Jointly Identifying Entities and Extracting Relations in Encyclopedia Text via A Graphical Model Approach*

Xiaofeng YU **Wai LAM**

Information Systems Laboratory
Department of Systems Engineering & Engineering Management
The Chinese University of Hong Kong
{xfyu,wlam}@se.cuhk.edu.hk

Abstract

In this paper, we investigate the problem of entity identification and relation extraction from encyclopedia articles, and we propose a joint discriminative probabilistic model with arbitrary graphical structure to optimize all relevant subtasks simultaneously. This modeling offers a natural formalism for exploiting rich dependencies and interactions between relevant subtasks to capture mutual benefits, as well as a great flexibility to incorporate a large collection of arbitrary, overlapping and non-independent features. We show the parameter estimation algorithm of this model. Moreover, we propose a new inference method, namely collective iterative classification (CIC), to find the most likely assignments for both entities and relations. We evaluate our model on real-world data from Wikipedia for this task, and compare with current state-of-the-art pipeline and joint models, demonstrating the effectiveness and feasibility of our approach.

1 Introduction

We investigate a compound information extraction (IE) problem from encyclopedia articles, which consists of two subtasks — recognizing structured information about entities and extracting the relationships between entities. The most common approach to this problem is a pipeline architecture: attempting to perform different subtasks, namely, named entity recognition and relation extraction between recognized entities in several separate, and independent stages. Such kind of design is widely adopted in NLP.

The work described in this paper is substantially supported by grants from the Research Grant Council of the Hong Kong Special Administrative Region, China (Project No: CUHK4128/07) and the Direct Grant of the Faculty of Engineering, CUHK (Project Codes: 2050442 and 2050476). This work is also affiliated with the Microsoft-CUHK Joint Laboratory for Human-centric Computing and Interface Technologies.

The most common and simplest approach to performing compound NLP tasks is the 1-best pipeline architecture, which only takes the 1-best hypothesis of each stage and pass it to the next one. Although it is comparatively easy to build and efficient to run, this pipeline approach is highly ineffective and suffers from serious problems such as error propagation (Finkel *et al.*, 2006; Yu, 2007; Yu *et al.*, 2008). It is not surprising that, the end-to-end performance will be restricted and upper-bounded.

Usually, one can pass N-best lists between different stages in pipeline architectures, and this often gives useful improvements (Hollingshead and Roark, 2007). However, effectively making use of N-best lists often requires lots of engineering and human effort (Toutanova, 2005). On the other hand, one can record the complete distribution at each stage in a pipeline, to compute or approximate the complete distribution at the next stage. Doing this is generally infeasible, and this solution is rarely adopted in practice.

One promising way to tackle the problem of error propagation is to explore joint learning which integrates evidences from multiple sources and captures mutual benefits across multiple components of a pipeline for all relevant subtasks simultaneously (e.g., (Toutanova *et al.*, 2005), (Poon and Domingos, 2007), (Singh *et al.*, 2009)). Joint learning aims to handle multiple hypotheses and uncertainty information and predict many variables at once such that subtasks can aid each other to boost the performance, and thus usually leads to complex model structure. However, it is typically intractable to run a joint model and they sometimes can hurt the performance, since they

Coling 2008: Poster Volume, pages 1399–1407,
Beijing, August 2010

increase the number of paths to propagate errors. Due to these difficulties, research on building joint approaches is still in the beginning stage.

A significant amount of recent work has shown the power of discriminatively-trained probabilistic graphical models for NLP tasks (Lafferty *et al.*, 2001; Sutton and McCallum, 2007; Wainwright and Jordan, 2008). The superiority of graphical model is its ability to represent a large number of random variables as a family of probability distributions that factorize according to an underlying graph, and capture complex dependencies between variables. And this progress has begun to make the joint learning approach possible.

In this paper we study and formally define the joint problem of entity identification and relation extraction from encyclopedia text, and we propose a joint paradigm in a single coherent framework to perform both subtasks simultaneously. This framework is based on undirected probabilistic graphical models with arbitrary graphical structure. We show how the parameters in this model can be estimated efficiently. More importantly, we propose a new inference method — collective iterative classification (CIC), to find the maximum a posteriori (MAP) assignments for both entities and relations. We perform extensive experiments on real-world data from Wikipedia for this task, and substantial gains are obtained over state-of-the-art probabilistic pipeline and joint models, illustrating the promise of our approach.

2 Problem Formulation

2.1 Problem Description

This problem involves identifying entities and discovering semantic relationships between entity pairs from English encyclopedic articles. The basic document is an article, which mainly defines and describes an entity (known as *principal entity*). This document mentions some other entities as *secondary entities* related to the principal entity. Clearly, our task consists of two subtasks — first, for entity identification, we need to recognize the secondary entities (both the boundaries and types of them) in the document [1]. Second, after all the secondary entities are identified, our goal for relation extraction is to predict what relation, if any, each secondary entity has to the principal entity. We assume that there is no relationship between any two secondary entities in one document.

