Structured Literature Image Finder

Amr Ahmed[1,6], Andrew Arnold[1], Luís Pedro Coelho[2,3,4], Joshua Kangas[2,3,4],
Abdul-Saboor Sheikh[3], Eric Xing[1,5,6], William Cohen[1], Robert F. Murphy[1,2,3,4,5,7*]

Abstract

The Structured Literature Image Finder tackles
two related problems posed by the vastness of the
biomedical literature: how to make it more accessi-
ble to scientists in the field and how to take advan-
tage of the primary data often locked inside papers.
Towards this goal, the SLIF project developed an in-
novative combination of text and image processing
methods.

Images from papers are classified according to
their type (fluorescence microscopy image, gel, . . .)
and their caption is parsed for biologically relevant
entities such as protein names. This enables tar-
geted queries for primary data (a feature that a
user study revealed to be highly valued by scien-
tists). Finally, using a novel extension to latent
topic models, we model papers at multiple levels
and provide the ability to find figures similar to a
query and refine these findings with interactive rel-
evance feedback.

SLIF is most advanced in processing fluorescent
microscopy images which are further categorised ac-
cording to the depicted subcellular localization pat-
tern.

The results of SLIF are made available to the
community through a user friendly web interface
(http://slif.cbi.cmu.edu).

1   Introduction

Biomedical research worldwide results in a very
high volume of information in the form of publica-
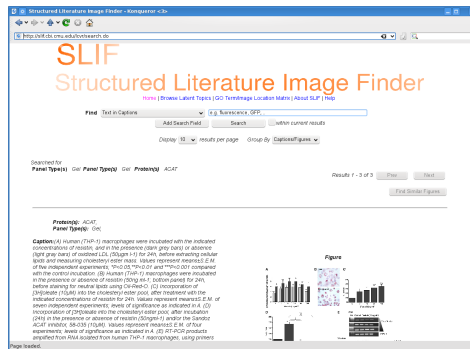tions. Biologists are faced with the daunting task of

Figure 1: Screenshot of the SLIF search engine
showing the results of a search.

querying and searching these publications to keep
up with recent developments and to answer specific
questions.

In the biomedical literature, data is most of-
ten presented in the form of images. A fluores-
cent micrograph image (FMI) or a gel is sometimes
the key to a whole paper. Compared to figures
in other scientific disciplines, biomedical figures are
often a stand alone source of information that sum-
marizes the findings of the research under consid-
eration. A random sampling of such figures in
the publicly available PubMed Central database
reveals that in some, if not most of the cases, a
biomedical figure can provide as much information
as a normal abstract. The information-rich, highly-
evolving knowledge source of the biomedical liter-
ature calls for automated systems that would help
biologists find information quickly and satisfacto-
rily. These systems should provide biologists with
a structured way of browsing the otherwise unstruc-
tured knowledge in a way that would inspire them
to ask questions that they never thought of before,
or reach a piece of information that they would have
never considered pertinent to start with.

Relevant to this goal, we developed the first sys-
tem for automated information extraction from im-
ages in biological journal articles (the "Subcellular
Location Image Finder," or SLIF, first described in

---

[1]Machine Learning Department, [5]Department of Bi-
ological Sciences, and [7]Department of Biomedical Engi-
neering, Carnegie Mellon University; [2]Joint Carnegie Mel-
lon University–University of Pittsburgh Ph.D. Program in
Computational Biology; [3]Center for Bioimage Informatics,
Carnegie Mellon University; [4]Lane Center for Computa-
tional Biology, Carnegie Mellon University; [6]Language Tech-
nologies Institute, Carnegie Mellon University. * to whom
correspondence should be addressed.

2001 [8]). Since then, we have made major enhancements and additions to the SLIF system [2, 7, 6], and now report not only additional enhancements but the broadening of its reach beyond fluorescent microscopy images. Reflecting this, we have now rechristened SLIF as the "Structured Literature Image Finder."

SLIF reached the final stage in the Elsevier Grand Challenge (4 out of 70), a contest sponsored by Elsevier to "improve the way scientific information is communicated and used."

2   Overview

SLIF provides both a pipeline for extracting structured information from papers (illustrated in Figure 2) and a web-accessible searchable database of the processed information (depicted in Figure 1).

