Extracting and Matching Authors and Affiliations in Scholarly Documents

Huy Hoang Nhat Do[1,2] Muthu Kumar Chandrasekaran[2] Philip S. Cho[2]
Min–Yen Kan[1,3*†]
[1]Department of Computer Science, School of Computing
[2]Asia Research Institute
[3]NUS Interactive and Digital Media Institute
National University of Singapore
{a0079636,a0092669,aripcss,dcskmy}@nus.edu.sg

ABSTRACT

We introduce *Enlil*, an information extraction system that discovers the institutional affiliations of authors in scholarly papers. Enlil consists of two steps: one that first identifies authors and affiliations using a conditional random field; and a second support vector machine that connects authors to their affiliations. We benchmark Enlil in three separate experiments drawn from three different sources: the ACL Anthology, the ACM Digital Library, and a set of cross-disciplinary scientific journal articles acquired by querying Google Scholar. Against a state-of-the-art production baseline, Enlil reports a statistically significant improvement in F_1 of nearly 10% (p « 0.01). In the case of multidisciplinary articles from Google Scholar, Enlil is benchmarked over both clean input ($F_1 > 90\%$) and automatically-acquired input ($F_1 > 80\%$).

We have deployed Enlil in a case study involving Asian genomics research publication patterns to understand how government sponsored collaborative links evolve. Enlil has enabled our team to construct and validate new metrics to quantify the facilitation of research as opposed to direct publication.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries

General Terms

Algorithms, Experimentation

Keywords

Metadata Extraction, Logical Structure Discovery, Conditional Random Fields, Support Vector Machine, Rich Document Features

*This work was supported in part by the National University of Singapore - Global Asia Institute Research Grant [AC-2010-1-004] and the Fetzer Franklin Trust project on Culture and Cognition.
†Contact Author.

JCDL'13, July 22–26, 2013, Indianapolis, Indiana, USA.

1. INTRODUCTION

Quantifying scholars' performance is a difficult problem. While there are many instruments to use to argue one's scientific prowess, it is generally accepted that others' use of a scholar's publications represents key evidence. Measuring "use" is none the easier, but has often been interpreted in terms of citation count or aggregate citation metrics such as *h-index* [15], *g-index* [9], or more recently, *ħ-index* [16].

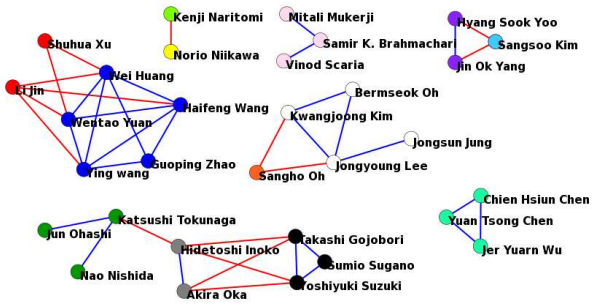

Figure 1: An example of a co-author network coded with institutional affiliations. Like-colored nodes indicate identical affiliations and edges represent publications. Blue edges indicate co-authorship among authors in the same institution; red, across institutions.

While benchmarking individual scholars is important to their promotion and tenure, higher-level (i.e., aggregate) bibliometrics and scientometrics are perhaps even more important. For example, a researcher may publish little, but play a significant role over time by forging critical connections among others to facilitate research. We have recently addressed this issue and developed research facilitation or RF-metrics [6] for individuals and groups. Measuring the performance of whole academic units, institutions or countries is important for department heads and deans, and also a key knowledge management issue for the policymakers and funding agencies that support their research work.

The institutional affiliations[1] of authors of scholarly papers is an important piece of metadata for enabling such publication data aggregation. Research works in social sciences, scientometrics, bibliometrics and knowledge management such as [24] also use publi-

[1]We use "affiliations" and "institutions" interchangeably, although an *affiliation* properly refers to just the correspondence between an author and his institution.

cations records and associated metadata. The correct identification of author's affiliations is crucial for research works such as [26] that study the impact of location, geography on emergent behavior. In the current scholarly publication practice, this data is manually entered by publishers or authors, and thus a valuable asset to publishers. Commercial services such as Thompson Reuters' Web of Science, Elsevier's Scopus and Google Scholar sometimes provide inaccurate mappings between the different authors and their institutional affiliations, incomplete listings of affiliations for publications with many authors, or do not provide the authors' affiliation at all. Also, access to such metadata can be prohibitively expensive, restricting research and benchmarking exercises that rely on this metadata to well-provisioned institutions[2].

However, thanks to the triad of open access, self-archiving and institutional repositories, scholarly publications themselves are now easier to obtain from web sources. Armed with an automated system that extracts authors from digital copies and associates them with their affiliations, the public can now perform their own bibliographic research, case studies and comparative benchmarking.

As a case in point, Figure 1 shows the actual co-authorship network of a large set of authors involved in the Pan-Asian SNP Consortium from 2004, where edges represent jointly published works. This international consortium was sponsored by various national government funding bodies. In our umbrella project, we are interested to know whether such funding resulted in sustained collaborative links, after the sponsored funding finished. With a system that can reliably extract author and affiliation information, construction of such networks becomes easy and the subsequent knowledge management and analyses of such networks becomes possible.

We report on *Enlil*, a system that performs this task in a two-step process: first, the extraction of author and institutions from scholarly documents via supervised sequence labeling; and second, the matching of authors to their institutions via relation classification. The key contributions of this paper are: 1) describing the features used by our system and their efficacy on the problem, 2) performance that significantly outperforms the current state-of-the-art, 3) detailed analysis of the major error classes of our system and the difficulties in author and affiliation extraction and matching, and 4) a discussion of the knowledge management which Enlil enables.

2. RELATED WORK

Extracting authors, institutions and correlating them relates to the domain of automated document processing, of which the general problem of document structure analysis has been an active research area since the 1960's. Many of these works focus on the scholarly record, and more recently at those encoded in rich text formats, such as Portable Document Format (PDF). Klink *et al.* [18] and Kim *et al.* [17] introduce methods that use both the textual and the layout features of a document to discover its structure, however, these systems rely on hand-tuned rule sets and templates. Such methods are brittle and may fail when presented with documents that follow new or different publishing styles. Gao *et al.* [12] recently described SEB, a framework to detect the hierarchy and reading order of a document using weighted bipartite graphs but its fixed rules are again not flexible enough to capture document metadata in practical scenarios.

A common method to address the fragility of handcrafted rule-based systems is to employ statistical learning methods that automatically construct the model from annotated data. This approach removes the need for experts to construct the rule base. It also enables adaptability, as when new annotated data arrives, the learned model can be regenerated to fit the new distribution of examples. While general information retrieval now uses statistical learning techniques at its core, current research in document analysis has started to use statistical learning only recently. For the problem of general document structure inference, Belaid *et al.* [2] has employed neural networks, while the SectLabel system [22] adopted a Conditional Random Field (CRF) as their learning approach. We build our work upon SectLabel, as its broad categories and adequate overall performance provide a good base for improvement.

