Kleio: A Knowledge-enriched Information Retrieval System for Biology

Chikashi Nobata[1,2] , Philip Cotter[1,2] Naoaki Okazaki[3], Brian Rea[1,2], Yutaka Sasaki[1], Yoshimasa Tsuruoka[1], Jun'ichi Tsujii[1,2,3] Sophia Ananiadou[1,2]

1. School of Computer Science, The University of Manchester, Oxford Road, Manchester, M13 9PL, UK.
{Sophia.Ananiadou, Philip.Cotter, Chikashi.Nobata, Brian.Rea, Yutaka.Sasaki, Yoshimasa.Tsuruoka}@ manchester.ac.uk

2. National Centre for Text Mining (NaCTeM), Manchester Interdisciplinary Biocentre, 131 Princess Street. Manchester M1 7DN, UK.

3. Department of Computer Science, The University of Tokyo 7-3-1 Hongo, Bunkyo-ku Tokyo, Japan
{okazaki, tsujii}@is.s.u-tokyo.ac.jp

ABSTRACT
Kleio is an advanced information retrieval (IR) system developed at the UK National Centre for Text Mining (NaCTeM)[1]. The system offers textual and metadata searches across MEDLINE and provides enhanced searching functionality by leveraging terminology management technologies.

Categories and Subject Descriptors
J.3 [**Life and Medical Sciences**]: Medical information systems.

General Terms
Design.

Keywords
Information Retrieval, Named Entity Recognition, MEDLINE.

1. INTRODUCTION
Kleio draws upon a number of core technologies from the NaCTeM text mining tool kit to enhance automated detection and mark-up of biologically important terms appearing in text, such as gene/protein names. One of these tools is AcroMine [5], which disambiguates acronyms based upon the context in which they appear. This functionality plays a key role in searching large document collections by allowing users to expand their queries and to include synonymous acronyms without losing the specificity of the original query.

The rich variety of term variants is a stumbling block for information retrieval as these many forms have to be recognised, indexed, linked and mapped from text to existing databases [1]. Typically, most of the currently available information retrieval systems for the biomedical domain fail to deal with the problems of term ambiguity and variability. For example, the term *{2-}{(}3,4-dihydroxy{phenyl|benzene{)}}{-| }{acetate|acetic acid}* can be expressed as *2-(3,4-dihydroxyphenyl)acetic acid, 3,4-Dihydroxyphenyl acetate* and *3,4-Dihydroxybenzeneacetate*. Kleio addresses this problem by using our text mining technology for reducing the diversity of term variation.

Another key innovation of Kleio is dealing with the variety of names (terms) for denoting the same concept. To map these forms (e.g. IL2, IL-2 and Interleukin-2) to biological databases we use machine learning based term normalisation techniques which reduce term variation (e.g. il2). An advantage of applying term normalisation is to permit efficient look-up and to discover ambiguous and variant terms in the resources. The novelty of our work lies in using existing resources to automatically learn term variation patterns.

2. TEXT MINING MODULES DRIVING KLEIO
Figure 1 illustrates a simplified diagram of the system architecture. The power of Kleio lies in advanced terminology management, i.e. linking term variants for indexing and query processing. Its key components are listed below.

1. Acronym recognition and disambiguation: AcroMine [5] recognises acronyms (e.g *DEAE*) and their definitions (e.g. *diethylaminoethyl*) from the whole of MEDLINE. It also disambiguates isolated acronyms using their context and maps them into corresponding definitions. Figure 2 shows an example of acronym disambiguation performed by AcroMine. It handles not only a local acronym (e.g. "extracellular matrix (ECM)") but also a global acronym (e.g. "VEGF") by utilizing abbreviation definitions in thewhole MEDLINE abstracts.

[1] http://www.nactem.ac.uk

SIGIR'08, July 20–24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.

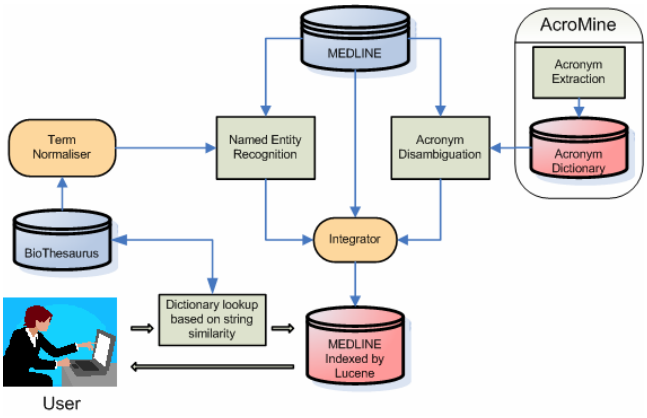

Figure 1. Kleio architecture.

