SemText: a semantically enriched information retrieval system for biology

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

1. School of Computer Science, The University of Manchester, UK
2. National Centre for Text Mining (NaCTeM), Manchester, UK
3. Department of Computer Science, The University of Tokyo, Japan
* Sophia.Ananiadou@manchester.ac.uk

Overview

Semantic Text (SemText) 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.

SemText 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[2] 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 (PubMed[3]) 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*. SemText addresses this problem by using our text mining technology for reducing the diversity of term variation.

The conceptual approach to IR realised by SemText brings novel and original functionality to meet the growing interest in the biosciences looking for solutions to literature mining [2].

---

1 http://www.nactem.ac.uk
2 http://www.nactem.ac.uk/software/acromine/
3 http://www.ncbi.nlm.nih.gov/sites/entrez?db=PubMed

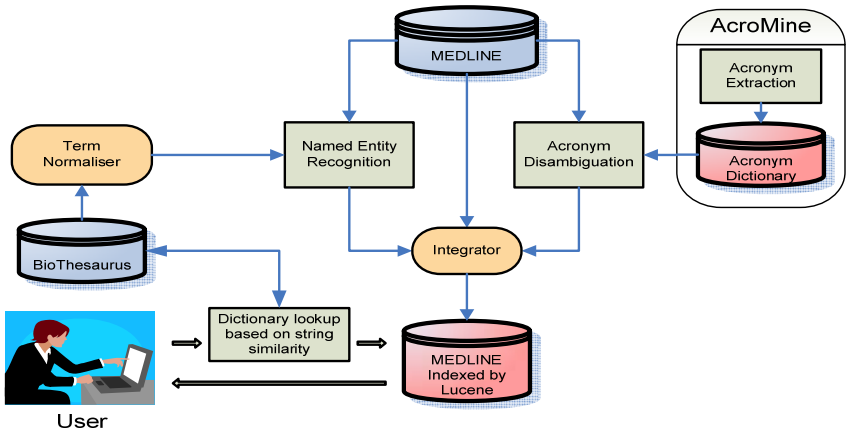

Figure 1: SemText architecture.

Text mining modules driving SemText

Figure 1 illustrates a simplified diagram of the system architecture. The power of SemText 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 [3] 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.

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 [4]. 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 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 named entities using a combination of conditional random fields and maximum entropy models to filter out false positives [5]. The dictionaries for this named entity recognition process are provided by step 2.

4. Indexing of terms: At the indexing stage, we link named entities and acronyms with the original text using Lucene [6], 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.

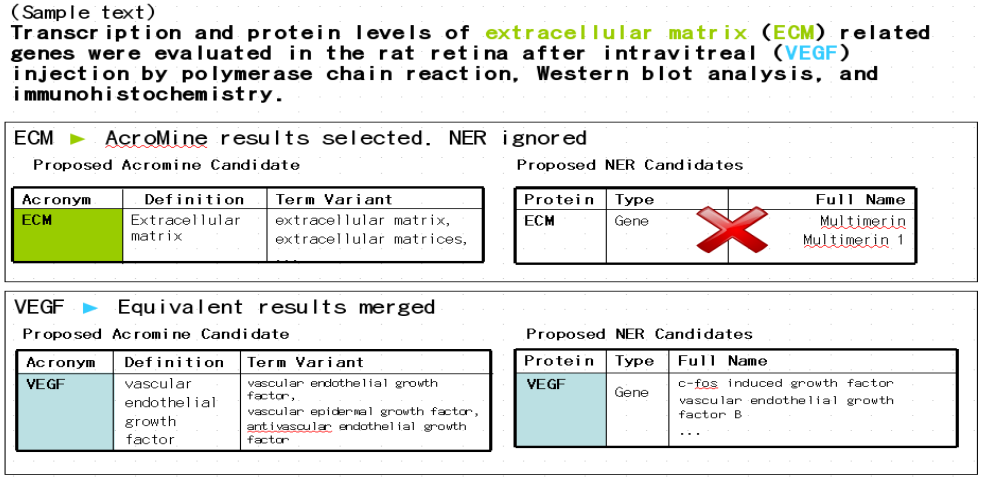

Figure 2: Named Entity-Acronym Integration.

References

[1] Ananiadou, S., Kell, D.B. and Tsujii, J. (2006) Text Mining and its potential applications in systems biology in *Trends in Biotechnology*, 12, 571-579

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

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

[4] 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*, doi:10.1093/bioinformatics/btm393

[5] 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.

[6] http://lucene.apache.org/java/docs/