As an illustrative example, Figure 1 shows the task of entity identification and relationship extraction from encyclopedic documents. Here, *Abraham Lincoln* is the principal entity. Our task consists of assigning a set of pre-defined entity types (e.g., PER, DATE, YEAR, and ORG) to segmentations in encyclopedic documents and assigning a set of pre-defined relations (e.g., birth_day, birth_year, and member_of) for each identified secondary entity to the principal entity.

2.2 Problem Formulation

Let $\mathbf{x}$ be an observation sequence of tokens in encyclopedic text and $\mathbf{x} = \{x_1, \cdots, x_N\}$. Let s_p be the principal entity (we assume that it is known or can be easily recognized), and let $\mathbf{s} = \{s_1, \cdots, s_L\}$ be a segmentation assignment of observation sequence $\mathbf{x}$. Each segment s_i is a triple $s_i = \{\alpha_i, \beta_i, y_i\}$, where α_i is a start position, β_i is an end position, and y_i is the label assigned to all tokens of this segment. The segment s_i satisfies $0 \leq \alpha_i < \beta_i \leq |\mathbf{x}|$ and $\alpha_{i+1} = \beta_i + 1$. Let r_{pn} be the relation assignment between principal entity s_p and secondary entity candidate s_n from the segmentation $\mathbf{s}$, and $\mathbf{r}$ be the set of relation assignments for sequence $\mathbf{x}$.

Let $\mathbf{y} = \{\mathbf{r}, \mathbf{s}\}$ be the pair of segmentation $\mathbf{s}$ and segment relations $\mathbf{r}$ for an observation sequence $\mathbf{x}$. A valid assignment $\mathbf{y}$ must satisfy the condition that the assignments of the segments and the assignments of the relations of segments are maximized simultaneously. We now formally define this joint optimization problem as follows:

Definition 1 ***(Joint Optimization of Entity Identification and Relation Extraction)****: Given an observation sequence* $\boldsymbol{x}$*, the goal of joint optimization of entity identification and relation extraction is to find the assignment* $\boldsymbol{y}^* = \{\boldsymbol{r}^*, \boldsymbol{s}^*\}$ *that has the maximum a posteriori (MAP) probability*

$$\boldsymbol{y}^* = \arg\max_{\boldsymbol{y}} P(\boldsymbol{y}|\boldsymbol{x}), \quad (1)$$

[1] Since the topic/title of an article usually defines a principal entity (e.g., a famous person) and it is easy to identify, in this paper we only focus on secondary entity identification.

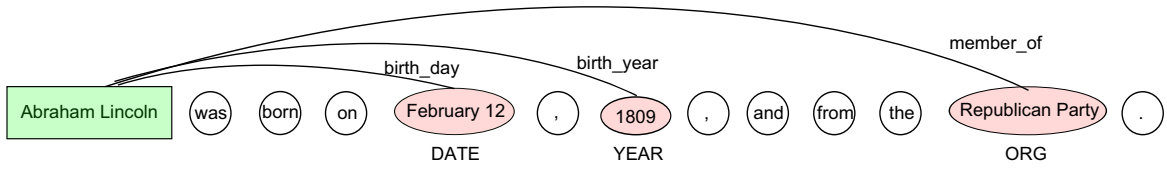

Figure 1: An example of entity identification and relation extraction excerpted from our dataset. The secondary entities are in pink color and labeled. The semantic relation of each secondary entity to the principal entity *Abraham Lincoln* (in green color and we assume that it is known or can be easily recognized) is also shown.

where $\mathbf{r}^*$ *and* $\mathbf{s}^*$ *denote the most likely relation assignment and segmentation assignment, respectively.*

Note that this problem is usually very challenging and offers new opportunities for information extraction, since complex dependencies between segmentations and relations should be exploited.

3 Our Proposed Model

3.1 Preliminaries

Conditional random fields (CRFs) (Lafferty *et al.*, 2001) are undirected graphical models trained to maximize the conditional probability of the desired outputs given the corresponding inputs. Let $\mathcal{G}$ be a factor graph (Kschischang *et al.*, 2001) defining a probability distribution over a set of output variables $\mathbf{o}$ conditioned on observation sequences $\mathbf{x}$. $C = \{\Phi_c(\mathbf{o}_c, \mathbf{x}_c)\}$ is a set of factors in $\mathcal{G}$, then the probability distribution over $\mathcal{G}$ can be written as

$$P(\mathbf{o}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{c \in C} \Phi_c(\mathbf{o}_c, \mathbf{x}_c) \qquad (2)$$

where Φ_c is a potential function and $Z(\mathbf{x}) = \sum_{\mathbf{o}} \prod_{c \in C} \Phi_c(\mathbf{o}_c, \mathbf{x}_c)$ is a normalization factor. We assume the potentials factorize according to a set of features $\{f_k(\mathbf{o}_c, \mathbf{x}_c)\}$ as $\Phi_c(\mathbf{o}_c, \mathbf{x}_c) = \exp(\sum_k \theta_k f_k(\mathbf{o}_c, \mathbf{x}_c))$ so that the family of distributions is an exponential family. The model parameters are a set of real-valued weights $\Theta = \{\theta_k\}$, one weight for each feature. Practical models rely extensively on parameter tying to use the same parameters for several factors.

