Computer Science Department Faculty Publication Series

Computer Science

2003

A Note on the Unification of Information Extraction and Data Mining using Conditional-Probability, Relational Models

Andrew McCallum
University of Massachusetts - Amherst

David Jensen
University of Massachusetts - Amherst

A Note on the Unification of Information Extraction and Data Mining using Conditional-Probability, Relational Models

Andrew McCallum

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

David Jensen

Department of Computer Science
University of Massachusetts Amherst
Amherst, MA 01003 USA
jensen@cs.umass.edu

Abstract

Although information extraction and data mining appear together in many applications, their interface in most current systems would better be described as serial juxtaposition than as tight integration. Information extraction populates slots in a database by identifying relevant subsequences of text, but is usually not aware of the emerging patterns and regularities in the database. Data mining methods begin from a populated database, and are often unaware of where the data came from, or its inherent uncertainties. The result is that the accuracy of both suffers, and significant mining of complex text sources is beyond reach.

This position paper proposes the use of unified, relational, undirected graphical models for information extraction and data mining, in which extraction decisions and data-mining decisions are made in the same probabilistic "currency," with a common inference procedure—each component thus being able to make up for the weaknesses of the other and therefore improving the performance of both. For example, data mining run on a partially-filled database can find patterns that provide "top-down" accuracy-improving constraints to information extraction. Information extraction can provide a much richer set of "bottom-up" hypotheses to data mining if the mining is set up to handle additional uncertainty information from extraction.

We outline an approach and describe several models, but provide no experimental results.

1 Introduction

Data mining gives us the ability to see patterns, predict the future, and make informed decisions based on the evidence in large databases. For example, data mining of categorical and numerical consumer shopping data allow a retailer to understand which items are bought by the same customers, predict sales of seasonal items, and more efficiently manage its inventory.[1] Over the past decade, the use of data mining

[1]While in some circles, data mining indicates "unsupervised discovery of patterns," here we include classification and other supervised learning tasks within the scope of data mining.

techniques has revolutionized many commercial and government enterprises by enabling more accurate decision making in such areas as industrial control [Wang, 1999], fraud detection [Fawcett and Provost, 1997], inventory management [Agrawal *et al.*, 1993], and customer relationship management [Domingos and Richardson, 2001].

There is already much data in the necessary "database-form," (with fields and records), but there is also a vast amount of important information available only in natural language text, such as Web pages, publications, corporate memos, research findings, government reports and other documents. To be accurately mined, these data must first be be first organized and normalized into database-form.

Information extraction aims to do just this—it is the process of filling the fields and records of a database from unstructured text. Its traditional intended use is as the first step of a pipeline in which unstructured text is converted into a structured database, and then data mining produces predictive models from this database. Historically information extraction has most often been studied for news articles [Appelt *et al.*, 1995], but more recently has been applied to many textual formats, including Web pages [Soderland, 1997; Craven *et al.*, 1998; Blei *et al.*, 2002], government reports [Pinto *et al.*, 2003], scientific articles [Lawrence *et al.*, 1999; McCallum *et al.*, 2000b; Ray and Craven, 2001] and legal documents [Bruninghaus and Ashley, 2001]. Also recently there has been somewhat of a revolution in the use of statistical and machine learning methods for information extraction, *e.g.* [Bikel *et al.*, 1997; McCallum *et al.*, 2000a; Lafferty *et al.*, 2001; Carreras *et al.*, 2002; Roth and Yih, 2002; Ray and Craven, 2001; Klein *et al.*, 2003].

However, in spite of the improved results of these machine learning methods, and in spite of a surge of over-anxious commercial ventures claiming success, information extraction with sufficient accuracy to dump directly into data mining remains elusive, and the promise of mining from textual sources is largely unfulfilled. Although there has been much discussion about combining information extraction and data mining, there are few examples of successful pipelining of the two technologies on anything but simple problems.

This position paper proposes *extraction-mining random fields*—a family of models for improving our ability to data mine information in unstructured text by using information extraction and data mining methods that have such tight integration that the boundaries between them disappear, and they

can be accurately described as a *unified framework* for extraction and mining. This framework uses rich, intertwined undirected graphical models in which extraction decisions and data-mining decisions are made with a common inference procedure—the evidence for an outcome being the result of inference both "bottom up" from extraction, and "top down" from data mining. Thus (1) intermediate hypotheses from both extraction and data mining can be easily communicated between extraction and data mining in a closed loop system, (2) mutually-reinforcing evidence and uncertainty will have the opportunity to be properly marshaled, (3) and accuracy and confidence assessment will improve.

Our focus in both areas is on relational data—data about entities and links that is better described by graphs than by the flat attribute-value representations used in much of machine learning. The edges (or hyper-edges) in such graphs represent binary (or n-ary) relations between entities, such as familial relationships among people or hyperlink relations among web pages. In terms of probabilistic models, individual relations or chains of multiple relations help structure the probabilistic dependencies among entities. More formally, in addition to having graph structure, we define a relational task as one in which the system's outputs have several components, $\mathbf{y} = \{y_1, ...\}$, and not all the components are independent from each other given the inputs, $\mathbf{x}$; thus $\exists i, j$ such that $P(y_i|\mathbf{x}) \neq P(y_i|y_j, \mathbf{x})$.

Our proposed models are all trained to maximize conditional probability of the outputs given the inputs. Such models have the advantage of not requiring explicit representation of dependencies among the features of the input. This is especially advantageous when using complex, overlapping and multi-granularity features, as is common in work with natural language text [McCallum *et al.*, 2000a; Lafferty *et al.*, 2001].

2 The Task and Problem

Data mining has enabled a revolution in planning, decision making and organizational efficiency in many areas of industry and government. A similar revolution could be brought about in many additional areas if it were possible to mine the vast amount of information currently locked in unstructured text. In many domains, there is far more information in documents and other text than there is in structured databases.