The pipeline begins by finding all figure-captions pairs. Each caption is then processed to identify biological entities (e.g., names of proteins and cell lines) and these are linked to external databases. Pointers from the caption to the image are identified, and the caption is broken into "scopes" so that terms can be linked to specific parts of the figure.

The image processing module begins by splitting each figure into its constituent panels, and then identifying the type of image contained in each panel. The patterns in FMIs are described using a set of biologically relevant image features [8], and the subcellular location depicted in each image is recognized.

The last step in the pipeline is to discover latent topics that are present in the collection of papers. These topics serve as the basis for visualization and semantic representation. Each topic consists of a triplet of distributions over words, image features, and proteins (possibly extended to include gene ontology terms and subcellular locations). 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.

All results of processing are stored in a database, which is accessible via a web interface or SOAP queries. The results of queries always include links back to the panel, figure, caption and the full paper.

Users can query the database for various information appearing in captions or images, including specific words, protein names, panel types, patterns in figures, or any combination of the above. Using the latent topic representation, we built an innovative interface that allows browsing through figures by their inferred topics and jumping to related figures from any currently viewed figure.

3   Caption Processing

In order to identify the protein depicted in an image, we look for protein names in the caption. The structure of captions can be complex (especially for multipanel figures). We therefore implemented a system for processing captions with three goals: identifying 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 that refer to an individual panel, color, or the entire figure, and recognizing protein and cell types.

Errors in optical character recognition can lead to low accuracy in matching image pointers to panel labels. Using regularities in the arrangement of the labels (e.g., if the letters A through D are found as image pointers and the panel labels are recognized as A,B,G and D, then the G should be corrected to a C) corrects some of the errors [6]. Using a test set from PNAS, the precision of the final matching process was found to be 83% and the recall to be 74% [4].

Recognition of named entities (such as protein and cell types) in free text is a difficult task that may be even more difficult in condensed text such as captions. We have implemented two schemes for recognizing protein names. The first (which is also used for cell type recognition) uses prefix and suffix features along with immediate context to identify candidate protein names. This approach has a low precision but an excellent recall (which is useful to enable database searches on abbreviations or synonyms that might not be present in structured protein databases). The second approach [5] uses a dictionary of names extracted from protein databases in combination with soft match learning methods to obtain a recall and precision above 70%. The occurrences of the names found in the captions are stored as being associated either with a panel or a figure, depending on the scope in which the protein name was found. The system also assigns subcellular locations to proteins using lookup of GO terms
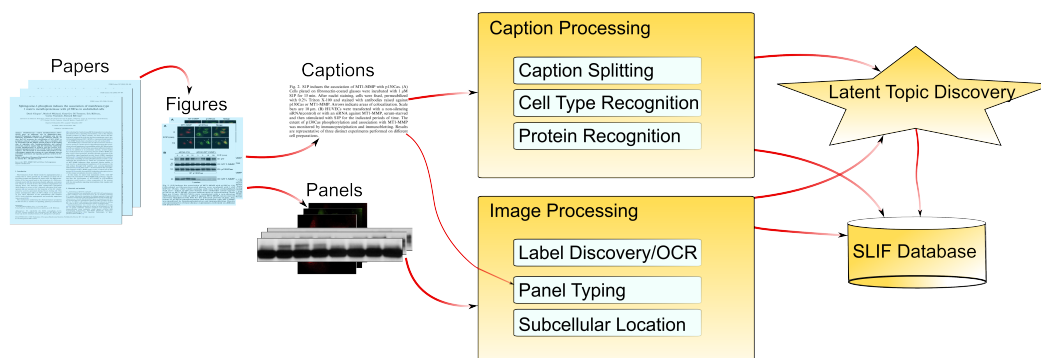
2

Figure 2: SLIF Pipeline. This figure shows the general pipeline through papers are processed.

in the Uniprot database, making it possible to find images depicting particular subcellular patterns.

Finally, the task of simply segmenting a paper and extracting the caption, even without named entity recognition or panel scoping, has proven very useful to our users, allowing easy search of free text which can be limited to the captions, and therefore the figures, of a paper.

4   Image Processing

Since, in most cases, figures are composed of multiple panels, the first step in our image processing pipeline is to divide the figures into panels. We employ a figure-splitting algorithm that recursively finds constant-intensity boundary regions in between panels, a method which we have previously shown can effectively split figures with complex panel layouts [8].

