Structured Literature Image Finder: Parsing Text and Figures in Biomedical Literature

Amr Ahmed[a,b], Andrew Arnold[a], Luis Pedro Coelho[c,d,e], Joshua Kangas[c,d,e], Abdul-Saboor Sheikh[d], Eric Xing[a,b,c,d,e,f], William Cohen[a,b,c,d,e], Robert F. Murphy[a,c,d,e,f,g]

[a] *Machine Learning Department, Carnegie Mellon University*
[b] *Language Technologies Institute, Carnegie Mellon University*
[c] *Joint Carnegie Mellon University–University of Pittsburgh Ph.D. Program in Computational Biology*
[d] *Center for Bioimage Informatics, Carnegie Mellon University*
[e] *Lane Center for Computational Biology, Carnegie Mellon University*
[f] *Department of Biological Sciences, Carnegie Mellon University*
[g] *Department of Biomedical Engineering, Carnegie Mellon University*

Abstract

The SLIF project combines text-mining and image processing to extract structured information from biomedical literature.

SLIF extracts images and their captions from published papers. The captions are automatically parsed for relevant biological entities (protein and cell type names), while the images are classified according to their type (e.g., micrograph or gel). Fluorescence microscopy images are further processed and classified according to the depicted subcellular localization.

The results of this process can be queried online using either a user-friendly web-interface or an XML-based web-service. As an alternative to the targeted query paradigm, SLIF also supports browsing the collection based on latent topic models which are derived from both the annotated text and the image data.

The SLIF web application, as well as labeled datasets used for training system components, is publicly available at http://slif.cbi.cmu.edu.

1. Introduction

Biomedical research results in a very high volume of information in the form of publications. Researchers are faced with the daunting task of querying and searching these publications to keep up with recent developments and to answer specific questions.

In the biomedical literature, data are most often presented in the form of images. A fluorescence micrograph image (FMI) or a gel is sometimes the key to a whole paper. Literature retrieval systems should provide biologists with a structured way of browsing the otherwise unstructured knowledge in a way

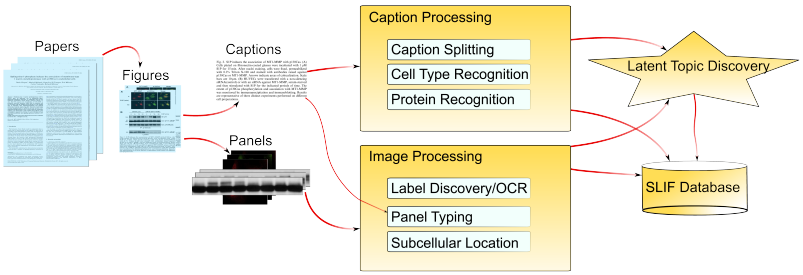

Figure 1: SLIF Pipeline. This figure shows the paper processing pipeline.

that inspires them to ask questions that they never thought of before, or reach a relevant piece of information that they would have never have explicitly searched for.

Relevant to this goal, our team developed the first system for automated information extraction from images in biological journal articles (the "Subcellular Location Image Finder," or SLIF, first described in 2001 [1]). Since then, we have reported a number of improvements to the SLIF system [2, 3, 4].

In response to the opportunity to participate in the Elsevier Grand Challenge, we have made major enhancements and additions to the system. In part reflecting this, we rechristened SLIF as the "Structured Literature Image Finder." The new SLIF provides both a pipeline for extracting structured information from papers and a web-accessible searchable database of the processed information. Users can query the database for information appearing in captions or images, including specific words, protein names, panel types, patterns in figures, or any combination of the above. We have also added a powerful tool for organizing figures by topics inferred from both image and text, and have provided a new interface that allows browsing through figures by their inferred topics and jumping to related figures from any currently viewed figure.

2. Overview

SLIF consists of a pipeline for extracting structured information from papers and a web application for accessing that information. The SLIF pipeline is broken into three main sections: caption processing, image processing and latent topic discovery, as illustrated in Figure 1.

The pipeline begins by finding all figure-caption pairs and creating database entries for each. Each caption is then processed to identify biological entities (names of proteins and cell lines) and these are linked to external databases.

The image processing section begins by splitting each figure into its constituent panels, and then identifying the type of image contained in each panel. The original SLIF system was trained to recognize only those panels containing fluorescence microscope images (FMIs), but as part of the work for the Elsevier Challenge we have extended SLIF to recognize other types of panels. The patterns in FMIs are then described using a set of biologically relevant image features [1], and the subcellular location depicted in each image is recognized.

