Extracting Formulaic and Free Text Clinical Research Articles Metadata using Conditional Random Fields

Sein Lin[1] **Jun-Ping Ng**[1] **Shreyasee Pradhan**[2]
Jatin Shah[2] **Ricardo Pietrobon**[2] **Min-Yen Kan**[1]

[1]Department of Computer Science, National University of Singapore
justin@seinlin.com, {junping,kanmy}@comp.nus.edu.sg

[2]Duke-NUS Graduate Medical School Singapore
{shreyasee.pradhan,jashstar}@gmail.com, rpietro@duke.edu

Abstract

We explore the use of conditional random fields (CRFs) to automatically extract important metadata from clinical research articles. These metadata fields include formulaic metadata about the authors, extracted from the title page, as well as free text fields concerning the study's critical parameters, such as longitudinal variables and medical intervention methods, extracted from the body text of the article. Extracting such information can help both readers conduct deep semantic search of articles and policy makers and sociologists track macro level trends in research. Preliminary results show an acceptable level of performance for formulaic metadata and a high precision for those found in the free text.

1 Introduction

The increasing number of clinical research articles published each year is a double-edged sword. As of 2009, PubMed indexed over 19 million citations, over which 700,000 were added over the previous year[1]. While the research results further our knowledge and competency in the field, the volume of information poses a challenge to researchers who need to stay up to speed. Even within a single clinical research area, there can be hundreds of new clinical research results per year. Policy makers, who need to decide which clinical research proposals to fund and fast-track, and which proposals could tag onto existing research and cost share, have equally daunting information synthesis issues that have both monetary and public health implications (Johnston et al., 2006).

[1]http://www.nlm.nih.gov/bsd/bsd_key.html

Systematic reviews – secondary publications that compile evidence and best practices from primary research results – partially address these concerns, but can take years before their final publication, due to liability and administrative overheads. In many fast-paced fields of clinical practice, such guidelines can be outdated by the time of publication. Researchers and policy makers alike still need effective tools to help them search, digest and organize their knowledge of the primary literature.

One avenue that researchers have turned to is the use of automated information extraction (IE). We distinguish between two distinct uses of Information Extraction: 1) extracting regular, formulaic fields (*e.g.*, author names, their institutional affiliation and email addresses), and 2) extracting free text descriptions of key study parameters (*e.g.*, longitudinal variables, observation time periods, databases utilized).

Extracting such formulaic fields helps policy makers determine returns on health-care investments (Kwan et al., 2007), as well as researchers in large scale sociological studies understand macroscopic trends in clinical research authorship and topic shifts over time (Cappell and Davis, 2008; Lin et al., 2008). But due to the wide variety of publication venues for clinical research, even performing the seemingly simple task of author name extraction turns out to be difficult, and published studies thus far have relied on manual analysis and extraction.

Proposals to extract values of key study parameters may have more profound effects. Deeper characterization of research artifacts will enable more semantically-oriented searches of the clinical literature. Further programmatic access allows and encourages data sharing of raw clinical trial results, databases and cohorts (Piwowar and Chap-

Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents, pages 90–95,
Los Angeles, California, June 2010.

man, 2008) that may result in cost sharing across on-going studies, saving funds for other deserving clinical trials.

On one hand, the medical community has been proactive in using natural language processing (NLP) and information extraction technology in analyzing their own literature. Many approaches to metadata extraction have used regular expressions or baseline supervised machine learning classifiers. However, these techniques are not considered state-of-the-art.

On the other hand, much of the work from the NLP community applied to biomedical research has been on in-depth relationship extraction, such as the identification of gene pathways and protein-protein interaction (PPI). While certainly difficult and worthwhile problems to solve, there is room for contribution even at the basic IE level, to retrieve both regular and free form metadata fields.

We address this need in this paper. We apply a linear-chain Conditional Random Field (CRF) (Lafferty et al., 2001) as our methodology for extracting metadata fields. CRFs are a sequence labeling model that has shown good performance over a large number of information extraction tasks. We conduct experiments using basic token features to assess their efficacy for metadata extraction. While preliminary, our results indicate that CRFs are suitable for identifying formulaic metadata, but may need additional deeper, natural language processing features to identify free text fields.

