Connecting Text Mining and Pathways using the PathText Resource

Sætre, Kemper, Oda, Okazaki[a], Matsuoka[b], Kikuchi[c], Kitano[d], Tsuruoka, Ananiadou, Tsujii[e]

[a]Computer Science, University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo 113-0033 Japan,
[b]The Systems Biology Institute, 6-31-15 Jingumae M31 6A, Shibuya-ku, Tokyo 150-0001 Japan,
[c]Mitsui Knowledge Industry Co., Ltd., 2-7-14 Higashinakano, Nakano-Ku, Tokyo 164-8555 Japan,
[d]Okinawa Institute of Science and Technology, 7542 Onna, Onna-Son, Kunigami, Okinawa 904-0411 Japan,
[e]National Centre for Text Mining (NaCTeM) 131 Princess Street, Manchester M1 7DN United Kingdom
[a]rune.saetre@is.s.u-tokyo.ac.jp, [b]myukiko@symbio.jst.go.jp,
[c]kikuchi-norihiro@mki.co.jp, [d]kitano@sbi.jp, [e]Sophia.ananiadou@manchester.ac.uk

Abstract

Many systems have been developed in the past few years to assist researchers in the discovery of knowledge published as English text, for example in the PubMed database. At the same time, higher level collective knowledge is often published using a graphical notation representing all the entities in a pathway and their interactions. We believe that these pathway visualizations could serve as an effective user interface for knowledge discovery if they can be linked to the text in publications. Since the graphical elements in a Pathway are of a very different nature than their corresponding descriptions in English text, we developed a prototype system called PathText. The goal of PathText is to serve as a bridge between these two different representations. In this paper, we first describe the overall architecture and the interfaces of the PathText system, and then provide some details about the core Text Mining components.

1. Introduction

One of the main challenges in biomedical Text Mining (TM) is the identification of terminology, which is a key factor for accessing and integrating the information stored in literature. Several approaches have been suggested to automatically integrate and map between resources, but the problems of extensive variability of lexical representations and ambiguity have been revealed (Nenadic et al., 2006). After the entities (proteins, genes, lipids, ions, carbohydrates, etc.) have been discovered, the natural next step is to find out what relationships exist between the entities. One type of such relationships is Protein-Protein Interaction (PPI), and it has received a lot of attention lately, for example in the Second BioCreative Challenge Evaluation Workshop (Krallinger, 2007). A related field, which has not received as much attention, is the creation of *Pathways* or *diagrams*. These Pathways capture the interaction between all the entities involved in a specific biological process. For example, a PPI usually involves more entities than just two proteins, and this creates new challenges for biomedical TM. In this paper we present a new system, PathText, and show how it can be used to browse a corpus of text that describes the entities in the graph.

2. PathText

PathText is a new system developed to connect Natural Language Processing (NLP) technology to the graphs and diagrams that are so often used by biologists. It uses the graphical interface of Payao (described below), to let the user quickly find related text snippets and articles related to the different parts of the Pathway. Figure 1 shows the overall structure of the PathText system. There are two interacting user interfaces: Payao (in the top-left corner), is used to show biologically meaningful diagrams to the user, and the PathText interface (bottom-left corner) connects to the corresponding language resources shown on the right side: KLEIO (Nobata et al., 2008), FACTA (Tsuruoka et al., 2008) and MEDIE (2008). In the rest of this paper, we first introduce the biological interface and modules, and then the information extraction and language processing modules.

2.1. Payao (*"Pa-ya-o"*)

Payao is a Web 2.0 community tagging system for biological networks, enabling a community to work online on the same models simultaneously. It provides an interface for users to add tags and comments to the models, and an Application Programming Interface (API) that lets our Text Mining (TM) tools automatically add more information to the models (Kitano et al., 2007b). The API provides a list of all the entities and interactions in the model, and the TM tool searches the existing or newly published literature for texts that describe these entities in the context of the given reactions. Automatically generated comments are added to the entities and reactions in the model, and they contain clickable links that take the users back to the PathText web page to show a highlighted version of the corresponding text, and more information about the selected entity or interaction. Payao relies on several other enabling technologies like the Systems Biology Markup Language, SBML (Hucka and et al., 2003), and Graphical Notation, SBGN (Kitano et al., 2007a), and it builds on the CellDesigner program (Funahashi et al., 2007) to show the graphical models.

