|  |  |  |
| --- | --- | --- |
| UAI 2004 | WELLNER ET AL. | 593 |

An Integrated, Conditional Model of

Information Extraction and Coreference

with Application to Citation Matching

Ben Wellner∗†, Andrew McCallum∗, Fuchun Peng∗, Michael Hay∗

∗University of Massachusetts Amherst   
Amherst, MA 01003 USA   
{wellner, mccallum, fuchun, mhay}@cs.umass.edu

†The MITRE Corporation 202 Burlington Road   
Bedford, MA 01730 USA

|  |  |  |
| --- | --- | --- |
| Abstract | “top down” from data mining. | Thus (a) intermedi- |

ate hypotheses from both extraction and data mining

Although information extraction and coref-erence resolution appear together in many applications, most current systems perform them as independent steps. This paper describes an approach to integrated infer-ence for extraction and coreference based on conditionally-trained undirected graphical models. We discuss the advantages of condi-tional probability training, and of a corefer-ence model structure based on graph parti-tioning. On a data set of research paper cita-tions, we show significant reduction in error by using extraction uncertainty to improve coreference citation matching accuracy, and using coreference to improve the accuracy of the extracted fields.

can be easily communicated between extraction and data mining in a closed loop system, (b) mutually-reinforcing evidence from multiple sources will have the opportunity to be properly marshaled, (c) and ac-curacy and confidence assessment should improve.

In particular, we advocate creating these joint models as conditional random fields (CRFs) (Lafferty et al., 2001) that have been configured to represent rela-tional data by using parameter tying in repeated pat-terns based on the structure of the data—also known as relational Markov networks (Taskar et al., 2002). In natural language processing, conditionally-trained rather than generatively-trained models almost al-ways perform better because they allow more free-dom to include a large collection of arbitrary, overlap-ping and non-independent features of the input with-out the need to explicitly represent their dependen-cies, e.g., (Lafferty et al., 2001; Carreras et al., 2002;

|  |  |  |
| --- | --- | --- |
| 1 | Introduction | Pinto et al., 2003). In relational modeling, undirected |
| graphical models allow greater freedom to represent |

Although information extraction (IE) and data mining appear together in many applications, their interface in most current systems would be better described as loose serial juxtaposition than as tight integration. In-formation extraction populates slots in a database by identifying relevant subsequences of text, but is usu-ally unaware of the emerging patterns and regularities in the database. Data mining begins from a populated database, and is often unaware of where the data came from, or their inherent uncertainties. The result is that the accuracy of both suffers, and significant mining of complex text sources is beyond reach.

To address this problem we have previously advocated (McCallum & Jensen, 2003) the use of joint probabilis-tic models that perform extraction and data mining in an integrated inference procedure—the evidence for an outcome being the result of simultaneously mak-ing inferences both “bottom up” from extraction, and

auto-correlation and other relations without concern for avoiding circularities (Taskar et al., 2002; Nevil-lle et al., 2004). Both these modeling choices are in contrast to other related work in using directed, generatively-trained probabilistic models for informa-tion extraction (Marthi et al., 2003).

This paper presents a model, inference and learning procedure for a preliminary case of this extraction and data mining integration—namely information extrac-tion and coreference on research paper citations.1Ex-traction in this context consists of segmenting and la-beling the various fields of a citation, including title, author, journal, year, etc. Coreference (also known as identity uncertainty, record linkage or object consol-

1We currently avoid calling this work a joint model of extraction and coreference, because we have not yet “closed the loop” by repeatedly cycling between extraction and coreference inference.

|  |  |  |
| --- | --- | --- |
| 594 | WELLNER ET AL. | UAI 2004 |

idation) is a key problem in creating databases cre-ated from noisy data. For example, without properly resolving “Stuart Russell,” “S. Russell,” and “Stuart Russel” to the same entity in a database, relational connections will be missing, and subsequent data min-ing will not find the patterns it should.

Using a data set of citations from CiteSeer (Lawrence et al., 1999), we present experimental results indi-cating that the type of integration we advocate does indeed hold promise—we show that modeling uncer-tainty about extraction of citation fields can improve coreference (in the face of different field orderings, ab-breviations and typographic errors), and that leverag-ing predicted coreference can improve the extraction accuracy of these fields. Measurements of best-case scenarios show that there is yet further gain available to be found through this integration. Certainly fur-ther gains are expected from experiments that close the loop between extraction and coreference, rather than the limited, separate bi-directional results pro-

within the entity attributes is performed exactly by exhaustive search. Across the three sub-structures, ap-proximate inference is accomplished by variants of it-erated conditional modes (ICM) (Besag, 1986). More precisely, approximate inference in the entire model proceeds as follows: (1) for each citation, N segmenta-tions with highest probability (N-best lists) are found by a variant of Viterbi and provided to coreference; (2) coreference decisions are inferred by approximate graph partitioning, integrating out uncertainty about the sampled N segmentations; (3) these coreference decisions are used to infer the attributes of each paper by searching over all combinations of values suggested by each citation’s segmentations; and finally (4) in-ference of citation segmentations are revised to make themselves more compatible with their corresponding entity attributes.

