Knowledge Representation and Management:
Transforming Textual Information into Useful Knowledge

A.-M. Rassinoux, Section Editor for the IMIA Yearbook Section on Knowledge Representation and Management
Department of Imaging and Medical Informatics, Geneva University Hospitals, Geneva, Switzerland

Summary
Objectives: To summarize current outstanding research in the field of knowledge representation and management.
Method: Synopsis of the articles selected for the IMIA Yearbook 2010.
Results: Four interesting papers, dealing with structured knowledge, have been selected for the section knowledge representation and management. Combining the newest techniques in computational linguistics and natural language processing with the latest methods in statistical data analysis, machine learning and text mining has proved to be efficient for turning unstructured textual information into meaningful knowledge. Three of the four selected papers for the section knowledge representation and management corroborate this approach and depict various experiments conducted to .extract meaningful knowledge from unstructured free texts such as extracting cancer disease characteristics from pathology reports, or extracting protein-protein interactions from biomedical papers, as well as extracting knowledge for the support of hypothesis generation in molecular biology from the Medline literature. Finally, the last paper addresses the level of formally representing and structuring information within clinical terminologies in order to render such information easily available and shareable among the health informatics community.
Conclusions: Delivering common powerful tools able to automatically extract meaningful information from the huge amount of electronically unstructured free texts is an essential step towards promoting sharing and reusability across applications, domains, and institutions thus contributing to building capacities worldwide.

Keywords
Structured knowledge, natural language processing (NLP), text mining, knowledge extraction, knowledge representation

Yearb Med Inform 2010: 64-7

Introduction

The explosive growth of health information in textual form has created a great demand for powerful tools, able to turn textual information into useful knowledge. In order to carry out this challenging and intricate task, latest state-of-the-art researches are focusing towards exploiting the newest techniques in computational linguistics and natural language processing (NLP), and combining them with the latest methods in statistical data analysis, machine learning and text mining [1, 2, 3].

The field of NLP, focusing on the analysis of narrative documents stored in the electronic health records (EHRs), has undergone decades of intensive research as reflected in the extensive overview of Meystre et al. [4], published in the IMIA Yearbook 2008. NLP offers a means to extract, encode, and structure information in free texts in order to be used by subsequent computerized applications [5]. When applied to patients' clinical documents, specific attentions must be paid to deal with misspellings, ad-hoc abbreviations and acronyms, as well as medical domain-specific vocabulary and language use.

Text mining is a more recent field whose application domains are mainly named entity recognition, text classification, terminology extraction, relationship extraction and hypothesis generation [6]. It applies techniques primarily developed in the areas of statistics, information retrieval and machine learning. Text mining aims at extracting patterns from natural language texts just as data mining aims at extracting patterns from structured databases. Its goal is not to understand all or even a large part of a text but rather to detect facts or patterns, as well as connections between different factors which influence medical care and health. NLP might benefit text mining by annotating documents with any useful parsing results that will be further exploited by the information extraction tools.

"Building Capacity Worldwide" is the topic of the IMIA Yearbook 2010 and one way of contributing to this issue, in the field of knowledge representation and management, is to extend mechanisms to extract, collect, categorize and disseminate relevant information around the world. In order to reach this goal, it is necessary to add some sort of structured form to the significant content of the continually growing body of clinical and biomedical documents as well as web pages, as already highlighted in the IMIA Yearbook 2008 [7]. Two major international initiatives, showing valuable insights and guidance for structuring textual information, are worth noting.

First, a collaborative effort has been led since a decade by the W3C (World Wide Web Consortium), for building the Semantic Web [8]. This project aims at extending the current state of the World Wide Web, by bringing structure (semantics) to the meaningful content of Web pages. Within the Semantic Web framework, new standards were proposed, most notably RDF (Resource Description Framework), a standard syntax for describing data and OWL (Web Ontology Language), a language for ontology specification.