Other research has tried to apply more fine-grained methods on smaller parts of the document to discover various types of relationship between different entities. Systems by both Gao *et al.* [12] and Lopez *et al.* [21] correlate figures to their captions, enabling better document figure retrieval.

With respect to scholarly document metadata, Cortez *et al.* [7], Councill *el al.* [8], aim to recover the citation network by recognizing and parsing the reference strings that appear in scholarly documents' reference section, using a knowledge-based approach and a CRF model, respectively. Chen *et al.* [4] integrated both knowledge-based and data driven approaches into a single architecture for the same problem. The BibAll system [10] integrates several existing systems such as ParsCit while also utilizing new digital library services including Google Scholar to provide a complete extraction of all the fields in a citation string. The CEBBIP system by Gao *et al.* [11] also demonstrated that a CRF works for the same problem for written Chinese. While all do identify and delimit author names in reference strings present in a paper's bibliography, they do not perform the recognition, delimitation and extraction of author names on the title page (i.e., they do not identify the authors of the document itself, but the authors of the documents which the document cites).

Our work here continues the automation process where previous work has left off. In particular, we focus our attention on the title page of a scholarly work, delimiting the document's authors and their institutions and building their correspondence. We observe that good performance on such extractions from title pages often requires knowledge of the spatial layout and font information in addition to the text itself. This is in contrast to author processing in reference strings in the bibliography that largely depends only on lexical information.

Most closely related to our work is work by the CiteSeer team at Penn State. Initial work by Han *et al.* [14] uses a Support Vector Machine (SVM) to extract document metadata. The work is comprehensive but limited by two important factors. First, only the textual features of the document header are used, which leads to issues in chunk identification in names that are delimited only by whitespace. Second, as the dataset contains only computer science research papers, its performance outside of this domain is unknown. Later work in SeerSuite [25] includes exactly our problem scope in this work as one component: SVM Header Parser. This system is closest to Enlil in terms of functionality: accepting PDF input and extracting and matching authors with affiliations. As such, we use SVM Header Parser as a comparative baseline in our system's evaluation. Our results show the improvement of our system over SVM Header Parser in both cross-domain and computer science datasets.

3. SYSTEM OVERVIEW

Our author and affiliation extraction and matching system, Enlil[3], comprises of two steps run in serial – an extraction step and a

[2]Thompson Reuters' Web of Science and Elsevier's Scopus can each cost between 10K to 100K USD per year [20].

[3]Enlil is the name of the air god in the Sumerian myth of the Huluppu tree, the World Tree that connects heaven and earth.

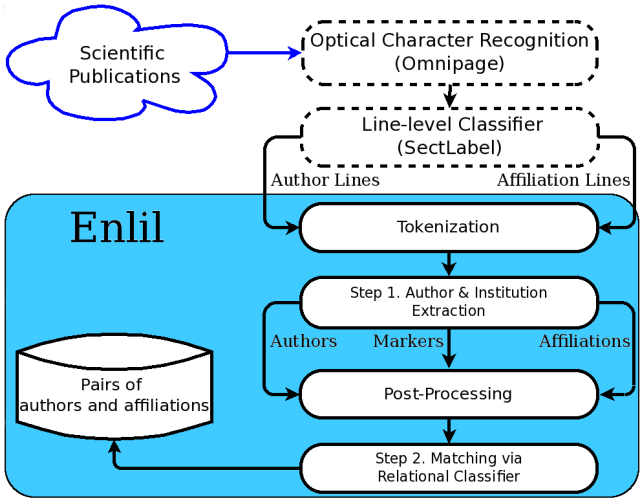

Figure 2: Enlil's system architecture. The rounded rectangle indicates the system modules detailed in this paper; modules outlined by dashes indicate preprocessing steps.

matching step. Figure 2 shows the overview of Enlil's modules and their interaction. The first step is modeled as sequence labeling tasks that use a CRF as the supervised learning model, while an SVM is used in the second step for relational classification. A significant preprocessing pipeline provides our system with the necessary textual and spatial input needed to complete our processing, which we detail first.

OCR. Portable Document Format (PDF), while now an open standard and the most common method for disseminating published scholarly work, composes the rendered or printed page in several different ways. While PDFs can be semantically rich, complete with document logical structure (e.g., indicating which text are headers, page numbers, titles), they can also lack the basic textual content shown in the image of the pages rendered – especially when scanned without an accompanying character recognition step. To produce a pipeline for author–affiliation matching that works under all such circumstances, we choose to treat (pages of) PDF documents as images, and use a commercial off-the-shelf optical character recognition software, Nuance OmniPage[4] to canonicalize any input PDF into a standard format. We use OmniPage to output an XML version of the document that provides both the textual and spatial information (i.e., the X- and Y-coordinates) for each word that appears on each page.

We then run SectLabel [22], an open-source module that takes this type of input, to assign one of 23 semantic classes to each line of text, including *Author* and *Affiliation*. Note that SectLabel does not further process individual lines to delimit authors or institutions. Enlil collects all reported author and affiliation lines as its input for processing, discarding the rest of the input document, as they are irrelevant to our problem. At this point, the system has the required input for starting – two sets of lines of text marked as either Author or Affiliation.

Tokenization. While individual lines of text classified as authors or institutions are the input to Enlil, there still remains significant processing needed to delimit the author and institutions correctly. Lines marked as *Authors* may contain multiple author names, markers that signify a correspondence between authors and footnotes or affiliations (hereafter called "markers"). Lines marked by the SectLabel preprocessor as *Affiliation* may also contain phone numbers

[4]Versions 16 and 17.

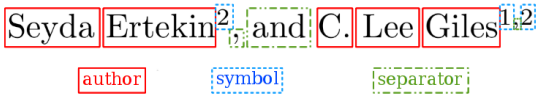

Figure 3: An example of name splitting.

or email addresses. Enlil's first step is to correctly tokenize and delimit the author and institutional information in the input. Tokenization is needed as author, marker and author-separating tokens are commonly concatenated into a single printed token without internal spacing, as shown in Figure 3. To perform this tokenization, we apply a short list of rules, starting with delimiting tokens by spaces. In addition to the use of spaces, we also use the change in formatting between characters (e.g., change to superscript or subscript) or the presence of punctuation marks to force the tokenizer to break the token into smaller components. For example, in the *Author* line, the signifier "$Seyda\ \ Ertekin^2,$" can be split into four tokens "$Seyda$", "$Ertekin$", "2" and "$,$".

3.1 Author and Institution Extraction

Once tokens are identified, we use supervised sequence labeling to assign each of the tokens in both author or affiliation input lines to one of three classes: *name*, *symbol*, and *separator*. These classes correspond to an author or institution token, a correspondence marker, and any separating token, respectively.

We use a linear Conditional Random Field (CRF) [19] to tackle each of these two problems, as it has been shown to work well on related sequence labeling tasks, such as reference string parsing, as well as part-of-speech tagging. We use the freely-available CRF++ implementation[5], which permits us to design feature classes that describe the traits of hidden nodes, as well as the relationships between them. We use the same feature sets for both of the problems of labeling author lines as well as affiliation lines, but use different training data. Two CRF models are learned from labeled training data sets (described later) – one for author line labeling and one for affiliation line labeling.