Joint parameter estimation in this complex model is intractable, and thus, as in inference, we perform pa-rameter estimation somewhat separately for each of

|  |  |  |
| --- | --- | --- |
| vided here. | the three sub-structures. | In all cases, estimation is |

Building on earlier work in coreference that assumes perfect extraction (McCallum & Wellner, 2003), we cast coreference as a problem in graph partitioning based on Markov random fields (Boykov et al., 1999; Bansal et al., 2002). The graphical model has cliques for pairs of citations, the log-clique-potential of which is the edge weight in the graph to be partitioned. These edge weights may be positive or negative, and thus the optimal number of partitions (equivalent to number of cited papers in this case) falls out naturally from the optimization function of the max-flow-min-cut partitioning. Later in this paper we provide statis-tical correlation results indicating that the redundancy in this “fully-connected graph partitioning” approach to coreference is more robust than a graphical model in which a “prototype yields observations.”

In the model introduced in this paper, the graphical model consists of three repeated sub-structures: (1) a linear-chain representing a finite-state segmenter for the sequence of words in each citation (2) a boolean variable in a clique between each pair of segmented ci-tations, representing graph-partitioning-style citation coreference decisions, (3) a collection of attribute vari-ables once for each paper entity (that is, one for each partition in the coreference graph), noting that the

iterative, consisting of BFGS quasi-Newton steps on a maximum a posteriori conditional likelihood, with a zero-mean spherical-variance Gaussian prior on the parameters. The parameters of the linear-chain are set to maximize the conditional likelihood of the correct label sequence, in the traditional fashion for linear-chain CRFs. The parameters for the distance function in graph partitioning are set to maximize the product of independent conditional likelihoods for each pair-wise coreference decision. The parameters for the en-tity attributes are set by pseudolikelihood to maximize the likelihood of correct placement of edges between highest-accuracy citation segmentations and their true entity attributes.

We present experimental results on the four sections of CiteSeer citation-matching data (Lawrence et al., 1999). Using our integrated model, both extraction and coreference show significant reductions in error—by 25-35% for coreference and by 6-14% for extraction. We also provide some encouraging best-case experi-ments showing substantial additional potential gain that may come from more integrated joint inference and creation of additional features that leverage the capabilities of conditional probability models.

|  |  |  |
| --- | --- | --- |
| number of these entity sub-structures is determined at | 2 | Model |
| inference time. Thus, this model is a special case of |

Model 1 in McCallum and Wellner (2003).

Inference within the linear chain is performed exactly by dynamic programming; inference within the fully-connected coreference is performed approximately by a simple graph partitioning algorithm, and inference

This paper presents a method for integrated informa-tion extraction and coreference based on conditionally-trained, undirected graphical models—also known as conditional random fields (Lafferty et al., 2001). The model predicts entities and their attributes condi-tioned on observed text.

UAI 2004 WELLNER ET AL. 595

The model contains three types of repeated sub- indicating the membership of a citation word to a field

structures with tied parameters. These three sub- (such as author, title or year). As a convenience, we

structures are responsible for (1) information extrac-tion, in the form of segmentation and labeling of word sequences in order to find (the fields of) each mention of an entity, (2) coreference among the mentions to dis-cover when two mentions are referring to the same un-derlying entity, (3) representing the attributes of each entity and the dependencies among those attribute val-ues. The attributes of each entity correspond to the canonical values that could be entered into database record fields, and the dependencies allow the model to represent expectations about what combinations of attributes would be expected in the world.

In general, conditional random fields (CRFs) are undi-rected graphical models that encode a conditional probability distribution using a given set of features. CRFs are defined as follows. Let G be an undirected

also define citation fields c = {c1, ...cK}, where ci is a collection of variables containing the complete string value for each of the various fields of xi, determinis-tically agglomerated from the label sequence si. Let y = {y1,2, ..yi,j, ...yK−1,K}, i < j be a set of boolean coreference variables indicating whether or not cita-tion xi is referring to the same paper as citation xj; (note y here is more specific than in Eq 1). Finally, let a = {a1, ...aM} be the set of attributes of each paper (“entities”), where M is the number of underlying re-search paper entities. Here, entity attributes are field values, such as title and year, but canonicalized from their noisy appearance in the multiple coreference ci-tation mentions.

