Bootstrapping an interactive information extraction system for FlyBase curation

Ted Briscoe, Caroline Gasperin, Ian Lewin, Andreas Vlachos
Computer Laboratory
University of Cambridge

26th March, 2008

We describe an adaptive information extraction (IE) system designed to aid the curation of papers about fruit fly genomics for incorporation into FlyBase. FlyBase employs a team of about eight curators who fill in prespecified IE templetes (called proformas) for each gene and allele discussed in a given paper with curatable information associated with it. The normal approach to curation is to load the PDF of the paper into a tool such as Acroread and to use the 'Find' function to search for repeated mentions of an entity of interest. The relevant information is then typed into the appropriate template fields. Templates are then checked for consistency and automatically integrated into the database.

We have developed PaperBrowser, a tool designed to make it easier for curators to locate relevant information. The tool takes the PDF version of the paper as input and rerenders it as SciXML, a standard developed at Cambridge for representing the logical structure of scientific articles in a fashion amenable to text mining. The basic SciXML is augmented by a gene name recogniser and anaphora resolution module so that PaperBrowser is able to highlight gene names in the paper and to provide a navigation bar which allows the curator to jump to specific mentions of a given gene in the various sections of the paper. Alternatively, the curator can select a specific gene mention and the browser will highlight all the noun phrases which are anaphorically linked to that gene mention. These anaphoric links can either be coreferential, or associative to the gene's products or components, such as proteins or RNA.

User-based evaluation of PaperBrowser in comparison to the use of Acroread, with FlyBase curators undertaking the task of finding the set of genes and alleles for which templates should be constructed, has demonstrated that curation is 20% faster at no cost to accuracy when using PaperBrowser. PaperBrowser uses a conditional random field model to perform gene name recognition bootstrapped from training data derived automatically via information in FlyBase. The anaphora resolution algorithm is unsupervised but uses information from the Sequence Ontology augmented with lexemes from UMLS to identify noun phrases referring to gene products and components. The PDF extraction tool uses a commercial OCR package augmented with a seed-based machine learning technique to learn the mapping from font and format information to the logical structure of the paper. Papers describing the complete processing pipeline, intrinsic evaluation of the individual components and user-based experiments, along with test datasets are available from the FlySlip Project website (http://www.wiki.cl.cam.ac.uk/rowiki/NaturalLanguage/FlySlip).

Dagstuhl Seminar Proceedings 08131
Ontologies and Text Mining for Life Sciences : Current Status and Future Perspectives
http://drops.dagstuhl.de/opus/volltexte/2008/1508