The first two sections result in panel-segmented, structurally and multimodally annotated figures. The last step in the pipeline is to discover a set of latent themes that are present in the collection of papers. These themes are called topics and serve as the basis for visualization and semantic representation. For instance, a topic about "tumorigenesis" is expected to give high probability to words like ("tumor", "positive", "h1b") and proteins like ("Caspase", "Actin") which are known to be related to tumorigenesis. Each figure in turn is represented as a distribution over these topics, and this distribution reflects the themes addressed in the figure. This representation serves as the basis for various tasks like image-based retrieval, text-based retrieval, and multimodal-based retrieval. Moreover, these discovered topics provide an overview of the information content of the collection and structurally guide its exploration. For instance, the user might ask for articles that have figures in which the "tumorigenesis" topic is highly represented.

3. Database Access

The results of processing papers are stored in a searchable database and are made available to the user through an interactive web-interface. A user can query the database for any combination of: text within captions, proteins extracted by protein name annotators, different properties of the image panels (panel type or pixel resolution), or images depicting a particular subcellular location (either inferred from the image or retrieved from a protein annotation database). The user can also view or browse the latent topics discovered from figures and captions.

Results can be presented at multiple levels (panel, figure, or paper level) and the user can switch between these presentation options from within the current results. A link is always provided to the original publication.

From the results of a search, users can also view the underlying papers or the UniProt record corresponding to an extracted protein name. They can also further refine the search results by adding more conditions. Alternatively, using latent topics, users can structurally browse the otherwise unstructured collection by giving relevance feedback to the system (interactively flagging certain results as relevant) to guide the system to show the user targeted results.

We also make the results available via a web service architecture. This enables other machines to consume SLIF results in automated fashion. For a set of processed results, we publish a WSDL (Web Services Description Language) document on the SLIF server that declares the database query procedure for clients in a standard XML based description language. Clients can query SLIF using an XML-based query submitted as a SOAP (Simple Object Access Protocol) message. Results are sent back a message in an XML-based format.

4. Caption Processing

The initial version of SLIF focused on finding micrographs that depicted a particular pattern, but could not associate that pattern with a specific protein.

The current system parses the caption for that information.

Information on the protein depicted in a given figure should be provided in its caption, but the structure of captions can be quite complex (especially for multipanel figures). We therefore identify the "image pointers" (e.g., *(A)* or *(red)*) in the caption that refer to specific panel labels or panel colors in the figure [2], dividing the caption into fragments (or "scopes") that refer to an individual panel, color, or the entire figure.

The next step is to match the image pointers to the panel labels found during image processing. We correct errors in optical character recognition by using regularities in the arrangement of the labels (if the letters A through D are found as image pointers and if the panel labels are recognized as A,B,G and D, then the G should be corrected to a C). The precision of the final matching process was found to be 83% and the recall to be 74% [5].

The recognition of named entities (such as protein and cell names) in free text is a difficult task that may be even more difficult in condensed text such as captions. In the current version of SLIF, we have implemented two schemes for recognizing protein names. The first uses prefix and suffix features along with immediate context to identify candidate protein names. This approach has a low precision but a good recall (which is useful to enable database searches on abbreviations or synonyms that might not be present in structured protein databases) [6]. The second approach uses exact matching to a dictionary of names extracted from protein databases. The protein names found by this approach can be associated with a supporting protein database entry.

5. Image Processing

In our image processing pipeline, we start by dividing the extracted figures into their constituent components, since, in a majority of the cases, the figures are comprised of multiple panels. For this purpose, we recursively break images along vertical or horizontal boundary regions. We have previously shown that the algorithm can effectively split figures with complex panel layouts [1].

SLIF was originally designed to process only FMI panels. As part of our work for the Elsevier Challenge, we expanded the classification to other panel types. This mirrors other systems that have appeared since the original SLIF which include more panel types [7, 8, 9].

We have manually labeled circa 700 panels into six panel classes: (1) FMI, (2) gel, (3) graph or illustration, (4) light microscopy, (5) X-ray, or (6) photograph using an active learning scheme [10] to optimise our labeling effort.

We decided to focus first on creating a high-quality classifier for the *gel* class, given its importance to the working scientist. Using a decision tree learning algorithm based both on textual and image features, we obtained very high precision (91%) at the cost of moderate recall (66%). When neither the FMI nor the gel detector were positive, we used a general purpose image-feature classifier for the other classes (accuracy: 69%).