As described above, the model consists of three re-peated sub-structures: (1) a linear-chain on the ele-

|  |  |  |
| --- | --- | --- |
| model over sets of random variables y and x. | If | ments of the si sequences conditioned on the xi se- |

C = {{yc, xc}} is the set of cliques in G, then CRFs define the conditional probability of an output label-ing, y given the observed variables, x as:

quences, for finite-state citation segmentation and la-beling (information extraction); (2) a fully-connected graph on the xi’s, with the binary coreference deci-

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| pθ(y|x) = | 1 | c∈C | Φ(yc, xc), | (1) | sion on the mention pair (xi, xj) indicated by yij; note |
| also that formally the graphical model requires po- |
| Z(x) |
| tentials over all triples (yi, yj, yk) in order to enforce |

transitivity, but these potentials never actually have to

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| where | Φ | is | a | potential | function | and | Z(x) | = | be instantiated in inference by graph partitioning; (3) |

y c∈CΦ(yc, xc) is a normalization factor. We as-sume the potentials factorize according to a set of

features {fk}, which are given and fixed, so that

Φ(yc, xc) = exp( kλkfk(yc, xc)). The model param-eters are a set of real-valued weights Λ = {λk}, one

potentials measuring the compatibility between each mention’s segmentation ci and the attributes of its cor-responding entity ak. This model is closely related to a combination of Models 1 and 2 in McCallum and Wellner (2003), where further details and background

|  |  |  |
| --- | --- | --- |
| weight for each feature. | may be found. | Inference in the model aims to find |

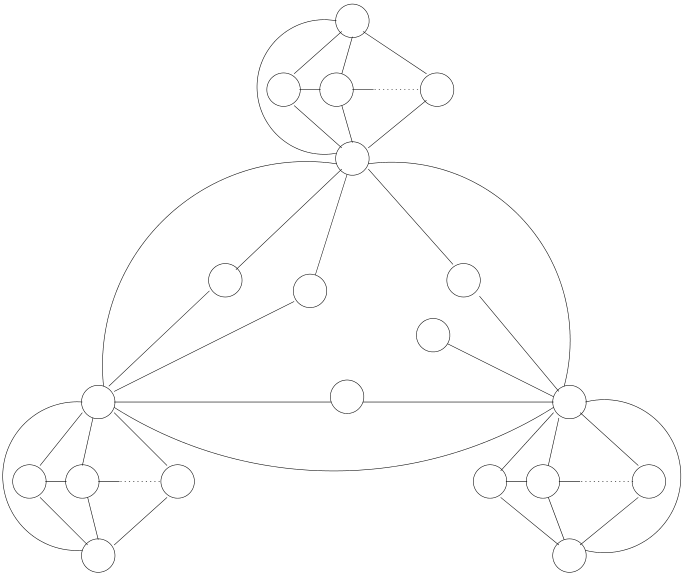
CRFs have shown recent success in a number of do-mains especially in sequence modeling for natural lan-guage tasks (Lafferty et al., 2001; Sha & Pereira, 2003; Pinto et al., 2003; McCallum & Li, 2003; Sutton et al., 2004), often outperforming their generative counter-

the mode of P(a|x) = c,s,yP(a, y, c, s|x), where the summed term is defined in Equation 2 by a product of potentials on cliques of the graphical model. Figure 1 shows an example graphical model with two coreferent citations and a singleton.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| parts. Their strength lies primarily in their ability to accommodate multiple, overlapping, non-independent features. By training the model to maximize the condi-tional probability of the output labels given the input values, CRFs avoid having to generate the observed variables or model their dependencies. | | P(a, y, c, s|x) = 1 Zx | | | |  | | K | | | φ0(yij, yik, yjk)   | |
| i=1,j>i,k>j | | | | |
|  | | M | K | φ1(ai, cj) | | | | K | | φ2(cj, ck, yjk)   |
| i=1 | | | j=1 | k>j | |
| 2.1 | A CRF for Citation Extraction and |  | | | | | K | |si| | φ3(si(t−1), sit, ci, xi)  | | | |
| Coreference  We now describe in detail a CRF for the integrated task of extracting fields and performing coreference among research paper citations. Let x = {x1, ...xK} | | i=1 | | | | | | t=1 |
| (2) | | | | | | | | | | |
| 2.2 | Inference | | | | | | | | | |

be a set of observed citations (“mentions”), where each xi = (xi1, xi2, ...) is a sequence of words forming the text of the citation. Let s = {s1, ...sK} be the corre-sponding set of label sequences, each label sequence, si

Exact inference in this model is clearly intractable. We have, however, some clearly defined sub-structures within the model, and there is considerable previous work on inference by structured approximations—that