2 Related Work

Many researchers have recognized the utility of the application of IE on biomedical text. These works have focused mainly on the application of well-known machine learning algorithms to tag important biomedical entities such as genes and proteins within biomedical articles. (Tanabe and Wilbur, 2002) uses a Naïve Bayes classifier, while (Zhou et al., 2004) uses use a Hidden Markov Model (HMM).

The Conditional Random Field (CRF) learning model combines the strengths of two well known methods: the Hidden Markov Model (HMM), a sequence labeling methodology, and the Maximum Entropy Model, a classification methodology. It models the probability of a class label y for a given token, directly from the observable stream of tokens x in direct (discriminative) manner, rather than as a by-product as in generative methods such as in HMMs. A CRF can model arbitrary dependencies between observation and class variables, but most commonly, a simple linear chain sequence is used (which connects adjacent class variables to each other and to their corresponding observation variable), making them topologically similar to HMMs.

Since their inception in 2001, (linear chain) CRFs have been applied extensively to many areas, including the biomedical field. CRFs have been used for the processing and extraction of important medical entities and their relationships among each other. (He and Kayaalp, 2008) reports on the suitability of CRFs to find biological entities, combining basic orthographic token features with features derived from semantic lexicons such as UMLS, ABGene, SemRep and MetaMap. In a related vein, CRFs have been applied to gene map and relationship identification as well (Bundschus et al., 2008; Talreja et al., 2004).

In a different domain, digital library practitioners have also studied how to extract formulaic metadata to enable more comprehensive article indexing. To extract author and title information, systems have used both the Support Vector Machine (SVM) (Han et al., 2003) and CRFs (Peng and McCallum, 2004; Councill et al., 2008). These works have been applied largely to the computer science community and have not yet been extensively tested on biomedical and clinical research articles.

Our work differs from the above by making use of CRFs to extract fields in clinical text. Similar lexical-based features are employed, however in addition to regular author metadata, we also attempt to extract domain-specific fields from the body text of the article.

3 Method

External to the scope of the research presented here, our wider project goal focuses on constructing a knowledge base of clinical researchers, databases, instruments and expertise in the Asia-Pacific region.

Dataset. In this pilot study, we created a gold standard dataset consisting of freely-available articles available from PubMedCentral. These arti-

cles focused on health services research in the Asia-Pacific region. In particular, we selected open-access full-text literature documenting oncological and cardio-vascular studies in the region, over a three year period from 2005 to 2008.

By constructing an appropriate staged query with PubMed, we obtained an initial listing of 260 articles. From an initial analysis, we determined that a significant portion ($\sim$1/3) of the retrieved full-text were not primary research, but reviews, case studies, editorials or descriptions. After eliminating these, the remaining 185 articles were earmarked to be manually tagged by clinicians affiliated with the project. Since the resulting corpus compiles articles across different journals and other publication venues, their presentation of even the formulaic author metadata varied.

The clinicians were given rich text (RTF) versions of the original HTML documents retrieved from PubMed. They identified and extracted only the sections of the articles that had pertinent data classes to tag. This process excluded most introductory, discussion and result sections, preserving the sections that described the study and results at a high level (*e.g.*, *Demographics* and *Methods*).

After an initial training session, each clinician used a word processor to manually insert opening and closing XML tags for the tagset for a particular subsection of the 185-article corpus. Due to the high cost of clinician time, we chose to emphasize coverage, rather than have the clinicians multiply annotate the same articles. As a result, we could not calculate annotation agreement, but feel that the repeatability of the annotation was addressed by the initial training. At the time of writing, 93 articles have been completely tagged and sectioned, with the remainder in progress. The average length of the documents is about 1300 words. Once the dataset has been completed, we plan to release the annotated data offsets to the public, to encourage comparative evaluation.

The clinicians annotated the following *Formulaic Author Metadata* (3 classes):

- **Author (Au):** The names of the authors of the study;
- **E-mail (Em):** The email addresses of the corresponding authors of the study;
- **Institution (In):** The names of the institutions that the authors are from.