2.2. Systems Biology Graphical Notation (SBGN)

PathText (through Payao and CellDesigner), uses SBGN to visualize the Protein Interaction Pathways. The goal of the SBGN effort is to help standardize a graphical notation for diagrams of computational models in systems biology. Such a standard notation will have broad impact, for example by bringing rigor and consistency to the usually ad hoc diagrams accompanying research articles today, as well as the user interfaces of different software tools and

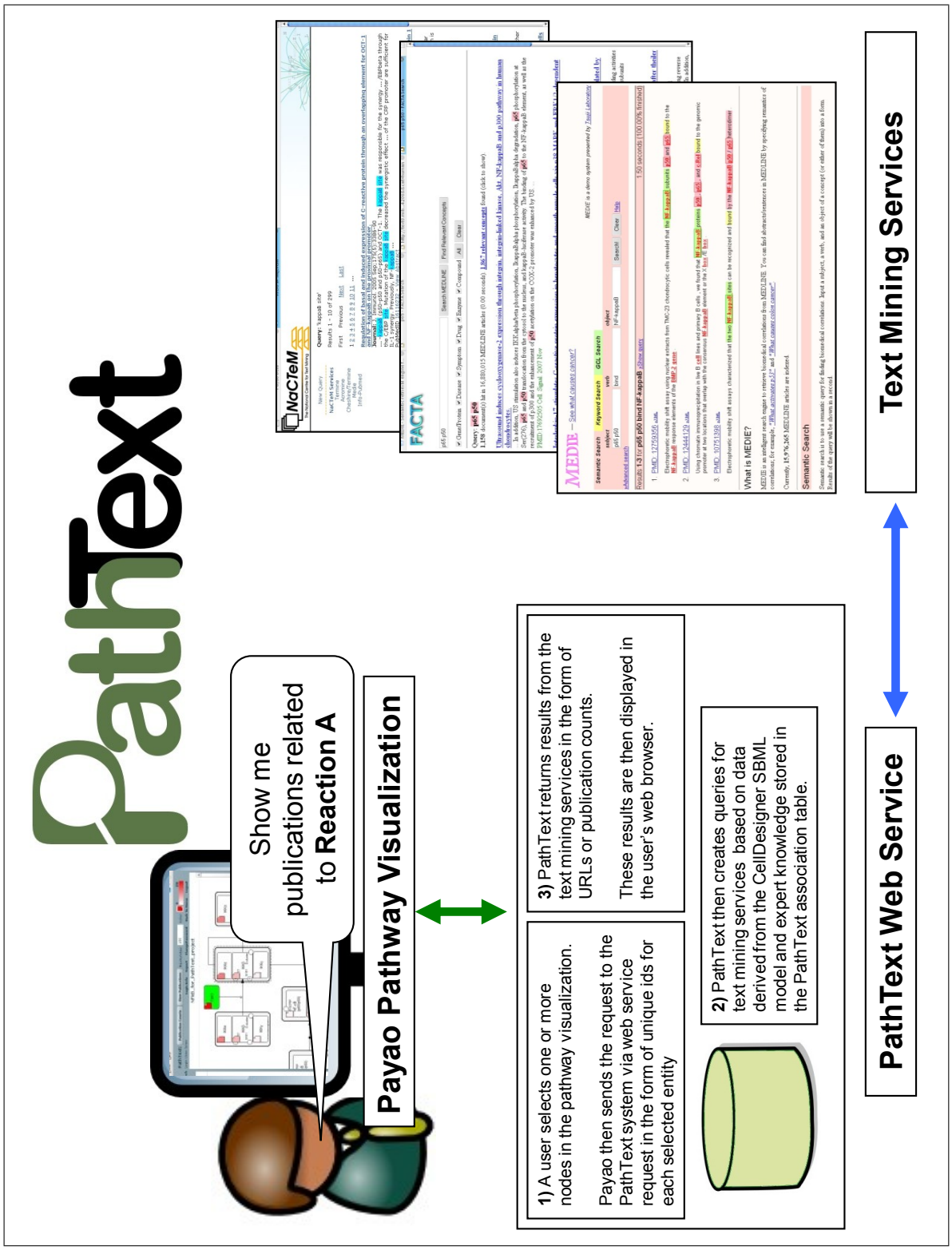

Figure 1: PathText System Overview

databases in systems biology. The real payoff will come when a common visual notation becomes as familiar to biologists as circuit schematics are to computer engineers. When researchers are saved the time and effort of familiarizing themselves with idiosyncratic notations, they can spend more time thinking about the underlying networks being depicted (Kitano et al., 2007a).

2.3. Text Mining Tools

This subsection describes the three Text Mining tools that have already been connected to PathText: KLEIO, FACTA and MEDIE. The connection is done by replacing the original user interface of the tool by a given Pathway map from Payao. The query created by Payao can be a simple protein name, or a complex interaction, involving several entities

acting as agents, targets or regulators/modifiers. Depending on the neighborhood of the selected query in the graph, different filters can be applied to the Text Mining results, to make sure that only relevant text is collected and presented to the user.

2.3.1. KLEIO

KLEIO is an advanced information retrieval system developed at the UK National Centre for Text Mining (www.nactem.ac.uk). It is one of the core components in the PathText system, as it offers textual and metadata searches across MEDLINE, with enhanced searching functionality by leveraging terminology management technologies. 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 the text, such as genes, proteins and other substance names (Nobata et al., 2008).

