Knowledge Representation and Management: Benefits and Challenges of the Semantic Web for the Fields of KRM and NLP

A.-M. Rassinoux, Section Editor for the IMIA Yearbook Section on Knowledge Representation and Management
Information Systems Division, Geneva University Hospitals, Geneva, Switzerland

Summary

Objectives: To summarize excellent current research in the field of knowledge representation and management (KRM).

Method: A synopsis of the articles selected for the IMIA Yearbook 2011 is provided and an attempt to highlight the current trends in the field is sketched.

Results: This last decade, with the extension of the text-based web towards a semantic-structured web, NLP techniques have experienced a renewed interest in knowledge extraction. This trend is corroborated through the five papers selected for the KRM section of the Yearbook 2011. They all depict outstanding studies that exploit NLP technologies whenever possible in order to accurately extract meaningful information from various biomedical textual sources.

Conclusions: Bringing semantic structure to the meaningful content of textual web pages affords the user with cooperative sharing and intelligent finding of electronic data. As exemplified by the best paper selection, more and more advanced biomedical applications aim at exploiting the meaningful richness of free-text documents in order to generate semantic metadata and recently to learn and populate domain ontologies. These later are becoming a key piece as they allow portraying the semantics of the Semantic Web content. Maintaining their consistency with documents and semantic annotations that refer to them is a crucial challenge of the Semantic Web for the coming years.

Keywords

Semantic Web, Natural language processing (NLP), knowledge extraction, semantic annotation

Yearb Med Inform 2011: 121-4

Introduction

This year, the topic of the IMIA Yearbook is "Towards Health Informatics 3.0". Health Informatics 3.0 offers new perspectives for heath care workers by combining various information sources, ranging from clinical or laboratory data to literature or guidelines, based on semantic tags. Semantic tagging together with context-based filtering of information appears as one of the fundamental features of the third generation of Internet-based services called Web 3.0 from which the Semantic Web is issued. The purpose of the Semantic Web [1] is to enhance the previously text-based web with machine-interpretable semantics, in order to allow automated processing and integration of the huge amount of electronically unstructured information.

The Semantic Web community has brought many benefits to the field of knowledge management and representation (KRM), as well as to the field of natural language processing (NLP). On the one hand, the World Wide Web Consortium (W3C) has adopted powerful formats that promote the interoperability and sharing of data resources [2] (see RDF: Resource Description Framework, SPARQL: RDF Query Language, and OWL: Web Ontology Language). On the other hand, collaborative efforts were undertaken to provide automated semantic tools able to recognize and extract textual information in order to learn and populate semantic networks. These semantic analysis methods involve more and more NLP techniques.

NLP is a continuously evolving field which has matured over the last decade through its use in a variety of robust and scalable applications as exemplified by the collection of selected papers for the KRM section of the Yearbook 2011. With the advent of the Semantic Web, NLP techniques were specifically trained to enable relationships between concepts to be identified from sentences thus producing fine-grained semantic markups in various domains and languages. This task, often referred as information extraction (IE), is realized by the successive application of NLP components of which the main ones are: tokenization, word and sentence segmentation, named entity and term tagging, part-of-speech (POS) tagging, syntactic parsing and finally semantic concept and relation annotation. The growing interest towards these parsing techniques has contributed to the development of open source solutions for text processing that deliver scalable, reusable and robust NLP components. To name only a few, the Stanford Parser [3] as well as the NLP toolkit platforms GATE (General Architecture for Text Engineering) [4], or UIMA (Unstructured Information Management Architecture) [5], initially developed for the general domain, demonstrate a good performance in the biomedical domain.

Best Paper Selection

The above development and research trends were already highlighted throughout the topics of the KRM section of the previous IMIA Yearbooks whose titles were: "Structuring knowledge for better access" [6], "Towards interoperable medical terminologies" [7] or "Trans

forming textual information into useful knowledge" [8]. This year, the best paper selection (see Table 1) pursues the semantic structuring approach and corroborates that knowledge extraction from free-text documents is an active and promising research area where NLP techniques play a significant and growing role. The real-time discovery of critical related information is particularly appealing and challenging to the biomedical domain. This is due to the complexity of the biomedical language and the dramatic change in the types and amount of electronic data available to researchers. A short summary for each paper can be found in the appendix of this synopsis.

This year's selection mainly focuses on the design of biomedical information extraction tools. The first paper [9] aims at extracting causal relations on HIV drug resistance from PubMed abstracts. The study, carried out by Cao et al. [10], focuses on extracting topics and keywords from complex clinical questions. Coulet et al. [11] report on a system that builds semantic networks from the pharmacogenomics knowledge extracted from MEDLINE abstracts. Denny et al. [12] describe an approach to detect timing and status descriptors for colonoscopy testing from electronic medical records. Finally, the last paper [13] highlights a biomedical knowledge extraction and visualization framework evaluated on the GENIA corpus.