Such metadata can be used to build an author citation network for macro trend analysis. Note that this data is obtained from the article's title page itself, and not from any references to source articles, which have been the target of previous studies on CRF-based information extraction (Peng and McCallum, 2004; Councill et al., 2008). The clinicians also annotated the following *Key Study Parameters* (10 classes):

- **Age Group (Ag):** The age range of the subjects of the study (*e.g.*, *45 and 80 years*, *21-79 years*);
- **Data Analysis Name (Da):** The name of the method or software used in the analysis of data collected for the study (*e.g.*, *proportional hazards survival models*, *SAS package*);
- **Data Collection Method (Dc):** The data collection methods for the study (*e.g.*, *medical records*, *review of medical records and linkage to computerized discharge abstracts*);
- **Database Name (Dn):** The name of any biomedical databases used or mentioned in the study (*e.g.*, *Queensland Cancer Registry*, *National Death Index*, *population-based registry*);
- **Data Type (Dt):** The type of data involved in the study (*e.g.*, *Cohort study*, *retrospectively*);
- **Geographical Area (Ga):** The names of the geographical area in which an experiment takes place or the subjects are from (*e.g.*, *Pune*, *Switzerland*);
- **Intervention (Iv):** The name of medical intervention used in the study (*e.g.*, *surgery*, *radiotherapy*, *chemotherapy*, *radio-frequency ablation*);
- **Longitudinal Variables (Lv):** Data collected over the observation period (*e.g.*, *subjects*);
- **Number of Observations (No):** The number of cases or subjects observed in the study (*e.g.*, *158 Indigenous*, *84 patients*);

• **Time Period (Tp):** The duration of an experiment or observation in the study (*e.g.*, *1997–2002*, *between January 1988 and June 2006*).

As can be seen from the examples, the tagging guidelines loosely define the criteria for tagging. For some classes, clinicians tagged entire noun phrases or clauses, and for others, only numeric values and modifiers were tagged. This variability arises from the difficulty in tagging these free text fields.

Features. The CRF model requires a set of binary features to serve as a representation of the text. A simple baseline is to use the presence/absence of particular tokens as features. The CRF software implementation we utilized is CRF++[2], which compiles the binary features automatically from a context window centered the current labeling problem instance.

We first preprocess an input article from its RTF representation and convert it into plain text. This is a lossy transformation that discards font information and corrupts mathematical symbols that could be helpful in the detection task. We take hyphenated token forms (*e.g.*, *2006-2007*) and convert them into individual tokens. The plain text is processed to note the specific locations of the XML tags for the learning process. The bulk of the words in each article were not tagged by clinicians, and for these words, we assigned a **Not Applicable (NA)** tag. We list the simple inventory of feature types that we use for classification.

• **Vocabulary:** Individual word tokens are stemmed with Porter's stemming algorithm (Porter, 1980) and down-cased to collapse orthographic variants. Each word token is then used as an individual feature. This feature alone was used to compile baseline performance as discussed later in evaluation.

• **Lexical:** Lists of keywords were compiled to lend additional weight for specific classes. In particular, we compiled lists of months, common names, cue words that signaled observations, institution names and data analyses methods. For example, a list of common given and surname names is useful for the **Au** field; while a list of months and their abbreviated forms help to identify **Tp**. Each list constitutes a different feature. As an example, in the case of human names, the names *Alice* and *Auburn* are on the list. If a word token corresponds to any of the words in the list, the corresponding feature is turned on (*e.g.*, isApersonName).

• **Position:** If a word token is within the first 15 lines of an article, this feature is turned on. This specifically caters to limit the scope of the formulaic author metadata fields, to match them only at the beginning of the article.

• **Email:** We create a specific feature for email addresses that is turned on when a particular word token is matched by a handwritten regular expression.

• **Numeric:** For some free text classes, such as **Ag**, **No** and **Tp**, the tagged text often contains numeric data. This can be present in both numeric and word form (*e.g.*, *23* versus. *twenty-three*). We turn this feature on for a token solely containing digits or numeric word forms.