However, the traditional fashion of CRFs can only deal with single task, they lack the capability to represent more complex interaction between multiple subtasks. In the following we will describe our joint model in detail for this problem.

3.2 A Joint Model for Entity Identification and Relation Extraction

Following the notations in Section 2.2, let L and M be the number of segments and number of relations for sequence $\mathbf{x}$, respectively. We define a joint conditional distribution for segmentation $\mathbf{s}$ in observation sequence $\mathbf{x}$ and segment relation $\mathbf{r}$ in undirected, probabilistic graphical models. The nature of our modeling enables us to partition the factors C of $\mathcal{G}$ into three groups $\{C_S, C_R, C_\nabla\}$=$\{\{\phi^S\}, \{\phi^R\}, \{\phi^\nabla\}\}$, namely the segmentation potential ϕ^S, the relation potential ϕ^R, and the segmentation-relation joint potential ϕ^∇, and each potential is a clique template whose parameters are tied. The potential function $\phi^S(i, \mathbf{s}, \mathbf{x})$ models segmentation $\mathbf{s}$ in $\mathbf{x}$, the potential function $\phi^R(r_{pm}, r_{pn}, \mathbf{r})$ $(m \neq n)$ represent dependencies (e.g., long-distance dependencies, relation transitivity, etc) between any two relations in the relation set $\mathbf{r}$, where r_{pm} is the relation assignment between the principal entity s_p and the secondary entity candidate s_m from $\mathbf{s}$, and similarly for r_{pn}. And the joint potential $\phi^\nabla(s_p, s_j, \mathbf{r})$ captures rich and complex interactions between segmentation $\mathbf{s}$ for secondary entity identification and relation $\mathbf{r}$ between each secondary entity candidate s_j to the principal entity s_p. According to the celebrated Hammersley-Clifford theorem (Besag, 1974), the joint conditional distribution $P(\mathbf{y}|\mathbf{x}) = P(\{\mathbf{r}, \mathbf{s}\}|\mathbf{x})$ is factorized as a product of potential functions over cliques in the graph $\mathcal{G}$ as the form of an exponential family:

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \Big(\prod_{C_S} \phi^S(i, \mathbf{s}, \mathbf{x}) \Big) \Big(\prod_{C_R} \phi^R(r_{pm}, r_{pn}, \mathbf{r}) \Big) \Big(\prod_{C_\nabla} \phi^\nabla(s_p, s_j, \mathbf{r}) \Big) \qquad (3)$$

where $Z(\mathbf{x}) = \sum_{\mathbf{y}} \prod_{C_S} \phi^S(i, \mathbf{s}, \mathbf{x}) \prod_{C_R} \phi^R(r_{pm}, r_{pn}, \mathbf{r}) \prod_{C_\nabla} \phi^\nabla(s_p, s_j, \mathbf{r})$ is the normalization factor of the joint model.

We assume the potential functions ϕ^S, ϕ^R and ϕ^∇ factorize according to a set of features and a corresponding set of real-valued weights. More specifically, $\phi^S(i, \mathbf{s}, \mathbf{x}) = \exp(\sum_{i=1}^{|\mathbf{s}|} \sum_{k=1}^{K} \lambda_k g_k(i, \mathbf{s}, \mathbf{x}))$. To effectively capture properties of segmentation, we relax the first-order Markov assumption to semi-Markov such that each segment feature function $g_k(\cdot)$ depends on the current segment s_i, the previous segment s_{i-1}, and the whole observation sequence $\mathbf{x}$, that is, $g_k(i, \mathbf{s}, \mathbf{x}) = g_k(s_{i-1}, s_i, \mathbf{x}) = g_k(y_{i-1}, y_i, \alpha_i, \beta_i, \mathbf{x})$. And transitions within a segment can be non-Markovian.

Similarly, the potential $\phi^R(r_{pm}, r_{pn}, \mathbf{r}) = \exp(\sum_{m,n}^{M} \sum_{w=1}^{W} \mu_w q_w(r_{pm}, r_{pn}, \mathbf{r}))$ and $\phi^\nabla(s_p, s_j, \mathbf{r}) = \exp(\sum_{j=1}^{L} \sum_{t=1}^{T} \nu_t h_t(s_p, s_j, \mathbf{r}))$, where W and T are number of feature functions, $q_w(\cdot)$ and $h_t(\cdot)$ are feature functions, μ_w and ν_t are corresponding weights for them. The potential $\phi^R(r_{pm}, r_{pn}, \mathbf{r})$ allows long-range dependency representation between different relations r_{pm} and r_{pn}. For example, if the same secondary entity is mentioned more than once in an observation sequence, all mentions probably have the same relation to the principal entity. Using potential $\phi^R(r_{pm}, r_{pn}, \mathbf{r})$, evidences for the same entity segments to the principal entity are shared among all their occurrences within the document. The joint factor $\phi^\nabla(s_p, s_j, \mathbf{r})$ exploits tight dependencies between segmentations and relations. For example, if a segment is labeled as a *location* and the principal entity is *person*, the semantic relation between them can be *birth_place* or *visited*, but cannot be *employment*. Such dependencies are essential and modeling them often leads to improved performance. In summary, the probability distribution of the joint model can be rewritten as:

