Expose for the Studienarbeit

Integration and visualization of heterogeneous annotation results of the BioCreAtIvE Meta-Server

Author: Johannes Starlinger
Supervisor: Prof. Ulf Leser

Wintersemester 2008

Introduction

As most digitized content available (on the web) today is in the form of fulltext, the task of automatic extraction of information from this content is becoming increasingly important. One aspect of such extraction is the recognition of named entities of a certain domain in the text (Named Entity Recognition, NER) and the mapping of such entities to a given standardized form (Named Entity Normalization, NEN).
As part of the BioCreAtIvE II challenge [3] NER and NEN in the biological domain were addressed by the tasks of finding gene names in MEDLINE abstracts and mapping those names to EntrezGene identifiers. The results found by the participating groups were, though similar, quite heterogeneous. Yet the quality of the results set a new landmark in this field.

To allow resarchers to easily utilize the tools created in the course of the BioCreAtIvE II challenge, the BioCreAtIvE Meta-Server [1, 2] was developed. It provides a centralized interface to several annotation servers, giving the user the possibility to search for PubMed abstracts and to view the annotations made by these servers. Heterogeneous results are nicely integrated in the annotation view of the BC-Metaserver, and thus the inconsistencies between the annotation servers become immediately apparent. It remains to be seen whether the overall quality of the annotations can be further improved by integrating them using a suitable consensus. This motivates close examination.

Goals

The main goal of the project lies in the examination of various methods for finding a consensus between different annotation servers for gene names. The

infrastructure provided by the BC-Metaserver is accessed to obtain the annotations.

For this purpose a web-based user interface will be created that allows for transparent interactive use of the implemented consensus finding methods and provides different ways of analysing and visualizing the results.

Approach

Outline

The user interface is divided into three parts: a search area that lets the user choose PubMed abstracts for further analysis, an annotation view that displays the results returned by the annotation servers (AS) and those calculated by the consensus finding methods (CM) and finally an analysis unit that facilitates detailed comparative examination of these results.

The **search area** will allow the user to not only search for single PubMed IDs but to also find multiple abstracts by entering search terms. The result of the query is then displayed as a list giving the user the option of selecting one or more abstracts for further processing by putting them into a shopping cart style abstract queue. Placement of an abstract in this queue initiates an asynchronous request to the server for annotation and analysis. Upon successful completion of the request the abstract becomes available to the other two components.

In the **annotation view** (Fig.1) the user can select single abstracs from the processing queue for which the tagging results of the different ASs shall be displayed. Besides a list view of gene names and normalizations found by the ASs, the core feature here is the display of the annotated abstract. It shows the annotation achieved by the user selected CM as text markup and additionally visualizes the tagging of each single AS in a track view below each line of text that can be switched on and off.

The **analysis unit** (Fig.2) lets the user interactively investigate the annotations. It shows a matrix comparing the individual ASs and CMs with each other by providing percentage values for pairwise annotation agreement. Furthermore, the user has the possibility to select an AS or CM as the 'gold standard' to be used for computing precision, recall and f-measure of the other members. The comparison of the ASs and CMs can also be viewed as a dendrogram. Other diagrammatic views are thinkable. The analysis can be carried out at two different levels. Both single and multiple abstracts from the processing queue can be used as the basis for calculations. A third level may be added through server-side storage to show long term comparison results.

In addition to using the ASs reachable through the BC-Metaserver, the user is given the option of passing their own annotations to the application. For legal reasons the user shall be provided with a tool for client-side extraction of

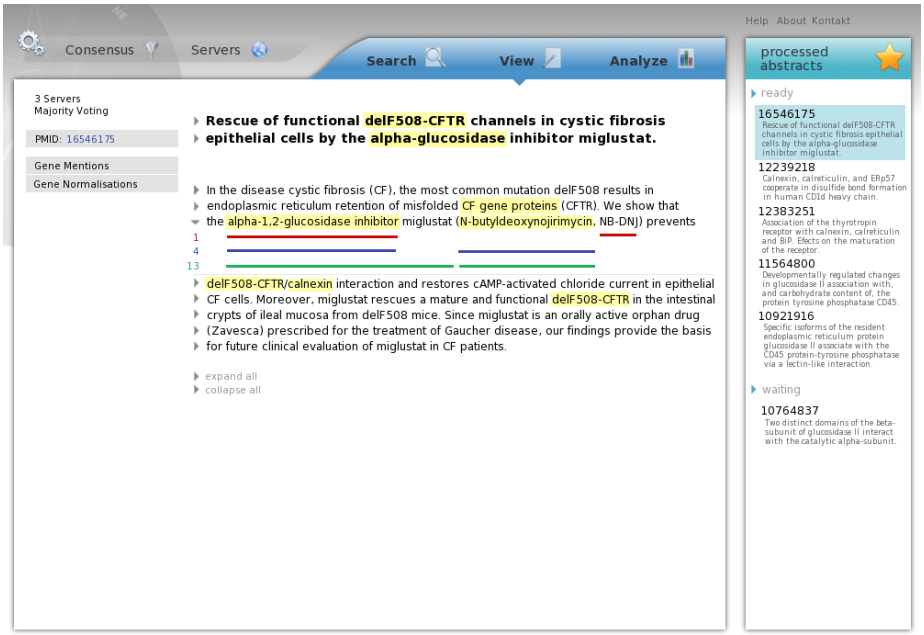

Figure 1: Annotation view mockup. The inline markup of the abstract reflects the currently selected CM. Additionaly each AS's tagging is displayed as a track below the text of the abstract. The track view can be switched on and off separately for each line of text. In the mockup the track view is expanded only for one line.

