Biomedical Data Mining for the Greek language

A Vagelatos[1], E Mantzari[2], G Orphanos[2], C Tsalidis[2], M Pantazara[2],
C Kalamara[3], C Diolis[1]

[1]Computer Technology Institute, Athens, Greece
[2]Neurosoft S.A., Athens, Greece
[3]Athens' Euroclinic, Greece

Abstract

The project IATROLEXI (http://www.iatrolexi.gr) aims at the creation of the critical infrastructure 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, text mining, question answering systems, etc. To accomplish this, a number of essential tools and resources will be constructed for the Greek language, which will allow better management and processing of the information in the biomedical field. This will be made possible through the compilation of a representative corpus of biomedical texts and the construction of NLP tools for structural, lexical and semantic annotation of those texts.

1. Introduction

Natural Language Processing (NLP) has been applied to biomedical text for decades, in fact, soon after computerized clinical record systems were introduced in the mid 1960s. In recent years, research has continued to focus on text indexing and document coding to allow powerful and meaningful retrieval of documents. Document indexing uses terms from a glossary or ontology (MeSH, Gene Ontology, Galen4) or text features such as words or phrases. Most NLP systems in clinical medicine work with text from patient records such as discharge summaries and diagnosis reports. NLP systems in bioinformatics use mostly articles or abstracts from the scientific medical literature.

The expected output of the project "Iatrolexi" (www.iatrolexi.gr) are tools that address directly the final user of the biomedical information, such as a spelling checker of Greek medical terms, and also tools that will mainly assist processing of the Greek biomedical texts and improve search and retrieval of biomedical data, such as a tagger for morphosyntactic annotation appropriately tuned to the particularities of the biomedical sublanguage and an ontology of the Greek biomedical terminology.

2. Methods

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.

3. Results

From the entire set of Greek NLP components specified in the plan of IATROLEXI, some components were available at the commencement of the project (as partners' contributions), some were developed at the early project stages and others are under development; the latest will become available progressively till the end of the project. This led us to divide the document annotation process into two phases: a) production of basic annotations (through utilisation of available components) and b) production of advanced annotations (through utilisation of forthcoming components). We call "basic annotations" the annotations that represent the outcome of tokenisation, sentence splitting, morphosyntactic tagging and biomedical word identification. Annotations that characterise a document globally (e.g. title, author(s), affiliation(s), source) are also basic annotations. We call "advanced annotations" the annotations that represent the outcome of multi-word term recognition and semantic tagging.

The software implementation platform of all NLP components is Java v 1.5. To integrate these components into the annotation process, we adopted the Apache UIMA platform. UIMA stands for Unstructured Information Management Architecture; it was developed by teams from IBM Research and IBM Software Group and is now released to the open-source community as an Apache project. Among the many useful (and sophisticated) features of UIMA, we were mainly attracted by a) its pretty straightforward mechanism of composing document analysis engines from primitive NLP components that cooperate via well-defined interfaces, and b) its powerful annotation representation model, called Common Annotation Structure (CAS), which borrows many ideas from the object-oriented world: annotations are objects; object types may be related to each other in a single-inheritance hierarchy; a sufficient set of basic types is already defined (in accordance with the primitive data types and data structures of programming languages, i.e. integer, real, boolean, string, array, list, structure); the developer can extend these types and define an arbitrarily rich type system.

In UIMA parlance, NLP components are called *annotators*. One annotator is combined with other annotators in a document processing *flow*. During runtime, each annotator processes one CAS at a time; the CAS contains the document text along with annotations produced by preceding annotators (if any). The annotator examines the document text and/or the available annotations and produces new annotations; it then adds the new annotations to the CAS and returns it to the UIMA runtime environment so as to be delivered to the next annotator in the flow.

4. On-going work

Currently, a part of our efforts focuses on the completion of the multi-word term recogniser. In subsection 5.5 we presented the extraction of candidate multi-word terms from the corpus, based on linguistic knowledge. To automatically decide upon real multi-word terms, we have to exploit some type of statistical evidence which will help us to compute a term-validity metric.

In parallel, we are developing a bilingual biomedical dictionary, by aligning Greek biomedical terms (collected from biomedical dictionaries and from the corpus) with American biomedical terms found in the UMLS. So far, we have coded a critical mass of 17.000 terms. Mapping a Greek term to a UMLS term is very important, as through UMLS we gain access to other very significant pieces of information about the term (of course in English): its classification (semantic type) in the Semantic Network, its definition, its synonyms and its relations with other terms from over 100 vocabularies contained in the Metathesaurus. We envisage (at least) three applications of the bilingual biomedical dictionary:

a) Semantic tagging. Any term found in the dictionary can receive an annotation that encodes its semantic type and thus links the term with the UMLS Semantic Network.
b) Bilingual term searching. A Greek term can be translated to its American equivalent(s) and then searched in American texts, and vice-versa.

Ontology-based query expansion. A query that contains a term of a specific semantic type can be enriched with other terms of the same semantic type or with terms of narrower semantic types.

5. Discussion and 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.

Acknowledgements

"IATROLEXI" is a Greek R&D project. It is partially funded by General Secretariat of Research & Development (www.gsrt.gr) (project code: 9) within Measure 3.3 of "Information Society" Operational Program.

References

[1] Alexander, U. Methods in Biomedical Ontology. Journal of Biomedical Informatics, Vol. 9 (2006) 252--266

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

[3] Eysenbach, G. The Semantic Web and healthcare consumers: a new challenge and opportunity on the horizon?. International Journal of Healthcare Technology and Management, Vol. 5, No.3/4/5 (2003) 194 - 212

[4] Spyns, P. Natural Language Processing in medicine: an overview. Methods Inf. Med. Vol. 35 (1996) 285--301

Address for correspondence

Name: Aristides Vagelatos
Full postal address: RACTI, Davaki 10, 11526 Athens, Greece
E-mail address: vagelat at cti dot gr