Enlil has two logically separate sets of feature classes that it uses for this extraction step. The first set of features are *content* features – ones related to the textual content of the tokens. However, we have found that relying solely on punctuations and token identity features that can be found in raw text fails to demarcate author and affiliation sections well. This is due to the fact that many publication formats depend on the visual layout of a title page for this task. A human reading the formatted rich text can easily detect different logical units by their font sizes or positions. For this reason, Enlil also utilizes a second set of features based on the text *layout* – these features capture font size or spatial position. We will see that this second set of features is especially important and it significantly improves performance (Section 4).

Our system only uses binary-valued features. Note that all but the first layout feature depends on rich text input that is provided by our OCR / line classification preprocessing pipeline. We list the feature classes used by the CRF for both author and affiliation line token labeling, listing the content features first.

— **Token Identity** (*content*): We encode separate features for each token in three different forms: 1) as-is, 2) lowercased, and 3) lowercased, stripped of punctuation. For this feature and the following n-gram feature, the feature sets are the binary presence or absence of the token.
— **N-gram prefix/suffix** (*content*): The first as well as the last 1-4 characters of the token in lowercase format.

[5]`http://crfpp.sourceforge.net/`

Table 1: Gazetteers used by Enlil and their sources.

	Dictionary Type	# of Entries
Person	First name (Male)[a]	3,901
	First name (Female)[a]	4,955
	Last name[b]	88,799
	Last name (Romanized Chinese)	286
Location	Countries and cities[c]	50,000

[a] `ftp://ftp.funet.fi/pub/doc/dictionaries/DanKlein/`
[b] `http://www.census.gov/genealogy/names/`
[c] `http://world-gazetteer.com`

— **Length** (*content*): The length of the token is also an important feature, as separators and markers are often fixed in length. Enlil counts the total number of characters in the token and turns on only ("1") of the following features: *1Char*, *2Char*, *3Char*, or *4+Char*.
— **Punctuation** (*content*): Any punctuation present in the token is detected and categorized into one of three groups: *continuingPunctuation* (e.g. commas, semicolons), *stopPunctuation* (e.g. periods), *others* (e.g. apostrophe), or *noPunctuation*. The first two groups usually serve as separators that demarcate authors' names, while the third can serve as a part of a name (e.g. "O'Reilly").
— **Number** (*content*): We identify numbers present in the token using simple regular expressions, assigning the numeric sequence to one of four mutually-exclusive categories – *1Digit*, *2Digit*, *3Digit* or *4+Digit*; otherwise, the feature gets the value *noDigit*. The number feature helps to distinguish numeric markers, and parts of addresses present in affiliations.
— **Gazetteers** (*content*): We incorporate several gazetteers of common tokens in place and people names into our system. Each token is checked for its presence on each of the gazetteers. Table 1 details the source and demographics of each.
— **First word in line** (*layout*): We check if the token is the first word in an input line, as an author name is usually short and contained within a single line, so the first word in a line may mark the start of a new author.
— **Source Section** (*layout*): Rich text formats (as provided by the OCR XML output format) propagate high-level document structures such as paragraphs and columns. We use this information to decide whether different tokens are in the same section. Two tokens cannot belong to the same logical unit if they are in two different sections of the text, which may occur when reading inputs line by line. Figure 4 shows an example of three lines of text logically containing six separate lines, due to the two-column format.
— **Orthographic case** (*layout*): We assign the token one of the following values: *Initialcaps*, *MixedCaps*, *ALLCAPS*, or *others*.
— **Sub–Superscript** (*layout*): Tokens have one of these two features turned on if it is a subscript or a superscript. This feature is especially useful for detecting affiliation correspondence markers, such as in Figure 3.
— **Font format** (*layout*): We use three features – *Bold*, *italic*, and *underline* – to indicate the tokens' respective format.
— **Font size** (*layout*): Different logical portions of the document may be rendered in different font sizes. Author names are usually in the same font size or in a larger font than affiliations and rarely smaller. This feature captures the regularities in size. However, as different documents may use different sizes, it is the relative difference between the base font size in a document and the target token that we encode using this feature class. The values for our feature are *small*, *normal*, and *large*.
— **Format change** (*layout*): This is a differential feature which is conditioned on two consecutive labels in the CRF model. It captures the change of layout format when moving from one token to the next. This feature encodes whether two consecutive tokens have the same format. Here, "format" is defined as the combination of "*Font format*", "*Sub–superscript*", and "*Font size*" features. The "*Format change*" is assigned the value "*fbegin*" if any of these three component features changes its value from the previous token to the current; otherwise, it is assigned the value "*fcontinue*".

Post-Processing then constructs author and institution names from the labeled tokens. As author names may occur in various format, such as "*S. Ertekin*" or "*Ertekin, S.*", each name is converted to its normalized form "*S Ertekin*" using regular expressions to detect locations of any separator characters, and transposing the internal results as needed. We define the normalized form to be the first initial, any middle initial, followed by last name (e.g., "S Ertekin"). This form is used when Enlil is evaluated under *relaxed* match conditions (See Section 4.1).

We then group consecutive tokens with the same class together to form a list of author names and a list of affiliations together with their markers. In Figure 3, Enlil generates "$Seyda\ \ Ertekin$" as the author name and "2" as the correspondence marker. As we use the change in formatting between characters in the tokenization step, consecutive tokens in the same word need to be concatenated differently. One common example is when an author name is written using small capitals, such as "ERTEKIN", there are two tokens "*E*" and "*RTEKIN*". In such cases, we need to ensure that no extra space is inserted between them to form the correct name.

In the previous steps, Enlil does not make use of the actual distance between tokens in rich text formats where larger spaces usually serve as implicit separators. For example, in Figure 4, the second affiliation line contains two occurrences of the string "*National University of Singapore*", separated only by a large gap with no explicit separator, hence, appending them together leads to an incorrect result. In post-processing, we measure the actual distance between consecutive words and check for tab characters if present. When a larger than usual space appears inside a string, it is split. This processes the authors and affiliations in Figure 4 correctly. Identifying the complete list of authors and affiliations ends the first step.

3.2 Author Matching via Relational Classifier

The second step then matches affiliations to authors, also using a supervised classification model. While quadratic in complexity ($O(nm)$, where n is the number of authors, and m the number of institutions), in practice this process is efficient as the total number of authors and affiliations is usually small. However, exceptions do occur in certain consortium papers common in fields such as High-Energy Physics, where hundreds of authors and tens of institutions are common.

While it is easy to decide the correct affiliations using simple heuristic rules with accompanying affiliation markers for the majority of cases, there are complications that we observed that led us to use a more sophisticated supervised learning approach. This is because the format of scholarly documents is diverse and many formats do not contain markers. Moreover, the input extraction process is not error-free and a learning system can deal better with noise than absolute, heuristic rules. Hence, our system also takes the distance between a candidate author and a candidate institution as evidence to improve accuracy. One limitation of this approach is that the spatial distance is only available in rich text formats. Fortunately, our processing pipeline provides this evidence source.

