Poster K-21

Frame-based Analysis of Complex Relationships in Biomedical Text: Annotating and Extracting Protein Transport Information in GeneRIFs

Authors:
Zhiyong Lu (*Center for Computational Pharmacology, School of Medicine, University of Colorado*)
William Baumgartner Jr. (*Center for Computational Pharmacology, School of Medicine, University of Colorado*)
Michael Bada (*Center for Computational Pharmacology, School of Medicine, University of Colorado*)
James Firby (*Center for Computational Pharmacology, School of Medicine, University of Colorado*)
Philip Ogren (*Center for Computational Pharmacology, School of Medicine, University of Colorado*)
Kevin Cohen (*Center for Computational Pharmacology, School of Medicine, University of Colorado*)
Lawrence Hunter (*Center for Computational Pharmacology, School of Medicine, University of Colorado*)

Short Abstract: We present a novel concept recognition system named OpenDMAP (Open Access Direct Memory Access Parser), which was used to extract the biological entities and relations pertinent to protein transport, the biological process of moving proteins from one cellular location to another.

Long Abstract:
Protein transport, the biological process of moving proteins from one cellular component to another, is essential to all living organisms. The primary publications in this area grow quickly and are spread across hundreds of journals.

The number of articles and journals makes it difficult for researchers to keep pace with new findings in the literature. Thus, there is a pressing need for developing an automatic system to extract statements about protein transport from biological texts.

In order to develop and evaluate such a system, two human experts manually annotated over a thousand of GeneRIFs [1] that mention protein transport by using Knowtator [2], a general-purpose text annotation tool that was developed as a Protégé [3] plug-in and thus takes advantage of Protégé's knowledge representation capabilities. An inter-annotator agreement of 83% was achieved when overlapped spans were taken into consideration.

Next, we implemented a novel concept recognition system named OpenDMAP (Open Access Direct Memory Access Parser), which was then utilized to extract the biological entities that are being transported and are assisting in the transport as well as the cellular origin and destination of the transport event. Our experiments on a subset of 701 GeneRIFs showed over 80% F-measure when comparing automatically extracted results to manually curated

gold standard annotations [4].

Both the data annotation and the information extraction steps are built on the same knowledge model, in which each concept pertinent to protein transport is described as a frame with a name and attributes that can be filled by values of predefined types.

In this presentation, we will discuss in detail the knowledge model, the annotation efforts, and the OpenDMAP algorithm.

[1] Gene Reference Into Function.
[http://www.ncbi.nlm.nih.gov/projects/GeneRIF/GeneRIFhelp.html]
[2] Knowtator: [http://bionlp.sourceforge.net/Knowtator]
[3] Protégé: [http://protégé.stanford.edu]
[4] Lu et al., Ontology-driven analysis of complex relationships in biomedical text: Extracting protein transport information in GeneRIFs, submitted