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left\{ \sum_{i=1}^{|\mathbf{s}|} \sum_{k=1}^{K} \lambda_k g_k(i, \mathbf{s}, \mathbf{x}) + \sum_{m,n}^{M} \sum_{w=1}^{W} \mu_w q_w(r_{pm}, r_{pn}, \mathbf{r}) + \sum_{j=1}^{L} \sum_{t=1}^{T} \nu_t h_t(s_p, s_j, \mathbf{r}) \right\} \quad (4)$$

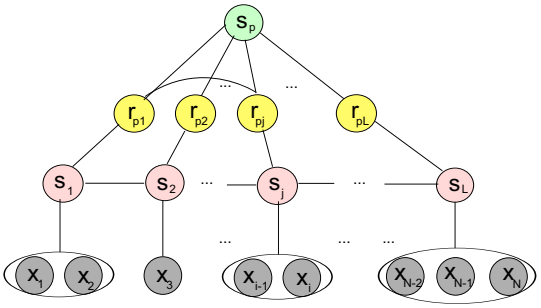

Figure 2: Graphical representation of the probabilistic joint model. The gray nodes represent sequence tokens $\{x_1, \cdots, x_N\}$. Each ellipse represents a segment consisting of several consecutive sequence tokens. The pink nodes represent segmentation assignment $\{s_1, \cdots, s_L\}$ of sequence. The yellow nodes represent relation assignment $\{r_{p1}, \cdots, r_{pL}\}$ between the principal entity s_p (in green color) and secondary entity segments.

As illustrated in Figure 2, our model consists of three sub-structures: a semi-Markov chain on the segmentations $\mathbf{s}$ conditioned on the observation sequences $\mathbf{x}$, represented by ϕ^S; potential ϕ^R measuring dependencies between different relations r_{pm} and r_{pn}; and a fully-connected graph on the principal entity s_p and each segment s_j for their relations, represented by ϕ^∇.

While several special cases of CRFs are of particular interest, and we emphasize on the differences and advantages of our model against others. Linear-chain CRFs (Lafferty *et al.*, 2001) can only perform single sequence labeling, they lack the ability to capture long-distance dependency and represent complex interactions between multiple subtasks. Skip-chain CRFs (Sutton and Mccallum, 2004) introduce skip edges to model long-distance dependencies to handle the label consistency problem in single sequence labeling and extraction. 2D CRFs (Zhu *et al.*, 2005) are two-dimensional conditional random fields incorporating the two-dimensional neighborhood dependencies in Web pages, and the graphical representation of this model is a 2D grid. Hierarchical CRFs (Liao *et al.*, 2007) are a class of CRFs with hierarchical tree structure. Our probabilistic model for joint entity identification and relation extraction has distinct graphical structure from 2D and hierarchical CRFs. And this modeling has sev-

eral advantages over previous probabilistic graphical models by using semi-Markov chains for efficient segmentation and labeling, by representing long-range dependencies between relations, and by capturing rich and complex interactions between relevant subtasks to exploit mutual benefits.

4 Learning the Parameters

Given independent and identically distributed (IID) training data $\mathcal{D} = \{\mathbf{x}^i, \mathbf{y}^i\}_{i=1}^{\mathbb{N}}$, where $\mathbf{x}^i$ is the i-th sequence instance, $\mathbf{y}^i = \{\mathbf{r}^i, \mathbf{s}^i\}$ is the corresponding segmentation and relation assignments. The objective of learning is to estimate $\Lambda = \{\lambda_k, \mu_w, \nu_t\}$ which is the vector of model's parameters. Under the IID assumption, we ignore the summation operator $\sum_{i=1}^{\mathbb{N}}$ in the log-likelihood during the following derivations. To reduce over-fitting, we use regularization and a common choice is a spherical Gaussian prior with zero mean and covariance $\sigma^2 I$. Then the regularized log-likelihood function $\mathcal{L}$ for the data is

$$\mathcal{L} = \log\big[\Phi(\mathbf{r}, \mathbf{s}, \mathbf{x})\big] - \log\big[Z(\mathbf{x})\big] - \sum_{k=1}^{K} \frac{\lambda_k^2}{2\sigma_\lambda^2} - \sum_{w=1}^{W} \frac{\mu_w^2}{2\sigma_\mu^2} - \sum_{t=1}^{T} \frac{\nu_t^2}{2\sigma_\nu^2} \quad (5)$$

where $\Phi(\mathbf{r}, \mathbf{s}, \mathbf{x}) = \exp\{\sum_{i=1}^{|\mathbf{s}|} \sum_{k=1}^{K} \lambda_k g_k(i, \mathbf{s}, \mathbf{x}) + \sum_{m,n}^{M} \sum_{w=1}^{W} \mu_w q_w(r_{pm}, r_{pn}, \mathbf{r}) + \sum_{j=1}^{L} \sum_{t=1}^{T} \nu_t h_t(s_p, s_j, \mathbf{r})\}$, $Z(\mathbf{x}) = \sum_{\mathbf{y}} \prod \Phi(\mathbf{r}, \mathbf{s}, \mathbf{x})$, and $1/2\sigma_\lambda^2$, $1/2\sigma_\mu^2$, $1/2\sigma_\nu^2$ are regularization parameters.