One of the KLEIO tools is AcroMine 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, since these many forms have to be recognized, indexed, linked and mapped from text to existing databases. Typically, most of the currently available information retrieval systems (like PubMed[1]) fail to deal with the problems of term ambiguity and variability. For example, the same term can be expressed as "2-(3,4- dihydroxyphenyl)acetic acid", "3,4-Dihydroxyphenyl acetate" or "3,4-Dihydroxybenzeneacetate" in English text. KLEIO addresses this problem by using Text Mining technology to reduce the diversity of term variation.

The conceptual approach to information retrieval realized by KLEIO brings novel and original functionality to meet the growing interest in the biosciences looking for solutions to literature mining. The core components are:

Acronym recognition and disambiguation AcroMine recognizes acronyms (e.g. DEAE) and their definitions (e.g. diethylaminoethyl) from the whole Medline Abstracts Database. It also disambiguates isolated acronyms using their context and maps them into corresponding definitions.

Normalization of biology terms A computationally efficient algorithms for term normalization, based on a combination of exact and soft string matching methods is used (Tsuruoka et al., 2007). An advantage of applying term normalization over such large scale dictionaries is to permit efficient look-up and to discover ambiguous and variant terms in the resources. The novelty of the approach lies in using existing resources to learn term variation patterns in a fully automatic manner.

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. KLEIO is using a combination of conditional random fields and maximum entropy models to filter out false positives. The dictionaries for the named entity recognition process are provided by the processing steps mentioned in the previous paragraph.

Indexing of terms At the indexing stage, named entities and acronyms are linked with the original text using Lucene, 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 and excluded as non-gene/protein names if their definitions are not gene or protein names. When a user enters a query containing any one of the surface forms, the results for all of the term variants are returned ensuring maximal expansion across the document collection.