For example, CiteSeer [Lawrence *et al.*, 1999] mines the Web for research papers, extracts title, authorship and citation information, and thus enables analysis of the citation graph for finding seminal and survey papers. This service has had significant impact on the the practice of computer science research. However, the variety of fields and relations it extracts is small, and the limited accuracy of its existing relations constrains the ability to perform more sophisticated data mining. For example, Pasula *et al.* [2002] note that CiteSeer contains records of over 30 separate AI textbooks written by Russel and Norvig, when actually there is only one.

Unfortunately, the complex data mining of rich unstructured text is not feasible with current methods: extraction is often inaccurate, co-reference resolution is often poor, and data mining is not able to recover from a noisy database.

2.1 Inaccurate extraction

State-of-the-art precision and recall for extracting named entities (such as people, organizations and locations) is in the low- to mid-90's percent for many systems and domains—including BBN's IdentiFinder on news wire articles [Bikel *et al.*, 1997], Cora's hidden Markov models on research paper headers [McCallum *et al.*, 2000b], and WhizBang Lab's extractors on web page data [McCallum, 2002]. The winners of the CoNLL-2002 named entity competition [Carreras *et al.*, 2002] reached only about 80% precision and recall on Spanish newswire text. One of the most recent research papers on named entity extraction from Web pages reached precision and recall in the high 80s [Collins, 2002]. Reaching about 90% precision and recall may seem good until one realizes that this means that more than one in ten fields in the database are either incorrect or missing.

When we consider the accuracy of database records (or "relations") instead of individual fields, the state-of-the-art is even worse. For a relation to be correct, all its constituent fields and its relation-type categorization must be correct. Even if a system had 95% accuracy in extracting individual fields and categorizing relations, the overall accuracy of a three-slot relation would be only 80%. This happens because each automated decision in the formation of a relation is performed independently, and the errors compound. For example, the top performer in the 2002 DARPA ACE evaluation had entity extraction precision and recall scores of about 80%, but binary relation extraction scores of only roughly 60% [DARPA, 2002].

A better solution should not treat the components of a relation independently, but should make coordinated decisions and model them together. For example, the model could know that a person graduates from a university, not from another person, and use this to coordinate its extraction of a person name, a university name, and its categorization of the relation. If done correctly, relations should actually provide constraints that help improve overall extraction accuracy, not hurt it. This idea is one component of our proposed approach, and is expanded in the section 3.

2.2 Poor coreference resolution

One of the key problems in current systems that work on unstructured text is recognizing when two string names are referring to the same entity. For example, "Colin Powell," "Powell," "U.S. Secretary of State," "the Secretary of State" are not string-identical, but in some context may all refer to the same person. If they get separate entries in the database, relational connections will be missing, and data mining will not find the patterns it should [Jensen, 1999].

Coreference (also known as de-duplication, or record matching) is also a difficult problem in traditional databases. There, some of the most successful approaches bring to bear a multitude of evidence from different fields of each record, *e.g.* [Borthwick *et al.*, 2000; Bilenko and Mooney, 2002]. However, the problem is especially difficult in text domains where the original data is unstructured, the availability of some fields is questionable, and the collection of fields into records has not yet been performed.

Often some amount of coreference resolution must happen in order to gather all fields of a record because the infor-

mation is dispersed across multiple sentences, paragraphs or documents. Thus we have a difficult chicken-and-egg problem: to perform accurate coreference we need a multitude of evidence from different fields of a record, but to gather all the fields of a record we rely on coreference resolution. Coreference resolution and record (relation) building should happen simultaneously in a coordinated effort.

Part of the reason coreference has historically been so problematic in text domains is that it sits on the boundary between extraction and data mining. Formation of the fields and records is addressed by extraction; record de-duplication is usually seen as a database issue. However, as we have just pointed out, they rely on each other in highly intertwined ways. They cannot be deeply solved separately. This is particularly true of cross-document coreference, an extremely important problem that has received little attention.

Early work on relational coreference resolution includes Pasula *et al.* [2002] and McCallum and Wellner [2003]; the later is briefly described in section 3.4.

2.3 Fragile data mining

One might hope that data mining techniques could compensate for the errors introduced by inaccurate extraction and poor coreference resolution. Research in data mining has a long history of constructing accurate models using combinations of many features. Work with decision trees, Bayesian classifiers, support-vector machines, and ensemble methods, has produced methods that combine large numbers of (potentially noisy) features into a single model that can "damp out" high levels of noise and allow accurate predictions.

Unfortunately, this existing work on high-accuracy classifiers presumes propositional instances, each of which has large numbers of features. In contrast, data produced by information extraction has a rich relational structure, but each entity and relation has relatively few features. This obviates the strategies used to such great effect in propositional learners, and can often result in brittle, inaccurate models. Some relational learning techniques attempt to overcome this difficulty by constructing relational features to supplement the relatively small number of intrinsic features present in the raw data. However, such calculations rely simultaneously on both extracted relations (the most error prone element of extracted data) and extracted features, so they suffer from the combined errors of both types of data.

Fortunately, relational graphical models can leverage two sources of added power to compensate for the relative lack of high-quality features. First, these models can incorporate information about the uncertainty of the underlying data to influence how strongly specific features influence predictions. By using uncertainty estimates on extracted entities, relations, and features, the models can "play to the strengths" available in extracted data. Second, these models can use the relational structure of the data themselves so that high-confidence inferences about some entities can be used to aid inferences about related entities. We discuss this approach in more detail in section 3.3.

2.4 Consequences of Problems

The consequence of these problems is that little or no data mining is conducted on databases produced through extrac-

tion from unstructured text.

A few preliminary research-level exceptions are discussed in section 4. Two larger-scale exceptions are FlipDog.com (a database of job openings populated mostly through extraction), and CiteSeer [Lawrence *et al.*, 1999] (a database of research papers and citations populated through various automatic methods). However, FlipDog makes significant concessions in recall to obtain higher precision, and also relies on non-trivial amounts of human verification to clean its database [McCallum, 2002]. In CiteSeer, the extraction of research paper references is significantly easier than most kinds of named entity extraction from less structured data, and CiteSeer still makes many significant errors in extraction and coreference (as described in the "Russell and Norvig" example in section 2).

