, have obtained most linguistic information from the linguistic material itself. The sentences types they present provide a kind of a text writing protocol in the particular application they investigate. Therefore the textual organization could be implicitly retrieved and used to complete information about patients, and thus help users in their queries. Chapter XIV is a mirror to Chapter XI. The latter focuses on the task (question answering) whereas Chapter XIV focuses on the need. Both are important pragmatic aspects. They could have been in the same section since sentence to text organization appears as a continuum in which the user is more and more concerned. Here we drop out of knowledge management, out of linguistics, to enter the outside world.

Section IV named, “NLP Software for IR in Biomedicine”, is dedicated to software dealing with information or knowledge extraction or verification (Chapters XV to XIX). It is a good thing for readers to know the software state of the art and compare it with more theoretical and methodological studies, in order to assess the gap between semi-automatic and automatic productions.

Chapter XV called, “Identification of Sequence Variants of Genes from Biomedical Literature: The OSIRIS Approach”, by Laura I. Furlong and Ferran Sanz from the Research Unit on Biomedical Informatics at Barcelona, Spain, deals with the problem of retrieving variants of genes described by literature. Their system, OSIRIS, integrates a cognitive module in its new version. This tries to relate more knowledge, NLP and algorithmics as a whole. Osiris can be used to link literature reference to biomedical databases entries, thus reducing the terminological variation.

Chapter XVI, “Verification of Uncurated Protein Annotations”, by Francisco M. Couto, Mário J. Sylva (both from the Universidade de Lisboa, Portugal) and Vivian Lee, Emily Dimmer, Evelyn Camon, Rolf Apweiler, Harald Kirsch, and Dietrich Rebholz-Schuhmann (European Bioinformatics Institute, Cambridge, UK) describes a tool that annotates literature with an available ontological description, GOAnnotator. Its aim is to assist the verification of uncurated protein annotations. In a way, it is a kind of a symmetrical tool to Osiris as a task (verification versus identification). GoAnnotator is a free tool, available on the Web, and is of a precious help to domain researchers.

Chapter XVII titled, “A Software Tool for Biomedical Information Extraction (And Beyond)”, by Burr Settles, from the University of Wisconsin-Madison, USA, is dedicated to the description of ABNER, a biomedical named entity recognizer, which is an open-source software tool for mining in the molecular biology literature. Like its fellow tools, ABNER deals with molecular biology where literature is abundant, knowledge is volatile and research interest definitely keen. The ‘named entity issue’ is one of the most typical of the difficulties encountered in literature, since it is related to the ‘unknown word or phrase’ basic NLP issue. As the author describes it, entity recognition is the natural first step in information management. Unlike Osiris and GOAnnotator, ABNER does not necessarily relate to database or ontology labels. The strings it deals with could be real words, acronyms, abbreviated words, and so forth. This chapter is particularly interesting from an evaluation point of view. The author provides a very educational insight on techniques in named entities recognition and their comparison. So, beyond ABNER, there is the crucial issue of the accurate model for a given task, and this chapter is definitely informative.

Chapter XVIII, “Problems-Solving Map Extraction with Collective Intelligence Analysis and Language Engineering”, by Asanee Kawkatrul (Kasetsart University and Ministry of Technology in Thailand), Chaveevari Petchsirim (University of Bangkok, Thailand), Sachit Rajbhandari (Food and Agriculture
Organization of the United Nations, Rome, Italy), and Frederic Andres (National Institute of Informatics Tokyo, Japan), is dedicated to the engineering of a framework aiming at reducing format heterogeneity in biological data. Their framework considers a collective action in knowledge management (extraction, enhancement, operation), and the issues they address are partly related to some of the other tools: Annotation (like Chapter XVI) but here for collaborative answers in question answering systems, named entity recognition (as in Chapter XVII), but also elements that could be related to our higher levels in NLP, discourse relations. This chapter is another bridge between AI and NLP. Ontologies and terminology are in the knowledge representation field of AI and the lexical level in NLP. Collective intelligence is at the social network, multi-agent level in AI and at the discourse and pragmatics level in NLP. Symmetry is preserved at every stage between language and intelligence.

