IMPLEMENTATION OF A BIOMEDICAL ONTOLOGY FOR THE GREEK LANGUAGE

Aristides Vagelatos

Computer Engineer, PhD
R.A. Computer Technology Institute, Athens, Greece

E-mail: vagelat@cti.gr

Elena Mantzari

Linguist, PhD candidate
Neurosoft S.A., Athens, Greece

E-mail: emantz@tee.gr

Mavina Pantazara

Linguist, PhD
Neurosoft S.A., Athens Greece

E-mail: mavina@neurosoft.gr

Vassilis Papapanagiotou

Medical Doctor, Cardiologist
Iaso General, Athens, Greece

E-mail: papapanagiotouv@yahoo.gr

Chryssoula Kalamara

Medical Doctor, Radiologist
Athens' Euroclinic, Athens, Greece

E-mail: kalamach@otenet.gr

ABSTRACT

This paper presents the design and implementation of terminological and specialized textual resources that are produced in the framework of the Greek national R&D project "IATROLEXI" (www.iatrolexi.gr). The aim of the project is to create the critical infrastructure for the Greek language, i.e. linguistic resources and tools, to be used in high level Natural Language Processing (NLP) applications in the domain of Biomedicine. The project builds upon existing resources that have been developed by the project partners, i.e. a Greek morphological lexicon of about 100.000 words, and language processing tools such as a lemmatizer and a morphosyntactic tagger, and it will further develop new resources such as a specialized corpus of biomedical texts and an ontology of medical terminology. Emphasis in this paper is given on the design and development of a biomedical ontology for the Greek language.

KEYWORDS: Biomedical Ontology, Natural Language Processing, Medical Informatics, Biomedical data mining

INTRODUCTION

In order to access information, medical practitioners, researchers, patients, or other interesting parts in the medical market are usually provided with unsophisticated tools, such as simple search engines which are seriously limited by their reliance on keyword-matching. These search mechanisms are unable to find information described by different terms and they often return information results that use the same words with a different meaning, while they are unable to combine information from diverse sources. These problems can be alleviated if search engines no longer search for matching keywords but for matching semantic concepts that underlie the information in web pages. The lack of high level language tools to facilitate accuracy and precision in accessing and retrieving the relevant information is harder in a less-used language like Greek, due to the limited research funding and the restricted interest by the medical industry, and also due to the intrinsic particularities of the Greek language morphology.

The project IATROLEXI ("IATROLEXI" project is being partially funded by General Secretariat of Research & Development within Measure 3.3 of "Information Society" Operational Program) aims at the creation among other tools, of a biomedical ontology for the Greek language which will constitute the groundwork for advanced NLP applications in the domain of biomedicine: i.e. text indexing, information extraction and retrieval, data mining, question answering systems, etc.

The methodology that is used for the construction of the ontology, combines bottom-up and top-down approaches for the determination of the semantic/conceptual framework to be used for the knowledge representation of the biomedical domain (i.e. selection of a conceptual hierarchy, semantic classes, relations between concepts, etc.) and the selection of the relevant biomedical terms that designate and instantiate the concepts of those hierarchy nodes. The UMLS semantic network is used as a frame basis for expressing the IATROLEXI's ontology.

REVIEWING DEFINITIONS OF THE TERM “ONTOLOGY”

The idea of capturing knowledge in a structured way is attributed to Aristotle, who first focused in a systematic way on the practical problem of representing the structure of reality. Even if since then philosophy has adopted a body of new analytical tools for the examination of ontological problems, a lot of ideas and terms of the ontological theory, such as the notions of *category* and *hierarchy* can fairly be attributed to Aristotle.

The philosophical ontology has taken a lot of forms and different philosophical schools have followed different approaches, however central role in the philosophical ontology remains the explicit and exhaustive classification of all entities. In Smith, 2003, the philosophical ontology is determined as the "science of what is, of the types and the structures, the objects, the attributes, the incidents, the processes and the relations in each domain of reality”. On the other hand, Guarino et al., 2004 adopts a *cognitive* approach that considers the categories as cognitive artifacts that they are dependent on the human perception.

Within the area of Artificial Intelligence (AI) the term "ontology" is referred to a category of artifacts that they are the products of the ontology engineering. The *ontology engineering* is determined by the Gomez-Perez et all., 2004, as "the total of activities that concerns the activity of growth of ontology, the circle of life of ontology, the methods and the methodologies for building ontologies, as well as the tools and the languages that support them". The

statement of Gruber, 1993 that "an ontology is a specification of conceptualization" constitutes the first effort to define the term in the framework of AI. This definition was criticized and led to different interpretations. Guarino, 1998 attempted to clarify the use of the ontology by the philosophical and the AI community. In the philosophical sense, the ontologies are systems of categories that account for the particular way of reading the world (it is what Guarino defines as conceptualization). On the other hand, the AI reading of ontology is referred to an artifact that is formulated via a logically organized vocabulary used to describe a certain reality via a set of *statements* that they determine the meaning of words in the vocabulary.