Second, a more recent initiative, instigated by researchers at both the Mayo Clinic and IBM, allows NLP tools to be freely available for the clinical and research communities through the Open

Health Natural Language Processing (OHNLP) Consortium [9]. As part of this consortium, a set of UIMA [10] annotators (programs) and pipelines (execution sequences) were released, currently including the Mayo's cTAKES pipeline (clinical Text Analysis and Knowledge Extraction System) for extracting entities from clinical texts [11] and the IBM's medKAT/P pipeline (medical Knowledge Analysis Tool/Pathology) for extracting cancer characteristics from pathology reports [12]. These modularized annotators work step-by-step (i.e. the output of one component providing context for the next one) towards identifying meaningful entities and relationships from unstructured texts. Customized and scalable NLP systems can then be built from this set of incremental and evolutionary annotators in order to address other particular clinical research studies or clinical trials.

Best Paper Selection

Following a comprehensive review process, four papers were selected this year for the section knowledge representation and management of the IMIA Yearbook 2010 (see Table 1). A brief content summary of each elected article can be found in the appendix of this synopsis. All these papers deal with the way of identifying, extracting and representing meaningful pieces of information, whether such information is embedded into unstructured textual documents or already described through terms into a terminology system. The aforementioned issues are also argued for each paper.

A knowledge representation model (CDKRM) for capturing cancer related information, in particular its characteristics and disease progression, is introduced by Coden et al. in the first selected paper [12]. In order to instantiate elements of this model from unstructured free-text pathology reports, a Medical Text Analysis System MedTAS/P is used. The later is based on the IBM's medKAT/P open-source framework of OHNLP. It relies on NLP principles and contains both rule-based and machine-learning based components. In addition, MedTAS/P provides a mechanism to ingest, process, and use external resources, such as terminologies and ontologies. This approach highlights the benefit of mixing NLP techniques and text mining as well as integrating existing knowledge such as terminologies. In the second article, Miyao et al. [2] provide a valuable comparative evaluation of current trends in NLP tools focusing on the task of identifying protein-protein interactions from biomedical papers. In order to choose the right parser and its appropriate representation, as a component in a biomedical text mining system focusing on a specific task, it is necessary to know and evaluate its capabilities in term of advantages and disadvantages. Further investigations should be conducted on combining various parsers and parse representations but applied to other meaningful bio-medical tasks as well as other corpora. Continuing with text mining experiments, the third paper [3] addresses a concrete study focusing on identifying hypothesis for a particular biomolecular mechanism. Both issues, mentioned in the previous section, are exploited in this experiment. Indeed, Semantic Web technologies are used to structure and store knowledge in a machine readable format, while a workflow extracts knowledge from Medline texts. Minimal proto-ontologies in OWL are designed for capturing different aspects of the text mining experiment thus allowing the semantic interoperability between them. Harmonizing concept representation, languages and methodologies was already brought to light in the last year IMIA Yearbook [13]. Moreover, using a workflow as the knowledge extraction procedure makes possible to repeatedly run the same workflow or adaptations of it in order to perform posterior analyses over the results from several experiments. Finally, the last paper by Rosenbloom and al. [14] compares SNOMED CT with two interface terminologies. The mapping and alignment of existing terminologies have long been carried out in the ontology community as a way to evaluate their usefulness, scalability and reusability across various applications and community boundaries [15]. This paper outlines that the structure of modern terminologies must evolve in order to formally integrate concept meanings and relationships. This is of paramount interest as more and more often ontologies enhance information discovery (see [12]), modeling, and semantic interoperability [13]. Moreover, an attempt to reformulate the SNOMED's logical formalism into the OWL standard has already been scrutinized [16, 13], in order to improve its accessibility and internationalization.

Table 1 Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2010 in the section 'Knowledge Representation and Management'. The articles are listed in alphabetical order of the first author's surname.