We believe that extraction and data mining should be able to help each other through close coordination rather than each failing separately. We describe our approach in some detail in the next section.

3 A Solution

Our approach to both information extraction and data mining is based on statistical machine learning and probabilistic models. These methods have had a high degree of success in each of the two fields recently. There are also strong benefits to using models of IE and data mining that are tightly compatible with each other—with both of them speaking the language of probabilities, they will share a common, low-level communication medium.

In fact, we propose a model that is so tightly integrated that the boundaries between IE and data mining disappear. Our proposed unified system can be understood as a single, large conditionally-trained undirected graphical model. This is a type of probabilistic model that excels at capturing highly interdependent, relational data in which strict causality among events is not necessarily apparent—a set of circumstances appearing both in low-level text data and higher-level relational data mining.

In the next subsections we describe how recent research in both information extraction and data mining have independently arrived at undirected graphical models, and then describe our proposed unification, the advantages of our approach, and several specific models.

3.1 Models for information extraction

Finite state machines are the dominant model for information extraction both in industry and research. There was significant early work with hand-tuned finite state transducers, *e.g.* [Jerry *et al.*, 1996], but more recent work is with finite state machines whose parameters are set by machine learning—most commonly hidden Markov models [Bikel *et al.*, 1997; Leek, 1997; Freitag and McCallum, 1999; Ray and Craven, 2001].

Hidden Markov models have parameters for state-to-state transition probabilities and per-state observation emission probabilities. From these one can easily calculate the probability that the model would have generated a particular state sequence associated with a particular observation symbol sequence. When used for extraction, the emission symbols are

typically natural language words, and states are associated with different extraction fields. For example, to extract person names, the hidden Markov model may have two states, one for person-names, and one for other. To perform extraction on a particular word sequence, one uses the Viterbi algorithm to find the state sequence most likely to have generated the given the observed word sequence, and then designates as person names any words Viterbi that claims were generated while in the person-name state.

A disadvantage of hidden Markov models is that, being generative models of the observation sequence, they are limited in their ability to represent many non-independent, overlapping features of the sequence. In other words, since the observations are *generated* by the model, the model must represent any correlations between features in order to faithfully reproduce them. When there are many correlated features, or complex dependencies among them, (or a desire to capture features at multiple levels of granularity and features of the past and future), this modeling is prohibitively difficult, (and in many cases impossible).

The ability to use arbitrary features is important because often significant features of the observation sequence include not just the identity of the words, (e.g. the word "Wisniewski" was observed), but also other features of the word and its context—for example, it is capitalized, it ends in "ski," it is in bold face, left justified, it is a member of a list of last names from the U.S. Census, the previous word is a recognized first name, and the next word is "said". All of these are powerful pieces of evidence that the word is a person's last name— especially useful evidence if the word "Wisniewski" does not appear anywhere in the labeled training data, (a typical circumstance in the common case of limited labeled data).

Furthermore, and highly significant to our approach, we also want an information extraction model that provides a place for data mining to inject arbitrary "top-down" information that could improve extraction accuracy. A simple, yet powerful interface between data mining and extraction is for the extraction model to see the output of data mining essentially as additional features—top-down features instead of bottom-up word features. Details and variations are discussed in the following subsections.

Maximum entropy Markov models (MEMMs) [McCallum *et al.*, 2000a] and conditional random fields (CRFs) [Lafferty *et al.*, 2001] are two conditional-probability finite state models that—because they are conditional instead of generative— afford the use of arbitrary features in their modeling of the observation sequence.

Conditional Markov models have provided strong empirical success. They extracted question-answer pairs from Frequently-Asked-Question lists with double the precision of an HMM [McCallum *et al.*, 2000a]. They reduced part-of-speech tagging errors on unknown words by 50% over an HMM [Lafferty *et al.*, 2001]. They have achieved world-class results in noun phrase segmentation [Sha and Pereira, 2003a]. They found tables in government reports significantly more accurately than previous methods [Pinto *et al.*, 2003]. They remain an extremely promising area for new research.

3.2 Models for data mining

Work on data mining has traditionally relied on a common family of techniques for learning statistical models from propositional data. For example, algorithms that learn decision trees [Quinlan, 1993; Breiman *et al.*, 1984], linear models [McCullagh and Nelder, 1989], and simple Bayesian classifiers [Mitchell, 1997] are typical parts of many data mining systems. More recently, work has focused on how to combine simple models into more complex models such as ensembles learned through bagging [Breiman, 1996] and boosting [Schapire, 1999]. Finally, the use of graphical models of propositional data [Jordan, 1998] has become widespread, often incorporating simple classifiers such as decision trees to estimate conditional probability distributions.

Unfortunately, attempting to adapt these propositional learners to relational data can lead to serious statistical errors. Over the past two years, the second author has identified several ways in which the structure of relational data can cause significant bias in learned models. For example, many relational data sets exhibit autocorrelation among the features of relational entities (e.g., most coauthors of a paper tend to be employed by a single type of organization). This autocorrelation can be useful for prediction, but it can also systematically bias naive learning algorithms toward features with the *least* supporting evidence [Jensen and Neville, 2002]. More recently, we have also discovered that correlation between the feature values and the structure of relational data can cause naive learners to produce models with invalid structure [Jensen *et al.*, 2003a]. We have found solutions to both these problems [Jensen and Neville, 2003; Jensen *et al.*, 2003a] and incorporated them into our own relational learning algorithms.

Another failing of many traditional data mining techniques is that they do not use uncertainty information on data items. Although we know the probability of correct extraction for a given entity or relation, most data mining models cannot use that information during learning or inference. Notable exceptions are the techniques for learning and inference in graphical models.