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 596 | s | x | Observed Citation | WELLNER ET AL. | UAI 2004 |
| tegrating out the uncertainty in segmentation”, that | |
| c | Segmentation |
| is, summing over all combinations of sampled label se- | |
| quences, si and their corresponding citation fields ci. | |
| Citation Fields |
| (Note that is is exactly where uncertainty in extrac- | |

tion is integrated into coreference.) Thus, the graph-

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| y | a | Entity  Attributes | a | y | | partitioning edge weights between citations ci and cj are set to | | | | |
| c | y | | | | c | wij = log | 0  @ ci,cj,si,sj X | φ2(ci, cj, yij) | | |
| Pairwise  Coreference | | | | | |
| s | s | | | | |
| x | x | | | | | 0 |si|  @ k=1 Y | φ3(si(t−1), sit, ci, xi) 1  A | | 0 |sj |  @ t=1 Y | φ3(sj(t−1), sjt, cj, xj) 1 1  A A |
| Figure 1: A model instance with three citations, two | | | | | | (3) | | | | |

of which are co-referent.

Joint inference over all coreference decisions involves finding the mode of

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| is, performing some form of inference separately in | P(y|x) = 1 Zx |  | K | φ0(yij, yik, yjk)   |  | K | ewij |  |
| different substructures and then (iteratively) integrat- |
| i=1,j>i,k>j | | i,j>i | |  |
| ing these results (Saul & Jordan, 1996; Yedidia et al., | (4) |
| 2000; Wiegerinck, 2000). |

Here we experiment with a particularly simple form of approximate inference: structured variants on iter-ated conditional modes (ICM) (Besag, 1986). In ICM, inference is performed by maximizing posterior prob-ability of a variable, conditioned on all others, and re-peatedly iterating over all variables. In its structured form, (possibly exact) inference may be performed on entire sub-structures of the model rather than a single variable, e.g. (Ying et al., 2002). In this paper we also use a variant we term iterated conditional sam-pling, in which, rather than selecting the single as-signment of variables with maximum probability, sev-eral assignments are sampled (although not necessarily randomly) from the posterior and made available to subsequent inference. We expect that doing so makes the procedure less sensitive to local minima. These samples can also be understood as a strong compres-sion of the exponentially-sized conditional probability tables that would have been sent as messages in struc-tured belief propagation.

Inference in our model is performed as follows. First exact inference is performed independently for each label sequence si, conditioned on its corresponding ci-tation word sequence xi using the Viterbi algorithm. Rather than selecting the single highest probability si (mode), however, we find the N-best list of label se-quences (sample). The citation fields ci are set deter-ministically from the sampled label sequences si.

Then coreference is accomplished by approximate in-ference via a greedy graph partitioning algorithm on the yij’s conditioned on the citation fields ci, as de-scribed in McCallum and Wellner (2003), except that the edge weights in the graph are determined by “in-

This problem is an instance of correlation clustering, which has sparked recent theoretical interest (Bansal et al., 2002; Demaine & Immorlica, 2003). Here we use a different approach to graph partitioning that bears more resemblance to agglomerative clustering — suit-able for larger graphs with a high ratio of negative to positive edges. We search through this space of possible partitionings using a stochastic beam search. Specifically, we begin with each citation in its own clus-ter and select k pairs of clusters to merge with proba-bility proportional to the edge-weight between the two clusters. We examine each of these k possible merges until one is found that results in an increase in the ob-jective function. We repeat this until we reach a stage where none of the k candidate merge pairs would result in an increased objective function value.

The final step to consider in our ICM-based inference is estimation of the attributes on entities, a, and a revis-itation of citation segmentation given coreference and these entity attributes. Although our current experi-ments do not use entity attributes to affect coreference directly, they could do so by creating entity variables on the fly as coreference decisions are hypothesized.

We seek to maximize

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| P(c, s, a|x) = 1 Z | | | |  | M | K | φ1(ai, cj)   | (5) |
|  | i | j |
|  | K | |sj| | φ3(sj(t−1), sjt, cj, xj) | | | | |
|  | j | t=1 |

The segmentations s and citation fields c are selected only among the N-best segmentations, and the entity

|  |  |  |
| --- | --- | --- |
| UAI 2004 | WELLNER ET AL. | 597 |

attributes are selected among the N-best citation fields of the coreferent citations for a given entity. Fortu-nately, there are few enough combinations that exact

common authors, publication venues, publishers, etc. The size of coreferent citation clusters has a skewed distribution; in three of the four subsets, at least 70%

|  |  |  |
| --- | --- | --- |
| inference can be performed here. In our experiments, | of the citations are singletons. | The largest citation |

there are a maximum of 13 citation fields, (although typically much fewer), 2-21 citations in a coreferent cluster, and N was 5 or less. Given the attributes of an entity, we compute scores for all (entity, segmen-tation) pairs. The segmentation with highest score is selected as the best segmentation for the citation. These entity attributes are also scored by summing these highest scores for all citations. In the end, the entity attributes with the highest score are chosen as the canonical citation for this cluster and best segmen-tations are selected based on these attributes.

cluster consists of 21 citations to the same paper.