Section Knowledge Representation and Management
▪ Coden A, Savova G, Sominsky I, Tanenblatt M, Masanz J, Schuler K, Cooper J, Guan W, de Groen PC. Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model. J Biomed Inform. 2009 Oct;42(5):937-49.
▪ Miyao Y, Sagae K, Saetre R, Matsuzaki T, Tsujii J. Evaluating contributions of natural language parsers to protein-protein interaction extraction. Bioinformatics. 2009 Feb 1;25(3):394-400.
▪ Roos M, Marshall MS, Gibson AP, Schuemie M, Meij E, Katrenko S, van Hage WR, Krommydas K, Adriaans PW. Structuring and extracting knowledge for the support of hypothesis generation in molecular biology. BMC Bioinformatics 2009; 10(Suppl 10):S9
▪ Rosenbloom ST, Brown SH, Froehling D, Bauer BA, Wahner-Roedler DL, Gregg WM, Elkin PL. Using SNOMED CT to Represent Two Interface Terminologies. J Am Med Inform Assoc. 2009 Jan-Feb;16(1):81-8.

Conclusions and Outlook

The collection of selected papers, for the section knowledge representation and management of the IMIA Yearbook 2010, mainly deals with NLP techniques, text mining, as well as standardized representation of information in terminologies. Nowadays, more and more studies are addressing the challenge of combining powerful NLP techniques with text mining in order to turn latent textual information into useful knowledge. The years to come should confirm this trend and extend the work of Miyao et al. [2] in order to provide a sound assessment of what NLP techniques are beneficial for what text mining application types and corpora. Finally, another major point emerging from the 2010 best paper selection is the need to share component-based architecture and standards for NLP activities, as well as standardized representation formalisms in order to take advantage of external resources such as models, dictionaries and ontologies [17].

Acknowledgement

I greatly acknowledge the support of Martina Hutter and of the reviewers in the selection process of the IMIA Yearbook.

References

1. Popowich F. Using text mining and natural language processing for health care claims processing. SIGKDD Exploration Newsletter, 2005;7(1):59–66.
2. Miyao Y, Sagae K, Saetre R, Matsuzaki T, Tsujii J. Evaluating contributions of natural language parsers to protein-protein interaction extraction. Bioinformatics. 2009 Feb 1;25(3):394-400.
3. Roos M, Marshall MS, Gibson AP, Schuemie M, Meij E, Katrenko S, et al. Structuring and extracting knowledge for the support of hypothesis generation in molecular biology. BMC Bioinformatics 2009;10(Suppl 10): S9.
4. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform. 2008:128-44.
5. Friedman C, Johnson S. Natural language and text processing in biomedicine. In: Shortliffe E, Cimino JJ, editors. Biomedical Informatics Computer Applications in Health Care and Biomedicine. 2006.
6. Cohen AM, Hersh WR. A survey of current work in biomedical text mining. Brief Bioinform 2005 Mar;6(1):57-71.
7. Rassinoux AM. Decision Support, Knowledge Representation and Management: Structuring Knowledge for Better Access. In: Geissbuhler A, Kulikowski C, editors. IMIA Yearbook of Medical Informatics 2008. Methods Inf Med 2008; 47 Suppl 1:80-2.
8. Feigenbaum L, Herman I, Hongsermeier T, Neumann E, Stephens S. The Semantic Web in Action. Scientific American Magazine 2007;297: 90-7.
9. https://cabig-kc.nci.nih.gov/Vocab/KC/index.php/OHNLP.
10. Savova G, Kipper-Schuler KC, Buntrock J, Chute C. UIMA-based clinical information extraction system. Language Resources and Evaluation Conference 2008 (LREC). Towards enhanced interoperability for large HLT systems: UIMA for NLP; Marrakech, Morocco; 2008.
11. Pakhomov J, Buntrock J, Duffy P. High throughput modularized NLP system for clinical text. In: Proceedings of the Association for Computational Linguistics (ACL'05), 2005: 25–8.
12. Coden A, Savova G, Sominsky I, Tanenblatt M, Masanz J, Schuler K, et al. Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model. J Biomed Inform 2009 Oct;42(5):937-49.
13. Rassinoux AM. Decision Support, Knowledge Representation and Management: Towards Interoperable Medical terminologies. In: Geissbuhler A, Kulikowski C, editors. IMIA Yearbook of Medical Informatics 2009. Methods Suppl 2009:99-102.
14. Rosenbloom ST, Brown SH, Froehling D, Bauer BA, Wahner-Roedler DL, Gregg WM, et al. Using SNOMED CT to Represent Two Interface Terminologies. J Am Med Inform Assoc 2009 Jan-Feb;16(1):81-8.
15. Yu AC. Methods in biomedical ontology J Biomed Inform 2006;39:252-66.
16. Rector AL, Brandt S. Why do it the hard way? The case for an expressive description logic for SNOMED. J Am Med Inform Assoc 2008;15(6):744-51.
17. Chen ES, Maloney FL, Shilmayster E, Goldberg HS. Laying the groundwork for enterprise-wide medical language processing services: architecture and process. AMIA Annu Symp Proc 2009 Nov 14;2009:97-101.