A final failing of traditional models learned through data mining is that they make predictions for each instance (e.g., each document) individually, independent of any other. These approaches typically "propositionalize" the data, by flattening complex relational data into a single table. Such approaches miss the potential opportunity to correct for errors on some instances based on higher-confidence predictions about related instances.

Fortunately, a small but growing body of researchers is exploring new methods for relational data mining that overcome these difficulties. These techniques move beyond naive adaptations of methods for propositional learning, and they take seriously the unique opportunities and challenges presented by relational data. One excellent example is the work by Getoor *et al.* [2001] on learning probabilistic relational models (PRMs), a form of directed graphical model that learns the interdependence among features of related entities. PRMs have been applied to learning relationships among movies and their actors, among tuberculosis patients and their contacts, and among Web pages.

Despite their power, PRMs are unable to express many of

the types of mutual dependence among features because a PRM must be a *directed acyclic* graph. For example, the acyclicity constraint makes it nearly impossible to express autocorrelation [Jensen and Neville, 2002], a nearly ubiquitous feature of relational data sets. Autocorrelation can be used to greatly improve model accuracy through the natural feedback of probabilistic inference.

Undirected graphical models, however, remove the acyclicity constraint, and some of the most advanced work in relational learning has focused on these models in the past two years. These models combine the benefits of traditional graphical models, including understandability and incorporation of uncertainty, with the advantages of full inferential feedback. Studies of *relational* or *collective* classification with undirected models [Taskar *et al.*, 2002; Neville and Jensen, 2000] have shown impressive gains in accuracy. Based on the our preliminary work, undirected graphical models of relational data are poised to produce substantial accuracy gains in almost all cases, analogous to the type of gains seen with ensemble classifiers [Breiman, 1996; Schapire, 1999] and for the same reasons—substantial reductions in variance because of an increase in the evidence used for each inference [Jensen *et al.*, 2003b].

3.3 A Unified Model

Thus, conditionally-trained, undirected graphical models are at the heart of recent work in two fields: one examining data at word level for information extraction, and the other examining data at the entity level for data mining. Even though they provide modeling at different levels of abstraction, they meet each other at the entity level, and are fundamentally providing models of the same data—one "bottom up," the other "top down."

The two models are entirely compatible with each other. An undirected graphical model of information extraction can be combined with an undirected graphical model of data mining in one grand, unified graphical model—a unified probabilistic model, with a unified representation of data and outcomes, a unified set of parameters, unified inference procedures, and unified learning procedures.

Seen in this light, information extraction and data mining are not separate processes, but a single collective whole. No hard, brittle decisions need to be made at one stage of a pipeline in order to be passed to the next stage—the subtlest and most uncertain of hypotheses can be communicated back and forth between extraction and data mining, each helping the other converge to an agreed upon conclusion.

For example, consider the following scenario. Word-level features alone might leave ambiguous whether an appearance of the word "Tracy" on a university Web page is a person name or a project name. An appearance of "Beth Smith" on the same page might more certainly be hypothesized to be a person name. Through initial coreference analysis, we might find Beth Smith's home page, and her relations to some other people. These patterns of relations (in combination with the words on her home page) might cause the model to decide that Beth Smith is likely a professor. Knowing this might help provide just enough additional evidence to the extraction model running in the context of the original page that it is able to hypothesize a Principal-Investigator-Of relation

between "Beth Smith" and "Tracy". Since the data mining model parameters indicate that this relation only occurs between a person and a project, it can be correctly deduced that the word "Tracy" must be a project name here, not a person name. And furthermore an appearance of the person name "Tracy Jones" on a different Web page can correctly be said not to be co-referent with the project "Tracy" on the first page. All of these constraints are communicated in subtle shades of probability that work themselves out through the statistically principled methods of inference.

3.4 Conditional Random Fields

In this section we define conditional random fields and describe how they may be used to create unified models for information extraction and data mining—illustrating this framework with several specific examples.

Conditional Random Fields [Lafferty *et al.*, 2001] are undirected graphical models (also known as *random fields*) used to calculate the conditional probability of values on designated output variables given values assigned to other designated input variables.[2]

Let $\mathbf{X}$ be a set of input random variables, and $\mathbf{Y}$ be a set of output random variables. Then, by the fundamental theorem of random fields [Hammersley and Clifford, 1971], a conditional random field defines the conditional probability of values $\mathbf{y}$ given values $\mathbf{x}$ to be a product of potential functions on cliques of the graph,

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_{\mathbf{x}}} \prod_{c \in \mathcal{C}} \Phi_c(\mathbf{x}_c, \mathbf{y}_c),$$

where $Z_{\mathbf{x}} = \sum_{\mathbf{y}'} \prod_{c \in \mathcal{C}} \Phi_c(\mathbf{x}_c, \mathbf{y}_c)$ is the partition function (normalizer), $\mathcal{C}$ is the set of all cliques, $\Phi_c(\cdot)$ is the potential function for clique c, $\mathbf{x}_c$ is that sub-set of the variables in $\mathbf{x}$ that participate in clique c, and $\mathbf{y}_c$ is defined analogously. We calculate the potential functions as a log-linear combination of weighted features, $\Phi_c(\mathbf{x}_c, \mathbf{y}_c) = \exp(\sum_k \lambda_{kc} f_{kc}(\mathbf{x}_c, \mathbf{y}_c))$, where $f_{kc}(s_{t-1}, s_t, \emptyset, t)$ is an arbitrary feature function over its arguments, and λ_{kc} is a learned weight for each feature function.

Linear Chain

In the special case in which the designated output nodes of the graphical model are linked only by edges in a *linear chain*, CRFs make a first-order Markov independence assumption among output nodes, and thus correspond to finite state machines (FSMs), which have been shown to be suitable sequence models for information extraction, *e.g.* [Bikel *et al.*, 1997; McCallum and Li, 2003].