Taking derivatives of the function $\mathcal{L}$ over the parameter λ_k yields:

$$\frac{\partial \mathcal{L}}{\partial \lambda_k} = \sum_{i=1}^{|\mathbf{s}|} g_k(i, \mathbf{s}, \mathbf{x}) - \sum_{i=1}^{|\mathbf{s}|} g_k(i, \mathbf{s}, \mathbf{x}) P(\mathbf{y}|\mathbf{x}) - \sum_{k=1}^{K} \frac{\lambda_k}{\sigma_\lambda^2} \quad (6)$$

Similarly, the partial derivatives of the log-likelihood with respect to parameters μ_w and nu_t are as follows:

$$\frac{\partial \mathcal{L}}{\partial \mu_w} = \sum_{m,n}^{M} q_w(r_{pm}, r_{pn}, \mathbf{r}) - \sum_{m,n}^{M} q_w(r_{pm}, r_{pn}, \mathbf{r}) \times P(\mathbf{y}|\mathbf{x}) - \sum_{w=1}^{W} \frac{\mu_w}{\sigma_\mu^2} \quad (7)$$

$$\frac{\partial \mathcal{L}}{\partial \nu_t} = \sum_{j=1}^{L} h_t(s_p, s_j, \mathbf{r}) - \sum_{j=1}^{L} h_t(s_p, s_j, \mathbf{r}) P(\mathbf{y}|\mathbf{x}) - \sum_{t=1}^{T} \frac{\nu_t}{\sigma_\nu^2} \quad (8)$$

The function $\mathcal{L}$ is concave, and can be efficiently maximized by standard techniques such as stochastic gradient and limited memory quasi-Newton (L-BFGS) algorithms. The parameters λ_k μ_w and ν_t are optimized iteratively until converge.

5 Finding the Most Likely Assignments

The objective of inference is to find $\mathbf{y}^* = \{\mathbf{r}^*, \mathbf{s}^*\} = \arg\max_{\{\mathbf{r},\mathbf{s}\}} P(\mathbf{r}, \mathbf{s}|\mathbf{x})$ such that both $\mathbf{s}^*$ and $\mathbf{r}^*$ are optimized simultaneously. Unfortunately, exact inference to this problem is generally prohibitive, since it requires enumerating all possible segmentation and corresponding relation assignments. Consequently, approximate inference becomes an alternative.

We propose a new algorithm: collective iterative classification (CIC) to perform approximate inference to find the maximum a posteriori (MAP) segmentation and relation assignments of our model in an iterative fashion. The basic idea of CIC is to decode every target hidden variable based on the assigning labels of its sampled variables, where the labels might be dynamically updated throughout the iterative process. Collective classification refers to the classification of relational objects described as nodes in a graphical structure, as in our model.

The CIC algorithm performs inference in two steps, as shown in Algorithm 1. The first step, `bootstrapping`, predicts an initial labeling assignment for a unlabeled sequence $\mathbf{x}_i$, given the trained model $P(\mathbf{y}|\mathbf{x})$. The second step is the `iterative classification` process which re-estimates the labeling assignment of $\mathbf{x}_i$ several times, picking them in a sample set $\mathcal{S}$ based on initial assignment for $\mathbf{x}_i$. Here we exploit the sampling technique (Andrieu *et al.*, 2003). The advantages of sampling are summarized as follows. Sampling stochastically enables us to generate a wide range of inference situations, and the samples are likely to be in high probability areas, increasing our chances of finding the max-

imum, thus leading to more robust and accurate performance. The CIC algorithm may converge if none of the labeling assignments change during an iteration or a given number of iterations is reached.

Noticeably, this inference algorithm is also used to efficiently compute the marginal probability $P(\mathbf{y}|\mathbf{x})$ during parameter estimation (the normalization constant $Z(\mathbf{x})$ can also be calculated via approximation techniques). As can be seen, this algorithm is simple to design, efficient and scales well *w.r.t.* the size of data.