Correspondence to:
Anne-Marie Rassinoux, Ph. D.
University Hospitals of Geneva
Service of Medical Informatics
Unit of Clinical Informatics
4, Rue Gabrielle-Perret-Gentil
CH-1211 Geneva 14
Switzerland
Tel: +41 22 372 6293
Fax: +41 22 372 8680
E-mail: anne-marie.rassinoux@hcuge.ch

Appendix: Content Summaries of Selected Best Papers for the IMIA Yearbook 2010, Section Knowledge Representation and Management*

Coden A, Savova G, Sominsky I, Tanenblatt M, Masanz J, Schuler K, Cooper J, Guan W, de Groen PC

Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model

J Biomed Inform 2009 Oct; 42(5):937-49

Today, the main cause of death in the United States is cancer. In order to enable and facilitate cancer research, the medical community is striving towards a structured representation of cancers that gathers information scattered among structured and unstructured data sources. A solution is proposed by the authors who describe a Cancer Disease Knowledge Representation Model (CDKRM) and a Medical Text Analysis System MedTAS/P that automatically populates pertinent parts of the model with information extracted from unstructured free-text pathology reports. The CDKRM is an extensible and adjustable knowledge representation that allows defining additional concepts as well as relations between cancer characteristics. MedTAS/P is a modular system based on an open-source framework and its components use natural language processing principles, machine learning and rules to discover and populate elements of the model. A gold-standard corpus of manually annotated colon cancer pathology reports has been built up by four coding domain experts. It was used to validate the model as well as to measure the accuracy of MedTAS/P. Algorithms, for automatically populating the CDKRM from free-text pathology reports, were developed.

* The complete papers can be accessed in the Yearbook's full electronic version, provided that permission has been granted by the copyright holder(s).

The precision and recall of these algorithms were evaluated against the gold-standard annotations, and reported by the authors through F1-scores that ranged from 0.9 to 1.0 for most tasks. Finally, the fact that the MedTAS/P system is modular makes possible its adaptation to other knowledge representation models not necessarily linked to the medical domain.

Miyao Y, Sagae K, Saetre R, Matsuzaki T, Tsujii J

Evaluating contributions of natural language parsers to protein-protein interaction extraction