NLP technologies are exploited whenever possible during the extraction process. In particular, for POS tagging, the Stanford Parser [3] is used in all the aforementioned projects except in the system developed by Denny et al. [12]. This later applies a locally-developed and general-purpose biomedical NLP system.

It is also worth noting that advanced NLP applications that support entity and relation extraction, need to rely on established resources that describe the domain knowledge necessary to guarantee performance during syntactic analyses [14]. Domain-specific lexicons, terminological sources as well as existing or learned ontologies are dedicated for this purpose. They often require important acquisition efforts as illustrated in the selected papers. In [9], a list of drug names was collected from websites related to HIV treatment such as the Stanford HIVDB and RegaDB in order to help collecting relevant abstracts and candidate sentences for the extraction process. In [10], 4654 clinical questions were annotated by the investigator who recorded the question. The criticality of the consistency of these annotations for question analysis tasks is pointed out by the authors. Lexicons of pharmacogenomics key entities available from PharmGKB were used by the authors in [11] in order to retrieve sentences mentioning pairs of key entities. These related entities were extracted and organized into an OWL ontology. This later was subsequently applied to all raw relationship instances in the corpus in order to produce a normalized set of relationships representing the semantic content of the corpus. In [12], concept hierarchies derived from UMLS metathesaurus were exploited to identify relevant concepts pertaining to colonoscopy. Moreover, a lexicon of temporal phrases as well as a lexicon of base word forms for each status category, were developed by the authors in order to accurately detect these features from sentences. Finally, the BioKEVis framework described by Jahiruddin et al. [13] is integrated with a biomedical named entity recognizer called ABNER that identifies a subset of concepts related to the existing molecular biology ontology GENIA.

Conclusion and Outlook

The fields of KRM and NLP have greatly evolved since this last decade with the emergence of the Semantic Web. While documents are central to knowledge management, their automated parsing towards providing semantic annotations, offers new perspectives for collaborative sharing and intelligent search based on semantic content. This is also a crucial step towards a full automated document management system. The best paper selection corroborates the increasing use of the NLP technologies to assist in the extraction and categorization process of relevant textual information. The advanced state of NLP, including machine learning and retrieval systems, together with the growing demand for ontologies to power the Semantic Web, are fuelling researches based on (semi-) automatic ontology learning from free-text documents [15]. This active and promising field, depicted in the recent review paper of Liu et al. [16], sheds light on a new technical challenge for the distributed Semantic Web. It concerns the complex task of maintaining the consistency between evolving documents, evolving ontologies, and semantic annotations that refer to them [17, 18].

Acknowledgement

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

Table 1 Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2011 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
▪ Bui QC, Nualláin BO, Boucher CA, Sloot PM. Extracting causal relations on HIV drug resistance from literature. BMC Bioinformatics 2010,11:101.
▪ Cao YG, Cimino JJ, Ely J, Yu H. Automatically extracting information needs from complex clinical questions. J Biomed Inform 2010 Dec;43(6):962-71.
▪ Coulet A, Shah NH, Garten Y, Musen M, Altman RB. Using text to build semantic networks for pharmacogenomics. J Biomed Inform 2010 Dec;43(6):1009-19.
▪ Denny JC, Peterson JF, Choma NN, Xu H, Miller RA, Bastarache L, Peterson NB. Extracting timing and status descriptors for colonoscopy testing from electronic medical records. J Am Med Inform Assoc 2010 Jul-Aug;17(4):383-8.
▪ Jahiruddin, Abulaish M, Dey L. A concept-driven biomedical knowledge extraction and visualization framework for conceptualization of text corpora. J Biomed Inform 2010 Dec;43(6):1020-35.