Moreover, the term "ontology" is often used in a way that does not fit to any of the above definitions. It is used in order to simply refer to "controlled vocabularies", as in the case of the biomedical ontology GO (Gene Ontology), which only provides a practical framework for recording biological comments that are applied in the products of genes. The authors of GO were not further interested in software applications neither in formulation of a formal theory that describes these terms.

ONTOLOGY AND INFORMATION SCIENCE

Apart from the domain of philosophy, the last years the term ontology constitutes one of the central issues regarding the representation of knowledge in the domain of information technology. The first and basic reason for the appearance of "the new ontology" as it is named by Smith is due to what we call "the Babel Tower problem". Different teams of designers of knowledge database systems use particular terms and concepts in order to build the frameworks for the information representation. Different databases can use the same terms but with different meanings. Alternatively, the same meaning can be expressed via different terms.

Initially, such incompatibilities were solved individually. Then, it was recognized that a common reference ontology – a common ontology of entities – could offer important advantages against the "ad hoc" or "punctual" approach, and the term ontology was reborn by the computer specialists in order to describe the construction of an exemplary, f*ormal* and in logical *terms description* of this type. In this framework, ontology is deemed as *a dictionary of terms that they are formulated in a formal syntax accompanied by their accepted definitions*. This dictionary is designed with the aim to producing a lexicological or taxonomic framework of knowledge representation which will be shared by different communities of computer scientist systems.

Thus, the methods that are used in the design and the construction of ontologies, on the one hand can be traced back to previous initiatives of the databases management systems, and on the other they comprise methods similar to those applied by the philosophy, included the methods that are used by the Logic and the formal semantic theories.

BIOMEDICAL ONTOLOGIES

In biomedicine, there is a large amount of terminology collections (ontologies, taxonomies, semantic networks) that have been developed for *different purposes* (concept-based text indexing, information extraction, question/answering systems, electronic patients files, statistical reports etc.), for *different subdomains* (diseases, microorganisms, medical devices,

processes, drugs etc.), and by *different organizations* (World Health Organization, governmental representations, professional groups etc.).

The most well-known biomedical semantic network is the **Unified Medical Language System** (ULMS - http://umlsinfo.nlm.nih.gov/). It has been developed by the US National Library of Medicine (NLM) in order to unify different terminologies used by different health organizations. It consists of the *Metathesaurus* (a large, multilingual concept-based database including terms and concepts from 100 different dictionaries, thesauri and taxonomies) and the *Semantic Network* (a categorization of these concepts according to semantic types and sets of relationships that hold between them). The most recent edition of UMLS includes 135 semantic types (e.g. *organism, anatomic structure, biological function, event, object, concept* or *idea*) and 54 types of relations between them.

The **Foundational Model of Anatomy ontology** (FMA - http://sig.biostr.washington.edu/projects/fm/AboutFM.html) is one of the most important and one of the largest computer-based knowledge sources in the biomedical sciences.
This domain ontology is a representation of classes and relationships concerning the structure of the human body, based on a set of well-defined principles, top level schemes, Aristotelian definitions and frame-based formalism. At first, it was developed as an extension of the anatomic concepts included in the UMLS, but now it is considered as a reference ontology, necessary for correlating different points of view in the domain of anatomy and aligning existing and submerging ontologies in bioinformatics. It currently contains approximately 75,000 classes and over 120,000 terms; over 2.1 million relationship instances from 168 relationship types link the FMA's classes into a coherent symbolic model. The FMA is one of the largest computer-based knowledge sources in the biomedical sciences.

The **Gene Ontology** (GO - http://www.geneontology.org/) was developed to address the need for a consistent representation of the information about the genes products provided by different databases. Structurally, it is divided in three ontologies based on the logical relations is_a and part_of, whose top nodes are: the Cellular component, the Molecular function and the Biological process. These three structured networks are treated as separate ontologies, and no ontological relation is defined between them. One of the major problems of the GO is located in the management of the relations between concepts, especially in the is_a relation. Nevertheless, the GO is broadly used in the biologists' community, and they try to represent it in a formalism of description logic in order to improve its compatibility with computers use.

The **MeSH®** (Medical Subject Headings - http://www.nlm.nih.gov/mesh/) is currently the most reliable thesaurus of medical terms. Since its first edition in 1960, it has been developed and maintained by the US NLM, for indexing, recording and searching articles, books, periodicals etc. included in the catalogue MEDLINE®/PubMed® and other databases of NLM. It has been translated in several languages and is broadly used by libraries and other institutions worldwide for indexing and listing medical documents and information. The 2006 edition includes about 24.500 descriptors (i.e. preferred terms), structured in 16 tree categories (such as *anatomy, organisms, diseases, chemical substances* etc.). Despite its taxonomic structure of concepts and terms, MeSH cannot be used independently as an ontological resource in NLP applications. It is connected to the Metahesaurus of the UMLS, and the ontological information about the descriptors is accessible only via the UMLS.

REPRESENTATION FORMALISM

One of the crucial decisions when building an ontology is to decide on the formalism in which the ontology will be represented and implemented. Although many formalisms, such as Ontolingua and LOOM have been used in the last decades for this task, the growth of Internet has led to the creation of web-based ontology markup languages, such as RDF, RDFS, DAML +OIL and OWL, which seem to exploit better the particularities of the World Wide Web.