Let $\mathbf{x} = \langle x_1, x_2, ...x_T \rangle$ be some observed input data sequence, such as a sequence of words text in a document, (the values on n input nodes of the graphical model). Let $\mathcal{S}$ be a set of FSM states, each of which is associated with a label, $l \in \mathcal{L}$, (such as a label PERSON). Let $\mathbf{y} = \langle y_1, y_2, ...y_T \rangle$ be some sequence of states, (the values on T output nodes).

---

[2]The term "random field" has common usage in the statistical physics and computer vision communities. In statistics the same models are also known as "Markov networks." Thus *Conditional Markov Networks* [Taskar *et al.*, 2002] are identical to Conditional Random Fields.

CRFs define the conditional probability of a state sequence given an input sequence as

$$P_\Lambda(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_\mathbf{x}} \exp \left(\sum_{t=1}^{T} \sum_k \lambda_k f_k(y_{t-1}, y_t, \mathbf{x}, t) \right).$$

This model ties parameters $\Lambda = \{\lambda, ...\}$ across sequence positions, but this is just one possible type of tying. Various patterns of parameter tying may be based on arbitrary SQL-like queries [Taskar *et al.*, 2002]. Several specific patterns relevant to unification of extraction and data mining are described below. Many others in this framework are also possible.

Cross-referenced Linear Chain
The previous model captures dependencies between adjacent pairs of labels, but in some cases we may have reason to believe that other, arbitrarily-separated words have dependent labels. For example, capturing the fact that two identical capitalized words in the same document often should share the same label will help us know that "Green" is a last name when we have seen the phrase "David Green" elsewhere in the document. Such dependencies among selected pairs, $\mathcal{P}$, of arbitrarily-separated words can be represented with a *cross-referenced linear chain*,

$$P_\Lambda(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_\mathbf{x}} \exp \left(\sum_{t=1}^{T} \sum_k \lambda_k f_k(y_{t-1}, y_t, \mathbf{x}, t) \right.$$
$$\left. + \sum_{\langle t,t' \rangle \in \mathcal{P}} \sum_{k'} \lambda_{k'} f_{k'}(y_t, y_{t'}, \mathbf{x}, t) \right).$$

Note that the edges among the output variables now form loops, and inference is more difficult than before. Approximate inference methods are discussed below.

Factorial Linear Chain
When there are multiple dimensions of labels to be predicted—for example part-of-speech, phrase boundaries, named entities, and the classification of entities into categories (such as STUDENT and PROFESSOR)—these multi-dimensional labels can be simultaneously predicted and efficiently represented in a factorial model. Ghahramani and Jordan [1995] describe a factorial HMM. Factorial CRFs are detailed in [Rohanimanesh and McCallum, 2003], and define the probability of two label sequence factors, $\mathbf{y}$ and $\mathbf{y}'$, connected in a grid as

$$P_\Lambda(\mathbf{y}, \mathbf{y}'|\mathbf{x}) = \frac{1}{Z_\mathbf{x}} \exp \left(\sum_{t=1}^{T} \sum_k \lambda_k f_k(y_{t-1}, y_t, \mathbf{x}, t) \right.$$
$$+ \sum_{t=1}^{T} \sum_{k'} \lambda_{k'} f_{k'}(y'_{t-1}, y'_t, \mathbf{x}, t)$$
$$\left. + \sum_{t=1}^{T} \sum_{k''} \lambda_{k''} f_{k''}(y_t, y'_t, \mathbf{x}, t) \right).$$

Affinity- or Relationship-Matrix
When predicting entity coreference (or other types of relationships), rather than a sequence of labels, let the output be a matrix $\mathbf{w} = \{w_{11}, w_{12}, ...w_{tt'}, ...w_{TT}\}$ of labels on pairs of words (or entities), and forming a matrix of coreference decisions or other binary relationships. We define the distribution,

$$P_\Lambda(\mathbf{w}|\mathbf{x}) = \frac{1}{Z_\mathbf{x}} \exp \left(\sum_{t,t'} \sum_{k'} \lambda_{k'} f_{k'}(w_{tt'}, \mathbf{x}, t, t') \right.$$
$$\left. + \sum_{t,t',t''} \lambda_* f_*(w_{tt'}, w_{t't''}, w_{tt''}) \right).$$

This model, in which inference corresponds to graph partitioning, is further described in McCallum and Wellner [2003], where the need for dependencies among the w's in the second sum is also explained. Another variant described there also predicts attributes associated with entities. The matrix can be made sparse by approximation with *Canopies* [McCallum *et al.*, 2000c].

Factorial Chain and Relationship-Matrix
Entity extraction, classification of entities, coreference, and determination of other relationships among entities can all be performed simultaneously by a factorial model over chains and matrices. This is a model that could solve the "Tracy" problem described above. The equation for $P(\mathbf{y}, \mathbf{y}', \mathbf{w}|\mathbf{x})$ (which we omit to save space) includes a straightforward combination of the sums from the previous two models, plus additional desired dependencies among output variables. Other promising variations include the integration of hierarchical models corresponding to parse trees.

Inference and Parameter Estimation
Given an inference procedure, parameter estimation in all these models can be performed with standard optimization procedures such as conjugate gradient or approximate quasi-Newton methods [Malouf, 2002; Sha and Pereira, 2003b]. Inference for Linear Chain models can be performed efficiently with dynamic programming [Lafferty *et al.*, 2001]. The other models have loops among output variables, and thus we must resort to approximate inference. Approximate inference in the Affinity-Matrix models can be performed by randomized graph partitioning algorithms, as described in [McCallum and Wellner, 2003]. We have had considerable success performing inference in the Factorial Linear Chain [Rohanimanesh and McCallum, 2003] with Tree-based Reparameterization [Jaakkola *et al.*, 2001]. Improved methods of efficient approximate inference in these models remains an open area for research. Feature induction (which also corresponds to graphical structure induction for these models) is described in McCallum [2003].

