Extraction and Characterization of Biologically Relevant Relations in Biomedical Literature

Hugo Costa[1,2], Isabel Rocha[2] e Anália Lourenço[2]

[1] Department of Informatics / CCTC - University of Minho

[2] IBB - Institute for Biotechnology and Bioengineering

Centre of Biological Engineering - University of Minho

Campus de Gualtar, 4710-057 Braga - PORTUGAL

Abstract

The development of biomedical research in recent years has produced a large amount of information mostly spread throughout scientific literature that is not directly available for computational processing. Due to this wide range of scientific publications, it is hard for the community to perform tasks such as collecting, processing and analyzing relevant information in literature. Therefore, there is an important need for methods that contribute to the automation of data collecting from the literature, transforming dispersed natural language into useful knowledge, making it a valuable resource to research.

In this context, the role of Biomedical Literature Mining techniques is crucial to aid in filling the gaps between the information in literature and other data sources. Automated processes have gradually replaced manual curation, and tools for named entity recognition in text have already a considerable degree of maturity. Their application for relation extraction from the biomedical literature is the next step.

In this work, based on natural language processing systems, a conceptual model has been created containing the workflow for extraction and characterization of relations from biomedical texts. This model makes use of annotation schemes working with previously annotated texts where biological entities have been identified, combining syntactic and semantic layers.
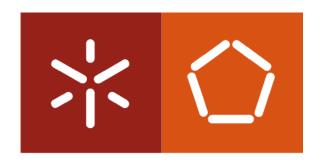
Considering this conceptual model we developed the Rel@tioN platform that combines natural language processes from the GATE platform and AIBench, a lightweight and non-intrusive workbench for the development of applications.

Under this platform, two case studies have been considered for evaluation: first, we used a corpus described in the literature, the GENIA corpus, related to human transcription factors; the second case study comes from our research group being related to the stringent response in Escherichia coli bacteria.

Detailed Program

Universidade do Minho