PREMIER REFERENCE SOURCE

INFORMATION RETRIEVAL IN BIOMEDICINE

Natural Language Processing for Knowledge Integration

VIOLAINE PRINCE & MATHIEU ROCHE

Detailed Table of Contents

Text mining provides the automated means to manage information overload and overlook. By adding meaning to text, text mining techniques produce a much more structured analysis of textual knowledge than do simple word searches, and can provide powerful tools for knowledge discovery in biomedicine. In this chapter, the authors focus on the text mining services for biomedicine offered by the United Kingdom National Centre for Text Mining.

Section I
Works at a Lexical Level: Crossroads Between NLP and Ontological Knowledge Management

The identification and mapping of terminology from large repositories of life science data onto concept hierarchies constitute an important initial step for a deeper semantic exploration of unstructured textual content. Accurate and efficient mapping of this kind is likely to provide better means of enhancing indexing and retrieval of text, uncovering subtle differences, similarities and useful patterns, and hopefully new knowledge, among complex surface realisations, overlooked by shallow techniques based on various forms of lexicon look-up approaches. However, a finer-grained level of mapping between terms as they occur in natural language and domain concepts is a cumbersome enterprise that requires various levels of processing in order to make explicit relevant linguistic structures. This chapter highlights some of the challenges encountered in the process of bridging free to controlled vocabularies and thesauri and vice versa. It investigates how the extensive variability of lexical terms in authentic data can be efficiently projected to hierarchically structured codes, while means to increase the coverage of the underlying lexical resources are also investigated.

This chapter argues for the integration of clinical knowledge extracted from medical ontologies in order to improve a Multi-Label Text Categorization (MLTC) system for medical domain. The approach is based on the concept of semantic enrichment by integrating knowledge in different biomedical collections. Specifically, the authors expand terms from these collections using the UMLS (Unified Medical Language System) metathesaurus. This resource includes several medical ontologies. The authors have managed two rather different medical collections: first, the CCHMC collection (Cincinnati Children's Hospital Medical Centre) from the Department of Radiology, and second, the widely used OHSUMED collection. The results obtained show that the use of the medical ontologies improves the system performance.

The present chapter discusses the use of terminological resources for Information Retrieval in the biomedical domain. It first introduces a number of example resources which can be used to compile terminologies for biomedical IR. The authors explain some of the common problems with such resources including redundancy, term ambiguity, insufficient coverage of concepts, and incomplete semantic organization of such resources for text mining purposes. They also discuss some techniques used to address each of these deficiencies, such as static polysemy detection as well as adding terms and linguistic annotation from the running text. In the second part of the chapter, the authors show how query expansion based on using synonyms of the original query terms derived from terminological resources potentially increases the recall of IR systems. Special care is needed to prevent a query drift produced by the usage of the added terms and high quality word sense disambiguation algorithms can be used to allow more conservative query expansion. In addition, the authors present solutions that help focus on the user's specific information need by navigating and rearranging the retrieved documents. Finally, they explain the advantages of applying terminological and semantic resources at indexing time. The authors argue that by creating a semantic index with terms disambiguated for the term's semantic types and larger chunks of text denoting entities and relations between them, they can facilitate query expansion, reduce the need for query refinement and increase the overall performance of Information Retrieval. Semantic indexing also provides support for generic queries for concept categories, such as genes or diseases, rather than singular keywords.

The automatic alignment between a specialized terminology used by librarians in order to index concepts and a general vocabulary employed by a neophyte user in order to retrieve medical information will certainly improve the performances of the search process, this being one of the purposes of the ANR VODEL project. The authors propose an original automatic alignment of definitions taken from different dictionaries that could be associated to the same concept although they may have different labels. The definitions are represented at different levels (lexical, semantic and syntactic), by using an original and shorter representation, which concatenates more similarities measures between definitions, instead of the classical one (as a vector of word occurrence, whose length equals the number of different words from all the dictionaries). The automatic alignment task is considered as a classification problem and three Machine Learning algorithms are utilised in order to solve it: a *k* Nearest Neighbour algorithm, an Evolutionary Algorithm and a Support Vector Machine algorithm. Numerical results indicate that the syntactic level of nouns seems to be the most important, determining the best performances of the SVM classifier.