2.3.2. FACTA

FACTA is an advanced text mining tool to help discover associations between biomedical concepts contained in MEDLINE articles. The user can navigate these associations and their corresponding articles in a highly interactive manner. The system accepts an arbitrary query term and displays relevant concepts on the spot. The quick responses are made possible by the pre-indexing of MEDLINE and efficient document/concept retrieval algorithms. A broad range of concepts are retrieved by the use of large-scale biomedical dictionaries containing the names of important concepts such as genes, proteins, diseases, and chemical compounds.

2.3.3. MEDIE

MEDIE is a system for accurate real-time retrieval of relational concepts from MEDLINE (Miyao et al., 2006). It uses off-line processing to pre-compute the semantic structures and on-line processing to search for the semantic structures that match a user's query.

Off-line processing: An HPSG parser (Miyao and Tsujii, 2005) and a term recognizer (Tsuruoka and Tsujii, 2004) are applied in order to create MEDLINE annotated with predicate argument structures and identifiers in ontology databases.

On-line processing: User input is converted into queries of extended region algebra, see (Masuda et al., 2003). A search engine then retrieves sentences having semantic annotations that match the queries.

Accurate retrieval of relational concepts is attained because we can precisely describe relational concepts using semantic annotations. In addition, real-time retrieval is possible because the semantic annotations are computed in advance.

3. Related Work

An important part of the PathText project is building a corpus that explicitly describes which articles that were used when creating a given relation in the graphical pathway. The corpus that we used for creating the PathText prototype system is described in more detail in Kim et al. (2008).

Several other recent publications have proposed similar systems to PathText. Most of them are automatic PPI (or

[1] http://www.ncbi.nlm.nih.gov/sites/entrez?db=PubMed

other relation) extraction systems combined with an ad hoc visualization system. As far as we know, this is the first attempt at connecting text mining tools to the emerging standard Systems Biology Graphical Notation (SBGN). An overview of the other existing systems can be found in (Hoffmann et al., 2005).

4. Future Work

We are currently developing a method to automatically display "recent publications" relevant to a pathway selected by a user. This system will also prioritize publications based on author, publisher or other user specified criteria.

The current pathway visualization is provided by Payao. We plan to expand upon this by adding other pathway visualization tools, in particular for metabolic pathways.

Also, the text mining services used by PathText will be broken down into highly specialized components with a standard API, allowing the services to work in conjunction to provide even more relevant search results.

This process will also include the ability to store user generated links between pathway components and publications and display these links for other users of the system at a later date.

And finally, as we learn more about the connections between text and diagrams, we can imagine reversing the process described in this paper. So instead of going from the manually created diagram to the relevant text, the system would be able to scan newly published text, and then create new nodes and connections in the relevant Pathways automatically. This will reduce the manual workload for the human users, who then only have to agree or disagree with the suggested additions.

A prototype of the PathText system will be publicly available online as soon as Payao is officially released. This is scheduled to happen in May 2008. Please check the NaCTeM or Tsujii-lab web pages to get an updated URL to the most recent version of the system.

5. Summary

We presented the PathText system, and showed how it can be used to browse a corpus consisting of graphical elements and corresponding scientific text. The PathText system uses existing Text Mining technology to help automate the process of creating the costly resources that are needed for the next generation of Text Mining tools.

6. Acknowledgements

This work was partially supported by "Grant-in-Aid for Specially Promoted Research" and the "Genome Network Project", both from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan. This work was also sponsored by Okinawa Institute of Science and Technology (OIST), Systems Biology Institute (SBI) and Sony Computer Science Laboratories, Inc.

7. References

Akira Funahashi, Yukiko Matsuoka, Akiya Jouraku, Haruka Sugimura, Yuichi Oikawa, Norihiro Kikuchi, and Hiroaki Kitano. 2007. CellDesigner 4.0beta: A Modeling Tool for Biochemical Networks. In *Proceedings of the Eight International Conference on Systems Biology (ICSB)*, page 56ff, October.