SLIF was originally designed to process only FMI panels, and subsequent systems created by others have included classifiers to distinguish other figure types [10, 3]. We have now expanded the classification to other panel types: (1) FMI, (2) gel, (3) graph or illustration, (4) photograph, (5) X-ray, or (6) light microscopy. Using active learning [9], we selected ca. 700 panels to label.

Given its importance to the working scientists, we focused on the *gel* class. Currently, the system proceeds through 3 classification levels: the first level, classifies the image into FMI or non-FMI using image based features (as previously reported); the second level, uses textual features to identify gels with high-precision (91%, and moderate recall: 66%); finally, if neither classifier has fired, a general purpose support vector machine classifier, operating on image-based features does the final classification (accuracy: 61%).

Perhaps the most important task that SLIF supports is the classification of FMI panels based on 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 selected through active learning.

We computed previously described features to represent the image patterns. If the scale is inferred from the image, then we normalize this feature value to square microns. Otherwise, we assume a default scale of $1\mu m$/pixel. 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.

5   Topic Discovery

The goal of the topic discovery phase 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* biomedical figures pose a set of new challenges to mainstream multimedia information management systems due to the hierarchical structure of the domain (panels contained within figures within papers) and the presence of both free form and annotated text in the caption (protein names should be handled differently from free text).

Our model, which we call multimodal factored, scoped, correspondence topic model, addresses the aforementioned challenges [1]. The input to the topic modeling system is the panel-segmented, structurally and multimodally annotated biomedical figures. The goal of our approach is to discover a

set of latent themes in the collection. These themes are called topics and serve as the basis for visualization and semantic representation. Each biomedical 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-based retrieval, text-based image retrieval, multimodal-based image retrieval and image annotation. We compared our model to various baselines over the aforementioned tasks with favorable results [1].

The latent topic representation facilitates the implementation of features such as finding similar objects to an example that the user has found as interesting (this can be done at any level: panel, figure, or paper).

6 Discussion

We have presented SLIF, a system which analyzes images in the biomedical literature. It processes both text and image, combining them through latent topic discovery. This enables users to browse through a collection of papers by looking for related topics or images that are similar to an image of interest.

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 that has been shown to yield usable results in real collections of papers.

The project is on-going and many avenues for improvement are being exploited. Among those are better semantic understanding of FMI data, more advanced image processing of gels, exploitation of the full-text, as well as a continuing improvement of all the components in the pipeline.

References

[1] A. Ahmed, E. Xing, W. Cohen, and R. F. Murphy. Modeling structurally-annotated figures using factored scoped correspondence topic models. In *ACM conference of Knowledge Discovery and Data Mining (submitted)*, 2009.

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

[3] J.-M. Geusebroek, M. Anh Hoang, J. van Gernert, and M. Worring. Genre-based search through biomedical images. volume 1, pages 271–274 vol.1, 2002.

[4] Z. Kou, W. Cohen, and R. F. Murphy. Extracting information from text and images for location proteomics. In Mohammed Javeed Zaki, Jason Tsong-Li Wang, and Hannu Toivonen, editors, *BIOKDD*, pages 2–9, 2003.

[5] Z. Kou, W. Cohen, and R. F. Murphy. High-recall protein entity recognition using a dictionary. In *ISMB (Supplement of Bioinformatics)*, pages 266–273, 2005.

[6] Z. Kou, W. Cohen, and R. F. Murphy. A stacked graphical model for associating sub-images with sub-captions. In R. B. Altman, A. K. Dunker, L. Hunter, T. Murray, and T. E. Klein, editors, *Pacific Symposium on Biocomputing*, pages 257–268. World Scientific, 2007.

[7] R. F. Murphy, Z. Kou, J. Hua, M. Joffe, and W. Cohen. Extracting and structuring subcellular location information from on-line journal articles: The subcellular location image finder. In *IASTED International Conference on Knowledge Sharing and Collaborative Engineering*, pages 109–114, 2004.

[8] R. F. Murphy, M. Velliste, J. Yao, and 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*, pages 119–128, Washington, DC, USA, 2001. IEEE Computer Society.

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

[10] H. Shatkay, N. Chen, and D. Blostein. Integrating image data into biomedical text categorization. In *ISMB (Supplement of Bioinformatics)*, pages 446–453, 2006.