AS OWL (Web Ontology Language, (http://www.w3.org/2004/OWL/) has become the standard ontology representation language of the WWW in general, and the Semantic Web in particular, the adoption of OWL for the needs of IATROLEXI is rather justified; not only due to the need of developing a biomedical ontology that will be WWW exploitable, but mainly due the semantic power of OWL's representation mechanism, which has been unanimously recognized among the knowledge engineering community. There is a vastly growing community working on OWL and new OWL tools emerge on a day to day basis. Among these, is Protégé (http://protege.stanford.edu), the open source ontology editor, which has an OWL plug-in that facilitates creating and reasoning with ontologies specified in OWL through a graphical user interface. Thus “Protégé” is also adopted for the ontology coding needs of IATROLEXI. It is not a coincidence that the developers of Protégé come from the area of medical informatics, since for many years now biomedical knowledge representation seeks to exploit formal and logic-based semantics offered by ontologies.

BIOMEDICAL ONTOLOGY FOR THE GREEK LANGUAGE

As the aim of IATROLEXI is to build a generic and application independent infrastructure for the language processing of the biomedical data written in Greek, the selection of the ontology according to which the terminological data will be built upon and structured should be content generic, multi-purpose in scope and application reusable. These essential characteristics are fairly attributed to the UMLS Knowlegde resources, namely UMLS Metathesaurus (MT) and UMLS Semantic Network (SN). Moreover, the UMLS is widely used in a variety of applications, serving information retrieval, text indexing, NLP automating indexing research and structured data entry.

The Metathesaurus is a large and multi-purpose database that contains information about biomedical concepts and terms, from about 100 diverse vocabularies, classifications and thesauri used for patient care, administrative health data, bibliographic and full-text databases, referred to as ``source vocabularies", providing, thus, a common and fully-unified structure for the representation of the biomedical knowledge. It is organized by concept, which is defined as a cluster of terms representing the same meaning, i.e. synonyms, variants and translation equivalents.

The UMLS MT includes more than 800,000 concepts and two million concept names in its source vocabularies. The UMLS SN offers the conceptualisation tools and assertions for the representation of the biomedical concepts and their linking mechanisms. The SN consists of: a) a set of 135 broad subject categories, the Semantic Types, that provide a consistent categorization of all concepts represented in the UMLS Metathesaurus, and b) a set of 54 useful and important relationships, the Semantic Relations, that exist between Semantic Types. The purpose of the Semantic Network is to provide a consistent categorization of all concepts represented in the UMLS MT and to provide a set of useful relationships between these concepts. The primary link in the SN is the “isa” relation. This establishes the hierarchy

of types within the SN and is used for deciding on the most specific semantic type for each concept in the MT. In addition, a set of non-hierarchical relations between the types are identified.

Up to now, the whole number of the SN semantic types and semantic relations have been translated into Greek, while both English and Greek versions of the SN have been fed into Protégé for further processing and evaluation by the two physicians-partners of the project. Adopting UMLS Semantic network as an initial top-level ontology, and mapping it into Greek, we achieve compliance with a resource which is considered reference ontology in the biomedical informatics domain, and at the same time we gain access to the conceptual information for some thousands of biomedical terms.

CONCLUSIONS

NLP infrastructure is a key element in the further development of informatics applications in several areas, such as data mining, knowledge-based decision support, terminology management, and systems interoperability and integration. A significant body of work now exists that report on experiences with various approaches in important problem areas of research. On the contrary in the biomedical field and especially for the Greek language, there is not much work implemented.

Project IATROLEXI aims to cover this certain gap by developing a number of NLP resources as well as application for the scientific community. On the one hand the scientist may use the outcomes of the project in his/her own way towards his/her special research needs. On the other hand, the user may look for information in texts or make searches with specific terms or combination of terms or relations that relate terms to each other.

REFERENCES

Bruijn, B., Martin J. (2002). Getting to the (c)ore of knowledge: mining biomedical literature. Int. Journal of Medical Informatics, Vol. 67, 7-18.

Gomez-Perez A, Fernandez-Lopez M, Corcho O. (2004). Ontological engineering. London: Springer-Verlag.

Gruber, T. R. (1993). Toward principles for the design of ontologies used for knowledge sharing, Technical Report KSL 93-04, Knowledge Systems Laboratory, Stanford University.

Guarino N, Welty C. (2004). An overview of OntoClean. In: Staab S, Studer R, editors. The handbook on ontologies. Berlin: Springer- Verlag, 151-72.

Guarino N. (1998). Formal ontology and information systems. In: Guarino N, editor. Proceedings of the first international conference on formal ontologies in information systems. FOIS'98. Trento, Italy: IOS Press, 3–15.

Kokkinakis, D. (2006). Developing resources for Swedish Bio-Medical text mining. Proceedings of the 2nd Int. Symposium on Semantic Mining in Biomedicine, Jena, Germany.

Smith B. (2003). Ontology: philosophical and computational. In: Floridi L, editor. The Blackwell guide to the philosophy of computing and information. Oxford: Blackwell Publishers.