Bioinformatics 2009 Feb 1;25(3):394-400

The explosive growth of biomedical literature has largely contributed to the development and popularity of text mining technologies. Nowadays, shallow text processing techniques are commonly used for tasks such as the identification of proteins and other entities in biomedical papers. However, more complex tasks, such as the identification of protein-protein interactions (PPI), require more advanced natural language processing (NLP) approaches that perform a syntactic and semantic analysis of the text structure. In order to help unskilled researchers selecting the appropriate NLP components, including the parser and its parse representation, this paper provides a comparative evaluation of current trends in NLP focusing on the task of identifying PPI information in biomedical papers. Eight representative parsers based on three different frameworks: dependency parsing, phrase structure parsing and deep parsing, as well as five representations are considered. The aim is to measure how each parser and its output representation, contributes to accuracy improvements of the PPI extraction task by incorporating the output of different parsers as statistical features in a machine learning classifier. Experiments on combining parsers and parse representations, while applying conversions between representations when necessary, as well as combining two parsers were carried out. The result showed that all parsers attained similar accuracy levels while differences in parsing speed were larger, the dependency parsers being much faster than the others. Finally, the best accuracy improvement of PPI extraction was obtained by combining two parsers as a component in a PPI system.

Roos M, Marshall MS, Gibson AP, Schuemie M, Meij E, Katrenko S, van Hage WR, Krommydas K, Adriaans PW

Structuring and extracting knowledge for the support of hypothesis generation in molecular biology

BMC Bioinformatics 2009;10 (Suppl 10):S9

Considering all potentially relevant facts while forming a hypothesis in molecular biology is increasingly challenging, due to the large scale of information embedded in the millions of biomedical documents available from PubMed. Automated support for extracting and managing hypotheses about biomolecular mechanisms is therefore a general requirement. The authors propose an original and controllable methodology that combines tools and expertise from the fields of Semantic Web, e-Science, as well as information retrieval and extraction. A workflow system, which is mainly composed of web services from the Adaptive Information Disclosure Application (AIDA) toolkit, is used for extracting knowledge from Medline literature. The Semantic Web, thanks to its web standards RDF and OWL, provides a way to structure and store knowledge in machine readable format that is amenable to machine inference. In particular, biological hypothesis, text and documents, text mining and workflow provenance are all captured into minimal distinct proto-ontologies in OWL. The output enriches a semantic model with putative biological relations and corresponding evidence. Moreover, this knowledge base is stored on myExperiment.org, thus allowing all resources relevant to an experiment to be shared on the web and therefore to perform posterior analyses over the results of multiple experiments.

Rosenbloom ST, Brown SH, Froehling D, Bauer BA, Wahner-Roedler DL, Gregg WM, Elkin PL

Using SNOMED CT to Represent Two Interface Terminologies

J Am Med Inform Assoc 2009 Jan-Feb;16(1):81-8

Clinical interface terminologies generally consist of a rich set of flexible and colloquial health care–related phrases also called terms. Their purpose is to support and ease the interactions between clinical users and structured medical information and as such, they are commonly used for documentation tasks within structured computer based documentation (CBD) systems. Besides, reference terminologies are typically designed to provide a more formal representation of medical knowledge, including entities and their inter-relationships and are optimized to support the storage, retrieval and classification of clinical data. The goal of the study described in this paper, is to evaluate how well the reference terminology SNOMED CT could map to and represent two interfaces terminologies, MEDCIN and the Categorical Health Information Structured Lexicon (CHISL). Automated processes, based on the Mayo Clinic's multi-threaded Clinical Vocabulary Server (MCVS), are applied to link the MEDCIN and CHISL interface terminologies to SNOMED CT. For both MEDCIN and CHISL, the MCVS presented 500 mappings between interface terms and reference concepts for subsequent human review. The current study demonstrated an excellent coverage (greater than 90%) of SNOMED CT for the concepts contained in random samples of the history and physical examination sections of the two interface terminologies, with however a better coverage for those from CHISL than from MEDCIN. CHISL also obtained a greater rate of agreement among reviewers for both concept mapping and semantic categorization. The authors explained these best results by the fact that CHISL is a smaller terminology that contains less complex and specialized terms. In conclusion, the experiment of mapping interface terminologies to reference terminologies covering the same domain knowledge allows bringing to light two issues. On the one hand, reference terminologies can provide ontological rigor and standardization to interface terminologies. On the other hand, reference terminologies can also be enriched with semantic linkages in order to improve representation of the external terms.