

## Accurate Information Extraction from Research Papers using Conditional Random Fields

Fuchun Peng  
Department of Computer Science  
University of Massachusetts  
Amherst, MA 01003  
fuchun@cs.umass.edu

Andrew McCallum  
Department of Computer Science  
University of Massachusetts  
Amherst, MA 01003  
mccallum@cs.umass.edu

### Abstract

With the increasing use of research paper search engines, such as CiteSeer, for both literature search and hiring decisions, the accuracy of such systems is of paramount importance. This paper employs Conditional Random Fields (CRFs) for the task of extracting various common fields from the headers and citation of research papers. The basic theory of CRFs is becoming well-understood, but best-practices for applying them to real-world data requires additional exploration. This paper makes an empirical exploration of several factors, including variations on Gaussian, exponential and hyperbolic- $L_1$  priors for improved regularization, and several classes of features and Markov order. On a standard benchmark data set, we achieve new state-of-the-art performance, reducing error in average F1 by 36%, and word error rate by 78% in comparison with the previous best SVM results. Accuracy compares even more favorably against HMMs.

### 1 Introduction

Research paper search engines, such as *CiteSeer* (Lawrence et al., 1999) and *Cora* (McCallum et al., 2000), give researchers tremendous power and convenience in their research. They are also becoming increasingly used for recruiting and hiring decisions. Thus the information quality of such systems is of significant importance. This quality critically depends on an information extraction component that extracts meta-data, such as title, author, institution, etc, from paper headers and references, because these meta-data are further used in many component applications such as field-based search, author analysis, and citation analysis.



Previous work in information extraction from research papers has been based on two major machine learning techniques. The first is hidden Markov models (HMM) (Seymore et al., 1999; Takasu, 2003). An HMM learns a generative model over input sequence and labeled sequence pairs. While enjoying wide historical success, standard HMM models have difficulty modeling multiple non-independent features of the observation sequence. The second technique is based on discriminatively-trained SVM classifiers (Han et al., 2003). These SVM classifiers can handle many non-independent features. However, for this sequence labeling problem, Han et al. (2003) work in a two stages process: first classifying each line independently to assign it label, then adjusting these labels based on an additional classifier that examines larger windows of labels. Solving the information extraction problem in two steps looses the tight interaction between state transitions and observations.

In this paper, we present results on this research paper meta-data extraction task using a Conditional Random Field (Lafferty et al., 2001), and explore several practical issues in applying CRFs to information extraction in general. The CRF approach draws together the advantages of both finite state HMM and discriminative SVM techniques by allowing use of arbitrary, dependent features and joint inference over entire sequences.

CRFs have been previously applied to other tasks such as name entity extraction (McCallum and Li, 2003), table extraction (Pinto et al., 2003) and shallow parsing (Sha and Pereira, 2003). The basic theory of CRFs is now well-understood, but the best-practices for applying them to new, real-world data is still in an early-exploration phase. Here we explore two key practical issues: (1) regularization, with an empirical study of Gaussian (Chen and Rosenfeld, 2000), exponential (Goodman, 2003), and hyperbolic- $L_1$  (Pinto et al., 2003) priors; (2) exploration of various families of features, including text, lexicons,

### Metadata

<b>Title:</b>	Accurate Information Extraction from Research Papers using Conditional Random Fields
<b>Author(s):</b>	Fuchun Peng, Andrew McCallum
<b>Venue:</b>	HLT-NAACL
<b>Year:</b>	2004

Title	Author(s)	Venue	Year
  Accurate Information Extraction from Research Papers using Conditional Random Fields	Fuchun Peng, Andrew McCallum	HLT-NAACL	2004