Algorithm 1: Collective Iterative Classification Inference

Input: A unlabeled sequence $\mathbf{x}_i$ and a trained model $P(\mathbf{y}|\mathbf{x})$
Output: The set of predicted assignment $y_i = \{r_i, s_i\}$
`// Bootstrapping`
foreach $y_i \in \mathcal{Y}$ **do**
  $\overline{y_i} \leftarrow \arg\max_{y_i} P(y_i|x_i)$;
end
`// Iterative Classification`
repeat
  Generate a sample set $\mathcal{S}$ based on initial label assignment $\overline{y_i}$ for sequence $\mathbf{x}_i$;
  foreach $\mathbb{s}_i \in \mathcal{S}$ **do**
    Assign new label assignment to sample $\mathbb{s}_i$;
  end
until *all labels have stabilized or a threshold number of iterations have elapsed* ;
return $y_i = \{r_i, s_i\}$

6 Experiments

6.1 Data

Our data comes from Wikipedia[2], the world's largest free online encyclopedia. This dataset consists of 1127 paragraphs from 441 pages from the online encyclopedia Wikipedia. We labeled 7740 entities into 8 categories, yielding 1243 *person*, 1085 *location*, 875 *organization*, 641 *date*, 1495 *year*, 38 *time*, 59 *number*, and 2304 *miscellaneous* names. This dataset also contains 4701 relation instances and 53 labeled relation types. The 10 most frequent relation types are *job_title*, *visited*, *birth_place*, *associate*, *birth_year*, *member_of*, *birth_day*, *opus*, *death_year*, and *death_day*. Note that this compound IE task involving entity identification and relation extraction is very challenging, and modeling tight interactions between entities and their relations is highly attractive.

6.2 Feature Set

Accurate entities enable features that are naturally expected to be useful to boost relation extraction. And a wide range of rich, overlapping features can be exploited in our model. These features include contextual features, part-of-speech (POS) tags, morphological features, entity-level dictionary features, clue word features. Feature conjunctions are also used. In leveraging relation extraction to improve entity identification, we use a combination of syntactic, entity, keyword, semantic, and Wikipedia characteristic features. More importantly, our model can incorporate multiple mention features $q_w(\cdot)$, which are used to collect evidences from other occurrences of the same secondary entities for consistent segmentation and relation labeling to the principal entity. The features $h_t(\cdot)$ capture deep dependencies between segmentations and relations, and they are natural and useful to enhance the performance.

[2]http://www.wikipedia.org/

6.3 Methodology

We perform four-fold cross-validation on this dataset, and take the average performance. For performance evaluation, we use the standard measures of Precision (P), Recall (R), and F-measure (the harmonic mean of P and R: $\frac{2PR}{P+R}$) for both entity identification and relation extraction. We conduct holdout methodology for parameter tuning and optimization of our model. We compare our approach with a series of linear-chain CRFs: **CRF+CRF** and a joint model **DCRF** (Sutton *et al.*, 2007): dynamic probabilistic models combined with factored approach to multiple sequence labeling. **CRF+CRF** perform entity identification and relation extraction separately. Relation extraction is viewed as a sequence labeling problem in the second CRF. All these models exploit standard parameter learning and inference algorithms

Table 1: Comparative performance of our model, CRF+CRF, and DCRF models for entity identification.

Entities	CRF+CRF			DCRF			Our model		
	P	R	F_1	P	R	F_1	P	R	F_1
person	75.33	83.22	79.08	75.96	83.82	79.70	82.91	84.26	**83.58**
location	77.03	69.45	73.04	77.68	70.13	73.71	82.94	80.52	**81.71**
organization	53.78	47.76	50.59	54.55	46.98	50.48	61.63	62.61	**62.12**
date	98.54	97.53	**98.03**	97.98	95.22	96.58	98.90	96.24	97.55
year	97.14	99.10	98.11	98.12	99.09	**98.60**	97.36	99.55	98.44
time	60.00	20.33	30.37	50.00	25.33	33.63	100.0	25.00	**40.00**
number	98.88	60.33	74.94	100.0	66.00	**79.52**	100.0	65.52	79.17
miscellaneous	77.42	80.56	78.96	79.81	83.14	81.44	82.69	85.16	**83.91**
Overall	89.55	88.70	89.12	90.98	90.37	90.67	93.35	93.37	**93.36**

in our experiments. To avoid over-fitting, penalization techniques on likelihood are performed. We also use the same set of features for all these models.

6.4 Experimental Results

Table 1 shows the performance of entity identification and Table 2 shows the overall performance of relation extraction [3], respectively. Our model substantially outperforms all baseline models on the overall F-measure for entity identification, resulting in an relative error reduction of up to 38.97% and 28.83% compared to **CRF+CRF** and **DCRF**, respectively. For relation extraction, the improvements on the F-measure over **CRF+CRF** and **DCRF** are 4.68% and 3.75%. McNemar's paired tests show that all improvements of our model over baseline models are statistically significant. These results demonstrate the merits of our approach by capturing tight interactions between entities and relations to explore mutual benefits. The pipeline model **CRF+CRF** performs entity identification and relation extraction independently, and suffers from problems such as error accumulation. For example, **CRF+CRF** cannot extract the *member_of* relation between the secondary entity *Republican* and the principal entity *George W. Bush*, since the organization name *Republican* is incorrectly labeled as a *miscellaneous*. By modeling interactions between two subtasks, enhanced performance is achieved, as illustrated by **DCRF**. Unfortunately, training a **DCRF** model with unobserved nodes (hidden variables) makes this approach difficult to optimize, as we will show below.