Fluorescent panels are further processed to identify the depicted subcellular localization. To provide training data for pattern classifiers, we hand-labeled

a set of images into four different subcellular location classes: (1) *nuclear*, (2) *cytoplasmic*, (3) *punctate*, and (4) *other*, again using active learning to select images to label. On the 3 main classes (nuclear, cytoplasmic, and punctate), we obtained 75% accuracy (as before, reported accuracies are estimated using 10 fold cross-validation and the classifier used was libSVM based). On the four classes, we obtained 61% accuracy.

Panels were associated with their scopes based on the textual information found in the panel itself and the areas surrounding the panels. Each figure is composed of a set of panels and a set of subimages which are too small to be panels. All of these subimages were analyzed using optical character recognition (OCR) to identify potential image pointers. The caption of each figure was parsed to find the set of associated image pointers. Image pointers in subimages and in the captions were matched. Each panel was matched to the nearest unique image pointer found in the figure using OCR. This enabled panels to be directly associated with the textual information found in a caption scope.

6. Topic Discovery

The goal of topic discovery is to enable the user to structurally browse the otherwise unstructured collection. This problem is reminiscent of the actively evolving field of multimedia information management and retrieval. However, *structurally-annotated* biological figures pose a set of new challenges [11]. First, figures can be comprised of structured multiple panels. Portions of the caption are associated with a given panel, while other portions of the caption are shared across all the panels and provide contextual information. Second, unlike most associated text-image datasets, the text annotation associated with each figure is free-form and not all of it is relevant to the graphical content of the figure. Finally, the figure's caption contains in addition to text, specific entities like protein names, or subcellular locations. To address these challenges, we developed what we call a structured correspondence topic model. For a full specification of the model, we refer the reader to [11].

The input to the topic modeling system is the panel-segmented, structurally and multimodally annotated biological figures. The goal of our approach is to discover a set of latent themes in the Elsevier paper collection. These themes are called topics and serve as the basis for visualization and semantic representation. Each figure, panel, and protein entity is then represented as a distribution over these latent topics. This representation serves as the basis for various tasks like image, text, or multimodal retrieval, and image annotation.

6.1. Structured Browsing and Relevance Feedback

Topic models endow the user with a bird's eye view over the paper collection by displaying a set of topics that summarize the themes addressed in the collection. If a topic interests the biologist, she can click on the browse button to see all panels (figures) that are relevant to this topic or all papers containing these figures.

Moreover, if the biologist has a focused search need, the system can confine the displayed topics to those topics associated with panels (figures) that interest the biologist. For instance, assume that the biologist searched for *high-resolution*, *FMI* panels that contain the protein *MT1-MMP*. The biologist can then click the *"view associated topics"* link below the displayed panel. The system will display only the topics addressed in this panel and if one of these *focused* topics interest the biologist, they can then browse for more panels that show the pattern(s) captured by this topic by clicking on the browse button (See [11, 12] for more details).

From the results of any SLIF query, a user can mark panels (or figures) as *interesting* and ask SLIF to retrieve panels (figures) similar to the marked ones. SLIF will then rank the panels (figures) in the database based on the similarity of their latent representations to the latent representation of the selected panels (figures). This process can be repeated recursively to refine the search outcome until a satisfactory result is reached.

7. User Study

We conducted a user study to validate the usability and usefulness of our technology. A detailed description of the study is given in [12]. Here, we only highlight the main aspects of the study.

Our target users were graduate students in the fields of biology, computational biology, and biomedical engineering. Each user was given an instruction sheet that described a set of tasks to be performed using both SLIF and a traditional search engine (which the user was free to choose). Examples of these tasks include searching for high-resolution images of a given protein, and papers with images related to a subcellular location. The user was given a short overall introduction to the goals of the project but no specific guidance on how to use the website as to best approximate real-world conditions.

The users were asked for feedback by answering questions related to the various tasks, as well as general feedback. Most answers were free-form in order to elicit comments that would allow us to improve the system.

When asked "Overall, how useful did you find SLIF?," six out of eight users considered SLIF useful and a seventh stated that the system had "great potential" (the question was free-form and we scored answers as positive or negative). To some extent, this mimics the results of Hearst et al. [13] who performed a user study on the viability of using caption searching to find relevant papers in the bioscience literature and found that "7 out of 8 [users] said they would use a search system with this kind of feature." Only one user found that the alternative search engine returned better results. Half found SLIF better and more relevant, and the other three thought the results were not directly comparable. Moreover, *six* out of the *eight* users said that using topic-models in organizing the information was very useful or interesting (a sample comment states that it was "useful in terms of depicting 'intuitive' relationships between various queries"). Negative remarks centered on the fact that a normal search engine returns more results than does SLIF, which is operating with a smaller

collection of papers (when compared to a search engine such as Google), as well as on particular points of the user interface (which were subsequently addressed in a revised interface).