2. Normalisation of biology terms: We have developed computationally efficient algorithms for term normalisation, based on a combination of exact and soft string matching methods [7]. An advantage of applying term normalisation over such large scale dictionaries is to permit efficient look-up and to discover ambiguous and variant terms in the resources. The novelty of our work lies in using existing resources to learn term variation patterns in a fully automatic manner.

3. Named entity recognition for gene/protein names: Named entity (NE) recognition is important to improve searching as it allows users to specify the entity type they want to retrieve e.g. protein, gene. Our method recognises NEs using a combination of conditional random fields and maximum entropy models to filter out false positives [4]. The dictionaries for this NE recognition process are provided by step 2.

Without NE recognition, false positive results occur due to the similarity with non-protein entities, and false negative results occur as search ignores synonym forms. This results in poor accuracy and noisy results of text retrieval. NE recognition enables a more focused query, i.e. only documents that include the annotated entity are returned. It also allows better integration with external protein databases and resources.

4. Indexing of terms: At the indexing stage, we link NEs and acronyms with the original text using Lucene[3], an open source information retrieval library. Before indexing, the extracted gene/protein names and acronyms are integrated into a unified set of terms. During the integration, acronym definitions are utilized to improve the precision of the gene/protein name recognition results. Abbreviations (e.g. CPR) are identified as non-gene/protein names if their definitions are not gene/protein names (Figure 2). When a user enters a query containing any of the surface forms, the results for all of the term variants are returned ensuring maximal expansion across the document collection. One of advantages of KLEIO over other systems such as PubMed[6] and CiteXplore [2] is that users can choose suitable ID numbers for query terms easily with dictionary lookup. Because annotated NEs are linked with unique ID numbers that are also indexed, searching with the ID number enables query expansion with synonyms in an efficient way.

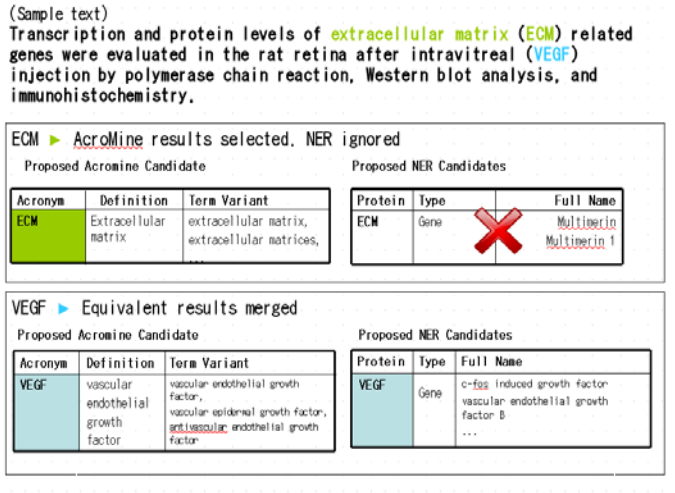

Figure 2. Named Entity-acronym integration.

REFERENCES

[1] Ananiadou, S. & McNaught, J. (Eds) (2006) Text Mining for Biology and Biomedicine, Artech House Books.

[2] CiteXplore. http://www.ebi.ac.uk/citexplore/

[3] Lucene. http://lucene.apache.org/java/docs/

[4] Okanohara, D., Miyao, Y., Tsuruoka, Y., and Tsujii, J. (2006) Improving the Scalability of Semi-Markov Conditional Random Fields for Named Entity Recognition. Proceedings of Coling/ACL 2006, Sydney, Australia.

[5] Okazaki, N. and Ananiadou, S. (2006) Building an abbreviation dictionary using a term recognition approach, *Bioinformatics*, 22(24), 3089-3095.

[6] PubMed. http://www.ncbi.nlm.nih.gov/pubmed/

[7] Tsuruoka, Y., McNaught, J., Tsujii, J. and Ananiadou, S. (2007) Learning string similarity measures for gene/protein name dictionary look-up using logistic regression. *Bioinformatics,* 23(20), 2768-2774.