[3]Due to space limitation, we only present the overall performance and omit the performance for 53 relation types.

Table 2: Comparative performance of our model, CRF+CRF, and DCRF models for relation extraction.

Model	Precision	Recall	F-measure
CRF+CRF	70.40	57.85	63.51
DCRF	69.30	60.22	64.44
Our model	72.57	64.30	**68.19**

The efficiency of different models is summarized in Table 3. Compared to the pipeline model **CRF+CRF**, the learning time of our model is only a small constant factor slower. Notably, our model is over orders of magnitude (approximately 15.7 times) faster than the joint model **DCRF**. The **DCRF** model uses loopy belief propagation (LBP) for approximate learning and inference. When the graph has large tree-width as in our case, the LBP algorithm in **DCRF** is inefficient, and is slow to converge. Using L-BFGS and the CIC approximate inference algorithms, both learning and decoding can be carried out efficiently.

Table 3: Efficiency comparison of different models on learning time (sec.) and inference time (sec.).

Model	Learning time	Inference time
CRF+CRF	2822.55	6.20
DCRF	105993.00	127.50
Our model	6733.69	62.75

Table 4 compares our CIC inference with two state-of-the-art inference approaches: Gibbs sampling (GS) (Geman and Geman, 1984) and the iterative classification algorithm (ICA) (Neville and Jensen, 2000) for our model. The CIC inference is shown empirically to help improve classi-

Table 4: Comparative performance of different inference algorithms for our model on entity identification and relation extraction.

Entity	Precision	Recall	F-measure
GS	92.45	92.15	92.30
ICA	92.19	91.98	92.08
CIC	93.35	93.37	**93.36**
Relation	Precision	Recall	F-measure
GS	71.22	63.29	67.02
ICA	71.58	63.68	67.40
CIC	72.57	64.30	**68.19**

fication accuracy and robustness over these two algorithms. When probability distributions are very complex or even unknown, the GS algorithm cannot be applied. ICA iteratively infers the states of variables given the current predicted labeling assignments of neighboring variables as observed information. Prediction errors on labels may then propagate during the iterations and the algorithm will then have difficulties to generalize correctly.

We mention some recently published results related to Wikipedia datasets (Note that it is difficult to compare with them strictly, since these results can be based on different experimental settings). Culotta *et al.* (2006) used a data set with a 70/30 split for training/testing and Nguyen *et al.* (2007) used 5930 articles for training and 45 for testing, to perform relatione extraction from Wikipedia. And the obtained F-measures were 67.91 and 37.76, respectively. Yu *et al.* (2009) proposed an integrated approach incorporating probabilistic graphical models with first-order logic to perform relation extraction from encyclopedia articles, with a F-measure of 65.66. All these systems assume that the golden-standard entities are already known and they only perform relation extraction. However, such assumption is not valid in practice. Notably, our approach deals with a fairly more challenging problem involving both entity identification and relation extraction, and it is more applicable to real-world IE tasks.

7 Related Work

A number of previous researchers have taken steps toward joint models in NLP and information extraction, and we mention some recently proposed, closely related approaches here. Roth and Yih (2007) considered multiple constraints between variables from tasks such as named entities and relations, and developed a integer linear programming formulation to seek an optimal global assignment to these variables. Zhang and Clark (2008) employed the generalized perceptron algorithm to train a statistical model for joint segmentation and POS tagging, and applied multiple-beam search algorithm for fast decoding. Toutanova *et al.* (2008) presented a model capturing the linguistic intuition that a semantic argument frame is a joint structure, with strong dependencies among the arguments. Finkel and Manning (2009) proposed a discriminative feature-based constituency parser for joint named entity recognition and parsing. And Dahlmeier *et al.* (2009) proposed a joint model for word sense disambiguation of prepositions and semantic role labeling of prepositional phrases. However, most of the mentioned approaches are task-specific (e.g., (Toutanova *et al.*, 2008) for semantic role labeling, and (Finkel and Manning, 2009) for parsing and NER), and they can hardly be applicable to other NLP tasks. Since we capture rich and complex dependencies between subtasks via potential functions in probabilistic graphical models, our approach is general and can be easily applied to a variety of NLP and IE tasks.

8 Conclusion and Future Work

In this paper, we investigate the compound IE task of identifying entities and extracting relations between entities in encyclopedia text. And we propose a unified framework based on undirected, conditionally-trained probabilistic graphical models to perform all relevant subtasks jointly. More importantly, we propose a new algorithm: CIC, to enable approximate inference to find the MAP assignments for both segmentations and relations. As we shown, our modeling offers several advantages over previous models and provides a natural formalism for this compound task. Experimental study exhibits that our model significantly outperforms state-of-the-art models while also running much faster than the joint models. In addition, the superiority of the CIC algorithm is also discussed and compared. We plan to improve the scalability of our approach and apply it to other real-world problems in the future.