• **Orthographic:** Orthographic features, such as the capitalization of a word token are useful to help identify proper nouns and names. If there are capital letters within a word token, this feature is turned on.

4 Evaluation

To ascertain the efficacy of our proposed solution, three-fold cross validation (CV) was first performed on a dataset comprising the 93 articles which have been completely annotated.

Baseline. For the purpose of comparison, we created a baseline system that utilizes the same CRF++ toolkit but uses only the vocabulary feature type with a five-word window (two previous tokens, the target token to be classified, and two subsequent tokens). The performance of this baseline system is shown in Table 1, where the standard classification performance measures of precision, recall and F_1 are given. *Count* measures the number of word tokens that are predicted as belonging to the stated field.

Discussion. We see that the overwhelming majority of tokens are not tagged (belonging to class **NA**).

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

The skewness of the dataset is not uncommon for IE tasks.

The baseline results show weak performance across the board. Clearly, significant feature engineering could help boost performance. Of particular surprise was the relatively weak performance on the formulaic metadata. From our manual analysis, it was clear that the wide range and variety of tokens present in names and institutions barred the system from achieving good performance on these classes. Comparative studies in citation and reference parsing usually peg classification performance of these classes at the 95% and above level.

Without suitable customization, detection of the key study parameters was also not possible. Only relatively common fields could be captured by the CRF, and when captured were more precise but lacked enough data to build a model with any acceptable level of recall.

Table 2 illustrates the improved results obtained by running CRF++ with all of the described features on the same dataset. The same five-word window size is used for the vocabulary feature. As seen, significant improvements over the baseline are obtained for all except four fields — **Da**, **Dc**, **Iv**, and **Lv**. These four fields were the classes with the most variability in annotation. For example, the data collection methodologies (**Dc**) and interventions (**Iv**) are often captured as long sentence fragments and hard to model with individual word cues.

The largest improvements occurred for the classes of age groups **Ag** and time periods **Tp**, both of which benefited from the addition of the numeric feature which boosted recognition performance.

5 Future Work

The work presented here is ongoing, and based on our current results, we are planning to re-examine the quality of the annotations and refine our annotation guideline and scheme. We discovered cases where the CRF tagger correctly annotated key study parameters which the annotators had missed or miskeyed. Drawing on lessons from the initial annotation exercise, a more comprehensive guideline is planned which will provide concise instructions with accompanying annotation examples.

We also plan to enrich the feature set. The current

Field	Prec.	Recall	F_1	Count
Formulaic Author Metadata				
Au	84.6	74.3	79.1	1818
Em	93.4	92.2	92.8	151
In	80.5	69.5	74.6	3906
Macro Avg.	86.2	78.7	82.3	
Key Study Parameters				
Ag	29.0	40.4	33.8	334
Da	61.0	39.0	47.6	708
Dc	8.3	3.2	4.6	48
Dn	35.9	15.1	21.2	92
Dt	52.8	26.8	35.5	36
Ga	7.3	4.5	5.6	41
Iv	4.6	1.4	2.1	22
Lv	15.4	20.0	17.4	13
No	14.4	5.8	8.3	125
Tp	73.6	55.8	63.5	261
Macro Avg.	30.2	21.2	24.0	
NA	97.1	98.5	97.8	119998

Table 1: Baseline aggregated results over 93 tagged articles under 3 fold cross validation.

Field	P.	Recall	F_1	Count
Formulaic Author Metadata				
Au	89.0	85.3	87.1	7312
Em	100.0	97.3	98.6	154
In	91.3	78.0	84.1	4515
Macro Avg.	93.4	86.6	89.9	
Key Study Parameters				
Ag	64.3	35.4	45.7	240
Da	79.3	37.2	50.6	2296
Dc	20.0	1.6	2.9	125
Dn	42.5	10.5	16.8	219
Dt	70.0	19.7	30.7	71
Ga	43.7	10.4	16.8	62
Iv	40.0	2.7	5.1	73
Lv	0.0	0.0	0.0	10
No	43.4	10.7	17.1	308
Tp	82.7	69.4	75.5	344
Macro Avg.	48.5	19.7	26.1	
NA	97.5	99.3	98.4	120430

Table 2: Aggregated results using the full feature set under 3 fold cross validation.

set employed is still simplistic and serves as a developmental platform for furthering our feature engineering process. For example, the vocabulary, position and word lists features can be further modified to capture more fined-grained information.

