Biomedical Data Mining for the Greek language

AUTHORS: Vagelatos A 1, Mantzari E 2, Oprhanos G 2, Tsalidis C 2, Pantazara M 2, Kalamara C 3, Diolis C 1

INSTITUTIONS: 1 RA Computer Technology Institute, 2 Neurosoft S.A., 3 Athens' "Euroclinic"

BACKGROUND:
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.

OBJECTIVE: 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.

METHODS: The project 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, data mining, question answering systems, etc. To accomplish this, a number of essential tools and resources are under construction for the Greek language that will allow better management and processing of the digitally encoded information in the biomedical field.

RESULTS: The project is at the mid of its duration. Until now, a) a Greek biomedical corpus has been collected, b) the initial top-level taxonomy has been implemented, which is the first step for the ontology construction, c) the collection of biomedical terms from the corpus has been completed and the terms have been incorporated included in the already developed Greek morphological lexicon. Additionally a number of tools have been implemented (a crawler, a morphosyntactic tagger, a noun phrase chunker) to support the whole process.
Till the end of the project, the biomedical ontology remains to be fully developed as well as a number of applications that will use the abovementioned infrastructure (ontology browser, spelling checker, intelligent search engine).

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.

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