We present results on two different sets of experiments. First, we consider the coreference component of the model, which takes as input a sample of the N-best segmentations of each observation. We compare its performance to coreference in which we assume per-fect labeling and in which we use no labeling at all. Second, we consider the segmentation performance of the model, which takes as input the citation clusters produced by the coreference component. We compare its accuracy to the baseline performance consisting of

|  |  |  |  |
| --- | --- | --- | --- |
| 2.3 | Parameter Estimation | the top Viterbi segmentation from a linear-chain CRF. | |
| Ideally we would perform parameter estimation by nu- | | 3.1 | Coreference Results |

merically climbing the gradient of the full, joint like-lihood. This approach is not practical because com-plete inference in this model is intractable. In addi-tion, our previous experience with coreference (Mc-Callum & Wellner, 2003) indicates that learning pa-rameters by maximizing a product of local marginals, i<jP(yij|xi, xj), provides equal or superior accuracy to stochastic gradient ascent on an approximation of the full joint likelihood.

Good segmentation of author, title and other fields en-ables features that are naturally expected to be useful to accurate coreference. The pair-wise coreference po-tentials are a function of a wide range of rich, overlap-ping features. The features largely consider field-level similarity using a number of string and token-based comparison metrics.2   
Briefly, these metrics include various string edit distance measures, TFIDF over to-kens, TFIDF over character n-grams as well hybrid

Following this success, here we train each sub- methods that combine token TFIDF and string edit

structure of the model separately, either as struc- distance. We also used feature conjunctions (e.g. a

tured pseudo-likelihood, or simply independently. feature that combines the author and title similarity

The parameters of the linear-chain CRF’s potentials,φ3(si(t−1), sit, cixi), are set to maximize the joint

measures). Some specialized features were developed for matching and normalizing author name fields as

|  |  |  |
| --- | --- | --- |
| probability of the correct label sequence, P(si|xi). | well as conference proceedings. | Finally, we also in- |

We employ feature induction as part of this train-ing (McCallum, 2003). The parameters of the coref-erence potentials on pairs of citations, φ2(ci, cj, yij),

cluded “global” features, based on string and token-based distance metrics, that looked at the entire cita-tion. The features are a mix of real-valued and binary-

are set to maximize the product of local likelihoods valued functions.

of each pair, i<jP(yij|xi, xj). The parameters of

entity-attribute/citation potentials, φ1(ai, cj), are set

by pseudolikelihood to maximize the likelihood of cor-

rect placement of edges between citations and their

true entity attributes. A spherical Gaussian prior with

zero mean is used in all cases.

We measure coreference performance at the pair-level and cluster-level. We report pairwise F1, which is the harmonic mean of pairwise precision and pairwise re-call. Pairwise precision is the fraction of pairs in the same cluster that are coreferent; pairwise recall is the fraction of coreferent pairs that were placed in the

|  |  |  |
| --- | --- | --- |
| 3 | Experiments and Results | same cluster. We also measure cluster recall, which |
| is simply the ratio of the number of correct clusters to |

the number of true clusters. Note that cluster recall

To evaluate our model, we apply it to a citation dataset from CiteSeer (Lawrence et al., 1999). The dataset contains approximately 1500 citations to 900 papers. The citations have been manually labeled for corefer-ence and manually segmented into fields, such as au-thor, title, etc. The dataset has four subsets of cita-tions, each one centered around a topic (e.g. reinforce-ment learning). Within a section, many citations share

gives no credit for a cluster that is partially correct.

Table 1 summarizes coreference performance in terms of pair-wise F1 and cluster recall respectively. Results reported are on the indicated test section; the model was trained on the other three sections. We report the

2We used the Secondstring package, some of the func-tions of which are described in (Cohen et al., 2003)

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 598 | Reinforce | Face | Reason | WELLNER ET AL. | | | | UAI 2004 | | | |
| Constraint 0.907  0.961  0.960  0.971  0.971 | conclusive—better in some cases, worse in others—and larger more interesting datasets (more and larger cita-tion clusters to leverage) may provide more interesting insights. Further leveraging conditional models’ facil-ity with feature engineering and induction may also | | | | | | |
| NoSeg | 0.836 | 0.879 | 0.801 |
| N=1 | 0.972 | 0.974 | 0.946 |
| N=3 | 0.95 | 0.979 | 0.961 |
| N=7  N=9 | 0.948  0.982 | 0.979  0.967 | 0.951  0.960 |
| Labeled | 0.956 | 0.965 | 0.964 | 0.971  0.988 | prove helpful. | | | | | | |
| Optimal | 0.995 | 0.992 | 0.994 |
| NoSeg | 0.787 | 0.931 | 0.883 | 0.892 | Reinforce | | | Face | Reason | | Const. |
| N=1 | 0.913 | 0.971 | 0.920 | 0.931 | Phrase Match | 0.79 | | 0.94 | 0.86 | | 0.89 |
| N=3 | 0.933 | 0.976 | 0.933 | 0.927 | RPM + MCMC | | 0.94 | 0.97 | 0.96 | | 0.93 |
| N=7  N=9 | 0.933  0.947 | 0.976  0.969 | 0.937  0.937 | 0.950  0.951 | CRF-Seg (N=9) | | 0.947 | 0.969 | | 0.937 | 0.951 |
| Labeled | 0.932 | 0.975 | 0.937 | 0.941  0.976 | Table 2: A comparison of cluster recall performance. The Phrase Matching and RPM + MCMC results are | | | | | | |
| Optimal | 0.98 | 0.996 | 0.993 |