Once we exhaust the development of basic features, our future work will attempt to harness deeper, semantic features, making use of part-of-speech tags, grammar parses, and named entity recognition for example. The incorporation of these features will likely be useful in improving the performance of the CRF learner. We also plan to use both clinical research and general medical ontologies (*e.g.*, UMLS) to gain additional insight on individual terms that have special domain-specific meanings.

6 Conclusion

We have developed a CRF-based information extraction system that targets two different types of metadata present in clinical articles. Our work in progress demonstrates that formulaic author metadata can be effectively extracted using the CRF methodology. By further performing feature engineering, we were able to extract key study parameters with a moderate level of success. Our post evaluation analysis indicates that more careful attention to annotation and feature engineering will be necessary to garner acceptable performance of such important clinical study parameters.

Acknowledgments

We like to express our gratitude to the reviewers whose insightful comments and pointers to additional relevant studies have helped improve the paper.

References

M. Bundschus, M. Dejori, M. Stetter, V. Tresp, and H.P. Kriegel. 2008. Extraction of semantic biomedical relations from text using conditional random fields. *BMC bioinformatics*, 9(1):207.

Mitchell S. Cappell and Michael Davis. 2008. A significant decline in the american domination of research in gastroenterology with increasing globalization from 1980 to 2005: An analysis of american authorship among 8,251 articles. *The American Journal of Gastroenterology*, 103:1065–1074.

Isaac G. Councill, C. Lee Giles, and Min-Yen Kan. 2008. ParsCit: An open-source CRF reference string parsing package. In *Proceedings of the Language Resources and Evaluation Conference (LREC 08)*, Marrakesh, Morrocco.

Hui Han, C. Giles, E. Manavoglu, H. Zha, Z. Zhang, and Ed Fox. 2003. Automatic document meta-data extraction using support vector machines. In *Proceedings of Joint Conference on Digital Libraries*.

Ying He and Mehmet Kayaalp. 2008. Biological entity recognition with conditional random fields. In *Proceedings of the Annual Symposium of the American Medical Informatics Association (AMIA)*, pages 293–297.

S.C. Johnston, J.D. Rootenberg, S Katrak, Wade S. Smith, and Jacob S Elkins. 2006. Effect of a us national institutes of health programme of clinical trials on public health and costs. *Lancet*, 367:13191327.

Patrick Kwan, Janice Johnston, Anne Fung, Doris SY Chong, Richard Collins, and Su Lo. 2007. A systematic evaluation of payback of publicly funded health and health services research in hong kong. *BMC Health Services Research*, 7(1):121.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the International Conference on Machine Learning*, pages 282–289.

JM Lin, JW Bohland, P Andrews, Burns GA, CB Allen, and PP Mitra. 2008. An analysis of the abstracts presented at the annual meetings of the society for neuroscience from 2001 to 2006. *PLoS ONE*, 3(e2052).

F. Peng and A. McCallum. 2004. Accurate information extraction from research papers using conditional random fields. In *Proceedings of Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 329–336.

Heather A. Piwowar and Wendy W. Chapman. 2008. Identifying data sharing in biomedical literature. In *Proceedings of the Annual Symposium of the American Medical Informatics Association (AMIA)*.

M.F. Porter. 1980. An Algorithm For Suffix Stripping. 14(3):130–137.

R. Talreja, A. Schein, S. Winters, and L. Ungar. 2004. GeneTaggerCRF: An entity tagger for recognizing gene names in text. Technical report, Univ. of Pennsylvania.

L. Tanabe and W.J. Wilbur. 2002. Tagging gene and protein names in biomedical text. *Bioinformatics*, 18(8):1124.

G. Zhou, J. Zhang, J. Su, D. Shen, and C. Tan. 2004. Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics*, 20(7):1178–1190.