4 Related Work

There has been a large amount of previous separate work on information extraction and data mining, some of which has been referenced and described previously in this paper.

4.1 Relational extraction and data mining

There is also a new and growing body of work in extraction of *relational* data, as well as separate work in data mining of *relational* data. In extraction, the association of entities into relations has traditionally been performed by classification of entity pairs independently from each other. For example, noun coreference can be decided by the output of a binary maximum entropy classifier indicating whether or not the two nouns in the pair are co-referent [Morton, 1997]. The binary classifiers can also be quite sophisticated, for example using SVMs with complex kernels [Zelenko *et al.*, 2003].

However, these methods perform entity extraction completely independently from association (causing errors to compound), and also make coreference and relation-formation decisions independently from each other (allowing decisions to be inconsistent and errorful). For example, one classifier might decide that "Mr. Smith" is co-referent with "Smith," and another classifier might incompatibly decide that this "Smith" is co-referent with "she." An alternative approach is to extract and build relations in a single augmented finite state machine [Ray and Craven, 2001] or parsing model [Miller *et al.*, 2000], however this only operates over relations formed within one sentence. Other work [Roth and Yih, 2002] recognizes and models the dependencies across multiple entity classifications and relations; however it relies on entity extraction having already been performed. Recent work in coreference analysis also explicitly models the dependencies among coreference decisions on multiple pairs of pre-extracted entities [Pasula *et al.*, 2002; McCallum and Wellner, 2003].

As described in section 3.2, there has been a recent surge of research on relational data mining. Particularly notable is work based on undirected graphical models [Taskar *et al.*, 2002], (and also indirectly [Neville and Jensen, 2000]). The former involves experiments on data mining of academic entities, although it does so through Web page and hyperlink classification, not through full information extraction (which would involve extracting multiple sub-segments of text on a page, and more difficult coreference and relation-building analysis).

4.2 Early work in integration of extraction and data mining

There has still been relatively little work on integration between extraction and data mining. Most current work is better characterized as serial juxtaposition, (e.g. [Ghani *et al.*, 2000]), or mining raw text data (such as documents, web sites, hyperlinks, or web logs), (e.g. [Hearst, 1999; Craven *et al.*, 1998; Taskar *et al.*, 2002; Kosala and Blockeel, 2000; Anderson *et al.*, 2002]), but not mining a rich database resulting from information extraction, (that is, sub-segments of text on a page, each referring to different entities—which is significantly more difficult).

One interestingly different approach does not aim to extract a correct database, but instead attempts to data mine a "soft database" consisting of the raw text of each mention, (without any coreference analysis having been performed, and perhaps with extraction boundary errors) [Cohen and Hirsh, 1998]. New database operations, such as "soft joins" may merge records based on TF-IDF similarity instead of exact matches—doing so on the fly in response to a particular query. This approach is intriguing, but it seems only to delay the inevitable difficulties. Much noise and error remains in these soft joins, and this approach could not support complex relational data mining.

Some of the most truly integrated work in extraction and data mining has been done by Ray Mooney's group at UT Austin. For example, in one project [Nahm and Mooney, 2000], twelve fields of data are extracted from USENET computer-related job ads using a rule learner. The fields include programming-language, hardware-platform, application-area, etc. A second rule learner is applied to an imperfectly-extracted database to produce rules that will predict the value in each field given the others. Then these rules are used to fill in missing values and correct errors in extraction—a very nice example of "closing (one turn of) the loop." This work is a promising first beginning; there remains much additional work to do, especially in the use of stronger statistical machine learning methods, such as graphical models, that have provided world-class performance in other independent extraction and data mining problems. This is the approach we put forward in this paper.

5 Conclusions

We have presented motivation, problems and proposed solutions for a unified framework of extraction and data mining using conditionally-trained undirected graphical models. This approach addresses the three critical topics of integrating extraction and data mining:

Uncertainty management — The hypotheses of both extraction and data mining are represented in probability distributions on nodes of the graphical model. For example, in extraction sections of the model, a node might represent an individual word, and contain a probability distribution over the entity labels person, project, university, other, etc. In the data mining sections of the model, a node might represent a relation between two entities, and contain a probability distribution over the labels principal-investigator-of, adviser-of, project-colleague-of, etc.

With both extraction and data mining embedded in the same model, intermediate hypothesis are naturally communicated back and forth in the language of probabilities. Rather than being a problem, uncertainty becomes an opportunity—with the ability for the intermediate hypotheses of data mining to improve extraction, and vice-versa.

Inferential Feedback — Closed-loop feedback between extraction and data mining is a natural outcome of inference in the unified graphical model.

Note that there has been some previous work on feeding extracted data into data mining (see section 4), and performing inference on this noisy data. However, we are proposing models that actually "close the loop" by feeding results of data mining back into extraction, and looping back to data mining repeatedly. This closed-loop, bi-directional communication will allow subtle constraints to flow both directions, let sharper conclusions be formed by the agglomeration of multiple pieces of uncertain evidence, and help turn the communication of uncertainty into an advantage, not a disadvantage.

Relational Data — Relational data are straightforwardly modeled in undirected graphical models by using tied parameters in patterns that reflect the nature of the relation. Patterns of tied parameters are common in many graphical models, including finite state machines [McCallum *et al.*, 2000a; Lafferty *et al.*, 2001], where they are tied across different sequence indices, and by more complex patterns, as in Taskar *et al.* [2002]. Tied parameters use for extraction, classification, coreference and other relationships are described in section 3.4.

Acknowledgments

References

[Agrawal *et al.*, 1993] Rakesh Agrawal, Tomasz Imielinski, and Arun N. Swami. Mining association rules between sets of items in large databases. In Peter Buneman and Sushil Jajodia, editors, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, Washington, D.C., 26–28 1993.

[Anderson *et al.*, 2002] C. Anderson, P. Domingos, and D. Weld. Relational markov models and their application to adaptive web navigation. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002)*. ACM Press, 2002.