For humans, the relationship between an author and an institution is often easily detected using simple observations. First, these names are usually clearly separated from the other parts of the document, and are at a close proximity to each other. Second, there are

Figure 4: An example of names in different columns in a rich text format document.

usually markers that indicate the connection between them. We use two sets of features that can capture this evidence: the markers provided by the previous module and the distances between the author and the affiliation. There are two distance features in our system: *Euclidean distance* and *Logical distance*. The *Euclidean distance* between two strings is measured by calculating the difference in the $x - y$ coordinates of the strings' center (as measured by their bounding box). The *logical distance* measures the difference in the document logical addresses of the strings' line. Each line can be assigned a unique logical address comprising of four components: its page, column, paragraph, and line index. For example, two lines are in the same paragraph if their page, column, and paragraph indices are equal. The logical distance can be computed in Enlil from the input XML that is provided by Omnipage, which structures its OCR output into such hierarchical text units.

Note that because we generate features for all possible pairs of authors and institutions, our matching problem is not an exclusive (hard) matching task: an author may have more than one affiliation and an institution may serve for more than one author's affiliation. A common scenario is when several authors have the same marker, thus, belonging to the same institution. In Enlil, we allow both author and institution to be associated with multiple markers, although we observe that it is rare for an institution to have more than one marker.

We use a standard SVM with a Gaussian kernel as the learning paradigm for this pointwise classification step, implemented using LibSVM [3], with the below binary features:

— **Signal symbol:** This feature is turned on when a candidate author and institution are both marked by an identical marker. All markers are stored in plain text with no additional format information as correspondence markers of an author and an affiliation may have different formatting. An author and an affiliation can also be associated with more than one marker. For example, an author "*C. Lee Giles*" in Figure 3 will be associated with two markers "*1*" and "*2*" that follow immediately after it. This one-to-many relationship helps finding the correct affiliations if there are more than one.
— **Logical distance:** This feature class is a collection of three features indicating whether the two strings are in the same *page*, *column*, and *paragraph*. We observe that the affiliated institution is usually in the same structure as the author.
— **Euclidean distance:** OmniPage assigns the x-y coordinates in pixels of the bounding box of every word in the document. Enlil uses these values to compute the coordinates of every author and affiliation strings. We then compute the distance between an author and all possible affiliations along the x- and y-axis and use the Euclidean distance to find the nearest affiliation, as the correct affiliation is often the nearest, as exemplified in Figure 4. To further favor author names that appear before a potential affiliation, we penalize affilations that appear before the author name by a factor of 2. If an affiliation is computed to be the nearest, value of the feature is *yes*, otherwise, the value is *no*. This feature is useful when there is no correspondence marker, as is typical in some publication styles (e.g., ACM proceedings).

4. EVALUATION

In evaluating Enlil, we have the following questions we want to answer:

Q1. How effective is Enlil at recovering exact author and institution strings from scholarly papers, when compared with a suitable baseline? Is Enlil also effective at matching authors to institutions to form correct affiliations?

Q2. Does performance vary when the documents come from different domains?

Q3. What type of errors does Enlil typically make and what sources cause them? How does performance vary when it is provided with real-world or clean input?

Q4. What impact do the different features have on author–affiliation matching performance?

We first describe the datasets used in our evaluation, experimental results and finish with a detailed analysis.

4.1 Datasets

We evaluate Enlil on three different datasets to illustrate both evaluation in depth (volume) and in breadth (variety): For volume, we used articles from the the ACM Digital Library, and a larger corpus from the Association for Computational Linguistics' (ACL) Anthology. For variety, we collected a set of cross-disciplinary scientific journal articles by querying Google Scholar.

The first two datasets are from specific technical domains. We sampled computer science papers and their metadata from the ACM Portal, representing 2,243 documents and 6,625 authors[6]. The metadata from the ACM contains a list of all authors and their affiliations, making it a useful gold standard for author extraction. However, it does not list the exact number of institutions shown as appearing in the actual PDF file. For example, it is common for two authors in a paper to be associated with the same institution. In this case there might be a single institution string for both authors, or two duplicated ones (one for each author, as in Figure 4). Hence, the ACM metadata does not represent a complete gold standard for institution extraction, and hence we omit institution extraction evaluation on this dataset.

The ACL Anthology Corpus is the largest of our datasets, containing 23,533 documents[7]. However, there is no ground truth so we need to manually extract and label the data. In evaluating Enlil's results on this dataset, we randomly selected 100 papers from this source, sampling mostly from its conference proceedings. Papers from the same conference usually use an identical conference template, so we made an assumption (discussed later) that the result would be indicative of Enlil's performance over the whole ACL Anthology Corpus. In addition, the processing time grows linearly with the number of articles and Enlil takes about 42 seconds on average to process one document in this corpus (cost is mostly based on the prerequisite OCR).

As Enlil aims to process scientific publications in general – primarily journal articles and conference proceedings – we must also test breadth-wise, on documents drawn from diverse sources. Enlil needs to perform well for both humanities and general sciences to support its downstream use cases, so we deemed the breadth-wise evaluation the most critical. To achieve this, we manually constructed and labeled a corpus of 800 documents from various scholarly journals across four large branches of science, as shown in Table 2. Documents were chosen from different fields and from

[6]We sampled Portal by retrieving all results returned by Google Scholar for the query "site:dl.acm.org".
[7]`http://www.aclweb.org/anthology`, visited on May 22th, 2013

different journals and publishers to ensure a wide variety of formats, indicative of the input we expect Enlil will need to handle when processing both science and humanities publications. Documents were also sampled uniformly – we drew each fourth of the corpus from four different major branches of science, extending the uniform sampling into the individual component fields. All documents were downloaded from our institution's library subscription or from freely-available Web sources. Documents were all uniformly in PDF and processed by our preprocessing pipeline to obtain the input content and layout information. We created the gold standard by manually performing both steps in the process: 1) extracting the authors and institutions from each document in the corpus, and 2) then matching author and institution pairs together to form affiliations. The dataset used in our experiments (minus the copyrighted PDFs) – including the gold standard are publicly available[8].

We trained a single pair of CRF (linear CRF with the window size of 2) and SVM (standard SVM with a Gaussian kernel) models for all of the results reported in the subsequent experiments. A small training dataset was separately (non-overlapping with the testing sets) compiled from articles from the same three sources. We used a total of 390 instances of author and institution extractions for the CRF model and 1,969 instances of author-affiliation matching for the SVM model. All the training data was annotated by the first author.

Within a document, we deem an author's name to be his unique identity. However, an author can have both his full name as well as a shortened version within the same paper (this occurs, for example, in Springer's proceedings template) and in such cases, we have chosen to retain the short name – as the full name can easily be converted to its short form for evaluation, described later. To maintain consistency in annotation we apply the following rules. First, titles (e.g., *PhD* or *MSc*) are removed if present; second, email addresses and phone numbers are discarded, but any physical address information in an affiliation is retained as part of the affiliation. Finally, the author and affiliation strings should be preserved in their original form if not otherwise affected (e.g. *National University of Singapore* is retained over its abbreviated form *NUS*, when both forms are present). Diacritics are also allowed to be represented by ASCII close equivalents (e.g., *université* for *universite*).