Chapter XIX titled, “Seekbio: Retrieval of Spatial Relations for System Biology”, by Christophe Jouis, Magali Roux-Rouquié, both from University Paris 3, Sorbonne Nouvelle and Jean-Gabriel Ganascia (University Paris 6) describes a software that retrieves spatial relations, i.e. topological relations between concepts (and their identifying terms), directly interesting system biology. The latter is an emerging field in which living organisms are considered as complex systemic devices, and it mainly focuses on dynamic interactions between the biological components of these organisms. If most terminological software retrieves static relations, Seekbio originality is its dedication to the dynamic aspect of knowledge relations system. The tool can be plugged at Pubmed, an important database in the medical field and provides the researcher in system biology with a promising insight for discovering new interactions.

Chapter II to XIX have described, through four main sections, the ongoing research, theoretical advances, and existing tools, produced by BioNLP, this crossroads field in which AI, NLP, statistics and application domains, biology and medicine, all interact.

Our last section, “Conclusion and Perspectives”, with its sole chapter, the twentieth titled, “Analysing Clinical Notes for Translation Research: Back to the Future”, by Jon Patrick and Pooyan Asgari from The University of Sydney, Australia, is meant not to be an end, but a window open on the future, already pending for most of the BioNLP community. The paper summarizes in its description several achievements that have been deeply investigated in this book through different applications. It has chosen, as a data field to plough, a type of text which is highly operational: No academic style, but something very heterogeneous, mixing images, sound, text, and in text, introducing abbreviations, acronyms, a particular grammar, a highly pragmatics-based way of organizing knowledge. This type of text is the patient folder (in health care institutions) and contains several units of clinical notes. Shifting from biology to medicine in this conclusion is also a way to shift from academic genre to a practitioner type of writing, in which general language is disturbed, noisy, because it is dedicated to a task in which time and goal are overwhelming: Caring of the sick and the dying. In this conclusion, we have tried, as editor to put a momentary final note to this volume (we could not go on for ever), but our baseline was to assess NLP presence and help as much in an intellectual research as in a day-to-day practice. Chapter XX could be a conclusion, or another chapter, but here we have chosen to show it as an opening to another world in which NLP is needed, with time and practice constraints.

HOW WE THINK THE BOOK IMPACTS THE FIELD

We emphasized the need for such a work and the originality of our collection of chapters in the first two sections of this preface. We described the different contributions in the light of our approach to NLP interaction with the application domains. The various authors of this volume chapters are very representative of the ongoing research in the field. They are spread all around the world, they belong to
multidisciplinary areas, some of them are famous and renowned in BioNLP, some other are less known but present original theories and outcomes. We believe that this book is a most appropriate gathering of what is to be said about the field (an exhaustive approach would have needed an encyclopedia format). It is also a door open to discussion, opinion debates, scientific confrontation and other healthy attitudes that science nowadays is in a bad need for. As editors, we chose not to contribute ourselves, although we are running research in NLP, and apply it to BioNLP. But we believe that editors should be the gatherers of variety, not the speakers of a given trend, and certainly not advertisers of their own work. Under the magnifying lens of several of these contributions, we feel that all new work in BioNLP, whether dedicated to biology and academic production, or to medicine and its multiplicity of genres, has to integrate the idea that it is not a simple application of a given technique to a given domain, but a lively interaction between two scientific fields, giving birth to a hybrid with its specific properties. Theories or techniques have to be adapted, and ‘dogmas’ need to be discarded. On the other hand, rough statistical approximations have to be refined, in order not to produce noise and wrong directions. Language multiple layers must not be forgotten, but their integration has to be as customized as possible. We hope that, after reading this book, the scientific community will get a picture as accurate as possible of the enthusiastic challenges of our domain, as well as of its limitations and drawbacks. But let that not hinder our eagerness to continue in improving BioNLP achievements for the best of all.

Violaine Prince and Mathieu Roche
University Montpellier 2, France & LIRMM-CNRS, France