[Appelt *et al.*, 1995] D. E. Appelt, Jerry R. Hobbs, J. Bear, D. Israel, M. Kameyama, A. Kehler, D. Martin, K. Myers, and M. Tyson. *Proceedings of the Sixth Message Understanding Conference (MUC-6)*. Morgan Kaufmann, 1995.

[Bikel *et al.*, 1997] Daniel M. Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. Nymble: a high-performance learning name-finder. In *Proceedings of ANLP-97*, pages 194–201, 1997.

[Bilenko and Mooney, 2002] Mikhail Bilenko and Raymond J. Mooney. Learning to combine trained distance metrics for duplicate detection in databases. Technical Report AI 02-296, Artificial Intelligence Lab, University of Texas at Austin, February 2002.

[Blei *et al.*, 2002] David Blei, Drew Bagnell, and Andrew McCallum. Learning with scope, with application to information extraction and classification. In *Uncertainty in Artificial Intelligence (UAI)*, 2002.

[Borthwick *et al.*, 2000] Andrew Borthwick, Vikki Papadouka, and Deborah Walker. The MEDD de-duplication project. In *Immunization Registry Conference*, 2000.

[Breiman *et al.*, 1984] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Chapman & Hall, New York, 1984.

[Breiman, 1996] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[Bruninghaus and Ashley, 2001] Stefanie Bruninghaus and Kevin D. Ashley. Improving the representation of legal case texts with information extraction methods. In *Proceedings of the 8th international conference on Artificial Intelligence and Law*, 2001.

[Carreras *et al.*, 2002] Xavier Carreras, Lluis Marques, and Lluis Padro. Named entity extraction using adaboost. In *Proceedings of CoNLL-2002*, pages 167–170, 2002.

[Cohen and Hirsh, 1998] William W. Cohen and Haym Hirsh. Joins that generalize: text classification using WHIRL. In Rakesh Agrawal, Paul E. Stolorz, and Gregory Piatetsky-Shapiro, editors, *Proceedings of KDD-98, 4th International Conference on Knowledge Discovery and Data Mining*, pages 169–173, New York, US, 1998. AAAI Press, Menlo Park, US.

[Collins, 2002] Michael Collins. Ranking algorithms for named-entity extraction: Boosting and the voted perceptron. In *ACL-02*, 2002.

[Craven *et al.*, 1998] M Craven, D DiPasquo, D Freitag, A McCallum, T Mitchell, K Nigam, and S Slattery. Learning to extract symbolic knowledge from the World Wide Web. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)*, pages 509–516, 1998.

[DARPA, 2002] DARPA. Darpa automatic content extraction program, 2002.

[Domingos and Richardson, 2001] P. Domingos and Matt Richardson. Mining the network value of customers. In *Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining*, pages 57–66, 2001.

[Fawcett and Provost, 1997] Tom Fawcett and Foster J. Provost. Adaptive fraud detection. *Data Mining and Knowledge Discovery*, 1(3):291–316, 1997.

[Freitag and McCallum, 1999] Dayne Freitag and Andrew Kachites McCallum. Information extraction with hmms and shrinkage. In *Proceedings of the AAAI-99 Workshop on Machine Learning for Informatino Extraction*, 1999.

[Getoor *et al.*, 2001] L. Getoor, N. Friedman, D. Koller, and A. Pfeffer. Learning probabilistic relational models. In S. Dzeroski and N. Lavrac, editors, *Relational Data Mining*. Springer-Verlag, 2001.

[Ghahramani and Jordan, 1995] Zoubin Ghahramani and Michael I. Jordan. Factorial hidden Markov models. In David S. Touretzky, Michael C. Mozer, and Michael E. Hasselmo, editors, *Proc. Conf. Advances in Neural Information Processing Systems, NIPS*, volume 8, pages 472–478. MIT Press, 1995.

[Ghani *et al.*, 2000] Rayid Ghani, Rosie Jones, Dunja Mladenic, Kamal Nigam, and Sean Slattery. Data mining on symbolic knowledge extracted from the web. In *KDD-2000 Workshop on Text Mining*, 2000.

[Hammersley and Clifford, 1971] J. Hammersley and P. Clifford. Markov fields on finite graphs and lattices. Unpublished manuscript, 1971.

[Hearst, 1999] Marti Hearst. Untangling text data mining. In *Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics*, 1999.

[Jaakkola *et al.*, 2001] T. Jaakkola, M. Wainwright, and A. Willsky. Tree-based reparameterization for approximate estimation on graphs with cycles. In *Neural Information Processing Systems (NIPS)*, 2001.

[Jensen and Neville, 2002] D. Jensen and J. Neville. Linkage and autocorrelation cause feature selection bias in relational learning. In *Proceedings of the Nineteenth International Conference on Machine Learning (ICML2002)*, pages 259–266. Morgan Kaufmann, 2002.

[Jensen and Neville, 2003] D. Jensen and J. Neville. Randomization tests for relational learning. In *International Joint Conference on Artificial Intelligence (submitted)*, 2003.

[Jensen *et al.*, 2003a] D. Jensen, J. Neville, and M. Hay. Degree disparity leads to aggregation bias in relational models. In *International Conference on Machine Learning (submitted)*, 2003.

[Jensen *et al.*, 2003b] D. Jensen, M. Rattigan, and H. Blau. Misclassification errors and collective inference in relational data. In *Conference on Knowledge Discovery and Data Mining (submitted)*, 2003.

[Jensen, 1999] D. Jensen. Statistical challenges to inductive inference in linked data. In *Papers of the 7th International Workshop on Artificial Intelligence and Statistics*, 1999.

[Jerry *et al.*, 1996] H. Jerry, R. Douglas, E. Appelt, J. Bear, D. Israel, M. Kameyama, M. Stickel, and M. Tyson. Fastus: A cascaded finite-state transducer for extracting information from natural-language text, 1996.