While our preprocessing pipeline affords rich text format input, it is not perfect; the output from the preprocessing is not always clean and contributes significant amount of noise. In analyzing the page image to produce the textual and spatial features the commercial OCR package occasionally misrecognizes content. In our observation, this problem occurred most frequently with characters with diacritics (as shown in Figure 5), as the commercial OCR package we used is tuned for English input. The SectLabel preprocessing runs after OCR, to infer the document logical structure and assign lines to their semantic categories. However, some lines may be incorrectly identified or missed, as shown in Figure 6.

As the preprocessing is a significant source of error, we also wish to benchmark Enlil on both clean input data and real input data. To enable such analyses, we selected a subset of 160 articles from the cross-disciplinary corpus which produce correct OCR and Sectlabel results. We use this smaller collection (dubbed the *Clean* dataset from Google Scholar) to assess Enlil's performance without cascading errors that come from preprocessing, whereas the original 800 document collection (dubbed *Full*) is used to approximate Enlil's expected, real-world performance.

[8] `http://wing.comp.nus.edu.sg/downloads/enlilCollection/`

Table 2: Demographics on our cross-disciplinary dataset. The dataset comprises of 800 documents, manifesting 2,217 instances of author and 1,821 instances of institution names.

Branch	Field	Number of Journals / Documents	Number of Authors / Affiliations
Applied	Aerospace	2 / 20	879 / 507
	Agricultural	2 / 20	
	Biomedical	2 / 20	
	Dentistry	2 / 20	
	Forestry	2 / 20	
	Mechanical	2 / 20	
	Medical	2 / 20	
	Mining	2 / 20	
	Nursing	2 / 20	
	Robotics	2 / 20	
Formal	Computer	5 / 50	519 / 388
	Logic	5 / 50	
	Mathematics	5 / 50	
	Statistics	5 / 50	
Natural	Astronomy	4 / 40	813 / 516
	Biology	4 / 40	
	Chemistry	4 / 40	
	Ecology	4 / 40	
	Physics	4 / 40	
Social	Anthropology	4 / 40	470 / 410
	Criminology	4 / 40	
	Economics	4 / 40	
	History	4 / 40	
	Psychology	4 / 40	

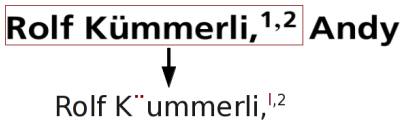

Figure 5: An example of incorrect character recognition from preprocessing by OmniPage. The umlaut (¨) is separated from its base *u* and the affiliation marker number *1* ("one") is misrecognized as the character l ("ell").

4.2 Experimental Results

We report extraction results using two schemes: exact match and relaxed match. *Exact* match judges an extraction as correct only when the author or affiliation name matches the ground truth with canonicalization – i.e., after minor, automatable rules, such as removing multiple spaces, trimming, lowercasing, and removing unnecessary punctuation, are applied. *Relaxed* match relaxes the exact match, by allowing small name variations (e.g., short name and full name, with or without middle initials). Figure 6 shows an example of author names in their short forms that can be matched with their full forms in relaxed mode. We do this by using the Jaro-Winkler string metric to compare the ground truth against the extracted output. We use a similarity threshold of 0.95 for author names and 0.85 for affiliation names. We chose a lower threshold for affiliation names, as observation showed that they are usually much longer and contain more variation as compared to author names. To be precise, relaxed matching performance will always be equal or greater to exact matching performance.

We use SVM Header Parser from SeerSuite [25] as the baseline in our experiments. As mentioned, this tool focuses only on the

computer and information science domains and related areas. To be expected, it does not work well with our cross-disciplinary collection and significantly underperforms Enlil in three of our cross-disciplinary datasets while performing equivalently only on the corpus of documents from the *Formal* domain.

We first review the results of the evaluation, examining each dataset in turn, to answer our evaluation Q1. We answer Q2–4 in the discussion immediately following.

Q1 (Performance vs. Baseline) – ACM Corpus. The evaluation results of the ACM Corpus show large improvements of Enlil over the current state-of-the-art SVM Header Parser in both author extraction and affiliation matching (Table 3). In relaxed mode, F_1 scores increase by 9.9% and 9.7% in these two experiments, respectively. When comparing two systems, both Enlil and SVM Header Parser can process PDF documents. However, SVM Header Parser only focuses on the textual content, while Enlil also accounts for the document's spatial layout and font information. Because of this loss of fidelity in its input representation, SVM Header Parser encounters some issues that do not occur in Enlil; for example, chunk identification in space-separated names, as illustrated by Figure 4.

ACL Anthology. Both the ACM Corpus and the ACL Anthology contain computer science research papers in well-defined templates so we had expected their evaluation results to be very similar. However, Table 3 shows that the affiliation matching performance of both systems drop noticeably in the ACL Anthology corpus. We believe two reasons are responsible for the drop. First, as we only sample a small number of documents from the ACL Anthology, failing to match authors and affiliations correctly in some documents greatly affect F_1. In order to get a better approximation, we need to build the gold standard for a larger number of documents; however, this task requires manual annotation, effort that we preferentially allocated to the cross-disciplinary dataset. Second, a number of the ACL Anthology articles are older records (dating from the 1970s and 80s, before digital desktop publishing), published with different layout formats. This is in contrast to the ACM Corpus whose papers are largely contemporary. These legacy documents are an important factor that reduces the performance of both Enlil and SVM Header Parser.

Cross Disciplinary Corpus. Table 4 shows the affiliation matching performance with the cross disciplinary corpus. Detailed, per-task performance for Enlil is detailed in Tables 5 and 6. Generally we can see that performance in terms of F_1 measure for both tasks is in the upper 80s for both tasks when the full automated pipeline is used. Extraction performance betters matching performance slightly as correct matching is preconditioned on correct extraction. In addition, invalid correspondence markers affect the matching result even though the extraction might be correct. Although the extraction of affiliation is much harder than author extraction due to its complex structure – for example some may use *Department, School, Address*, while others use *Institute* – name canonicalization is beyond the scope of the current work. Thus, there is little difference in performance between author and affiliation extraction in Enlil.

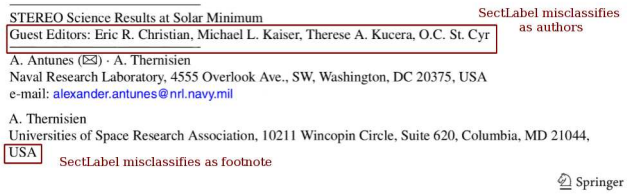

Figure 6: Examples of incorrect classifications by the second stage of preprocessing by Sectlabel, in which an *Editor* line is misclassified as *Author*, and a part of an institution is misclassified as *Footnote*.