References

1. Berners-Lee T, Hendler J, Lassila O. The Semantic Web. Sci Am, 2001: 34-43.
2. Feigenbaum L, Herman I, Hongsermeier T, Neumann E, Stephens S. The Semantic Web in action. Sci Am, 2007 Dec;297(6):64-71.
3. Klein D, Manning CD. Accurate unlexicalized parsing. Proc of the 41st Meeting of the Association for Computational Linguistics, 2003:423–30.
4. Bontcheva K, Tablan V, Maynard D, Cunningham H. Evolving GATE to Meet New Challenges in Language Engineering. Nat Lang Eng 2004,10: 349-73.
5. Ferrucci D, Lally A. UIMA: an architectural approach to unstructured information processing in the corporate research environment. Nat Lang Eng 2004, 10(3:4): 327-48.
6. 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.
7. 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.
8. Rassinoux AM. Decision Support, Knowledge Representation and Management: Transforming Textual Information into Useful Knowledge. In: Kulikowski C, Geissbuhler A, editors. Yearb Inform Med 2010:64-7.
9. Bui QC, Nualláin BO, Boucher CA, Sloot PM. Extracting causal relations on HIV drug resistance from literature. BMC Bioinformatics 2010, 11:101.
10. Cao YG, Cimino JJ, Ely J, Yu H. Automatically extracting information needs from complex clinical questions. J Biomed Inform 2010 Dec;43(6):962-71.
11. Coulet A, Shah NH, Garten Y, Musen M, Altman RB. Using text to build semantic networks for pharmacogenomics. J Biomed Inform 2010 Dec; 43(6):1009-19.
12. Denny JC, Peterson JF, Choma NN, Xu H, Miller RA, Bastarache L, Peterson NB. Extracting timing and status descriptors for colonoscopy testing from electronic medical records. J Am Med Inform Assoc 2010 Jul-Aug;17(4):383-8.
13. Jahiruddin, Abulaish M, Dey L. A concept-driven biomedical knowledge extraction and visualization framework for conceptualization of text corpora. J Biomed Inform 2010 Dec;43(6):1020-35.
14. Bodenreider O. Lexical, terminological and ontological resources for biological text mining. In: Ananiadou S, McNaught J, editors. Text mining for biology and biomedicine: Artech House; 2006:43-66.
15. Maynard D, Li Y, Peters D. NLP Techniques for Term Extraction and Ontology Population. In: Buitelaar P and Cimiano P, editors. Bridging the Gap between Text and Knowledge – Selected Contributions to Ontology Learning and Population. IOS Press, 2008:107-28.
16. Liu K, Hogan WR, Crowley RS. Natural Language Processing methods and systems for biomedical ontology learning. J Biomed Inform 2011;44:163-79.
17. Uren V, Cimiano P, Iria J, Handschuh S, Vargas-Vera M, Motta E, Ciravegna F. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. Journal of Web Semantics 2006:14-28.
18. Bada M, Hunter L. Desiderata for ontologies to be used in semantic annotation of biomedical documents. J Biomed Inform 2011;44(1): 94-101.

Correspondence to:
Anne-Marie Rassinoux, Ph. D.
University Hospitals of Geneva
Information Systems Division
4, Rue Gabrielle-Perret-Gentil
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 2011, Section Knowledge Representation and Management*

Bui QC, Nualláin BO, Boucher CA, Sloot PM

Extracting causal relations on HIV drug resistance from literature

BMC Bioinformatics 2010,11:101

In order to select the appropriate HIV treatment, virologists and medical doctors need to have up-to-date HIV drug resistance data since HIV shows a very high rate of mutation. However, these drug resistance data are so far extracted from scientific publications by experts in the field. To alleviate this time-consuming and often error-prone manual process, the authors present a novel method that automates the extraction and combination of relationships between HIV drugs and mutations in viral genomes.

First, the extraction phase uses natural language processing (NLP) techniques to produce grammatical relations between sentence constituents. A set of rules is then applied over these grammatical relations to extract causal relations between drugs and virus mutations. Second, the combination phase starts by grouping relations with the same mutations and drugs thus dealing with contradictory relations. Then, in each group, the relations are categorized into four subgroups according to their resistance properties: resistant, susceptible, responsive, and associated. Finally, a logistic regression classifier is used to combine these extracted relations in order to generate a unique resistance value for each drug-mutation pair.

The system shows promising results compared with the Stanford HIVDB dataset whose resistance data are manually gleaned from scientific publications. Indeed, for the ten most frequently occurring mutations, 85% agreement was observed between the two systems based on two levels of resistance and 76% based on three levels of resistance. The fact that the proposed method uses publicly available NLP tools makes it easily applicable to extract other types of relations such as gene-protein, gene-disease or disease-mutation.

Cao YG, Cimino JJ, Ely J, Yu H

Automatically extracting information needs from complex clinical questions

J Biomed Inform 2010 Dec;43(6):962-71

Clinicians have many questions when seeing patients but have limited time for searching and browsing the huge published biomedical literature. To address this issue, the authors are currently building a fully automated system called AskHERMES. It helps clinicians extract and articulate multimedia information from literature to answer their ad-hoc clinical questions within a time-frame that meets their requirements. Within the framework of this system, this paper reports on two natural language processing (NLP) models that together automatically and effectively extract information needs from complex clinical questions.

The first model, called automatic topic assignment, classifies clinical questions according to their general topics including etiology, procedure, diagnosis, prognosis, as well as treatment and prevention. Supervised machine-learning approaches have been explored during this phase and have achieved an average performance of 76% for F1 score. The second model, named keyword identifica-

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

tion, aims at identifying the main content of the question by extracting keywords that embed a semantic content. Both unsupervised and supervised approaches have been investigated during this step and have carried out an F1 score of 53%.