[Jordan, 1998] M. Jordan, editor. *Learning in Graphical Models*. MIT Press, Cambridge, 1998.

[Klein *et al.*, 2003] Dan Klein, Joseph Smarr, Huy Nguyen, and Christopher Manning. Named entity recognition with character-level models. In *Proceedings the Seventh Conference on Natural Language Learning*, 2003.

[Kosala and Blockeel, 2000] Kosala and Blockeel. Web mining research: A survey. *SIGKDD: SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining, ACM*, 2, 2000.

[Lafferty *et al.*, 2001] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. ICML*, pages 282–289, 2001.

[Lawrence *et al.*, 1999] Steve Lawrence, C. Lee Giles, and Kurt Bollacker. Digital libraries and Autonomous Citation Indexing. *IEEE Computer*, 32(6):67–71, 1999.

[Leek, 1997] Timothy R. Leek. Information extraction using hidden Markov models. Master's thesis, UC San Diego, 1997.

[Malouf, 2002] Robert Malouf. A comparison of algorithms for maximum entropy parameter estimation. In *Sixth Workshop on Computational Language Learning (CoNLL-2002)*, 2002.

[McCallum and Li, 2003] Andrew McCallum and Wei Li. Early results for named entity extraction with conditional random fields, feature induction and web-enhanced lexicons. In *Seventh Conference on Natural Language Learning (CoNLL)*, 2003.

[McCallum and Wellner, 2003] Andrew McCallum and Ben Wellner. Toward conditional models of identity uncertainty with application to proper noun coreference. In *IJCAI Workshop on Information Integration on the Web*, 2003.

[McCallum *et al.*, 2000a] Andrew McCallum, Dayne Freitag, and Fernando Pereira. Maximum entropy Markov models for information extraction and segmentation. In *Proceedings of ICML*, pages 591–598, 2000.

[McCallum *et al.*, 2000b] Andrew McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Automating the contruction of internet portals with machine learning. *Information Retrieval Journal*, 3:127–163, 2000.

[McCallum *et al.*, 2000c] Andrew McCallum, Kamal Nigam, and Lyle H. Ungar. Efficient clustering of high-dimensional data sets with application to reference matching. In *Knowledge Discovery and Data Mining*, pages 169–178, 2000.

[McCallum, 2002] Andrew McCallum. Personal experience at WhizBang Labs, Inc.

[McCallum, 2003] Andrew McCallum. Efficiently inducing features of conditional random fields. In *Nineteenth Conference on Uncertainty in Artificial Intelligence (UAI03)*, 2003.

[McCullagh and Nelder, 1989] P. McCullagh and J. Nelder. *Generalized Linear Models*. Chapman & Hall, New York, 1989.

[Miller *et al.*, 2000] Scott Miller, Heidi Fox, Lance Ramshaw, and Ralph Weischedel. A novel use of statistical parsing to extract information from text. In *6th Applied Natural Language Processing Conference*, 2000.

[Mitchell, 1997] T. Mitchell. *Machine Learning*. McGraw Hill, 1997.

[Morton, 1997] Thomas Morton. Coreference for NLP applications. In *Proceedings ACL*, 1997.

[Nahm and Mooney, 2000] Un Yong Nahm and Raymond J. Mooney. A mutually beneficial integration of data mining and information extraction. In *AAAI/IAAI*, pages 627–632, 2000.

[Neville and Jensen, 2000] J. Neville and D. Jensen. Iterative classification in relational data. In L. Getoor and D. Jensen, editors, *Learning Statistical Models From Relational Data: Papers from the AAAI Workshop*, pages 42–49, Meno Park, CA, 2000. AAAI Press.

[Pasula *et al.*, 2002] Hanna Pasula, Bhaskara Marthi, Brian Milch, Stuart Russell, and Ilya Shpitser. Identity uncertainty and citation matching. In *Advances in Neural Information Processing (NIPS)*, 2002.

[Pinto *et al.*, 2003] David Pinto, Andrew McCallum, Xen Lee, and W. Bruce Croft. Combining classifiers in text categorization. In *Submitted to SIGIR '03: Proceedings of the Twenty-sixth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003.

[Quinlan, 1993] R. Quinlan. *C4.5: Programs for machine learning*. Morgan Kaufmann, San Mateo, CA, 1993.

[Ray and Craven, 2001] S. Ray and M. Craven. Representing sentence structure in hidden markov models for information extraction. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence*. Morgan Kaufmann, 2001.

[Rohanimanesh and McCallum, 2003] Khashayar Rohanimanesh and Andrew McCallum. Factorial conditional random fields. Technical report, Department of Computer Science, University of Massachusetts, 2003.

[Roth and Yih, 2002] Dan Roth and Wen-tau Yih. Probabilistic reasoning for entity and relation recognition. In *COLING'02*, 2002.

[Schapire, 1999] Robert E. Schapire. Theoretical views of boosting. *Lecture Notes in Computer Science*, 1572:1–10, 1999.

[Sha and Pereira, 2003a] Fei Sha and Fernando Pereira. Shallow parsing with conditional random fields. Technical Report CIS TR MS-CIS-02-35, University of Pennsylvania, 2003.

[Sha and Pereira, 2003b] Fei Sha and Fernando Pereira. Shallow parsing with conditional random fields. In *Proceedings of Human Language Technology, NAACL*, 2003.

[Soderland, 1997] S. Soderland. Learning to extract text-based information from the world wide web. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD-97)*, 1997.

[Taskar *et al.*, 2002] B. Taskar, P. Abbeel, and D. Koller. Discriminative probabilistic models for relational data. In *Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI02)*, 2002.

[Wang, 1999] Xue Z. Wang. *Data Mining and Knowledge Discovery for Process Monitoring and Control*. Springer Verlag, 1999.

[Zelenko *et al.*, 2003] Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. Kernel methods for relation extraction. *Journal of Machine Learning Research (submitted)*, 2003.