Q2 (Domain Sensitivity). To answer the second question, we analyse the evaluation results of Enlil on different domains in Table 4. In general, the performance drops noticeably in the *Nature* and *Social* subsets. This indicates that the current models of Enlil fail to recognize some paper formats in these categories. This is a common problem of systems employing statistical learning methods when encountering new types of input data. We try to alleviate this by collecting training data from many major publishers, such as Elsevier or Springer. Their articles not only have well-defined structure and conform to a limited number of templates but also cover multiple scientific disciplines. The result is that Enlil can handle cross-disciplinary papers better than SVM Header Parser. In SVM Header Parser, there are large gaps between the *Formal* and other subsets. In our experiments, it always achieves the best performance in the *Formal* subset because its training data contained mostly computing-related papers which are most similar to this subset (Table 2). The t–test result in Table 4 supports this finding, showing significant differences between Enlil and SVM Header Parser in *Applied*, *Nature*, and *Social* datasets ($p < 0.01$).

Q3 (Clean vs. Noisy). We observe that cascading errors from OmniPage and SectLabel do significantly affect performance. When we remove these errors in the *Clean* dataset (Table 4), the average performance increases significantly ($p \ll 0.01$) with gain of 10.3% for exact match and 9.7% under the relaxed match scheme. These values indicate that there is significant difference in Enlil's performance when the dataset is clean than when it is not, and that proper clean input is important to ensure optimum performance. As in many other cases, the main difficulty comes from the fact that such tools also need to deal with a variety of input forms from various disciplines.

In particular, the effects of three minor types of error are alleviated when our system is in *relaxed* mode: 1) incorrect character recognition caused by OCR, 2) missing words in affiliations extracted by the SectLabel line classifier, and 3) occurrences of mixed use of both full and short names in a document. In Table 4, the average difference between two matching modes in *Full* dataset (3.4%) is slightly larger than this number in the *Clean* dataset (2.5%), because the first two errors are removed in the second dataset and only the third type remains (interleaved full/short names). This kind of mixture deserves special care as it affects our system in two ways: First, when the full name and the short name of the same author are stored as different entities, it reduces the system's precision. This is ignored in *relaxed* mode where all names are converted to a standard form, first initial, middle initial, and last name. However, the second effect, caused by the different positions of the full name and the short name within the document, still remain. Our system uses various types of distances between the author's name and the affiliation as features, and the correct affiliation is likely to be the nearest one. So, an author may be associated with incorrect affiliations because the nearest ones are usually not the same with respect to the locations of his full name and short name.

With respect to clean input, we would like to highlight two other systematic errors that are intrinsic to Enlil. First, when lines contain author or affiliation data but co-occur with other metadata (e.g., authors or email addresses concatenated with institution names), Enlil's extraction CRF may fail to correctly identify the data and discard the noise. Although we try to remove redundant parts of a line in pre-processing using regular expressions, this does not filter

Table 3: Performance comparison between Enlil and SVM Header Parser (SHP) on the ACM and ACL datasets. Double asterisks (**) indicates that F_1 performance was significantly better than SHP ($p < 0.01$).

Experiment	Corpus	Mode	Enlil			SVM Header Parser		
			Precision	Recall	F_1	Precision	Recall	F_1
Author Name Extraction	ACM	Exact	95.7	93.5	94.6**	84.1	73.5	78.4
		Relaxed	97.9	95.5	96.7**	93.2	81.3	86.8
	ACL	Exact	93.4	90.1	91.8**	84.8	72.7	78.3
		Relaxed	94.7	91.3	92.9**	92.2	79.1	85.1
Affiliation matching	ACM	Exact	89.6	88.2	88.9**	78.8	68.8	73.5
		Relaxed	91.4	89.9	90.6**	87.0	75.7	80.9
	ACL	Exact	84.5	82.8	83.6**	74.2	62.9	68.1
		Relaxed	85.7	84.0	84.8**	79.3	67.2	72.7

Table 4: Enlil's affiliation matching comparison between Enlil and SHP on the cross-domain full and clean datasets. Exact match results shown, with relaxed matching results shown in parentheses. Significant improvement in Enlil between the full and clean datasets denoted by "*" ($p < 0.05$) and "" ($p < 0.01$). "†" indicates significance ($p < 0.01$) of Enlil's performance over SHP.**

Dataset	Branch	Enlil			SVM Header Parser		
		Precision	Recall	F_1	Precision	Recall	F_1
Full	Applied	86.3 (89.7)	89.7 (90.9)	87.9 (90.3)†	38.1 (46.3)	7.99 (9.63)	13.2 (15.9)
	Formal	87.8 (90.1)	87.9 (89.5)	87.9 (89.8)†	57.6 (62.6)	41.1 (44.5)	47.9 (52.0)
	Natural	80.1 (80.7)	81.2 (81.8)	80.7 (81.3)†	55.7 (56.7)	28.3 (28.8)	37.5 (38.2)
	Social	70.4 (81.4)	83.4 (92.7)	76.3 (86.7)†	37.4 (44.9)	26.7 (32.1)	31.2 (37.4)
	Average	81.6 (85.5)	85.6 (88.7)	83.6 (87.0)†	47.2 (52.6)	26.0 (28.8)	32.5 (35.9)
Clean	Applied	95.9 (96.8)	98.1	97.0 (97.4)**†	41.3 (44.4)	12.1 (13.1)	18.8 (20.2)
	Formal	95.6 (95.6)	97.3	96.4 (96.4)**†	63.9 (68.6)	51.9 (55.7)	57.3 (61.5)
	Natural	95.4 (96.7)	96.7 (97.4)	96.1 (97.0)**†	47.1 (50.6)	26.5 (28.5)	33.9 (36.5)
	Social	80.4 (92.3)	90.9 (97.0)	85.3 (94.6)*†	38.7 (44.0)	29.3 (33.3)	33.3 (37.9)
	Average	92.0 (95.4)	95.9 (97.5)	93.9 (96.4)**†	47.8 (51.9)	29.9 (32.6)	35.8 (38.8)

Table 5: Author extraction comparison between the cross-domain full and clean dataset. Exact match results shown, with relaxed results in parentheses when different. Significant improvement denoted by "*" ($p < 0.05$) and "" ($p < 0.01$).**

Dataset	Branch	Precision	Recall	F_1
Full	Applied	81.9 (85.7)	95.9 (96.4)	88.3 (90.7)
	Formal	89.5 (94.1)	94.0 (94.6)	91.7 (94.3)
	Natural	85.5 (86.4)	94.6 (95.1)	89.8 (90.5)
	Social	73.8 (85.3)	95.3 (97.4)	83.2 (90.9)
	Average	82.7 (87.9)	95.0 (95.9)	88.4 (91.6)
Clean	Applied	97.5 (99.5)	100.0	98.8 (99.7)**
	Formal	96.4	100.0	98.1 **
	Natural	94.8 (96.7)	99.3	97.0 (97.9)**
	Social	80.5 (93.1)	100.0	89.2 (96.4)*
	Average	92.3 (96.4)	99.8	95.8 (98.0)**

Table 6: Institution name extraction comparison between the cross-domain full and clean dataset, using exact match. Relaxed results are identical. Significant improvement denoted by "*" ($p < 0.05$) and "" ($p < 0.01$).**

Dataset	Branch	Precision	Recall	F_1
Full	Applied	88.2	95.7	91.8
	Formal	90.6	89.7	90.2
	Natural	87.8	87.9	87.9
	Social	80.9	93.2	86.6
	Average	86.8	91.7	89.2
Clean	Applied	96.1	99.0	97.5**
	Formal	98.7	97.4	98.1*
	Natural	96.7	98.9	97.8**
	Social	87.8	95.2	91.3
	Average	94.8	97.6	96.2**

all of the noise. To address this, the CRF also needs to be trained to handle noisy input. As our current feature set caters only to characterize names, separators and markers, further work needs to encode features dedicated to marking such noise. A second error happens when the relationship between an author and his affiliations is expressed in the prose content, rather than relying on visual or spatial effects or stylistic conventions. Figure 7 shows an example of such error. Solving this problem requires understanding its content, beyond the scope of the current work.