Table 1: coreference performance measured by pair-wise F1 (upper part) cluster recall (lower part) us-

from (Pasula et al., 2003).

|  |  |  |
| --- | --- | --- |
| ing no segmentation (NoSeg), an average of the N - | 3.2 | Segmentation Results |
| best Viterbi segmentations, and hand-labeled segmen- |

tations (Labeled). The Optimal result represents an upper-bound where the optimal pair-wise potential is chosen by an oracle.

performance of our model using the N -best Viterbi segmentations for different values of N. Overall best

In leveraging coference to improve extraction, we use a combination of local (e.g. word contains digits), lay-out, lexicon membership features (e.g. membership in a database of Bibtex records). See (Peng & McCal-lum, 2004) for a description of features. Segmenta-tion performance is measured by the micro-averaged F1 across all fields, which approximates the accuracy

|  |  |  |
| --- | --- | --- |
| results occur at N = 9. | As a lower bound we also | of database fields. The segmentation component of the |

include the coreference performance when we include no segmentation information (NoSeg) and rely solely on the “global” features.

model was trained on a completely separate data set of citations (Peng & McCallum, 2004).

Table 3 shows the improved segmentation performance

Also included in the tables is the coreference perfor- using coreference information. The results reported

mance when the hand-labeled segmentation is pro- here only consider citations that were grouped to-

vided (Labeled). Note that the results using the N = 9 gether with at least one other citation (i.e. non-

Viterbi segmentations are comparable to or higher than those using the correctly labeled segmentations—indicating that neither segmentation performance nor our technique for incorporating segmentation uncer-tainty are the inhibiting factor in improving corefer-ence performance.

As an upper-bound experiment, we evaluate corefer-ence performance assuming the model always chooses

singletons), since these are the only citations whose segmentation we might hope to improve by using coreference. To test the significance of the improve-ments, we use McNemar’s test on labeling disagree-ments (Gillick & Cox, 1989). Table 3 summarizes the significance test results. At the 95% confidence level (p-value smaller than 0.05), the improvements on the four datasets are significant.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| the optimal pair-wise potential from among the N2 potentials. Thus, if the pair is coreferent, the poten-tial is set to the maximum potential; if the pair is not coreferent, the potential is set to the minimum poten-tial. The performance is near perfect (see Optimal in Tables 1).  Table 2 compares our best results to the results | | | Reinforce | | Face | Reason | Constraint |
| Baseline | .943 | .908 | .929 | .934 |
| W/ Coref | .949 | .914 | .935 | .943 |
| Err. Red. | .101 | .062 | .090 | .142 |
| P-value | .0442 | .0014 | .0001 | .0001 |
| Table 3: Comparison of segmentation performance on non-singleton citations using entity attributes gener-ated through coreference vs. baseline segmentation. | | | | |
| presented in | (Pasula et al., 2003). | As a base- |

line, we include the results of their implementa-

tion of the Phrase matching algorithm—a greedy ag-glomerative clustering algorithm where pair-wise ci-tation similarity is based on the overlap in words

We also explore the potential for improving segmenta-tion performance by selecting among the N-best seg-mentations. For a given list of N segmentations, we

and phrases (word bigrams). RPM + MCMC is aim to select the segmentation closest to the true seg-

their first-order, generatively-trained graphical model mentation. The results in Table 4 show the optimal

(see Related Work section). These results are not segmentation performance for different values of N.

|  |  |  |
| --- | --- | --- |
| UAI 2004 | WELLNER ET AL. | 599 |

We can see that there is further potential to improve segmentation based on optimally selecting segmenta-

We then compute the objective function values of these partitionings according to both the pair-wise and

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| tions. | Reinforce | Face | Reason | Constraint | entity-based models. The pair-wise model objection | |
| function is described above. The entity-based model’s | |
| objective function is the product of the potentials be- | |
| N=1 | 0.936 | 0.911 | 0.912 | 0.933 |
| tween attributes of the entity and the field values of | |
| N=3 | 0.958 | 0.937 | 0.940 | 0.969 |
| citations within a partition. | The value of each en- |
| N=5 | 0.962 | 0.948 | 0.946 | 0.975 |
| tity attribute was generated by selecting the medoid | |