The above outcomes are based on the evaluation of 4654 annotated clinical questions, which were collected in practice, and on which all machine-learning models were trained. The authors point out that the performance of both NLP models can be considerably improved if questions are steadily assigned with a significant amount of consistent annotations.

Coulet A, Shah NH, Garten Y, Musen M, Altman RB

Using text to build semantic networks for pharmacogenomics

J Biomed Inform 2010 Dec;43(6):1009-19

Pharmacogenomics (PGx) is a new and growing field, that studies how individual genomic variations influence drug-response phenotypes. Until now, this critical knowledge is largely concealed in the text of published studies. To make it available for automated computation, the authors describe a four steps method that uses first a syntactical parser to extract raw relationships expressing PGx knowledge and then a learned ontology to normalize these later.

The first step uses lexicons of PGx key entities, describing drugs, genes and phenotypes, to extract individual sentences from a large set of article abstracts. These sentences are then parsed by the Stanford statistical natural language parser which yields a Dependency Graph data structure. This syntactical structure is the starting point of the second step that aims at extracting the raw relationships between key entities themselves or other entities that they modify. During the third step, these raw relationships and the associated entities are hierarchically organized into an OWL ontology thus mapping diverse sentence structures and vocabularies to a common semantics. Finally, this ontology is applied to all relationship instances in the corpus in order to create a large set of normalized relationships. Pharmacogenomics networks, where nodes are PGx entities and edges are normalized relationships, are built during this last step thus representing the semantic content of the corpus.

This study was carried out on a large sample of MEDLINE abstracts resulting in 87 million sentences analyzed by the syntactic parser. More than 40000 raw relationships were extracted, among them, the 200 most frequent relations and their modified entities were used to create the ontology. This common semantics is fast becoming the foundation to manage the use and discovery of PGx knowledge thus guiding the curation of this new evolving field.

Denny JC, Peterson JF, Choma NN, Xu H, Miller RA, Bastarache L, Peterson NB

Extracting timing and status descriptors for colonoscopy testing from electronic medical records

J Am Med Inform Assoc 2010 Jul-Aug; 17(4):383-8

In this paper, the authors investigate the use of natural language processing (NLP) algorithms to detect completed colonoscopies within electronic medical records (EMR) documents. Indeed, the critical challenge is to be able to identify quickly and accurately patients who need a colorectal cancer (CRC) screening. To achieve this goal, the locally developed NLP system, called KnowledgeMap concept identifier (KMCI), was tuned with new algorithms to detect both temporal expressions and status indicators that are associated to colonoscopies described in EMR notes.

The general purpose temporal extraction algorithm, developed to assign dates to colonoscopies, acts in three steps: detection of time descriptors, conversion of these later into a standard representation of date and time, and finally linkage to the corresponding EMR CRC screening test concept. Beside, to determine accurately if a patient had undergone colonoscopy, six categories of status were defined and a lexicon of base word forms for each category was created. Then, the status detection algorithm uses part of speech and verb type approaches to assign the corresponding status to the right event.

From a sample of 29 770 total EMR notes, belonging to 200 randomly selected patients who were at least 50 years old, the NLP algorithms identified 147 of 157 possible completed colonoscopies out of a total of 1208 colonoscopy references, with a calculated recall and precision greater than 90%. The authors recognize that the NLP approach detected more references to completed colonoscopy tests than a billing records query alone. They conclude that NLP constitutes a useful adjunct to traditional methods of detecting CRC screening testing.

Jahiruddin, Abulaish M, Dey L

A concept-driven biomedical knowledge extraction and visualization framework for conceptualization of text corpora

J Biomed Inform 2010 Dec;43(6):1020-35

The design of a novel biomedical knowledge extraction and visualization system, called BioKEVis, is reported by Jahiruddin et al. in this paper.

The architecture of the overall system encompasses various modules which are centered around the information components that describe domain entities and their relationships extracted from PubMed documents. Linguistic analysis, latent semantic analysis as well as rules are applied to extract key entities and relation triplets in the form: (Subject, Relational verb, Object). These relation triplets, that highlight the role of a single entity in various contexts, are used to build the semantic net. This network provides a comprehensive view of the document collection and facilitates the user navigation through the piles of documents by allowing navigation over documents with similar information components. A query-processing module is also available and allows users to formulate queries in a guided way at different levels of specificity. The fact that the system is also integrated with a biomedical entity recognizer called ABNER, that identifies a subset of GENIA ontology concepts, helps in answering queries based on biological concepts rather than on particular entities only. Finally, a document ranking mechanism, which orders retrieved documents according to their relevance to the user query, is also implemented.

The overall system was evaluated on GENIA corpus, in which entity names are tagged with GENIA ontology concepts. The assessment shows a quite high precision value but a somewhat low recall value. This indicates that most of the extracted instances are correctly identified but several relevant elements are not extracted from the texts. A refinement of the relation extraction is forecasted by the authors to improve these outcomes.