Table 7: Effect of different feature sets – (Log)ical distance, (Euc)lidean distance and (Sig)nal symbol – on the *Full* and *Clean* dataset. Exact match results shown, with relaxed matching results shown in parentheses when different. "" show significant differences from the 'All' feature set ($p < 0.01$).**

Dataset	Branch	Features			
		Log+Euc**	Sig+Log**	Sig+Euc	All
Full	Applied	49.4 (51.2)	82.4 (83.6)	86.9 (89.2)	87.9 (90.3)
	Formal	73.1 (74.9)	68.0 (68.1)	87.9 (89.8)	87.9 (89.8)
	Natural	48.8 (58.9)	73.9 (74.3)	79.3 (79.6)	80.7 (81.3)
	Social	58.7 (68.9)	66.7 (68.4)	75.3 (85.5)	76.3 (86.7)
	Average	57.5 (63.5)	72.8 (73.6)	82.4 (86.0)	83.6 (87.0)
Clean	Applied	57.2 (58.5)	85.9 (86.4)	96.8 (97.2)	97.0 (97.4)
	Formal	78.7	71.6	96.4	96.4
	Natural	58.0 (58.9)	85.4	95.4 (96.4)	96.1 (97.0)
	Social	65.1 (73.3)	67.4 (69.8)	83.6 (92.2)	85.3 (94.6)
	Average	64.8 (67.4)	77.6 (78.3)	93.0 (95.6)	93.9 (96.4)

Q4 (Effectiveness of different features). As our contribution is primarily with using logical structure, we focus our discussion on the effectiveness of the layout features proposed for use in the author–affiliation matching task. In Enlil, we use a set of three features – *signal symbol*, *logical distance* and *Euclidean distance* – in the SVM model. We compare their effects in Table 7 by carrying out feature ablation testing.

It is clear that the *signal symbol* is the most important feature in both *exact* mode and *relaxed* mode. On average, the removal of this feature causes significant drop ($p < 0.01$) in Enlil's performance –

T. H. Yoon, E. J. Hwang, D. Y. Shin, and S. I. Park are with the Inter-university Semiconductor Research Center and School of Electrical Engineering, Seoul National University, Seoul 151-742, Korea.

S. J. Oh, S. C. Jung, and H. C. Shin are with the Department of Physiology, College of Medicine, Hallym University, Kangwon-do 200-702, Korea.

Figure 7: A complex scenario when Enlil fails.

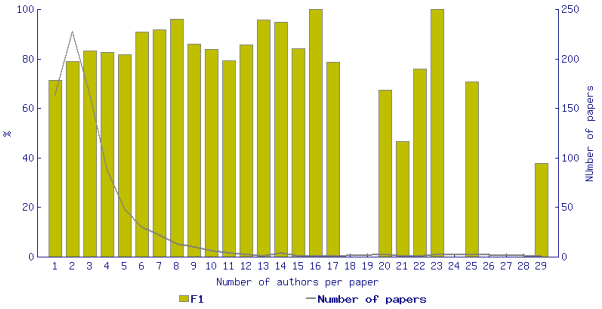

Figure 8: Enlil's exact match performance on the *Full* dataset, broken down by number of authors per paper. The dark line shows the number of papers represented by it (for e.g., in the leftmost cluster the F_1 score of .72 is the averaged performance of Enlil on 163 singly-authored papers).

26.1% in *exact* mode and 23.5% in *relaxed* mode – compared with the full feature set. These differences reflect the common use of correspondence markers in scholarly documents.

The second most important feature is the *Euclidean distance* between authors and affiliations. Its effect is less than that of *signal symbol* but still important. Enlil's performance decreases by 10.8% ($p < 0.01$) and 13.4% ($p < 0.01$) on average in *exact* and *relaxed* modes respectively. In some specific domains, such as *Formal* and *Social*, the effect of *Euclidean distance* may even outweigh the importance of *signal symbol*. We conjecture that in domains where the number of authors per publication is small, affiliations can be juxtaposed with authors' name. In such cases, correspondence markers are unnecessary and can be dismissed, which reduces the effect of *signal symbol* feature. The average numbers of authors per publication from *Social* and *Formal* domain are 2.35 and 2.60 respectively, which are smaller than the value of *Applied* domain (4.40) and *Natural* domain (4.07). This evidence strengthens our conclusion.

Logical distance is the feature which has the least significant ($p > 0.05$) effect on performance. In all cases, the performance gain of this feature is very small (from 1.0% to 1.2%). It is interesting to note that this feature is the only feature that can be obtained reliably from raw text – the other two are only available in rich text formats that encode layout information. This result implies that raw text input (provided by text extraction software *e.g. PDFBox*) will likely be problematic for any solution looking to perform author–affiliation matching.

Given that one of ten authors/institutions are incorrectly extracted and one in fifteen affiliations are incorrectly matched by Enlil, is this performance sufficient to support downstream analytics? We think the answer is yes. Extraction and matching failures happen in specific papers, and as such, errors are not uniformly distributed. Documents that do not follow regular conventions (e.g., Figure 7) will generate low-confidence results, and should be easy to detect. We note that problems also occur more frequently on scholarly documents with many authors and institutions (Figure 8), as these demonstrate more exceptions than documents with few authors.

5. APPLICATIONS

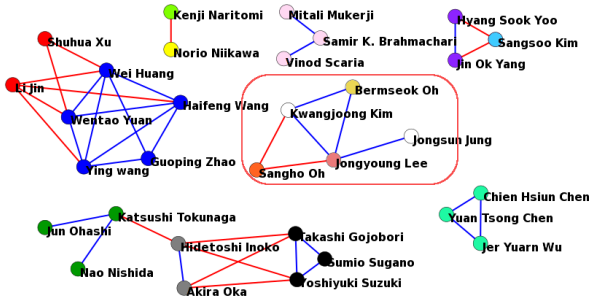

Figure 9: The same co-author network as in Figure 1 where the affiliation data is provided automatically by Enlil, compiled from freely-available publications. The circled group indicates a discrepancy in comparison to the ground truth in Figure 1.

Enlil has proven immediately useful in several scholarly analyses. Our umbrella project (from which Enlil's development was funded) aimed to define Asia via natural scholarly collaboration patterns rather than by geopolitical boundaries. By associating scholars to their institutions and countries via Enlil, we have developed novel metrics that discover nascent collaboration trends, and uncover institutions/countries that facilitate collaborations between other researchers (rather than direct publications) [6]. Cho *et al.* [5] used Enlil's author-affiliation output to aggregate the micro-patterns of collaboration among authors to macro-patterns of collaboration among Asian countries. As an example, Figure 9 shows the same network as the manually constructed version in Figure 1, but the latter uses data automatically acquired by Enlil. In this example, Enlil fails to process a publication of authors in a single marked group. Hence, all of them are assumed to have different affiliations. The rest of the network remains almost the same with only minor changes between closely related authors.