Table 4: Optimal segmentation improvement for dif-ferent values of N over all citations (including single-

field values from among the citations in the partition. (A medoid is the item in a cluster that has the min-

tons). imum dissimilarity to all other items in the cluster—

in this case, the similarity metric is a string-edit dis-

|  |  |  |  |
| --- | --- | --- | --- |
| 4 | Model Comparison | tance.) | The potentials between citation-entity pairs |
| were learned with exactly the same features and train- | |
| ing procedure as in the pair-wise citation model. | |

Our model consists of both explicit pair-wise coref-erence variables for each pair of citations, as well as explicit entity attribute variables for each group of co-referent citations. At inference time in our current ex-periments, however, coreference is driven solely by the pair-wise citation potentials. An alternative method would ignore the pair-wise potentials and consider only entity-citation potentials, creating entities at inference time as necessary.

If given a uniform prior over the number of entities, such a method would always chose to have a separate entity for each citation, and we must include a prior that prefers smaller numbers of entities (or equiva-lently, a penalty for generating each entity). Thus, the number of entities induced is a function of the “ten-sion” between the entity-citation potentials (which de-pend on the observed citation strings, be highly pa-rameterized, and be learned) and the prior (which do not). Intuitively, we wonder if this imbalance in ex-

For the randomly generated partitionings, we examine how well the probabilities for both models correlate (r2correlation) with the Pair F1 evaluation metric; see Table 5. Figure 2 shows scatter plots for the Con-straint data set illustrating the correlation between the model objective function values and Pair F1.

We hypothesize that the pair-wise model correlates better partly due to the fact that the noise in edge po-tentials is ameliorated by averaging over n potentials for each citation (one for each other citation) instead of only a single potential as is the case with the entity-based model. Perhaps even though the edge potentials in the entity-based model are expected to be less noisy (because the entity attributes are more “canonical”), averaging in the pair-wise model is still more robust. Further work is needed here to determine the effect of the quality of the generated entity attributes and learned potentials on the model’s performance.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| pressiveness and learnability provides less delineative power and robustness than a more balanced alterna-tive. | | | | | Reinforce | | Face | Reason | Constraint |
| Pair | 0.976 | 0.975 | 0.967 | 0.974 |
| Entity | 0.795 | 0.592 | 0.360 | 0.775 |
| By | contrast, | a | graph-partitioning, | correlational- | Table 5: Model objective function r2correlation with PairF1, in which the correct number of entities is pro-vided. | | | | |
| clustering approach to coreference includes potentials on citation pairs, and the number of entities emerges | | | | |

naturally from the tension between positive and nega-

|  |  |  |
| --- | --- | --- |
| tive edge weights in the graph—both of which can be | 5 | Related Work |
| conditionally trained and highly parameterized in their |

dependency on the observed citation strings. Also the

compatibility between a set of hypothesized coreferent mentions is represented by a “mixture” rather than a single “prototype.”

To explore these issues, we compare the two models us-ing a randomly-generated sample of 100 partitionings of our citation dataset. Here we focus on the robust-ness of clustering coreferent citations; to remove the issues of tuning a prior over the number of entities, all randomly-generated partitions were constrained to have the correct number of clusters.

This paper is an example of integrating information extraction and data mining, as discussed in McCallum and Jensen (2003). Additional previous work in this area includes Nahm and Mooney (2000), in which data mining association rules are applied to imperfectly-extracted job posting data, and then used to improve the recall of subsequent information extraction.

Our work here is most related to the work of Pasula et al. (2003), who describe a first-order probabilistic model for citation segmentation and coreference. From

UAI 2004 WELLNER ET AL. 601

not necessarily reflect those of the sponsor. We are also Milch, B., Marthi, B., & Russell, S. (2004). Blog: Re-

grateful to Charles Sutton and Brian Milch for helpful com- lational modeling with unknown objects. ICML 2004

ments on a previous draft and David Jensen for helpful Workshop on Statistical Relational Learning and Its

discussions. Connections to Other Fields.

Nahm, U. Y., & Mooney, R. J. (2000). A mutually benefi-

References

Bansal, N., Chawla, S., & Blum, A. (2002). Correlation clustering. The 43rd Annual Symposium on Foundations of Computer Science (FOCS) (pp. 238–247).

Besag, J. (1986). On the statistical analysis of dirty pic-tures. Journal of the Royal Statistics Society B, 48, 259–

cial integration of data mining and information extrac-tion. Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI). Austin, TX.