8. Discussion

We have presented a new version of SLIF, a system that analyses images and their associated captions in biomedical papers. SLIF demonstrates how text-mining and image processing can intermingle to extract information from scientific figures. Figures are broken down into their constituent panels, which are handled separately. Panels are classified into different types, with the current focus on FMI and gel images, but this could be extended to other types. FMIs are further processed by classifying them into their depicted subcellular location pattern. The results of this pipeline are made available through a either a web-interface or programmatically using SOAP technology.

A new addition to our system is latent topic discovery which is performed using both text and image. This enables users to browse through a collection of papers by looking for related topics. This includes the possibility of interactively marking certain images as relevant to one's particular interests, which the system uses to update its estimate of the users' interests and present them with more targeted results.

Although it is crucial that individual components achieve good results (and we have shown good results in our sub-tasks), good component performance is not sufficient for a working system. SLIF is a production system which working scientists in biomedical related fields have described as "very useful."

8.1. Acknowledgments

The SLIF project is currently supported by NIH grant R01 GM078622. L.P.C. is supported by Fundação para a Ciência e Tecnologia (SFRH/BD/37535/2007).

References

[1] R. F. Murphy, M. Velliste, J. Yao, G. Porreca, Searching online journals for fluorescence microscope images depicting protein subcellular location patterns, in: BIBE '01: Proceedings of the 2nd IEEE International Symposium on Bioinformatics and Bioengineering, IEEE Computer Society, Washington, DC, USA, 2001, pp. 119–128.

[2] W. W. Cohen, R. Wang, R. F. Murphy, Understanding captions in biomedical publications, in: KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, NY, USA, 2003, pp. 499–504.

[3] R. F. Murphy, Z. Kou, J. Hua, M. Joffe, W. W. Cohen, Extracting and structuring subcellular location information from on-line journal articles:

The subcellular location image finder, in: Proceedings of the IASTED International Conference on Knowledge Sharing and Collaborative Engineering, 2004, pp. 109–114.

[4] Z. Kou, W. W. Cohen, R. F. Murphy, A stacked graphical model for associating sub-images with sub-captions, in: Proceeding of Pacific Symposium on Biocomputing, World Scientific, 2007, pp. 257–268.

[5] Z. Kou, W. W. Cohen, R. F. Murphy, Extracting information from text and images for location proteomics, in: M. J. Zaki, J. T.-L. Wang, H. Toivonen (Eds.), Proceedings of BIOKDD, 2003, pp. 2–9.

[6] Z. Kou, W. W. Cohen, R. F. Murphy, High-recall protein entity recognition using a dictionary, Bioinformatics 21 (2005) i266–i273.

[7] J.-M. Geusebroek, M. A. Hoang, J. van Gernert, M. Worring, Genre-based search through biomedical images, in: Proceedings of 16th Int. Conf. on Pattern Recognition, Vol. 1, 2002, pp. 271–274.

[8] H. Shatkay, N. Chen, D. Blostein, Integrating image data into biomedical text categorization, Bioinformatics (2006) i446–453.

[9] B. Rafkind, M. Lee, S. Chang, H. Yu, Exploring text and image features to classify images in bioscience literature, in: Proceedings of the BioNLP Workshop on Linking Natural Language Processing and Biology at HLT-NAACL, Vol. 6, 2006, pp. 73–80.

[10] N. Roy, A. Mccallum, Toward optimal active learning through sampling estimation of error reduction, in: In Proc. 18th International Conf. on Machine Learning, Morgan Kaufmann, 2001, pp. 441–448.

[11] A. Ahmed, E. Xing, W. Cohen, R. F. Murphy, Structured correspondence topic models for mining captioned figures in biological literature, in: Proceeding of the ACM conference on Knowledge Discovery and Data Mining, 2009, pp. 39–48.

[12] A.-S. Sheikh, A. Ahmed, A. Arnold, L. P. Coelho, J. Kangas, E. P. Xing, W. W. Cohen, R. F. Murphy, Structured literature image finder: Open source software for extracting and disseminating information from text and figures in biomedical literature, Tech. rep., Carnegie Mellon University School of Computer Science, Pittsburgh, USA, CMU-CB-09-101 (2009).

[13] M. A. Hearst, A. Divoli, J. Ye, Exploring the efficacy of caption search for bioscience journal search interfaces, in: In ACL 2007 Workshop on BioNLP, 2007, pp. 73–80.