We also have created a public website[9] that allows students and laymen to perform their own scholarly network analyses through a simple four-step serial workflow. In the website, Enlil is a primary component on the backend pipeline processing. Users first input some seed authors of interest. They are then presented with a set of results harvested from Google Scholar, with which they can refine and visualize the resultant co-authorship network. The site allows the visualization of the network over time, showing how collaborations and network/citation metrics evolve. Previous scientometric work (e.g., [1]) has largely been the province of experts working with clean and costly data sources (such as Thomson Reuters' Web of Science). Our work in integrating Enlil into such a web-based frontend shows that such knowledge management tasks can now be relatively cost-free and taken back by citizen science.

We are currently using Enlil for tracking the migration of research networks. Based on publication co-authors and institutions, a researcher's changing institutional affiliations and collaborative network can be traced over time. This information can be used to construct large-scale patterns in the migration of research networks to elucidate the shifting centers of intellectual labor (Luukkonen *et al.* [23], Glanzel *et al.* [13]). Such information is vital for policy makers, economists, and social scientists to gauge how well investments into research capital and Big Science facilities have translated into the production of innovation hubs.

6. CONCLUSIONS

We have studied the extraction of author-affiliation relationship in scholarly documents and introduced *Enlil*, a complete system for

[9] `http://www.huluppu.net/`

matching authors and affiliations from publications in PDF format. Our system employs both sequence and pointwise machine learning models (CRF and SVM) to outperform the previous state-of-the-art in author/affiliation extraction and matching. The features in our system are also carefully selected and analyzed to make better use of the layout information in rich text documents.

Our evaluation over three datasets assessed Enlil's performance in breadth as well as depth. For computer science articles, Enlil significantly outperforms prior work by almost 10% in F measure. Enlil's performance gain widens when tested with articles representing a variety of domains. Analysis shows that Enlil's combined representation of both spatial and logical document structure significantly aids its performance, validating work [22] arguing in favor of rich document representations for document analysis input.

Enlil is but one component of a digital library and knowledge managment solution for scholarly publications. We have used Enlil in our various scientometric projects of which we have highlighted the PASNP case study. It has shown its value in creating network analyses in a simple and less costly manner than previous work that required clean, publisher-provisioned metadata and expert knowledge.

7. REFERENCES

[1] D. Aumüller and E. Rahm. Affiliation analysis of database publications. *SIGMOD Record*, 40(1), 2011.

[2] A. Belaïd and Y. Rangoni. Structure extraction in printed documents using neural approaches, 2006.

[3] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

[4] C.-C. Chen, K.-H. Yang, C.-L. Chen, and J.-M. Ho. Bibpro: A citation parser based on sequence alignment. *Knowledge and Data Engineering, IEEE Transactions on*, 24(2):236–250, 2012.

[5] P. S. Cho, N. Bullock, and D. Ali. The bioinformatic basis of pan-asianism. *East Asian Science, Technology, and Society*, 7(2), 2013.

[6] P. S. Cho, H. H. N. Do, M. K. Chandrasekaran, and M.-Y. Kan. Identifying research facilitators in an emerging asian research area: A case study of the pan-asian snp consortium. *Scientometrics*, 2013 Forthcoming.

[7] E. Cortez, A. S. da Silva, M. A. Goncalves, F. Mesquita, and E. S. de Moura. FLUX-CIM: flexible unsupervised extraction of citation meta-data. In *Proceedings of the ACM/IEEE-CS Joint conference on Digital libraries*, 2007.

[8] I. G. Councill, C. L. Giles, and M.-Y. Kan. ParsCit: An open-source CRF reference string parsing package. In *LREC*, May 2008.

[9] L. Egghe. Theory and practise of the g-index. *Scientometrics*, 69(1):131–152, 2006.

[10] L. Gao, X. Qi, Z. Tang, X. Lin, and Y. Liu. Web-based citation parsing, correction and augmentation. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, pages 295–304. ACM, 2012.

[11] L. Gao, Z. Tang, and X. Lin. Cebbip: A parser of bibliographic information in chinese electronic books. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, 2009.

[12] L. Gao, Z. Tang, X. Lin, Y. Liu, R. Qiu, and Y. Wang. Structure extraction from PDF-based book documents. In *Proceedings of the 11th ACM/IEEE-CS Joint conference on Digital libraries*, 2011.

[13] W. Glanzel, K. Debackere, and M. Meyer. 'triad' or 'tetrad'? on global changes in a dynamic world. *Scientometrics*, 74(1):71–88, 2008.

[14] H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox. Automatic document metadata extraction using support vector machines. In *In JCDL '03: Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 37–48, 2003.

[15] J. E. Hirsch. An index to quantify an individual's scientific research output. In *Proceedings of the National Academy of Sciences of the United States of America*, volume 102, 2005.

[16] J. E. Hirsch. An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship. *Scientometrics*, 85(3), 2010.

[17] J. Kim, D. X. Le, and G. R. Thoma. Automated labeling in document images. In *Proc. SPIE: Document Recognition and Retrieval VIII*, pages 111–122, 2001.

[18] S. Klink, A. Dengel, and T. Kieninger. Document structure analysis based on layout and textual features. In *Proc. of International Workshop on Document Analysis Systems, DAS2000*, pages 99–111, 2000.

[19] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289, 2001.

[20] C. LaGuardia. Scopus vs. web of science. http://www.libraryjournal.com/article/CA491154.html.

[21] L. D. Lopez, J. Yu, C. N. Arighi, H. Huang, H. Shatkay, and C. H. Wu. An automatic system for extracting figures and captions in biomedical pdf documents. In *IEEE International Conference on Bioinformatics and Biomedicine*, pages 578–581, 2011.

[22] M.-T. Luong, T.-D. Nguyen, and M.-Y. Kan. Logical structure recovery in scholarly articles with rich document features. *International Journal of Digital Library Systems*, 1(4):1–23, 2010.

[23] T. Luukkonen, O. Persson, and G. Sivertsen. Understanding patterns of international scientific collaboration. *Science, Technology & Human Values*, 17(1):101–126, 1992.

[24] B. U. Stefan Wuchty, Benjamin F. Jones. The increasing dominance of teams in production of knowledge. *Science*, 316(1036), 2007.

[25] P. B. Teregowda, I. G. Councill, R. J. P. Fernández, M. Khabsa, S. Zheng, and C. L. Giles. Seersuite: developing a scalable and reliable application framework for building digital libraries by crawling the web. In *Proceedings of the 2010 USENIX conference on Web application development*, pages 14–14. USENIX Association, 2010.

[26] C. S. Wagner and L. Leydesdorff. Network structure, self-organization, and the growth of international collaboration in science. *Research Policy*, 34(10):1608 – 1618, 2005.