Nevillle, J., Simsek, O., & Jensen, D. (2004). Autocorrela-tion and relational learning: Challenges and opportuni-ties. To appear in ICML Statistical Relational Learning Workshop. Banff, Canada.

302. Pasula, H., Marthi, B., Milch, B., Russell, S., & Shpitser,

Boykov, Y., Veksler, O., & Zabih, R. (1999). Fast approx-imate energy minimization via graph cuts. Proceedings of the Seventh IEEE International Conference on Com-puter Vision (ICCV) (1) (pp. 377–384).

Carreras, X., Marques, L., & Padro, L. (2002). Named entity extraction using AdaBoost. Proc. 6th Workshop on Computational Language Learning (CoNLL).

Cohen, W., Ravikumar, P., & Fienberg, S. (2003). A comparison of string metrics for matching names and records. KDD-2003 Workshop on Data Cleaning and Object Consolidation.

Demaine, E., & Immorlica, N. (2003). Correlation cluster-ing with partial information. 6th International Work-shop on Approximation Algorithms for Combinatorial Optimization Problems (APPROX 2003).

Gillick, L., & Cox, S. (1989). Some statistical issues in the compairson of speech recognition algorithms. Proceed-ings of the International Conference on Acoustics Speech and Signal Processing (ICASSP) (pp. 532–535).

Lafferty, J., McCallum, A., & Pereira, F. (2001). Condi-tional random fields: Probabilistic models for segment-ing and labeling sequence data. Proc. of The Eighteenth International Conf. on Machine Learning (ICML).

Lawrence, S., Giles, C. L., & Bollacker, K. (1999). Dig-ital libraries and autonomous citation indexing. IEEE Computer, 32, 67–71.

I. (2003). Identity uncertainty and citation matching. Advances in Neural Information Processing (NIPS).

Peng, F., & McCallum, A. (2004). Accurate information extraction from research papers using conditional ran-dom fields. Proceedings of Human Language Technology Conference and North American Chapter of the Associ-ation for Computational Linguistics(HLT-NAACL) (pp. 329–336).

Pinto, D., McCallum, A., Lee, X., & Croft, W. B. (2003). Table extraction using conditional random fields. Pro-ceedings of the 26th ACM SIGIR Conference on Re-search and Development in Information Retrieval.

Roth, D., & tau Yih, W. (2004). A linear programming for-mulation for global inference in natural language tasks. Proceedings of the Eighth Workshop on Computational Language Learning (CoNLL).

Saul, L. K., & Jordan, M. I. (1996). Exploiting tractable substructures in intractable networks. Advances in Neu-ral Information Processing Systems (NIPS).

Sha, F., & Pereira, F. (2003). Shallow parsing with con-ditional random fields. Proc. of Human Language Tech-nology Conf. and North American Chapter of the Asso-ciation for Computational Linguistics(HLT-NAACL).

Sutton, C., Rohanimanesh, K., & McCallum, A. (2004). Dynamic conditional random fields: Factorized prob- abilistic models for labeling and segmenting sequence data. Proceedings of the Twenty-First International

|  |  |  |
| --- | --- | --- |
| Marthi, B., Milch, B., & Russell, S. (2003). | First-order | Conference on Machine Learning (ICML). |

probabilistic models for information extraction. IJCAI 2003 Workshop on Learning Statistical Models from Re-lational Data.

Taskar, B., Abbeel, P., & Koller, D. (2002). Discriminative probabilistic models for relational data. Proceedings of the Eighteenth Conference on Uncertainty in Artificial

McCallum, A. (2003). Efficiently inducing features of con- Intelligence (UAI).

ditional random fields. Proceedings of 19th Conference on Uncertainty in Artificial Intelligence (UAI).

McCallum, A., & Jensen, D. (2003). A note on the uni- fication of information extraction and data mining us- ing conditional-probability relational models. IJCAI Workshop on Learning Statistical Models from Rela- tional Data.

McCallum, A., & Li, W. (2003). Early results for named entity recognition with conditional random fields, fea-ture induction and web-enhanced lexicons. Proceedings of the Seventh Workshop on Computational Language Learning (CoNLL).

McCallum, A., & Wellner, B. (2003). Toward conditional models of identity uncertainty. IJCAI Workshop on In-formation Integration and the Web.

Wiegerinck, W. (2000). Variational approximations be- tween mean field theory and the junction tree algorithm. The Sixteenth Conference on Uncertainty in Artificial Intelligence (UAI) (pp. 626–633).

Yedidia, J., Freeman, W., & Weiss, Y. (2000). Generalized belief propagation. Advances in Neural Information Pro-cessing Systems (NIPS) (pp. 689–695).

Ying, L., Frey, B., Koetter, R., & Munson, D. (2002). Anal-ysis of an iterative dynamic programming approach to 2-d phase unwrapping. IEEE International Conference on Image Processing.