References

Christophe Andrieu, Nando de Freitas, Arnaud Doucet, and Michael I. Jordan. An introduction to MCMC for machine learning. *Machine Learning*, 50(1):5–43, 2003.

Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society*, 36:192–236, 1974.

Aron Culotta, Andrew McCallum, and Jonathan Betz. Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In *Proceedings of HLT/NAACL-06*, pages 296–303, New York, 2006.

Daniel Dahlmeier, Hwee Tou Ng, and Tanja Schultz. Joint learning of preposition senses and semantic roles of prepositional phrases. In *Proceedings of EMNLP-09*, pages 450–458, Singapore, 2009.

Jenny Rose Finkel and Christopher D. Manning. Joint parsing and named entity recognition. In *Proceedings of HLT/NAACL-09*, pages 326–334, Boulder, Colorado, 2009.

Jenny Rose Finkel, Christopher D. Manning, and Andrew Y. Ng. Solving the problem of cascading errors: Approximate Bayesian inference for linguistic annotation pipelines. In *Proceedings of EMNLP-06*, pages 618–626, Sydney, Australia, 2006.

Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transitions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

Kristy Hollingshead and Brian Roark. Pipeline iteration. In *Proceedings of ACL-07*, pages 952–959, Prague, Czech Republic, 2007.

Frank R. Kschischang, Brendan J. Frey, and Hans-Andrea Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47:498–519, 2001.

John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML-01*, pages 282–289, 2001.

Lin Liao, Dieter Fox, and Henry Kautz. Extracting places and activities from GPS traces using hierarchical conditional random fields. *International Journal of Robotics Research*, 26:119–134, 2007.

Jennifer Neville and David Jensen. Iterative classification in relational data. In *Proceedings of the AAAI-2000 Workshop on Learning Statistical Models from Relational Data*, pages 42–49, 2000.

Dat P. T. Nguyen, Yutaka Matsuo, and Mitsuru Ishizuka. Relation extraction from Wikipedia using subtree mining. In *Proceedings of AAAI-07*, pages 1414–1420, Vancouver, British Columbia, Canada, 2007.

Hoifung Poon and Pedro Domingos. Joint inference in information extraction. In *Proceedings of AAAI-07*, pages 913–918, Vancouver, British Columbia, Canada, 2007.

Dan Roth and Wentau Yih. Global inference for entity and relation identification via a linear programming formulation. In Lise Getoor and Ben Taskar, editors, *Introduction to Statistical Relational Learning*. MIT Press, 2007.

Sameer Singh, Karl Schultz, and Andrew Mccallum. Bidirectional joint inference for entity resolution and segmentation using imperatively-defined factor graphs. In *Proceedings of ECML/PKDD-09*, pages 414–429, Bled, Slovenia, 2009.

Charles Sutton and Andrew Mccallum. Collective segmentation and labeling of distant entities in information extraction. In *Proceedings of ICML Workshop on Statistical Relational Learning and Its Connections to Other Fields*, 2004.

Charles Sutton and Andrew McCallum. An introduction to conditional random fields for relational learning. In Lise Getoor and Ben Taskar, editors, *Introduction to Statistical Relational Learning*. MIT Press, 2007.

Charles Sutton, Andrew McCallum, and Khashayar Rohanimanesh. Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. *Journal of Machine Learning Research*, 8:693–723, 2007.

Kristina Toutanova, Aria Haghighi, and Christopher D. Manning. Joint learning improves semantic role labeling. In *Proceedings of ACL-05*, pages 589–596, 2005.

Kristina Toutanova, Aria Haghighi, and Christopher D. Manning. A global joint model for semantic role labeling. *Computational Linguistics*, 34:161–191, 2008.

Kristina Toutanova. *Effective statistical models for syntactic and semantic disambiguation*. PhD thesis, Stanford University, 2005.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1:1–305, 2008.

Xiaofeng Yu, Wai Lam, and Shing-Kit Chan. A framework based on graphical models with logic for chinese named entity recognition. In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP-08)*, pages 335–342, Hyderabad, India, 2008.

Xiaofeng Yu, Wai Lam, and Bo Chen. An integrated discriminative probabilistic approach to information extraction. In *Proceedings of CIKM-09*, pages 325–334, Hong Kong, China, 2009.

Xiaofeng Yu. Chinese named entity recognition with cascaded hybrid model. In *Proceedings of HLT/NAACL-07*, pages 197–200, Rochester, New York, 2007.

Yue Zhang and Stephen Clark. Joint word segmentation and POS tagging using a single perceptron. In *Proceedings of ACL-08*, pages 888–896, Ohio, USA, 2008.

Jun Zhu, Zaiqing Nie, Ji-Rong Wen, Bo Zhang, and Wei-Ying Ma. 2D conditional random fields for Web information extraction. In *Proceedings of ICML-05*, pages 1044–1051, Bonn, Germany, 2005.