R. Hoffmann, M. Krallinger, E. Andres, J. Tamames, C. Blaschke, and A. Valencia. 2005. Text mining for metabolic pathways, signaling cascades, and protein networks. *Sci STKE*, 283:pe21.

Michael Hucka and et al. 2003. The Systems Biology Markup Language (SBML): A Medium for Representation and Exchange of Biochemical Network Models. *Bioinformatics*, 4(9):524–531, September.

Jin-Dong Kim, Tomoko Ohta, Kanae Oda, and Jun'ichi Tsujii. 2008. From text to pathway: Corpus annotation for knowledge acquisition from biomedical literature. In *Proceedings of the 6th Asia Pacific Bioinformatics Conference (APBC)*.

Hiroaki Kitano, Akira Funahashi, Michael Hucka, Nicolas Le Novère, Yukiko Matsuoka, and SBGN consortium. 2007a. Systems Biology Graphical Notation (SBGN). In *Proceedings of the Eight International Conference on Systems Biology (ICSB)*, page 58ff, October.

Hiroaki Kitano, Norihiro Kikuchi, Haruka Sugimura, and Yukiko Matsuoka. 2007b. "Payao": Web 2.0 community tagging system for biological networks. In *Proceedings of the Eight International Conference on Systems Biology (ICSB)*, page 57ff, October.

Martin Krallinger. 2007. The interaction-pair and interaction method sub-task evaluation. In Lynette Hirschman, Martin Krallinger, and Alfonso Valencia, editors, *Proceedings of the Second BioCreative Challenge Workshop*.

Katsuya Masuda, Takashi Ninomiya, Yusuke Miyao, Tomoko Ohta, and Jun'ichi Tsujii. 2003. A robust retrieval engine for proximal and structural search. In *Proceedings of HLT-NAACL 2003 Short papers*, pages 58–60, Edmonton, Canada, May.

MEDIE. 2008. Semantic retrieval engine for MEDLINE. http://www-tsujii.is.s.u-tokyo.ac.jp/medie/.

Yusuke Miyao and Jun'ichi Tsujii. 2005. Probabilistic disambiguation models for wide-coverage hpsg parsing. In *Proceedings of ACL 2005*, pages 83–90, Ann Arbor, Michigan, June.

Yusuke Miyao, Tomoko Ohta, Katsuya Masuda, Yoshimasa Tsuruoka, Kazuhiro Yoshida, Takashi Ninomiya, and Jun'ichi Tsujii. 2006. Semantic retrieval for the accurate identification of relational concepts in massive textbases. In *Proceedings of COLING-ACL 2006*, pages 1017–1024, Sydney, Australia, July.

Goran Nenadic, Naoki Okazaki, and Sophia Ananiadou. 2006. Towards a terminological resource for biomedical text mining. In *Proceedings of LREC-5*, Genoa, Italy, May.

Chikashi Nobata, Philip Cotter, Naoaki Okazaki, Brian Rea, Yutaka Sasaki, Yoshimasa Tsuruoka, Jun'ichi Tsujii, and Sophia Ananiadou. 2008. Kleio: a knowledge-enriched information retrieval system for biology. In *Proceedings of the ACM SIGIR Conference*, July.

Yoshimasa Tsuruoka and Jun'ichi Tsujii. 2004. Improving the performance of dictionary-based approaches in pro-

tein name recognition. *Journal of Biomedical Informatics*, 37(6):461–470.

Yoshimasa Tsuruoka, John McNaught, Jun'ichi Tsujii, and Sophia Ananiadou. 2007. Learning string similarity measures for gene/protein name dictionary look-up using logistic regression. *Bioinformatics*. doi: 10.1093/bioinformatics/btm393.

Yoshimasa Tsuruoka, Chikashi Nobata, Philip Baker, Douglas Kell, and S. Ananiadou. 2008. Facta: a web-based text mining system for finding associations between biomedical concepts. In *Genomes to Systems 2008*, Manchester, UK, March. The Consortium for Post-Genome Science.