BioKI:Enzymes — an adaptable system to locate low-frequency information in full-text proteomics articles

Sabine Bergler, Jonathan Schuman, Julien Dubuc, Alexandr Lebedev
The CLaC Laboratory
Department of Computer Science and Software Engineering
Concordia University, 1455 de Maisonneuve Blvd West, Montreal, Quebec, H3G 1M8
bioki@cs.concordia.ca

1 Goals

BioKI:Enzymes is a literature navigation system that uses a two-step process. First, full-text articles are retrieved from PubMed Central (PMC). Then, for each article, the most relevant passages are identified according to a set of user selected keywords, and the articles are ranked according to the pertinence of the representative passages.

In contrast to most existing systems in information retrieval (IR) and information extraction (IE) for bioinformatics, BioKI:Enzymes processes full-text articles, not abstracts. Full-text articles[1] permit to highlight low-frequency information—i.e. information that is not redundant, that does not necessarily occur in many articles, and within each article, may be expressed only once (most likely in the body of the article, not the abstract). It contrasts thus with GoPubMed (Doms and Schroeder, 2005), a clustering system that retrieves abstracts using PMC search and clusters them according to terms from the Gene Ontology (GO).

Scientists face two major obstacles in using IR and IE technology: how to select the best keywords for an intended search and how to assess the validity and relevance of the extracted information.

To address the latter problem, BioKI provides convenient access to different degrees of context by allowing the user to view the information in three different formats. At the most abstract level, the ranked list of articles provides the first five lines of the most pertinent text segment selected by BioKI (similar to the snippets provided by Google). Clicking on the article link will open a new window with a side-by-side view of the full-text article as retrieved through PMC on the left and the different text segments[2], ordered by their relevance to the user selected keywords, on the right. The user has thus the possibility to assess the information in the context of the text segment first, and in the original, if desired.

2 Keyword-based Ranking

To address the problem of finding the best keywords, BioKI:Enzymes explores different approaches. For research in enzymology, our users specified a standard pattern of information retrieval, which is reflected in the user interface.

Enzymes are proteins that catalyze reactions differently in different environments (pH and temperature). Enzymes are characterized by the substrate they act on and by the product of their catalysis. Accordingly, a keyphrase pattern has entities (that tended to recur) prespecified for selection in four categories: enzymes, their activities (such as *carbohydrate degrading*), their qualities (such as *maximum activity*), and measurements (such as *pH*). The provided word lists are not exhaustive and BioKI:Enzymes expects the user to specify new terms (which are not required to conceptually fit the category). The word lists are convenient for selecting alternate spellings that might be hard to enter (α-*amylase*) and for setting up keyphrase templates in a *profile*, which can be stored under a name and later reused. Completion of the keyword lists is provided through stemming and the equivalent treatment of Greek characters and their different transliterations.

The interface presents the user with a search window, which has two distinct fields, one to specify

[1] Only articles that are available in HTML format can currently be processed.

[2] We use TextTiler (Hearst, 1997) to segment the article.

Proceedings of the BioNLP Workshop on Linking Natural Language Processing and Biology at HLT-NAACL 06, pages 91–92,
New York City, June 2006.

the search terms for the PMC search, the other to specify the (more fine-grained) keywords the system uses to select the most relevant passages in the texts and to rank the texts based on this choice. The BioKI specific keywords can be chosen from the four categories of keyword lists mentioned above or entered. What distinguishes BioKI:Enzymes is the direct control the user has over the weight of the keywords in the ranking and the general mode of considering the keywords. Each of the four keyword categories has a weight associated with it. In addition, bonus scores can be assigned for keywords that co-occur at a distance less than a user-defined threshold. The two modes of ranking are a basic "and", where the weight and threshold settings are ignored and the text segment that has the most specified keywords closest together will be ranked highest. This is the mode of choice for a targeted search for specific information, like "pH optima" in a PMC subcorpus for *amylase*.

The other mode is a basic "or", with additional points for the co-occurrence of keywords within the same text segment. Here, the co-occurrence bonus is given for terms from the four different lists, not for terms from the same list. While the search space is much too big for a scientist to control all these degrees of freedom without support, our initial experiments have shown that we could control the ranking behavior with repeated refinements of the weight settings, and even simulate the behavior of an "and" by judicious weight selection.

3 Assessment and Future Work

The evaluation of a ranking of full-text articles, for which there are no Gold standards as of yet, is difficult and begins in the anecdotal. Our experts did not explore the changes in ranking based on different weight settings, but found the "and" to be just what they wanted from the system. We will experiment with different weight distribution patterns to see whether a small size of different weight settings can be specified for predictable behavior and whether this will have better acceptance.

The strength of BioKI lies in its adaptability to user queries. In this it contrasts with template-based IE systems like BioRAT (Corney et al., 2004), which extracts information from full-length articles, but uses handcoded templates to do so. Since BioKI is not specific to an information need, but is meant to give more control to the user and thus facilitate access to any type of PMC search results, it is important that the same PMC search results can be reordered by successively refining the selected BioKI keywords until more desirable texts appear at the top. This behavior is modeled after frequent behavior using search engines such as Google, where often the first search serves to better select keywords for a subsequent, better targeted search. This reranking based on keyword refinement can be done almost instantaneously (20 sec for 480 keyphrases on 161 articles), since the downloaded texts from PMC are cached, and since the system spends most of its runtime downloading and storing the articles from PMC. This is currently a feasibility study, targeted to eventually become a Web service. Performance still needs to be improved (3:14 min for 1 keyphrase on 161 articles, including downloading), but the quality of the ranking and variable context views might still entice users to wait for them.

In conclusion, it is feasible to develop a highly user-adaptable passage highlighting system over full-text articles that focuses on low-frequency information. This adaptability is provided both through increased user control of the ranking parameters and through presentation of results in different contexts which at the same time justify the ranking and authenticate keyword occurrences in their source text.

Acknowledgments

The first prototype of BioKI was implemented by Evan Desai. We thank our domain experts Justin Powlowski, Emma Masters, and Regis-Olivier Benech. Work funded by Genome Quebec.

References

D. P. A. Corney, B.F. Buxton, W.B. Langdon, and D.T. Jones. 2004. BioRAT: Extracting biological information from full-length papers. *Bioinformatics*, 20(17):3206–3213.

Andreas Doms and Michael Schroeder. 2005. GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Research*, 33:W783—W786. Web Server issue.

M.A. Hearst. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):34–64.