This chapter presents a simple yet efficient approach to translate automatically unknown biomedical terms from one language into another. This approach relies on a machine learning process able to infer rewriting rules from examples, that is, from a list of paired terms in two studied languages. Any new term is then simply translated by applying the rewriting rules to it. When different translations are produced by conflicting rewriting rules, the authors use language modeling to single out the best candidate. The experiments reported here show that this technique yields very good results for different language pairs (including Czech, English, French, Italian, Portuguese, Spanish and even Russian). They also show how this translation technique could be used in a cross-language information retrieval task and thus complete the dictionary-based existing approaches.

This chapter is concerned with lexical enrichment of ontologies, that is, how to enrich a given ontology with lexical information derived from a semantic lexicon such as WordNet or other lexical resources. The authors present an approach towards the integration of both types of resources, in particular for the human anatomy domain as represented by the Foundational Model of Anatomy and for the molecular biology domain as represented by an ontology of biochemical substances. The chapter describes their approach on enriching these biomedical ontologies with information derived from WordNet and Wikipedia by matching ontology class labels to entries in WordNet and Wikipedia. In the first case, they acquire WordNet synonyms for the ontology class label, whereas in the second case they acquire multilingual translations as provided by Wikipedia. A particular point of emphasis here is on selecting the appropriate interpretation of ambiguous ontology class labels through sense disambiguation, which they address by use of a simple algorithm that selects the most likely sense for an ambiguous term by statistical significance of co-occurring words in a domain corpus. Acquired synonyms and translations are added to the ontology by use of the LingInfo model, which provides an ontology-based lexicon model for the annotation of ontology classes with (multilingual) terms and their linguistic properties.

Ambiguity is a common phenomenon in text, especially in the biomedical domain. For instance, it is frequently the case that a gene, a protein encoded by the gene, and a disease associated with the protein share the same name. Resolving this problem, that is assigning to an ambiguous word in a given context its correct meaning is called word sense disambiguation (WSD). It is a pre-requisite for associating entities in text to external identifiers and thus to put the results from text mining into a larger knowledge framework. In this chapter, the authors introduce the WSD problem and sketch general approaches for solving it. They then describe in detail the results of a study in WSD using classification. For each sense of an ambiguous term, they collected a large number of exemplary texts automatically and used them to train an SVM-based classifier. This method reaches a median success rate of 97%. They also provide an analysis of potential sources and methods to obtain training examples, which proved to be the most difficult part of this study.

Section II
Going Beyond Words: NLP Approaches Involving the Sentence Level

Protein posttranslational modification (PTM) is a fundamental biological process, and currently few text mining systems focus on PTM information extraction. A rule-based text mining system, RLIMS-P (Rule-based LIterature Mining System for Protein Phosphorylation), was recently developed by our group to extract protein substrate, kinase and phosphorylated residue/sites from MEDLINE abstracts. This chapter covers the evaluation and benchmarking of RLIMS-P and highlights some novel and unique features of the system. The extraction patterns of RLIMS-P capture a range of lexical, syntactic and semantic constraints found in sentences expressing phosphorylation information. RLIMS-P also has a second phase that puts together information extracted from different sentences. This is an important feature since it is not common to find the kinase, substrate and site of phosphorylation to be mentioned in the same sentence. Small modifications to the rules for extraction of phosphorylation information have also allowed us to develop systems for extraction of two other PTMs, acetylation and methylation. A thorough evaluation of these two systems needs to be completed. Finally, an online version of RLIMS-P with enhanced functionalities, namely, phosphorylation annotation ranking, evidence tagging, and protein entity mapping, has been developed and is publicly accessible.

This chapter argues that in order to extract significant knowledge from masses of technical texts, it is necessary to provide the field specialists with programming tools with which they themselves may use to program their text analysis tools. These programming tools, besides helping the programming effort of the field specialists, must also help them to gather the field knowledge necessary for defining and retrieving what they define as significant knowledge. This necessary field knowledge must be included in a well-structured and easy to use part of the programming tool. The authors illustrate their argument by presenting a programming language, CorTag, which they have built in order to correct existing tags in a text, while trying to follow the informal specification given above.