annotation data from abstracts annotated in a widely used standard, i.e. IOB. This data can then be transmitted to the server to be sourced as an additional component in consensus finding and analysis.

Overview

- Consensus finding
  - methods: majority voting, union, intersection, threshold
  - each with strict / loose evaluation option
  - user definable weights for the individual ASs that can be included in computations
- Search area
  - search by PMID or search terms
  - search for PMIDs found in uploaded files containing annotation data
  - display of search result as list
  - selection of single or multiple abstracts for further processing
- Annotation view
  - consolidated list view of gene names an normalizations reported by the ASs

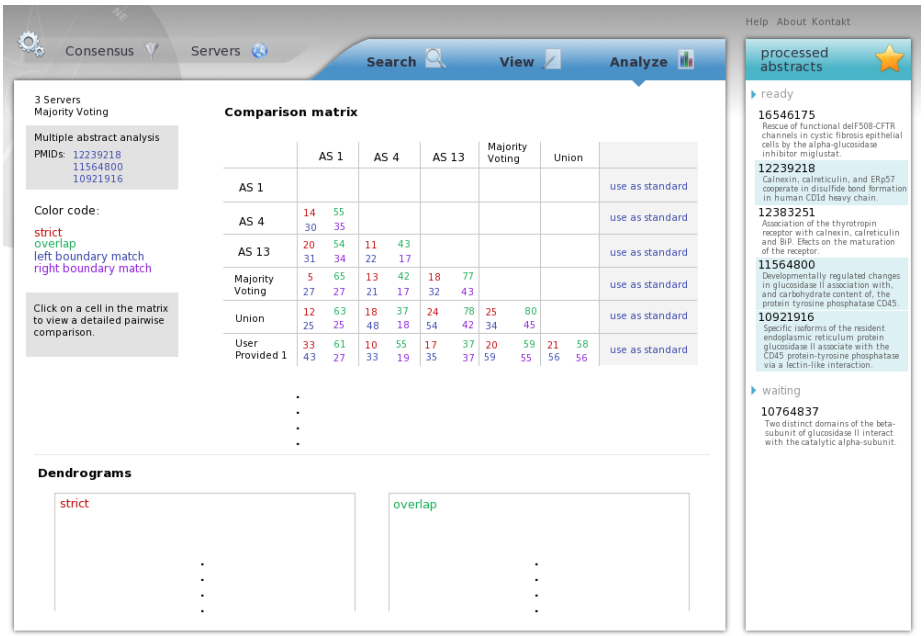

Figure 2: Analysis unit mockup with fictitious matrix values.

  - ASs to be included selectable by the user
  - CM to be used selectable by the user
  - display of the abstract with markup for tagged gene names
  - track view below each row of text showing the annotation spans of each AS

- Analysis unit
  - two levels: single abstract and multiple abstracts
  - possibly long term analysis through server-side storage
  - AS comparison matrix with pairwise annotation agreement distinguishing strict match, overlap, left boundary match, right boundary match
  - option to select one AS/CM as 'gold standard' for calculation of P/R/F-measure
  - separate calculation of micro-/macro-average when using multiple abstracts
  - dendrogram of AS comparison
  - possibly other diagrams

Technical aspects

The user interface will be built as a web application. Perl is the language chosen for server-side implementation, XHTML and JavaScript will be used on the client side. The server is responsible for finding abstracts corresponding to the users search query, communicating with the BC-Metaserver through its XML-RPC interface, computing the consensus and doing the calculations involved in the analysis of the data. Management of the user interface and integration of the results returned by the server is done on the client, including handling of the processing queue without server-side session management. The interaction of client and server shall be accomplished mostly through AJAX based communications. Where possible, results of requests are to be temporarily stored on the client to increase interface responsiveness and to take load of the server.

References

[1] BioCreAtIvE Meta-Server: http://bcms.bioinfo.cnio.es/

[2] F. Leitner et al., Introducing meta-services for biomedical information extraction,
Genome Biology 2008, 9(Suppl 2):S6

[3] BioCreAtIvE: http://biocreative.sourceforge.net/

[4] U. Leser, J. Hakenberg, What makes a gene name? Named entity recognition in the biomedical literature,
Briefings in Bioinformatics, Vol 6, No 4, 357-369, December 2005

[5] J. Bleiholder, F. Naumann, Data Fusion,
ACM Computing Survey, to appear