The task of question answering (QA) is to find an accurate and precise answer to a natural language question in some predefined text. Most existing QA systems handle fact-based questions that usually take named entities as the answers. In this chapter, the authors take clinical QA as an example to deal with more complex information needs. They propose an approach using semantic class analysis as the organizing principle to answer clinical questions. They investigate three semantic classes that correspond to roles in the commonly accepted PICO format of describing clinical scenarios. The three semantic classes are: the description of the patient (or the problem), the intervention used to treat the problem, and the clinical outcome. They focus on automatic analysis of two important properties of the semantic classes.

Section III
Pragmatics, Discourse Structures and Segment Level as the Last Stage in the NLP Offer to Biomedicine

from the diagnostic and treatment procedures. This chapter proposes a novel category set to cover a wide variety of semantics in the description of clinical case studies which distinguishes each unique patient case. A manually annotated corpus consisting of over 5,000 sentences from 75 journal articles of clinical case studies has been created. A sentence classification system which identifies 13 classes of clinically relevant content has been developed. A golden standard for assessing the automatic classifications has been established by manual annotation. A maximum entropy (MaxEnt) classifier is shown to produce better results than a Support Vector Machine (SVM) classifier on the corpus.

Section IV
NLP Software for IR in Biomedicine

SNPs constitute key elements in genetic epidemiology and pharmacogenomics. While data about genetic variation is found at sequence databases, functional and phenotypic information on consequences of the variations resides in literature. Literature mining is mainly hampered by the terminology problem. Thus, automatic systems for the identification of citations of allelic variants of genes in biomedical texts are required. The authors have reported the development of OSIRIS, aimed at retrieving literature about allelic variants of genes, a system that evolved towards a new version incorporating a new entity recognition module. The new version is based on a terminology of variations and a pattern-based search algorithm for the identification of variation terms and their disambiguation to dbSNP identifiers. OSIRISv1.2 can be used to link literature references to dbSNP database entries with high accuracy, and is suitable for collecting current knowledge on gene sequence variations for supporting the functional annotation of variation databases.

Molecular Biology research projects produced vast amounts of data, part of which has been preserved in a variety of public databases. However, a large portion of the data contains a significant number of errors and therefore requires careful verification by curators, a painful and costly task, before being reliable enough to derive valid conclusions from it. On the other hand, research in biomedical information retrieval and information extraction are nowadays delivering Text Mining solutions that can support curators to improve the efficiency of their work to deliver better data resources. Over the past decades, automatic text processing systems have successfully exploited biomedical scientific literature to reduce the researchers' efforts to keep up to date, but many of these systems still rely on domain knowledge that is integrated manually leading to unnecessary overheads and restrictions in its use. A more efficient approach would acquire the domain knowledge automatically from publicly available biological sources, such as BioOntologies, rather than using manually inserted domain knowledge. An example of this approach is GOAnnotator, a tool that assists the verification of uncurated protein annotations. It provided correct evidence text at 93% precision to the curators and thus achieved promising results. GOAnnotator was implemented as a web tool that is freely available at http://xldb.di.fc.ul.pt/rebil/tools/goa/.

ABNER (A Biomedical Named Entity Recognizer) is an open-source software tool for text mining in the molecular biology literature. It processes unstructured biomedical documents in order to discover and annotate mentions of genes, proteins, cell types, and other entities of interest. This task, known as *named entity recognition* (NER), is an important first step for many larger information management goals in biomedicine, namely extraction of biochemical relationships, document classification, information retrieval, and the like. To accomplish this task, ABNER uses state-of-the-art machine learning models for sequence labeling called *conditional random fields* (CRFs). The software distribution comes bundled with two models that are pre-trained on standard evaluation corpora. ABNER can run as a stand-alone application with a graphical user interface, or be accessed as a Java API allowing it to be re-trained with new labeled corpora and incorporated into other, higher-level applications. This chapter describes the software and its features, presents an overview of the underlying technology, and provides a discussion of some of the more advanced natural language processing systems for which ABNER has been used as a component. ABNER is open-source and freely available from http://pages.cs.wisc.edu/~bsettles/abner/

Valuable knowledge has been distributed in heterogeneous formats on many different Web sites and other sources over the Internet. However, finding the needed information is a complex task since there is a lack of semantic relations and organization between them. This chapter presents a problem-solving map framework for extracting and integrating knowledge from unstructured documents on the Inter-

Section V
Conclusion and Perspectives